Contracts with Interdependent Preferences

Debraj Ray and Marek Weretka[†]

March 2024

Abstract. A principal contracts with multiple agents, as in Lazear and Rosen (1981) and Green and Stokey (1983). The setup is classical except for the assumption that agents have interdependent preferences. We characterize cost effective contracts, and relate the direction of co-movement in rewards — "joint liability" (positive) or "tournaments" (negative) — to the assumed structure of preference interdependence. We also study the implications of preference interdependence for the principal's payoffs. We identify two asymmetries. First, the optimal contract leans towards joint liability rather than tournaments, especially in larger teams, in a sense made precise in the paper. Second, when the mechanism-design problem is augmented by robustness constraints designed to eliminate multiple equilibria, the principal may prefer teams linked via adversarial rather than altruistic preferences.

1. Introduction

Imagine yourself as an economist designing contracts for a plantation in a developing country. The workers are fruit pickers or tea pluckers, drawn from nearby villages so that many are friends or family, or have a social connection. Upon observing the prevalent payment structure, which consists of a fixed wage accompanied by a bonus tied to the quantity of individually harvested output, you — a firm believer in economic incentives — may be tempted to further enhance the incentive scheme. We don't mean a reduction in the base wage and a larger slope on harvested output; these may be restricted either because of limited liability or mandated minimum wage payments (Jayaraman et al. 2016). You might resort to more intricate, *inter-agent incentives*, where the payoff to one worker is somehow related to the performance of her compatriots; see Lazear and Rosen (1981), Holmstrom (1982), Green and Stokey (1983), Nalebuff and Stiglitz (1983) and Mookherjee (1984).

Examples of such mechanisms are tournaments and joint liability mechanisms. They are often used; see Wantchekon (1994) and Bandiera, Barankay and Rasul (2005) for instances of the former, and Banerjee, Besley and Guinnane (1994), Besley and Coate (1995), Ghatak (1999) or Ghatak and Guinnane (1999) for examples of the latter. Such arrangements involve interactions across agent payment. In a tournament, a worker is rewarded if she outperforms her colleague, creating

[†]Ray: New York University and University of Warwick, debraj.ray@nyu.edu; Weretka: University of Wisconsin-Madison and FAME/GRAPE. We thank Dilip Abreu, Yeon-Koo Che, Parikshit Ghosh, Navin Kartik, Meg Meyer, Dilip Mookherjee, Nick Netzer, Marco Ottaviani, David Pearce, Fernando Vega-Redondo, Lones Smith, Ennio Stacchetti and Eyal Winter for insightful comments. Ray acknowledges support under National Science Foundation Grant SES-2315720. Weretka acknowledges the support of National Center for Science (grant 2019/33/B/HS4/00151).

a competitive environment between them. On the other hand, joint liability mechanisms link higher worker compensation to better compatriot performance, giving rise to a more cooperative environment. Of course, one could envision other, less extreme, variants of these mechanisms, with standard pay for performance suitably augmented by winners' prizes or joint performance bonuses. We are going to study the relationship between *interdependent preferences* and optimal contracting.

Substantial empirical evidence suggests that agents not only prioritize material rewards but are also affected, positively or negatively, by the payoffs of their fellow agents. Sobel (2005) and Fehr and Gächter (2000) extensively review the empirical literature on interdependent preferences. The importance of social preferences in the workplace is specifically recognized in Bandiera et al. (2005), List (2009) and Luft (2016). Such interdependence is particularly relevant in teams drawn from a particular geographical location such as a village, where agents are more likely to know one another.

It is with this emphasis in mind that we introduce payoff-interdependence — altruistic or adversarial — into an otherwise standard team agency problem. As already noted, we are not the first to do this, but our focus is firmly on how competition and cooperation relate to the nature of preference interdependence, a theme on which we elaborate below. Evolutionary arguments suggest that both altruism and envy could emerge depending on the ambient environment, with independence perhaps being the exception rather than the rule; see Axelrod and Hamilton (1981), Kockesen, Ok and Sethi (1997), Trivers (2006), Marsh (2016), and Robson (2017). Apart from evolutionary arguments, it is intuitive that in settings where individuals know one another, preferences will naturally intertwine owing to ties of family and kinship, and additionally for reasons of socioeconomic similarity, friendship, shared experiences, and past interactions.

We study a principal who employs a number of risk-averse identical agents. Each of them chooses a binary effort (work, shirk). Our specification is extremely simple: we presume that each agent assigns symmetric weight to the welfare of other agents. This weight can be positive or negative, and it is commonly known to all parties. The principal's objective is to efficiently encourage unobservable effort, by offering a symmetric contract to all agents. The contract specifies a non-negative monetary transfer to the agent, conditional on the vector of observed idiosyncratic signals, interpreted here as agent outputs. We deliberately restrict the analysis by supposing that each agent's signal is uninformative about the efforts of others. For this reason, all cross-agent dependencies in payoffs will arise solely from payoff externalities and are not influenced by the informational considerations that have already been extensively studied in the literature. (Our analysis can be easily extended to incorporate such informational considerations, at the cost of blunting our desired focus.)

Incentivizing agents with interdependent preferences requires us to accommodate a central idea which can quickly turn quite complex. The offer of a contract not only affects an agent's incentives via her material payoff, but also has indirect repercussions — her actions may well affect the

payoffs of others. Designing an optimal contract necessitates that such payoff spillovers be taken fully into account. A technical contribution of our exercise is the introduction and use of a statistic Ψ that comfortably summarizes the total impact on an agent by incorporating direct and indirect payoff externalities. We derive Ψ through an appropriately weighted sum of likelihood ratios of signals, and use it to decouple agent behavior within a team, leading to a simpler analysis.

Of course, we do not wish to remain in the technical realm of simply linking optimal contracts to a particular statistic. We would like to use that linking device to say something about the qualitative features of optimal contracts. The key lessons are outlined below:

1. Optimal contracts span an entire spectrum, broadly ranging from competitive to cooperative. Tournaments, in which agents are rewarded for outperforming their compatriots, exemplify the former. Joint liability, in which agents are rewarded for their own successes *and* those of others, are characteristic of the latter. Proposition 1 links such contracts to adversarial and altruistic preferences, offering a "simple unified explanation" that accounts for a plethora of contract formats within the same model. At the same time, the model eliminates certain contracts, a theme that we develop in the next point.

2. There is an asymmetry between the extent of competitiveness and the extent of cooperation under optimal contractual solutions. This arises from the reinforcing nature of payoff interdependence, and the asymmetry is more pronounced in larger teams. For instance, in an example with 20 team members, a contract is highly cooperative when each agent assigns weight close to $\alpha = 4\%$ to the well-being of others. The resulting sensitivity of optimal payments to each compatriot's output (relative to the sensitivity to one's own output) is 14%. But a value of α around -4% results in a sensitivity of less than 2.3% to the performance of other agents. The asymmetry emerges starkly in the general model when passing to the two limits with maximally altruistic and maximally adversarial attitudes;[1] see Propositions 2 and 3. In the altruistic limit, every agent's contract converges to a pure joint liability contract, where payment could depend on the vector of individual performances but is exactly the same for every agent regardless of performance. In the corresponding adversarial limit, there is convergence to a pure tournament with winner taking all, *but only when there are just two team members*. With three or more agents, the limit contract pays an agent for her own performance even when that performance is not ranked the best. The prediction that cooperative rewards can more effectively motivate effort aligns with the patterns observed in labor markets; see Che and Yoo (2001) and Luft (2016).[2]

[1] These limits are defined by natural upper bounds on the degree of interdependence that prevent the interdependent utility system from becoming "explosive." The limits exclude situations in which agents can become infinitely happy (or unhappy) *simply* via the changing utilities of other agents. They remain fundamentally anchored by their *own* material payoffs. For derivations of these bounds and studies of their implications in other contexts, see Pearce (2008), Hori and Kanaya (1989), Bergstrom (1999) and Ray ⓡ Vohra (2020).

[2] For more in-depth discussion of the literature that documents these trends see Che and Yoo (2001). For the literature in the field of managerial studies, see Luft (2016).

3. Principals prefer interdependent preferences of either kind — altruistic or adversarial — to independent preferences (Proposition 4). Specifically, when incentive constraints are strong enough so as to render participation constraints irrelevant, a higher degree of preference interdependence benefits the principal, irrespective of the sign of the preference coefficient. That is, the principal's expected payout to the agents as a function of α follows an inverted U-shaped curve, peaking at independent preferences and monotonically decreasing with the absolute value of the interdependence parameter. Intuitively, the principal capitalizes on either form of interdependence by choosing either competitive or cooperative bonuses, which reduce the overall transfer made to agents, thereby benefiting the principal.

4. Altruism can be detrimental to the material payoffs of agents, and to their overall welfare, though the latter could be a problematic concept with changing interdependence (see Section 7.2). Normalize material payoffs so that they are zero at zero consumption and zero effort. Then an agent with independent preferences and a limited liability constraint at zero consumption cannot be driven below zero payoff — she can always set effort equal to zero. However, with interdependent preferences, agents could experience negative expected material payoffs even with limited liability in place. Altruistic agents may choose costly effort so as to protect their partners. Adversarial agents might do the same to *reduce* the likelihood of favorable outcomes for others. Thus, preference interdependence can be leveraged to relax the limited liability constraint. With additively interdependent preferences, this also means that an agent's overall welfare might sink to negative levels in equilibrium.

5. Finally, there are other asymmetries across the altruistic and adversarial scenarios. These have to do with the unique implementation of high-effort outcomes. Competitive contracts are typically submodular for adversarial agents, inducing a unique Nash equilibrium. Implementation of the high-effort equilibrium therefore comes with no strategic ambiguity. The same is not true when agents are altruistic. Joint liability typically generates supermodularity across efforts, raising the specter of multiple equilibria in the post-contract game across agents.[3] For unique implementation, the principal needs to dial back on team rewards, and only partially capitalize on available altruism; see Proposition 5. This observation might rationalize the use of competitive mechanisms even when agents are altruistic, especially in settings with a small number of agents. This consideration stands in contrast to the "cooperative bias" exhibited in item 2, so that the earlier arguments in favor of joint liability are attenuated when potentially bad equilibria need to be eliminated.

We conclude this Introduction by pointing out some limitations and possible extensions of our exercise. An especially stark assumption concerns the equality, both in magnitude and sign, of the

[3]On contracts that secure unique implementation, see Segal (1999, 2003), Winter (2004), Genicot and Ray (2006), Bernstein and Winter (2012), Halac, Kremer and Winter (2020), Halac, Lipnowski and Rappoport (2021), Halac, Kremer and Winter (2023), and Camboni and Porcellacchia (2023). For a specific instance of multiple equilibria in the joint liability setting, see Besley and Coate (1995).

interdependence parameter α across all agents. If we drop these assumptions, the specific contractual form is complicated and depends in a detailed way on the full vector of α's. Differences in the extent of altruism now matter, as also the possibility that there may well be friends and enemies within the same team. So the optimal contract will exhibit elements of competition and cooperation in complex ways. Such constructions rely on fine details that are unlikely to be available to the principal, who might have some overall idea of where her labor force is from (e.g., the same village), but not much more than that. That raises the interesting question of designing robust mechanisms in this setting.

Second, both the intensity and sign of the parameter α may well depend on the past experiences (and contractual settings) of team members. But the dependence is subtle. At first glance it seems natural that adversarial interdependence would arise in competitive environments, in which one person's gains are another's loss; see, e.g., Lanzetta and Englis (1989) and Zillman and Cantor (1977). But it is unclear that tennis players or students graded on the curve are any more adversarial than anyone else. After all, they *do* understand the rules of the game, and why their compatriots seek to outdo them. Indeed, it is in "cooperative" joint liability settings, in which low payments might be blamed on some shirkers within the team, that antagonism could more easily appear (see Ghatak and Guinnane 1999 for some empirical observations in the context of microfinance). It is even possible that an overall drift towards adversarial attitudes might come to dominate the present, regardless of the contractual environment of the past (Kockesen, Ok and Sethi 1997). There are potentially interesting dynamics involved here, but these are beyond the scope of the current paper.

2. Related Literature

The early theoretical literature on team agency emphasized the role of informational externalities among agents. Lazear and Rosen (1981), Holmstrom (1982), Green and Stokey (1983), Nalebuff and Stiglitz (1983) and Mookherjee (1984) offer an explanation for why with common informational shocks principals may resort to tournaments. The literature on joint liability (Stiglitz 1990, Ghatak 1999, or Ghatak and Guinnane 1999) shows how agent payoffs may co-move positively under optimal contracting to monitoring advantages. Both literatures ignore the payoff externalities imposed on others under optimal contracts via interdependent preferences. Our problem differs from these papers: in our framework, the output of an agent contains *no information* about her partner's effort. This allows us to identify the effects of interdependent preferences in isolation from the informational considerations that have already been studied extensively in the literature.

Our specific interest lies in the possible asymmetry across contracts with positive and negative co-movement of rewards. In this sense, we are in line with Meyer and Mookherjee (1987), Che and Yoo (2001) and DeMarzo and Kaniel (2023), who all contrast group rewards to tournament-like structures. Meyer and Mookherjee (1987) warn against the use of tournaments when the principal cares for equality ex post. Such ex post concerns should precipitate a social preference for

positively correlated agent payoffs, thereby reducing the use of competitive contracts but eroding agent incentives in the process. They conclude that "welfare-optimal compensation schemes in general depend separately on an equity and an incentive component that tend to correlate agent compensations in different directions." In contrast, Che and Yoo (2001) is set in a dynamic context and the incentive effect of positive correlation is different there. Cooperative rewards encourage peer monitoring of effort among team members and can be used effectively as punishments in the event of shirking — it is easier for a team member to shirk and lower the reward of a compatriot in a cooperative setting, as opposed to working hard to achieve the same end in a competitive setting.[4] DeMarzo and Kaniel (2023) study a setting in which agents assess their wage against a weighted average of wages paid to fellow agents. Their exercise emphasizes the impact of wage transparency and inter-principal competition on equilibrium wage structure.

Relative to these papers, our focus on interdependent preferences is distinct, as are the specific results we obtain. Our model is static, so the forces basic to the Che-Yoo argument are absent here. Indeed, the asymmetry we uncover relies on teams with at at least *three* agents, whereas Che and Yoo (2001) considers a two-agent model throughout. And unlike Meyer and Mookherjee (1987) or DeMarzo and Kaniel (2023), our principal or agents have no particular preference for *ex post* equality, which is of course central to their exercise.

Other contributions to contracting with interdependent preferences ask different questions (in different models). Letina, Liu and Netzer (2020) study multiple agents with a single intermediary — a "reviewer" — positioned between principal and agents. This reviewer observes agent performance and has social preferences towards them. The study explores the effect of reviewer bias; e.g., a leniency bias in which the reviewer hesitates to report instances of agent shirking to avoid subjecting them to contractual punishment. Itoh (2004) considers a two-state model with a principal and two risk-neutral but inequality-averse agents, each of whom compares their wages to those paid to the other. To some degree, inequality-aversion captures both altruism (no one wants to forge too far ahead) and antagonism (no one wants to be left behind). Depending on the magnitudes of these two components, derived contracts can be competitive or cooperative. Vásquez and Weretka (2021) study firms who hire workers with interdependent preferences, but they presume that all workers are paid an unconditional, uniform wage. Their goal is to study the equilibrium size of the work force, and performance-based pay plays no role in this exercise.

Our paper is distinct in from these contributions, in that the model we use is different.[5] But more importantly, we emphasize a different question altogether, which (as already noted) studies the *asymmetric* incidence of competitive and cooperative contracts, which none of these papers do.

[4] As a countervailing force, Che and Yoo (2001) also incorporate an informational externality in the form of a common shock to outputs, which tends to favor competitive contracts. Their study then highlights the optimality of joint liability mechanisms when the discount factor is large relative to the size of the informational externality, and the optimality of a tournament otherwise.

[5] For instance, in contrast to Letina et al. (2020), we examine the interplay of preferences among the agents themselves. In contrast to Itoh (2004), we use preference interdependence rather than inequality aversion, and the presence of three or more agents as well as their risk aversions play crucial roles in establishing our results.

This, in conjunction with the related question of the implications of interdependent preferences for principal payoffs, is a central theme of our paper.

3. Model

3.1. **Agent Payoffs and Outputs.** There are n agents, each of whom make a binary effort choice $e \in \{\ell, h\}$, at a cost of 0 for $e = \ell$ and $c > 0$ for $e = h$. For each effort level $e \in \{\ell, h\}$, a probability measure μ^e, identical for all agents, determines output y. The two measures have common support Y. We assume that μ^ℓ is absolutely continuous with respect to μ^h, so that the Radon–Nikodym derivative $[d\mu^\ell / d\mu^h](y)$, referred to here as the *likelihood ratio* at y, is always well-defined. As is well known, this reduces to the ratio of probabilities over $y \in Y$ when Y is discrete, and to the ratio of densities for distributions on Y, when those densities exist. Without any essential loss of generality[6] we assume that

$$\frac{d\mu^\ell}{d\mu^h}(y) \text{ is declining in } y \text{ on } Y. \tag{1}$$

Outputs are taken to be independent across agents. Interdependent outputs are not hard to incorporate, but our focus is on interdependent *preferences* instead. Specifically, each agent has symmetric interdependent preferences, with utility given by

$$U_i = [u(m) - c\mathbb{1}_i^h] + \alpha \sum_{j \neq i} U_j, \text{ for } i = 1, \ldots, n, \tag{2}$$

where $u(m)$ is a vNM felicity on money, $\mathbb{1}_i^h$ is an indicator for high effort, α is a measure of altruism or antipathy across agents, and $\{U_j\}$ are the utilities of the other agents. The felicity function is twice continuously differentiable, with $u(0) = 0$, $u'(m) > 0$, $u''(m) < 0$, $u'(0) = \infty$ and $u'(\infty) = 0$. The agent's material payoff is $u(m) - c\mathbb{1}_i^h$.

3.2. **Reduced-Form Payoff Representation.** A contract is offered. Then agents interact, each choosing effort independently. This interaction across agents is not formally a game. It becomes one, once we reduce all utility interactions in (2) to a collection of payoff functions that depend on agent *actions*. We follow Pearce (2008), Hori and Kanaya (1989), Bergstrom (1999) and Ray ⓡ Vohra (2020) to obtain a "coherent" utility representation of this interactive system (2) on the space of actions. The case studied in this paper is particularly simple. For every i, $U_i = [u(m_i) - c\mathbb{1}_i^h] + \alpha \sum_{j \neq i} U_j$, so that

$$(1 + \alpha) U_i = [u(m_i) - c\mathbb{1}_i^h] + \alpha S, \tag{3}$$

[6] In fact, our approach allows for signals $\mathbf{y}$ to be abstract outcomes in some probability space. Define $\lambda_i \equiv 1 - \frac{d\mu^\ell}{d\mu^h}(y_i)$ as we do below; then these objects are increasing in "good performance" and optimal contracts can be directly expressed as functions of $\boldsymbol{\lambda} = (\lambda_1, ..., \lambda_n)$ rather than $\mathbf{y}$. The optimal contracts for the original signals $\mathbf{y}$ can then be backed out as $m(\boldsymbol{\lambda}(\mathbf{y}))$ where $m(\boldsymbol{\lambda})$ is an optimal contract defined on realizations of $\boldsymbol{\lambda}$.

where $S \equiv \sum_j U_j$. Coherence asks that this sum be well-defined and move in the same way as the sum of material payoffs, the necessary and sufficient condition for which is $|\alpha(n-1)| < 1$ (see the references cited above). We can then add (3) over all i and divide by $1-\alpha(n-1)$ to obtain

$$S = \frac{\sum_j [u(m_j) - c\mathbb{1}_j^h]}{1-\alpha(n-1)}, \tag{4}$$

and using (4) in (3) and transposing terms, we see that for every i,

$$V_i \equiv (1+\alpha)\frac{1-\alpha(n-1)}{1-\alpha(n-2)}U_i = [u(m_i) - c\mathbb{1}_i^h] + \frac{\alpha}{1-\alpha(n-2)}\sum_{j\neq i}[u(m_j) - c\mathbb{1}_j^h], \tag{5}$$

where V_i will denote payoff, renormalized by $(1+\alpha)[1-\alpha(n-1)]/[1-\alpha(n-2)]$. While material payoffs affect this utility one for one, the combined weight of other individuals' material payoff is always less than one, given that $|\alpha(n-1)| < 1$.

3.3. **Contracts and Equilibirum.** A principal hires the agents and offers i a limited liability contract $m_i(\mathbf{y})$ with $m_i(\mathbf{y}) \geq 0$ for each $\mathbf{y}$. Contracts are assumed to be symmetric across agents. That is, each agent is paid the same way with respect to own output, and with respect to the vector of outputs produced by compatriots.[7] Technically, we use a weaker form of symmetry. Fix the agent order $1, \ldots, n$. For any $\mathbf{y}$ and agent i, a *rotation* $\mathbf{y}^{[i]}$ arises from a permutation which sends agent 1 to position i, and likewise shifts all other agents by the amount $(i-1)\text{mod}(n)$. Then there is a common function m on $\mathbb{R}_+^n$ such that $m_i(\mathbf{y}) = m(\mathbf{y}^{[i]})$ for all i and $\mathbf{y}$.

Symmetry is easy enough to motivate on legal grounds of "equal treatment," or from the assumption that discriminatory payment schemes would lead to an unacceptable loss of worker morale. But the assumption is not without loss of generality. A principal who is free to offer different contracts, even to identical agents, might benefit from that ability.

Throughout, we assume that the principal wants to implement high effort for all agents. We write the principal's problem as choosing a nonnegative function m, to be interpreted as described above, so as to

$$\text{minimize } \mathbb{E}(m|\mathbf{e}=\mathbf{h}), \text{ such that } \mathbf{e}=\mathbf{h} \text{ is incentive-compatible.} \tag{6}$$

The interpretation of "incentive compatibility" that we adopt here presumes that all efforts and all payments are revealed *ex-post* to every agent. Thus, when an agent contemplates a deviation, she understands that all agents will know about the new effort choice and reward payments, and experience their payoffs accordingly. Specifically, there are three sets of effects when i deviates from $e=h$ to $e=\ell$:

(i) i's lower effort changes her own material payoffs.

[7] That is, there is a function m on $\mathbb{R}_+^n$, invariant to permutations of its last $n-1$ entries, such that $m_i(\mathbf{y}) = m(\mathbf{y}^{[i]})$, where $\mathbf{y}^{[i]}$ is any vector with $y_1^{[i]} = y_i$ as its first entry and any permutation of $\mathbf{y}_{-i}$ for the rest.

(ii) i's lower effort changes the material payoffs to other agents $j \neq i$, because their contractual payment may well depend on i's output.

(iii) Each of these material changes echo through the utility system in (2), leading to a final payoff experience for i.

Other interpretations of incentive compatibility can also be studied (see Conclusions). That said, we always maintain that efforts are not verifiable by the principal, so no contractual payment can be conditioned on effort.

For most of the analysis that follows, we do not explicitly include a participation constraint for the agents. The limited liability constraint already constrains payments on the downside, so that our problem is non-trivial. That said, we study participation constraints in Section 6. Second, "implementing high effort" as described in (6) means that *some* equilibrium involves high effort for all. In contrast, "robust implementation" would require *every* second stage equilibrium to exhibit high effort. We return to this issue in Section 8.

4. Optimal Contracts

Recall that Y is the common support of outputs. Define $\mathbf{Y} \equiv Y^n$, where n is the number of agents. We write $d\mu_j^e$ for $\mu^e(dy_j)$, where dy_j could denote counting measure or Lebesgue measure (or some mix thereof) depending on the context. We set $\boldsymbol{d\mu}^{\mathbf{e}} \equiv \prod_j d\mu_j^{e_j}$ to indicate integration with respect to the joint probability of $\mathbf{y}$ when agents supply effort vector $\mathbf{e}$, and set $\boldsymbol{d\mu}^{\mathbf{e}}_{-i} \equiv \prod_{j\neq i} d\mu_j^e$.

Recall that we study symmetric contracts m, so that the monetary reward to i under $\mathbf{y}$ is given by $m(\mathbf{y}^{[i]})$, using the notation already introduced.

4.1. The Incentive Constraint.

Suppose that agent i's compatriots all choose $e_j = h$. Then by (5), if agent i chooses $e_i = h$, her (renormalized) payoff is

$$V_i^h = \left[\int_{\mathbf{Y}} u(m(\mathbf{y}^{[i]}))\boldsymbol{d\mu}^{\boldsymbol{h}} - c\right] + \frac{\alpha}{1-\alpha(n-2)}\sum_{j\neq i}\left[\int_{\mathbf{Y}} u(m(\mathbf{y}^{[j]})\boldsymbol{d\mu}^{\boldsymbol{h}} - c\right]. \tag{7}$$

In similar fashion, if our agent chooses $e = \ell$, her expected payoff V_i^ℓ is given by

$$V_i^\ell = \left[\int_{\mathbf{Y}} u(m_i(\mathbf{y}^{[i]}))d\mu_i^\ell \boldsymbol{d\mu}^{\boldsymbol{h}}_{-i}\right] + \frac{\alpha}{1-\alpha(n-2)}\sum_{j\neq i}\left[\int_{\mathbf{Y}} u(m(\mathbf{y}^{[j]}))d\mu_i^\ell \boldsymbol{d\mu}^{\boldsymbol{h}}_{-i} - c\right]. \tag{8}$$

Remembering that the principal wishes to implement $e = h$, the incentive constraint is given by $V_i^h \geq V_i^\ell$, or, using (7) and (8),

$$\int_{\mathbf{Y}} u(m(\mathbf{y}^{[i]}))\lambda(y_i)\boldsymbol{d\mu}^{\boldsymbol{h}} + \frac{\alpha}{1-\alpha(n-2)}\sum_{j\neq i}\left[\int_{\mathbf{Y}} u(m(\mathbf{y}^{[j]}))\lambda(y_i)\boldsymbol{d\mu}^{\boldsymbol{h}}\right] \geq c,$$

where $\lambda(y) \equiv 1 - [d\mu^\ell / d\mu^h](y)$ is well defined by the assumption of absolute continuity $\mu^\ell \ll \mu^h$. We now conduct a change of variables within the integrals in the summation above. First set $i = 1$ and note that $\mathbf{y}^{[1]} = \mathbf{y}$. Next, for each $j \neq 1$, we "rotate" the entries in $\mathbf{y}^{[j]}$ so that $[j]$ is replaced by $[1]$, with all other indices permuted accordingly, including the index $i = 1$ under the second integral, which will run the entire gamut of values $\{2, \ldots, n\}$ as different values of j are thus replaced. Because the contract is symmetric, that gives us an equivalent representation of the incentive constraint as

$$\int_{\mathbf{Y}} u(m(\mathbf{y}))\Psi(\mathbf{y})\boldsymbol{d\mu^h} \geq c, \tag{9}$$

where we've defined

$$\Psi(\mathbf{y}) \equiv \lambda(y_1) + \frac{\alpha}{1 - \alpha(n-2)} \sum_{j \neq 1} \lambda(y_j). \tag{10}$$

The function Ψ will play a central role in the design of optimal contracts with interdependent preferences. It is an "incentive metric" that measures the extent to which the provision of an additional util to an agent in event $\mathbf{y}$ encourages her effort (or discourages it if negative), appropriately accounting for the utility spillovers from other agents. Because $\lambda(y)$ is one minus the likelihood ratio at y, it must be that $\int \lambda(y) d\mu^h(y) = 0$, as can be verified by direct integration. So, in particular, we take note of the fact that $\int_{\mathbf{Y}} \Psi(\mathbf{y})\boldsymbol{d\mu^h} = 0$.

4.2. **The Principal's Objective.** The principal's expected wage bill under a symmetric contract m is given by

$$\int_{\mathbf{Y}} \sum_{j=1}^{n} m(\mathbf{y}^{[j]})\boldsymbol{d\mu^h} = n \int_{\mathbf{Y}} m(\mathbf{y})\boldsymbol{d\mu^h} \tag{11}$$

and she will seek to minimize this payout, subject to meeting the incentive constraint (9). The use of the incentive metric Ψ in that constraint enables us to effectively reduce a team agency setting with interdependent preferences to a single-agent problem.

4.3. **Independent Preferences.** As a warm-up, shut down utility interdependence by setting $\alpha = 0$. Then $\Psi(\mathbf{y}) = \lambda(y_1)$, and so (9) reduces to

$$\int_{\mathbf{Y}} u(m(\mathbf{y}))\lambda(y_1)\boldsymbol{d\mu^h} = \int_{Y} \left[\int_{Y^{n-1}} u(m(y_1, \mathbf{y}_{-1}))\boldsymbol{d\mu^h_{-1}} \right] \lambda(y_1) d\mu^h_1 \geq c,$$

and written thus, it is clear that conditioning m on $\mathbf{y}_{-1}$ is absurd when u is strictly concave. Calling the solution $m(y_1)$ by a slight abuse of notation, the incentive constraint is

$$\int_{Y} u(m(y_1))\lambda(y_1) d\mu^h_1 \geq c, \tag{12}$$

for agent 1, and by symmetry (12) applies to any agent i, once we replace 1 by i. Because the principal seeks to minimize $\int_Y m(y_1) d\mu^h_1$, it must be the case that the "marginal product" of money

across any set of states with $m(y_1) > 0$ must be constant, or more precisely:

$$\text{If } m(y_1) > 0, \text{ then } u'(m(y_1))\lambda(y_1) \geq u'(m(y_1'))\lambda(y_1') \text{ for all } y_1' \in Y, \tag{13}$$

which means that $u'(m(y_1))\lambda(y_1) = u'(m(y_1'))\lambda(y_1')$ for any (y_1, y_1') with $m(y_1) > 0$ and $m(y_1') > 0$. Under (1), the function $\lambda(y)$ is increasing. Combining this observation with (13), we see that $m(y_i)$ must be flat at 0 up to a threshold y^* that satisfies $\lambda(y^*) = 0$, and thereafter increases with y_i. This standard solution extends nicely to the case of interdependent preferences.

4.4. **Interdependent Preferences.** We use Lagrangean methods which can be easily formalized, despite the possibility that Y might be a continuum. For a reward function m, define the Lagrangean functional

$$\mathcal{L}(m) \equiv -\int_{\mathbf{Y}} m(\mathbf{y}) \boldsymbol{d\mu^h} + \nu^{\text{IC}} \left[\int_{\mathbf{Y}} u(m(\mathbf{y}))\Psi(\mathbf{y}) \boldsymbol{d\mu^h} - c \right], \tag{14}$$

where ν^{IC} is a multiplier for the incentive compatibility constraint, and $\Psi(\mathbf{y})$ is the incentive metric defined in (10).[8] The first order conditions are given by

$$-1 + \nu^{\text{IC}} u'(m(\mathbf{y}))\Psi(\mathbf{y}) \leq 0, \text{ with equality if } m(\mathbf{y}) > 0.$$

But $m(\mathbf{y})$ *must* be positive for some $\mathbf{y}$ otherwise no incentives can be provided. In particular, it follows that $\nu^{\text{IC}} > 0$, and so the first order condition above reduces to

$$u'(m(\mathbf{y}))\Psi(\mathbf{y}) \text{ is constant in } \mathbf{y} \text{ if } \Psi(\mathbf{y}) > 0, \text{ and } m(\mathbf{y}) = 0 \text{ otherwise.} \tag{15}$$

where we've used the end-point conditions $u'(0) = \infty$ and $u'(\infty) = 0$.

To uncover m, unpack Ψ as described in (10). If y_1 alone increases, then $d\mu_1^\ell / d\mu_1^h$ falls and λ rises, increasing $\Psi(\mathbf{y})$. It follows that any agent's reward increases in her own output, provided that it is strictly positive to begin with. As for the cross-effects, they depend on whether the agents are altruistic ($\alpha > 0$) or adversarial ($\alpha < 0$). When $\alpha > 0$, an increase in another agent j's output will raise λ, and by extension — using (10) — it will raise Ψ and therefore the reward to agent 1. This is in the spirit of joint liability. Exactly the opposite happens when $\alpha < 0$, which is in the spirit of a tournament. That establishes

Proposition 1. *Assume the likelihood condition (1). Then the reward to an agent is strictly increasing in own output at any point where $\Psi(\mathbf{y}) > 0$. In addition, it must co-move positively with partner output when $\alpha > 0$ and negatively when $\alpha < 0$. The former arrangement is cooperative; the latter is competitive.*

Given (15), the argument for Proposition 1 is simple. Equation (15) and the strict concavity of the function u imply that an agent's reward must positively co-move with the incentive metric Ψ.

4.5. **Competition and Cooperation.** The degree to which the contract is competitive across agents (at least locally) is fully summarized by the normalized gradient $\bar{\nabla} m \equiv \frac{\nabla m}{\partial m / \partial y_1}$. Equation (15)

[8] The non-negativity constraints on $m(\mathbf{y})$ will be handled by our arguments.

allows us to connect this gradient to the incentive metric $\Psi(\mathbf{y})$. The payment zone within $\mathbf{Y}$ where payments are positive is demarcated by the $\Psi|_{=0}$ manifold. Within this zone, the isoquants for the payment $m(\mathbf{y})$ and statistic $\Psi(\mathbf{y})$ are perfectly aligned, and ∇m is collinear with $\nabla\Psi$:

$$\nabla m = \left[\frac{k}{(|u''(m(\mathbf{y}))|[\Psi(\mathbf{y})]^2}\right] \times \nabla\Psi, \tag{16}$$

where k is constant within the payment zone. This unveils the key principle for contract design with interdependent preferences: *the payment should increase most rapidly in the direction where the incentive metric exhibits the steepest slope.* In turn, that metric is specifically affected by the interdependence coefficient α and the likelihood ratio function λ.

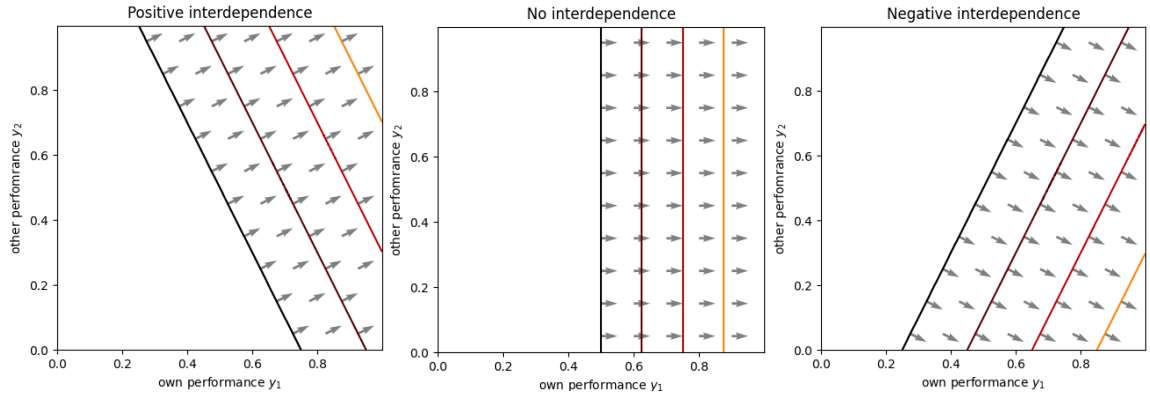

Figure 1. Isoquant maps of the Ψ metric and the coresponding gradient fields for different levels of α. We consider three types of interdependencies, altruistic preferences $\alpha = \frac{1}{2}$ (left panel), independent $\alpha = 0$ (middle) and adversarial $\alpha = -\frac{1}{2}$ (right panel).

Figure 1 illustrates the discussion for two individuals and signal distributions $\mu_i^h((0, y_i]) = y_i$ (with uniform density) and $\mu_i^\ell((0, y_i]) = 2y_i - (y_i)^2$ (with affine downward-sloping density) over the interval $Y = [0, 1]$. Then $\lambda(y) = -1 + 2y$, is uniformly distributed under high effort and the gradient of the incentive metric is given by

$$\nabla\Psi(\mathbf{y}) = [\lambda'(y_1), \alpha\lambda'(y_2)] = 2 \times (1, \alpha) \tag{17}$$

which points northeast (southeast) when α is positive (negative). As a result, the normalized payment gradient is constant everywhere and given by

$$\bar{\nabla} m = [1, \alpha]. \tag{18}$$

For any α, the $\Psi|_{=0}$ isoquant passes through $(y^*, y^*) = (\frac{1}{2}, \frac{1}{2})$. For independent preferences, this line is vertical; see middle panel. In this case, the normalized payment gradient $\bar{\nabla} m$ points due east, and the associated contract is neither competitive nor cooperative.

For positive interdependence, $\bar{\nabla} m$ points northeast, and the payment to agent 1 increases in both y_1 and y_2. So team rewards are naturally incorporated in the contract, though the payment response to own output exceeds the response to partner output. When $\alpha < 0$, the gradient points

southeast. The payment to 1 increases with y_1 and decreases with y_2. This scenario has the feel of a competitive environment with a relative performance component. Figure 2 depicts the optimal contract for $\alpha \in \{-\frac{1}{2}, 0, \frac{1}{2}\}$, likelihood ratio distributed uniformly under μ^h, and isoelastic felicity $u(m) = [m^{1-\theta} - 1]/(1-\theta)$, with $\theta = \frac{1}{2}$.

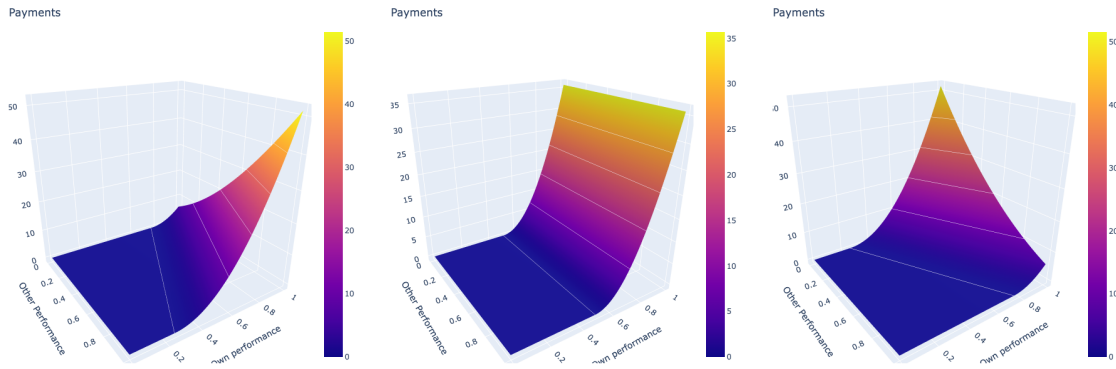

Figure 2. Optimal contract in an example with uniformly distributed likelihood ratio and risk aversion $\theta = \frac{1}{2}$, for $\alpha = \frac{1}{2}$ (left), independent preferences $\alpha = 0$ (middle) and $\alpha = -\frac{1}{2}$ (right).

4.6. **Concentration of Monetary Incentives.** Contracts are also distinguished by the extent to which they concentrate incentives on outputs associated with the highest values of the incentive metric Ψ. Such concentration is primarily driven by the risk attitudes of agents. For instance, with isoelastic felicity $u(m) = [m^{1-\theta} - 1]/(1-\theta)$, optimal payment is $(\max(\Psi/k, 0))^{1/\theta}$, which moves linearly with Ψ in the payment zone when utility is logarithmic ($\theta \simeq 1$). Payments turn increasingly convex with Ψ for risk-tolerant individuals. Figure 3 depicts the contract for $\theta = \frac{1}{10}$, so that agents are highly risk-tolerant. Payments are moderate for the majority of output realizations within the payment zone, sharply increasing only in the vicinity of outcomes that maximize the incentive metric. An altruistic agent receives meaningful rewards only when both achieve peak performance. An adversarial agent is rewarded significantly only when she attains the highest output values while the other agent's performance is very poor.

This concentration of monetary incentives becomes extreme for risk-neutral agents. An optimal contract (when one exists) requires payments to be made exclusively when Ψ attains a maximum, and never otherwise.[9] That is, for altruistic preferences, Ψ is maximized when — and *only* when — everyone produces the highest possible output. When that largest output has positive probability mass, as it would if Y were a finite set, optimal contracts can be shown to exist,[10] but display an

[9]Suppose that the assertion is false, so that the contract pays out over a set E of events of positive probability with Ψ not at its maximum. Create a new contract which removes these payments and shifts their expectation (conditional on E, and high effort) to a set E' with larger values of Ψ. Then the new contract has the same expected payout, but slackens the incentive compatibility constraint (9), because the likelihood ratio is more favorable at E'. That allows the principal to lower expected payment and still guarantee agent compliance.

[10]No optimal solution exists when, say, outputs have densities on some compact interval as in our example. It should be pointed out that non-existence could occur even with strictly concave utility.

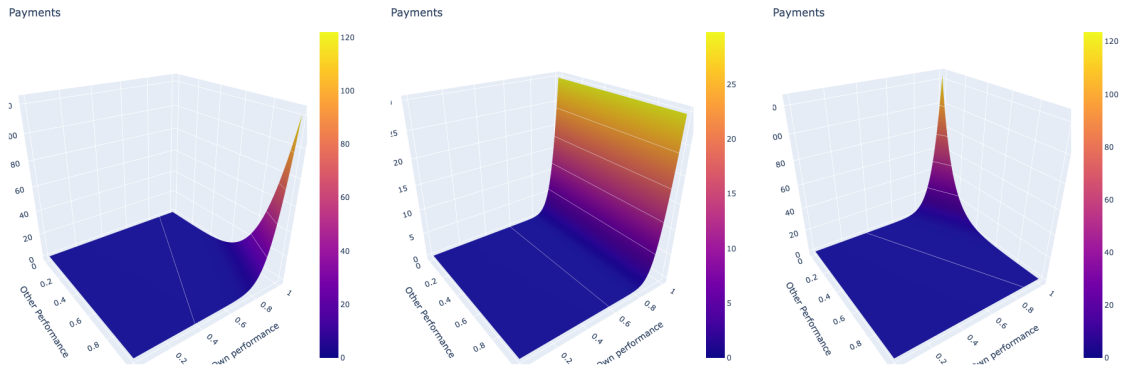

Figure 3. Optimal contract in the example with uniformly distributed likelihood ratio and risk aversion $\theta = \frac{1}{10}$, for $\alpha = \frac{1}{2}$ (left), independent preferences $\alpha = 0$ (middle) and $\alpha = -\frac{1}{2}$ (right).

extreme form of joint liability wherein every team member must *fully* excel for any of them to receive any compensation at all.

With adversarial preferences, Ψ attains its highest value when own output is maximal while for all other agents, output is minimal. The contract becomes an extreme tournament where, in order to be compensated, an agent must outperform *every* other agent by *the largest possible* margin, $\max Y - \min Y$. (Both extremal values of the support need to have positive probability mass to guarantee existence.) It is hard to imagine that contracts that assume such extreme forms are empirically relevant. Nevertheless, our discussion clarifies the role played by smoother payments for risk averse agents; the payments offered outside of the Ψ-extreme events are a form of insurance offered to agents by a risk neutral principal.

5. The Asymmetry of Altruistic and Adversarial Preferences

In the two-agent example introduced in Section 4.5, optimal contracts are symmetric across positive and negative interdependence. Specifically, with $\mathbf{Y} = [0,1]^2$, the promised payment for a given output realization $\mathbf{y} = (y_1, y_2)$ is identical to the payment at $\mathbf{y}' = (y_1, 1 - y_2)$ for interdependence coefficient $\alpha' = -\alpha$.[11] Figure 2 shows that the contracts are "mirrored" relative to each other along the hinge located at $y_1 = \frac{1}{2}$. This symmetry is fundamentally broken with three or more agents, as we shall now see.

Retain the parametric example of Section 4.5, but suppose $n = 3$. Now $\mathbf{Y} = [0,1]^3$, with α restricted to lie in $(-\frac{1}{2}, \frac{1}{2})$ to preserve coherence. We compare scenarios with both positive and negative interdependence coefficients, where the absolute value $|\alpha|$ is the same across the two scenarios. With the affine likelihood function of our example, the payment isoquants are hyperplanes with

[11] $\Psi(y_1, y_2) = \Psi'(y_1, 1 - y_2)$ where Ψ' is derived for $-\alpha$. So $m(y_1, y_2)$ is IC at α if and only if so does the symmetric contract for $-\alpha$. With uniformly distributed λ, the contracts have identical expected payment.

the gradient collinear to

$$\nabla\Psi = 2 \times \left(1, \frac{\alpha}{1-\alpha}, \frac{\alpha}{1-\alpha}\right), \tag{19}$$

and the $\Psi|_{=0}$ isoquant intersects the midpoint realization $(y^*, y^*, y^*) = (\frac{1}{2}, \frac{1}{2}, \frac{1}{2})$.

When $\alpha = \frac{1}{3}$, the relative responsiveness of agent payment to compatriot outputs is characterized by $|\frac{\partial m}{\partial y_j} / \frac{\partial m}{\partial y_i}| = \frac{1}{2}$. However, in the mirror-image case with $\alpha' = -\frac{1}{3}$, this cross-person sensitivity diminishes to $-\frac{1}{4}$. Loosely speaking, t*he principal exploits altruistic motives to a greater degree than she does when preferences are adversarial*. This disparity is magnified in still larger teams. For any arbitrary value of n and a uniformly distributed likelihood under μ^h, the local competitiveness of the contracts can be expressed as:

$$\frac{\partial m}{\partial y_j} / \frac{\partial m}{\partial y_i} = \frac{\alpha}{1-\alpha(n-2)}.$$

In a team of 20 with $\alpha \approx 0.04$, the principal links payment quite strongly to the performance of each compatriot, resulting in $\frac{\partial m}{\partial y_j} / \frac{\partial m}{\partial y_i} \approx 0.142$. A similarly-sized *adversarial* coefficient can only explain a relative responsiveness to others' performance of around -0.023.

This effect turns into a qualitative change at extreme levels of preference interdependence. In the example with three agents and $\alpha \approx \frac{1}{2}$, the payment gradient determined by the incentive metric becomes aligned with the vector $(1, 1, 1)$, so that payment reacts equally to the performance of each agent, *including* own performance. This is a "pure" joint liability mechanism, in which the associated monetary transfer is entirely influenced by the average performance of the agents. This pure joint liability mechanism is optimal, even though each agent is far from being a symmetric utilitarian. They don't need to be: the third-party feedback effects sufficiently magnify their altruistic tendencies.

In sharp contrast, when agents display extreme adversarial preferences ($\alpha \approx -\frac{1}{2}$), third-party interactions attenuate preference interdependence. The limit gradient of payment becomes aligned with $(1, -\frac{1}{3}, -\frac{1}{3})$, which is certainly not the mirror image of its altruistic limit counterpart. The magnitude of payment responsiveness to the performance of others is of the order of $1:3$, notably smaller than the $1:1$ response at the altruistic limit. An agent's payment is now contingent on surpassing a benchmark set at two-thirds of the average compatriot performance. So she could potentially receive compensation even if her performance falls below that of the other two agents.

This qualitative difference between positive and negative inderdependence is general. It extends to all environments with three or more agents and arbitrary distributions. In what follows, we vary α towards both the altruistic or adversarial limit, subject to maintaining the coherence of preference representations; that is, we retain the condition $|\alpha(n-1)| < 1$. For altruistic preferences, we have:

Proposition 2. *In the altruistic limit where $\alpha \uparrow 1/(n-1)$, every agent's contract converges to a common contract that depends on the vector of individual performances but rewards every agent the same amount.*

Specifically, in this limit, the payment to every agent is zero for any $\mathbf{y}$ *such that* $\sum_j \lambda(y_j) \leq 0$, *and otherwise, if* $\sum_j \lambda(y_j) > 0$, *it is given by the common payment*

$$m(\mathbf{y}) = (u')^{-1}\left(k \Big/ \sum_j \lambda(y_j)\right) \tag{20}$$

to every agent when $\sum_j \lambda(y_j) > 0$, *where the constant* k *is chosen so that incentive constraint (9) binds. Under the likelihood condition (1), payments increase in team performance in the sense that they increase vector-wise in* $\mathbf{y}$, *once positive.*

The proof of Proposition 2 follows by passing to the limit in the formula (10) for Ψ. Doing so, we see that

$$\Psi(\mathbf{y}) \to \sum_j \lambda(y_j) \text{ for every vector of outputs } \mathbf{y}, \text{ as } \alpha \uparrow 1/(n-1),$$

Using this observation in (15), we must conclude that every agent receives the same team in the altruistic limit, *irrespective of their own production*. Clearly, (20) follows from (15).

In the altruistic limit, agents not only get a symmetric contract modulo permutations of $\mathbf{y}$, but they literally get the *same contract*, entirely conditioned on team performance. In the particular scenario with uniformly distributed λ the payment is conditioned on the team's overall output, $\sum_j y_j$ and becomes one of pure joint liability.[12]

The adversarial limit is attained as $\alpha \downarrow -1/(n-1)$. Using this in (10), we see that

$$\Psi(\mathbf{y}) \to \lambda(y_i) - \frac{1}{2n-3}\sum_{j\neq i}\lambda(y_j) \text{ for every vector of outputs } \mathbf{y},$$

which yields the following observation:

Proposition 3. *In the adversarial limit where* $\alpha \downarrow -1/(n-1)$, *an agent* i *receives a positive reward if and only if*

$$\lambda(y_i) > \frac{1}{2n-3}\sum_{j\neq i}\lambda(y_j), \tag{21}$$

and in that case her reward is given by

$$m(\mathbf{y}^{[i]}) = (u')^{-1}\left(k' \Big/ \lambda(y_i) - \frac{1}{2n-3}\sum_{j\neq i}\lambda(y_j)\right), \tag{22}$$

for some constant k' *chosen to make the incentive constraint (9) bind. Under condition (1), the payoff to an agent continues to increase in own performance and decrease in the performance of others whenever it is strictly positive, as it will be whenever (22) holds.*

[12] Equation (20) pins down how the common team reward varies over the domain $\mathbf{Y}$. But it leaves the "scaling" of those rewards indeterminate, and dependent on the value of k in equation (20). That value is determined by the need to make the incentive constraint hold with equality.

If Proposition 3 were a perfect antithesis of Proposition 2, it would predict a pure tournament in the adversarial limit, and this is true of the two-agent case (inspect (21) for $n = 2$). But with three or more agents, the tournament is never "pure". An agent will be rewarded for her own performance even when it is not ranked the best. Indeed, she is paid even if her output is below the *average* output of her compatriots. If we define "performance" by the value of $\lambda(y_i)$, we see that agent i gets a positive payment if and only if her performance exceeds $(n-1)/(2n-3)$ of the average performance of her compatriots. ("Performance" in this sense reduces to output when λ is linear.) For $n = 3$, this is only two-thirds the average performance of the others. As n grows large, she is paid if her performance is better than just half the average performance of the others, driving home the asymmetry between the altruistic and adversarial limits.

The literature recognizes that competitive structures are not extensively employed, whereas team-based bonuses are more commonly favored. For instance, Lazear (1989) notes that despite the theoretical support for rank-order tournaments, their use is infrequent. Taken together, our discussion regarding local contract competitiveness as well as Propositions 2 and 3 shed some light on this phenomenon. At the same time, we will see that there is a different force in our setting that pushes towards competition in smaller teams. We take up this observation in Section 8, where we study robust implementation.

6. Participation Constraints

Our analysis can be easily extended to include participation constraints of the form

$$\left[\int_{\mathbf{Y}} u(m(\mathbf{y}^{[i]}))\boldsymbol{d\mu}^{\boldsymbol{h}} - c\right] + \frac{\alpha}{1-\alpha(n-2)}\sum_{j\neq i}\left[\int_{\mathbf{Y}} u(m(\mathbf{y}^{[j]}))\boldsymbol{d\mu}^{\boldsymbol{h}} - c\right] \geq v^0$$

for every agent i, where v^0 is a (renormalized) outside-option utility, common to every agent. By the same "rotational" change of variables that we conducted to arrive at the incentive constraint (9), we can easily restate this participation constraint as

$$\int_{\mathbf{Y}} u(m(\mathbf{y}))\boldsymbol{d\mu}^{\boldsymbol{h}} \geq \frac{v^0[1-\alpha(n-2)]}{1+\alpha} + c \equiv w. \tag{23}$$

How does this additional constraint affect Propositions 1–3? Recall the Lagrangean used in the proof of Proposition 1 and modify it in the obvious way:

$$\mathcal{L}(m) \equiv -\int_{\mathbf{Y}} m(\mathbf{y})\boldsymbol{d\mu}^{\boldsymbol{h}} + \nu^{\mathrm{IC}}\left[\int_{\mathbf{Y}} u(m(\mathbf{y}))\Psi(\mathbf{y})\boldsymbol{d\mu}^{\boldsymbol{h}} - c\right] + \nu^{\mathrm{P}}\left[\int_{\mathbf{Y}} u(m(\mathbf{y}))\boldsymbol{d\mu}^{\boldsymbol{h}} - w\right], \tag{24}$$

where ν^{IC} and ν^{P} are multipliers attached to the incentive constraint (9) and the participation constraint (23) respectively. The first order conditions for this problem are given by

$$-1 + \nu^{\mathrm{IC}}u'(m(\mathbf{y}))\Psi(\mathbf{y}) + \nu^{\mathrm{P}}u'(m(\mathbf{y})) \leq 0, \text{ with equality if } m(\mathbf{y}) > 0,$$

for each $\mathbf{y}$. It can be seen that the incentive constraint continues to bind, so that $\nu^{\text{IC}} > 0$.[13] With this in mind, define $\nu \equiv \nu^{\text{P}}/\nu^{\text{IC}} \geq 0$; then the first order condition above reduces to

$$u'(m(\mathbf{y}))[\Psi(\mathbf{y}) + \nu] \text{ is constant in } \mathbf{y} \text{ if } \Psi(\mathbf{y}) > -\nu, \text{ and } m(\mathbf{y}) = 0 \text{ otherwise,} \tag{25}$$

where we've used $u'(0) = \infty$. Of course, the participation constraint may or may not be binding, so that ν could be positive or zero depending on the parameters of the problem. (In the latter case we're back to the baseline without participation constraints.)

The optimal contract has a common property as we range over the realizations of different $\mathbf{y}$-values. In any situation with $m(\mathbf{y}) > 0$, any change in $\mathbf{y}$ that raises $\Psi(\mathbf{y})$ must create a larger reward for our agent. This gives us an immediate analogue of Proposition 1:

Observation 1 (Extension of Proposition 1)**.** *Assume the likelihood condition (1) and introduce the participation constraint (23). Then Proposition 1 is qualitatively unchanged: at any* $\mathbf{y}$ *with* $m(\mathbf{y}) > 0$*, the reward to an agent must positively co-move with partner output when* $\alpha > 0$ *and negatively co-move when* $\alpha < 0$*. The former arrangement resembles joint liability; the latter, a tournament. Additionally, the local competitiveness of the contract in the payment zone,* $\bar{\nabla} m = \frac{\nabla \Psi}{\partial \Psi / \partial y_1}$*, is unaffected by the outside option.*

With a participation constraint, the boundary for the payment zone is attained on a lower isoquant, $\Psi|_{=-\nu}$, instead of $\Psi|_{=0}$. Consequently, the payment zone expands. As before, within this zone the isoquants for contract $m(\mathbf{y})$ are perfectly aligned with the corresponding curves for Ψ. Therefore, the introduction of an outside option does not alter the presence of cooperation or competition in the contract. It is possible, however, that payment *levels* adjust. In our example, the relevant demarcation lines for the payment zone (corresponding to three values of α in Figure 1) all move westward to a lower $\Psi|_{=-v}$ isoquant, and for any $\mathbf{y}$ the payment levels rise. In line with Observation 1, the local competitiveness of the contract remains unchanged in that example.

7. Payoffs Under Interdependent Preferences

7.1. **Principal Payoffs.** A managerial literature indicates that employers elicit preference interdependence among employees through various policies, which certainly suggests that a principal can benefit from tailoring working conditions and contract pay to such interdependence.[14] Hamilton et al. (2003) present evidence suggesting that productivity is higher under team-based pay compared to individualized performance-based bonuses alone. All that is certainly in line with our model, but we can ask the stronger question: are interdependent preferences *invariably* advantageous to a principal?

[13] Suppose on the contrary that $\nu^{\text{IC}} = 0$, then, because the outside option $v \geq u(0)$ and $c > 0$, we have $m(\mathbf{y}) > 0$ for some subset of $\mathbf{y}$-realizations. That implies $\nu^{\text{P}} > 0$. But then $m(\mathbf{y})$ equals a constant r^* for all $\mathbf{y}$, which reduces (9) to $\int_{\mathbf{Y}} \Psi(\mathbf{y}) d\boldsymbol{\mu}^{\boldsymbol{h}} \geq c/u'(r^*) > 0$, a contradiction, because the left hand side of this inequality is zero as we discussed in the text).

[14] As highlighted by Berman et al. (2002), more than 85% of managers within the United States proactively cultivate workplace friendships by arranging social gatherings for their staff. See also Cohen and Prusak (2002).

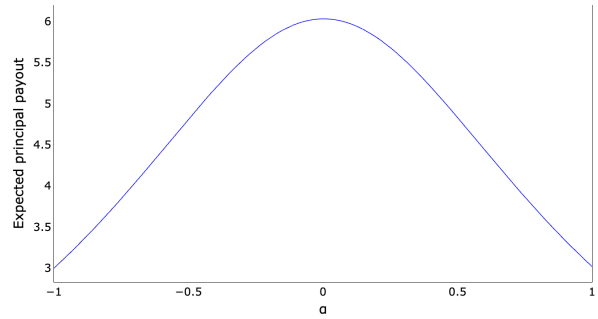

Figure 4. Expected payout by the principal as a function of α in the example with uniformly distributed likelihood ratio and $\theta = \frac{1}{2}$.

If preferences are additive in the utilities of others, it seems intuitive that the principal would weakly prefer *some* interdependence to none at all. The reason is that the principal can always feasibly offer an individualized contract which would precipitate the same effort as in the case of independent preferences. An optimal departure from such a contract to exploit any interdependence cannot hurt the principal. In support of this intuition, Figure 4 uses our example with uniformly distributed likelihood ratio to depict the expected payout by an optimizing principal as a function of α. In both settings the expected payout by the principal decreases as the absolute value of α increases, reaching its maximum at $\alpha = 0$. In particular, expected payment is minimized at the extremes. A principal would appear to thrive under both altruistic and antagonistic conditions in her team.[15]

This is actually a stronger statement than our intuitive assertion. It suggests a *single* local and global peak in payouts at $\alpha = 0$. For more intuition, observe that the willingness of agents to work is influenced by two types of incentives. First, they are motivated by monetary rewards when they achieve success. Second, they are influenced by social incentives that indirectly arise from internalizing the utility of others. With just two agents, the optimal contract for agent 1 strikes a balance between the direct incentives for that agent and the social effect on agent 2. The "social motivator" for agent 2 can be quantified as

$$\int_{\mathbf{Y}} \lambda(y_2)\alpha u(m(\mathbf{y}))\boldsymbol{d\mu^h}, \tag{26}$$

where m is the payment to player 1. This incentivizes 2 only when there is a positive correlation between 2's social experience $\alpha u(m(\mathbf{y}))$, and her own performance $\lambda(y_2)$, so that 2 is materially *and* socially rewarded when she performs. The optimal design ensures this correlation by aligning 1's payment with 2's performance (positively under altruism and negatively under antipathy), ensuring that the social motivator (26) is always positive.

[15] In the example, the effects of positive and negative interdependence are symmetric. They are the same whether the agents like or dislike each other, provided that they experience these attitudes with the same intensity. The principal equally prefers $\alpha = \frac{1}{2}$ and $\alpha = -\frac{1}{2}$ over $\alpha = 0$, and is indifferent between the two non-zero levels. This symmetry emphasizes our point, but is not general.

When this same contract is offered under a higher degree of interdependence α', i.e. $|\alpha'| > |\alpha|$, the incentive value of the social motivator increases. That amplification strengthens the incentives for exerting effort via social interaction. The principal can exploit those heightened incentives, *positive or negative*, thereby reducing her overall monetary payout. It turns out that this finding is general, at least when participation constraints do not bind:

Proposition 4. *Assume that the likelihood ratio condition (1) holds. Then the expected principal payout that implements high effort is decreasing in α when $\alpha > 0$ and increasing in α when $\alpha < 0$, reaching a maximum when $\alpha = 0$.*

Proof. Fix some contract m. For each index j, define a function σ_j by

$$\sigma_j(y_j) \equiv \int_{\mathbf{Y}^{n-1}} u(m(\mathbf{y}))\boldsymbol{d\mu}^{\boldsymbol{h}}_{-j},$$

describing the expected felicity of agent 1 as a function of y_j for any given $j \neq 1$, *after* integrating out over $\mathbf{y}_{-j}$.[16] We claim that if m is an optimal contract, then for every $j \neq 1$,

$$\int_Y \lambda(y_j)\sigma_j(y_j)d\mu^h_j \gtreqless 0 \text{ as } \alpha \gtreqless 0. \tag{27}$$

To establish (27), suppose first that $\alpha > 0$. By Proposition 1, σ_j is nondecreasing for every $j \neq 1$. Moreover, it is strictly increasing for every $y = y_j$ such that $\Psi(y_j, \mathbf{y}_{-j}) > 0$ under a positive $\boldsymbol{\mu}^{\boldsymbol{h}}_{-j}$ measure of $\mathbf{y}_{-j}$ realizations. (In those cases the payment to agent 1 will be sensitive to y_j.) Because these latter conditions always hold for a positive μ^h-measure of y_j, we conclude that σ_j is nondecreasing, and it is *strictly* increasing for a positive μ^h-measure of y_j. Because likelihood ratios integrate to 1, we know that $\int_Y \lambda(y_j)d\mu^h_j = 0$. Combining this observation with that for σ_j, we conclude that condition (27) must hold for $\alpha > 0$. A parallel argument applies when $\alpha < 0$.

To complete the proof, recall from (9) and (10) that the incentive constraint is given by

$$\int_{\mathbf{Y}} \left[\lambda(y_1)u(m(\mathbf{y})) + \frac{\alpha}{1-\alpha(n-2)}\sum_{j\neq 1}\lambda(y_j)u(m(\mathbf{y}))\right]\boldsymbol{d\mu}^{\boldsymbol{h}} \geq c.$$

Because outputs are conditionally independent, we can use σ_j to rewrite this inequality as

$$\int_Y \lambda(y_1)\sigma_1(y_1)d\mu^h_1 + \frac{\alpha}{1-\alpha(n-2)}\sum_{j\neq 1}\int_Y \lambda(y_j)\sigma_j(y_j)d\mu^h_j \geq c. \tag{28}$$

Now suppose that $\alpha \geq 0$, and increase its value to $\alpha' \in (\alpha, 1/(n-1))$. It is obvious that

$$\frac{\alpha'}{1-\alpha'(n-2)} > \frac{\alpha}{1-\alpha(n-2)}.$$

[16] Under an optimal symmetric contract, σ_j is the same over all $j \neq 1$, but we don't use that information here.

Combining this information with (27), so that $\sum_{j\neq 1}\int_Y \lambda(y_j)\sigma_j(y_j)d\mu_j^h > 0$, we must conclude that if m is optimal under α, then

$$\int_Y \lambda(y_1)\sigma_1(y_1)d\mu_1^h + \frac{\alpha'}{1-\alpha'(n-2)}\sum_{j\neq 1}\int_Y \lambda(y_j)\sigma_j(y_j)d\mu_j^h > c, \tag{29}$$

which means that the earlier contract is now "strictly feasible" under α'. Because the incentive constraint is always binding at the optimum, this must mean that the principal can strictly gain by adjusting the contract.

A parallel argument holds when $\alpha < 0$. If $\alpha' < \alpha$ (with $|\alpha'|(n-1) < 1$), then

$$\frac{\alpha'}{1-\alpha'(n-2)} < \frac{\alpha}{1-\alpha(n-2)}.$$

Because (27) holds with negative sign, (29) holds again, completing the proof. □

With binding participation constraints, Proposition 4 may not hold when preferences are antagonistic. Recall the participation constraint (23), which states that

$$\int_{\mathbf{Y}} u(m(\mathbf{y}))\boldsymbol{d\mu^h} \geq \frac{v^0(1-\alpha(n-2))}{1+\alpha} + c \equiv w.$$

and note that if it is binding to begin with, then it fails to hold under the old optimal contract once preferences become "more antagonistic," even as the incentive constraint (9) slackens, as shown in the proof of Proposition 4. Whether or not the principal benefits is then ambiguous. This is not surprising, as workers are entering a more toxic environment and must be suitably compensated for it, even as the principal casts about for new schemes that exploits that heightened toxicity. But there is no such tradeoff when altruism increases. The participation constraint must slacken, as must the incentive constraint. The new incentive-compatible contract must then involve lower expected payouts by the principal. In this sense, a principal might prefer agents linked by altruism, at least when participation constraints are binding. This adds another layer of asymmetry across the altruistic and adversarial scenarios, in addition to the limit arguments provided earlier.

7.2. **Agent Payoffs.** Our discussion also has implications for *agent* payoffs, though here we are on more delicate ground. After all, by changing α, we are also changing the utility function, and so must be careful in interpreting any changes in payoff.[17] But we can entertain these thought experiments by asking whether agents would prefer to form (otherwise identical) teams in groups of friends, strangers or enemies. Under this interpretation, there is no change in preferences as such, but only in the set of team members that the agent is interacting with.

Consider adversarial preferences. We've already noted that the principal will welcome a more toxic environment when participation constraints are not binding, but might be stymied when they are. That suggests that in the *absence* of participation constraints, agents will be hurt by their

[17]Such "cardinal" comparisons are conceptually similar to those in Kreps (1979), Akerlof and Dickens (1982), or Ely and Yilankaya (2001).

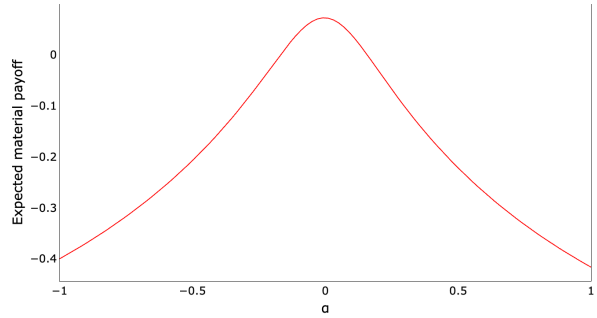

Figure 5. Expected material payoff of an agent as a function of α with uniformly distributed likelihood ratio and $\theta = \frac{1}{10}$.

heightened antagonism. This by itself is not surprising as the intrinsic environment has itself worsened, but the principal will additionally exploit that worsened environment in a detrimental way for the agents. (When participation constraints bind, the principal will be forced to adjust the contracts accordingly, and there will be no net effect.)

The case of altruistic preferences is more intriguing. Consider two altruistic agents with an isoelastic felicity function $u = m^{1-\theta}/(1-\theta)$ and a uniformly distributed likelihood ratio. We examine the case of $\theta = \frac{1}{10}$, so that agents have high risk tolerance. As highlighted in Section 4.6, the optimal mechanism involves approximate joint liability, wherein positive rewards are available only in the proximity of instances where both agents attain maximal outputs. Consider an agent's incentive constraint. If she shirks, she is very unlikely to produce that maximal output. She will not only let herself down, but also her compatriot, whose payoff she values. She is therefore less willing to shirk as her altruism climbs.

Understanding this, the principal can lower the reward to joint success, as that will be enough to satisfy the incentive constraint. In essence, both agents are motivated to work not for monetary rewards but to ensure that their partners do not lose out on compensation despite putting in effort. The equilibrium material payoff of an agent as a function of α is depicted in Figure 5. As the agent becomes sufficiently altruistic (or adversarial), her expected material payoff becomes negative, reflecting the argument above. *They would have been better off in material terms had they not cared for each other*. (With higher risk aversion, these effects are attenuated as the principal needs to offer insurance, resulting in a more evenly distributed material payoff across output realizations.)

8. Robustness Concerns

An older literature in implementation theory (Mookherjee 1984, Ma 1988, Ma, Moore and Turnbull 1988, Mookherjee and Reichelstein 1992, Bergemann and Morris 2009) argues that under an optimal contract, there could be multiple equilibria among agents, with varying payoffs for the principal; see Segal (1999, 2003), Genicot and Ray (2006), Winter (2004), Halac, Kremer and Winter (2020, 2023), or Halac, Lipnowski and Rappoport (2021). A central point in this literature is

that coordination on an equilibrium that favors the principal is by no means guaranteed — especially if there is another equilibrium that favors the agents. We now remark on a problem faced by a designer who fears the least-preferred equilibrium outcome. For simplicity, we ignore the participation constraint, though this can be tagged on in the same way as in Observation 1.

Our incentive constraint (9) requires that an agent weakly prefers working to shirking, assuming that all the other agents are working. However, that may not eliminate effort profiles that involve shirking by two or more agents, and might constitute potentially undesirable equilibria. To ensure robust implementation, we augment the principal's problem by adding constraints that ensure effort from (say) agent 1, for any subgroup of compatriots that might choose to shirk. As is standard in the literature on robust design, we consider only those equilibria satisfying an "indifference refinement" condition, under which workers choose to exert effort when they are indifferent between working and shirking.

8.1. **Robustness Constraint.** We proceed in parallel to our derivation of (9), making use of the assumed symmetry of the contract.[18] We set $n = 2$ and relegate a description of the $n \geq 2$ case to the Appendix. If agent 1 chooses $e_1 = h$, her (renormalized) payoff when agent 2 is a shirker is

$$\begin{aligned} V_1^h &= \int_{\mathbf{Y}} u(m(\mathbf{y}))[1-\lambda(y_2)]d\mu_1^h d\mu_2^h - c + \alpha \int_{\mathbf{Y}} u(m(\mathbf{y}^{[2]}))[1-\lambda(y_2)]d\mu_1^h d\mu_2^h \\ &= \int_{\mathbf{Y}} u(m(\mathbf{y}))[1-\lambda(y_2)]d\mu_1^h d\mu_2^h - c + \alpha \int_{\mathbf{Y}} u(m(\mathbf{y}))[1-\lambda(y_1)]d\mu_1^h d\mu_2^h, \end{aligned} \tag{30}$$

where in going from the first line to the second we've made a change of variables within the second integral, just as we did earlier. Likewise, if our agent chooses $e = \ell$, her expected payoff is given by

$$V_1^\ell = \int_{\mathbf{Y}} u(m(\mathbf{y}))[1-\lambda(y_2)]d\mu_1^\ell d\mu_2^h + \alpha \int_{\mathbf{Y}} u(m(\mathbf{y}))[1-\lambda(y_1)]d\mu_2^\ell d\mu_1^h. \tag{31}$$

Because the principal must implement $e = h$ even when the agent's partner is shirking, the resulting robustness constraint is given by $V_1^h \geq V_1^\ell$, or, using (30) and (31), by:

$$\int_{\mathbf{Y}} u(m(\mathbf{y}))[\Psi(\mathbf{y}) - \Psi^S(\mathbf{y})]\boldsymbol{d\mu^h} \geq c, \tag{32}$$

where $\Psi(\mathbf{y}) \equiv \lambda(y_1) + \alpha\lambda(y_2)$ just as before, and $\Psi^S(\mathbf{y}) \equiv (1+\alpha)\lambda(y_1)\lambda(y_2)$ is an additional term that arises when the partner is believed to be shirking. The left hand side of the old incentive constraint (9) measures the utility gain — not counting cost — of shifting from low to high effort when the partner agent chooses high effort, whereas the the left hand side of the new robustness constraint (32) does the same when the partner chooses low effort. So the additional term

$$S(\alpha) \equiv \int_{\mathbf{Y}} u(m(\mathbf{y}))\Psi^S(\mathbf{y})\boldsymbol{d\mu^h} = (1+\alpha)\int_{\mathbf{Y}} u(m(\mathbf{y}))\lambda(y_1)\lambda(y_2)\boldsymbol{d\mu^h} \tag{33}$$

[18] The assumption of symmetry continues to be strong when studying robust implementation. Asymmetric contracts that generate dominant strategy cascades might be cost effective.

is a measure of the supermodularity built into $m(y_1, y_2)$.[19] The sign of $S(\alpha)$ determines whether the robustness constraint (32) must bind. Specifically, suppose that $S(\alpha) \leq 0$. Then at an optimum, (9) always binds and (32) is slack. To see this, consider our original problem from Section 4.4 in which expected principal payoffs are minimized subject to just (9). Because $S(\alpha) < 0$, (32) must be slack at that optimum of that relaxed problem.

Conversely, if $S(\alpha) > 0$, (32) can never be slack at a robust optimum. For suppose it were slack; then the optimum is again achieved under the relaxed problem described above, with only (9) imposed. It is easy to see that (9) must then bind. But because $S(\alpha) > 0$, (32) cannot then hold at the optimum of the relaxed problem, a contradiction.

We return to these observations below.

8.2. **Optimal Robust Contract.** Cost-effective robust contracts must minimize the expression $2 \times \int_{\mathbf{Y}} m(\mathbf{y}) \boldsymbol{d\mu^h}$ copied from (11), subject to the incentive constraint (9) as well as the robustness constraint (32). As before, the solution to this problem can also be obtained with the Lagrangean method, provided we allow for both constraints. Write the Lagrangean as

$$\mathcal{L}(m) \equiv -\int_{\mathbf{Y}} m(\mathbf{y}) \boldsymbol{d\mu^h} + [\nu^{\text{IC}} + \nu^{\text{S}}] \left[\int_{\mathbf{Y}} u(m(\mathbf{y})) \Psi(\mathbf{y}) \boldsymbol{d\mu^h} - c \right] - \nu^{\text{S}} \int_{\mathbf{Y}} u(m(\mathbf{y})) \Psi^{\text{S}}(\mathbf{y}) \boldsymbol{d\mu^h},$$

where ν^{IC} is the multiplier on the original incentive constraint (9), and ν^{S} is the multiplier on the robustness constraint (32). Both are nonnegative. The first order conditions are given by

$$-1 + \left\{ [\nu^{\text{IC}} + \nu^{\text{S}}] \Psi(\mathbf{y}) - \nu^{\text{S}} \Psi^{\text{S}}(\mathbf{y}) \right\} u'(m(\mathbf{y})) \leq 0, \text{ with equality if } m(\mathbf{y}) > 0.$$

But $m(\mathbf{y})$ *must* be positive for some $\mathbf{y}$ otherwise no incentives can be provided. So it must be that $\nu^{\text{IC}} + \nu^{\text{S}} > 0$,[20] and therefore the first order condition above reduces to

$$u'(m(\mathbf{y}))[\Psi(\mathbf{y}) - \nu \Psi^{\text{S}}(\mathbf{y})] \text{ is constant in } \mathbf{y} \text{ if } \Psi(\mathbf{y}) - \nu \Psi^{\text{S}}(\mathbf{y}) > 0, \text{ and } m(\mathbf{y}) = 0 \text{ otherwise,} \tag{34}$$

where we've used the assumed end-point conditions on u, and defined $\nu \equiv \nu^{\text{S}} / (\nu^{\text{IC}} + \nu^{\text{S}})$. Observe that $\nu \in [0, 1]$. The extreme values of the interval are attained in situations for which only one of the two constraints is binding.

8.3. **Robustness Under Adversarial and Altruistic Interdependence.** As we've observed already, our baseline setting assures us that an agent is happy to exert effort *provided* she believes that her compatriot is doing the same. But then there is the emerging specter of a bad equilibrium — a profile in which all agents shirk — if the optimal contract generates effort supermodularity. In that case, the principal would need to adjust the original contract to achieve robust implementation.[21]

[19] That is, let $\Delta(e_2)$ represent the gain in expected utility for agent 1 when switching from shirking to working, when agent 2 is expected to provide effort e_2. Showing that $S(\alpha) = \Delta(e_2^h) - \Delta(e_2^\ell)$ is straightforward.

[20] Each multiplier is nonnegative, so if we presume that $\nu^{\text{IC}} + \nu^{\text{S}} = 0$, then $\nu^{\text{S}} = 0$ as well, but that contradicts the first order condition whenever $m(\mathbf{y}) > 0$, as it must be for some $\mathbf{y}$.

[21] That does not mean that the original incentive constraint is irrelevant — *both* (9) and (32) could well bite as they do, for instance, in the example.

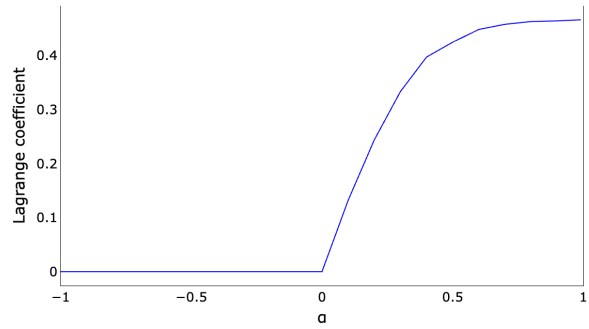

Figure 6. ν as a function of α in the example with uniformly distributed likelihood ratio and $\theta = \frac{1}{2}$.

Figure 6 depicts the Lagrangian coefficient ν as a function of α in the example with uniformly distributed likelihood ratio and CRRA utility with risk parameter $\theta = \frac{1}{2}$. When $\alpha < 0$, the incentive constraint binds in this example, and the robustness constraint remains slack (i.e., $\nu = 0$). Conversely, for $\alpha > 0$, the robustness constraint becomes binding; indeed, $\nu > 0$ and monotonically increases in α. That is, the optimal contract of Section 4.4 is robust to the additional constraint when $\alpha < 0$, but this is no longer true once $\alpha > 0$.

This pattern — supermodularity when agents are altruistic and submodularity when agents are adversarial — transcends the example of Figure 6. The next proposition provides three sets of sufficient conditions. To state it, define $f(z) \equiv u \circ u'^{-1}(1/z)$ for $z > 0$ and $f(0) \equiv 0$ otherwise. Our assumptions on u ensure that f is strictly increasing in z. Later, we interpret $f(\Psi/k)$ as equilibrium utility in the baseline model.

Proposition 5. (a) *There exists a threshold $\epsilon > 0$ such that whenever $|\alpha| < \epsilon$:*

(i) *The optimal contract with adversarial agents automatically satisfies the constraint (39), and is therefore robustly implementable with no change .*

(ii) *The optimal contract for altruistic agents must be adjusted to meet the robustness constraint (39), which is therefore always binding under robust implementation.*

Statements (i) *and* (ii) *also hold without the above restriction on α if*

(b) *The function $f(z)$ is strictly convex.*

(c) *λ is uniformly distributed when an agent exerts effort.*

The proof is in the Appendix, but some aspects of it illuminate well the problem at hand; we highlight those points here. Recall our solution to the baseline problem from (15):

$$u'(m(\mathbf{y}))\Psi(\mathbf{y}) = k \text{ for some } k > 0 \; \mathbf{y} \text{ if } \Psi(\mathbf{y}) > 0, \text{ and } m(\mathbf{y}) = 0 \text{ otherwise.}$$

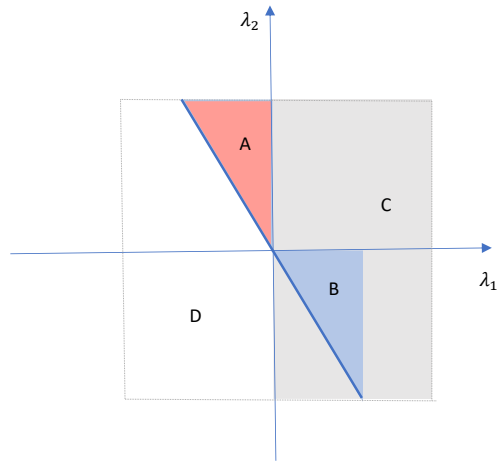

Figure 7. Partition of the domain (λ_1, λ_2) into four subsets. The blue line is the locus $\lambda_1 + \alpha\lambda_2 = 0$, which determines the boundary of the payment zone.

This solution reduces to $m(\mathbf{y}) = u'^{-1}\left(\frac{\Psi(\mathbf{y})}{k}\right)$, and so the level of offered utility becomes

$$f\left(\frac{\Psi(\mathbf{y})}{k}\right) \text{ for all } \mathbf{y}. \tag{35}$$

Using (35) in the expression (33) for supermodularity, we see that

$$S(\alpha) = (1+\alpha)\int_{\mathbf{Y}} f\left(\frac{\Psi(\mathbf{y})}{k}\right)\lambda(y_1)\lambda(y_2)\boldsymbol{d\mu^h} = (1+\alpha)\int_{-\infty}^{1}\underbrace{\left[\int_{-\alpha\lambda_2}^{1} f\left(\frac{\lambda_1+\alpha\lambda_2}{k}\right)\lambda_1 d\gamma^h(\lambda_1)\right]}_{\text{Weighting } W(\lambda_2)}\lambda_2 d\gamma^h(\lambda_2),$$

where we've simplified notation by changing variables to the likelihood values $\lambda_1 = \lambda(y_1)$ and $\lambda_2 = \lambda(y_2)$, noting that these ratios must have support contained in $(-\infty, 1]$, and writing γ^h for the induced distribution of λ under high effort. Part (a) of the Proposition is proved by first observing that $S(0) = 0$,[22] and then showing that $S'(0) > 0$, so that for small positive α, $S(\alpha) > 0$ and for small negative values of α, $S'(\alpha) < 0$.

The marked term $W(\lambda_2)$ in the expression for $S(\alpha)$ is a positive weighting function for the integral of λ_2, where we already know that the *unweighted* integral $\int \lambda_2 d\gamma^h(\lambda_2)$ equals zero. So if $W(\lambda_2)$ is also increasing, then $S(\alpha) > 0$, whereas if it is decreasing, $S(\alpha) < 0$. These properties are exploited for the proof of part (b) of the Proposition, using the assumption that realized utility f in (35) is strictly convex.

The convexity of f, particularly within the CRRA framework, is closely linked to the extent of risk tolerance. In the CRRA class with risk-aversion parameter θ, it is easy to see that $f(z) = z^{(1-\theta)/\theta}/(1-\theta)$. For individuals exhibiting high risk tolerance; specifically, with $\theta < \frac{1}{2}$, the function f is indeed strictly convex.

[22] $S(0) = \int_{-\infty}^{1}\left[\int_{0}^{1} f\left(\frac{\lambda_1}{k}\right)\lambda_1 d\gamma^h(\lambda_1)\right]\lambda_2 d\gamma^h(\lambda_2) = \left[\int_{0}^{1} f\left(\frac{\lambda_1}{k}\right)\lambda_1 d\gamma^h(\lambda_1)\right]\left[\int_{-\infty}^{1}\lambda_2 d\gamma^h(\lambda_2)\right] = 0.$

Finally, part (c) is established by presuming that γ^h is uniform, and then integrating the integrals defining S over the four subsets of the domain (λ_1, λ_2); see Figure 7. When $\alpha > 0$, for instance, the integral over domain D is zero and that over C is positive (Claim 1 in Appendix). Additionally, the absolute value of the negative integral over A is dominated by the positive value of the integral over B (Claim 2 in Appendix), thus ensuring that the integral over the entire domain is positive. For details of the arguments for all three parts of the proof, see the Appendix.

How robust are these patterns described in Proposition 5? In the general model, different information structures typically yield non-uniform distributions of λ over the support $(-\infty, 1]$. Despite this, the optimal contract (under altruism) engenders submodular incentives exclusively in region A, while strictly supermodular incentives are observed in regions B and C. It is important to acknowledge that certain distributions and utility functions exist where set $B \cup C$ does not dominate A and submodularity obtains under altruism.

As an example (see Appendix for details), consider a two-agent setting in which $u(c) = c^{1-\theta}/(1-\theta)$, with $\theta = \frac{9}{10}$, $\alpha = \frac{1}{2}$, and the cost of effort $e = 0.336$. The output distributions concentrate the entire probability mass on two possible output realizations, namely 0 and 1, making them binary with probabilities $\mu^h(\{0\}) = \frac{4}{5}$, and $\mu^\ell(\{0\}) = 1$. In this example, the relaxed optimal contract offered to altruistic agents generates effort *sub*modularity, thereby ensuring robust implementation under the baseline equilibrium. As suggested by part (b) of the Proposition, a specific aspect of the example that contributes to this property is the high degree of risk-aversion. The particular concentrations of the probability mass of λ in region A also play a role, though the binary nature of the signal structure is irrelevant. An analogous example can be constructed for adversarial agents, generating *super*modularity and the consequent non-robustness of the baseline optimal contract even though $\alpha < 0$.

As agent altruism decreases, the probability mass allocated to set A diminishes. Moreover, there exists a uniform upper bound on λ_1 that converges to zero over the shrinking set A as $\alpha \downarrow 0$. It is therefore reasonable to conjecture that for sufficiently small α, the perverse effects of integration over A will eventually be overshadowed by those over the set $B \cup C$. Indeed, that is precisely what part (a) of Proposition 5 demonstrates. So, this example notwithstanding, Proposition 5 is quite unequivocal for scenarios with risk-tolerant agents, uniformly distributed λ, or when preference interdependence is quantitatively small.

8.4. **The Structure of Robust Contracts With Altruism.** In this section, we briefly examine the direction of the contract adjustments that ensures robust implementation. Throughout, we will presume that the likelihood ratio moves smoothly with output, so that λ is differentiable. Drawing upon arguments analogous to those beginning with (34) and further elucidated in Section 4.5, we see that the gradient of the optimal contract with respect to outputs aligns itself perfectly with the augmented gradient $\nabla\Psi - \nu\nabla\Psi^S$. Given that the extent of competitivenessis governed by the slope of this augmented gradient, any alteration in its value relative to standard implementation is driven by the interplay of $\nabla\Psi^S$ and the multiplier ratio $\nu = \nu^S/(\nu^{IC} + \nu^S)$.

In our example with a uniformly distributed likelihood ratio and $\alpha = \theta = \frac{1}{2}$, the supermodularity term Ψ^S achieves extrema at the four corners of its domain. The two maxima (with value $1+\alpha$) are situated in the northeastern and southwestern corners, while the minima (with value $-(1+\alpha)$) are in the northwestern and southeastern corners. So payments in states in which the outputs of the agents exhibit the greatest misalignment yield the most potent submodular incentives. Along the vertical and horizontal lines passing through the center of the domain, Ψ^S equals zero. It also exhibits symmetry around the 45° line.

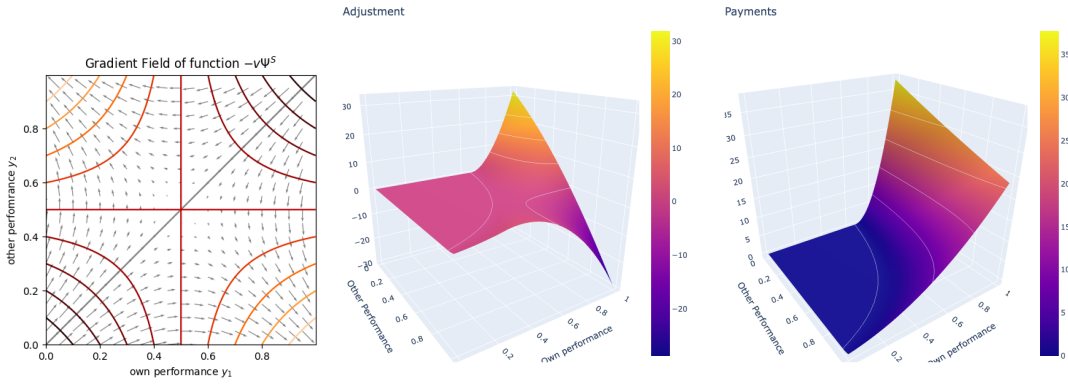

Figure 8. Optimal robust contract in the example with the linear likelihood ratio, risk aversion $\theta = \frac{1}{2}$ and preference interdependence $\alpha = \frac{1}{2}$. For such values, $\nu \approx 0.414$. The figure depicts the gradient field of the the *sub*modular component, $-\nu\Psi^S$ (left panel), the payment adjustment relative to the standard contract (middle panel) and the optimal robust contract (right panel).

In the left panel of Figure 8, we depict the gradient field of $-\nu\Psi^S$, which seamlessly translates into the competitiveness of the contract. For any signal realization falling below the diagonal line, the slope of the vector $\nabla\Psi - \nu\nabla\Psi^S$ is strictly smaller than that of $\nabla\Psi$. In instances where agent 1 outperforms the other, i.e., $y_1 > y_2$, the robust payment moves less positively with the success of the counterpart than it did before, rendering the contract more competitive. Conversely, for all realizations above the diagonal ($y_1 < y_2$), the augmented gradient becomes steeper, signifying a locally more cooperative contract.

The resulting adjustments in payment for altruistic agents and the robust contract are shown in the middle and right panels of Figure 8. The robustness constraint amplifies monetary rewards for individual success while providing a (small) consolation prize for an agent whose output realizations significantly lag behind those of their partner. These dual benefits come, however at the expense of reduced payment in the event when *both* agents succeed. This pattern remains robust in all settings with two agents as long as λ is strictly increasing.

It might appear counterintuitive that (absent a participation constraint) agent 1 receives positive compensation in the vicinity of $(y_1, y_2) = (0, 1)$. This happens even though the negative likelihood $\lambda(0) = -1$, statistically suggests shirking. The puzzle can be solved as follows. In this region, agents' outputs are most misaligned, making these signal realizations particularly potent

in providing submodular incentives, i.e., $\Psi^S(0,1) = (1+\alpha)\lambda(0)\lambda(1) = -\frac{3}{2}$. Because the robustness constraint is binding (with a multiplier value of $\nu = 0.414$), the overall effect of the payment in the neighborhood of the extreme event $(y_1, y_2) = (0,1)$ is beneficial to the principal, when both constraints are taken into account, i.e., $\Psi(0,1) - \nu\Psi^S(0,1) = -\frac{1}{2} + 0.414 \times \frac{3}{2} > 0$. Consequently, for sufficiently high α, the principal may find it optimal to offer a payment that appears as a consolation prize.

8.5. **The Cost of Robust Implementation.** Does the achievement of robust implementation entail additional costs for the principal? In the setting of Proposition 5, the answer is in the negative when agents are adversarial. The robustness constraint remains non-binding for negative values of α. Therefore, robust implementation with adversarial agents comes at no additional cost. However, in the case of altruistic agents, $\nu > 0$, and therefore the robustness implementation of high effort requires the principal to distort the original contracts. Consequently, robust implementation is indeed costly for the principal.

For instance, in the scenario featuring a uniformly distributed likelihood ratio, the expected payment curve ceases to be symmetric around $\alpha = 0$. Figure 9 illustrates the principal's payout as a function of α. One lesson from this exercise, that extends to all environments covered by Proposition 5, is that a principal seeking to robustly implement effort might be inclined to do so with teams of adversaries rather than friends. It is to be noted that such an inclination runs counter to the preference that we earlier described for friends, especially in larger teams. A final assessment on a principal's preferences for cooperative group contracts as opposed to competitive tournaments must therefore rest on these conflicting considerations.

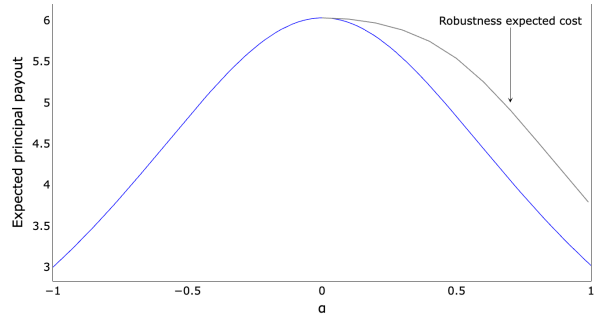

Figure 9. Expected payout by the principal that implement efforts in the unique equilibrium, as a function of α in the example with the linear likelihood ratio and $\theta = \frac{1}{2}$.

9. Future Directions

In this paper we characterized optimal contracts for teams with interdependent preferences. In doing so we make several strong assumptions. For instance, we consider incentive compatibility

conditions assuming that efforts and payments are revealed ex-post to every agent. These conditions are not the only ones that can be studied. A plausible alternative that one could consider is that effort choices are not observed *ex post*, but all payments still are. The second is that neither the effort choices nor the payments to other agents are observed. Our preferred interpretation that we use in the paper is, therefore, not devoid of qualification.

Additionally, the model is static and in particular presumes an exogenous interdependence coefficient that remains unaffected by the choices made by agents or by the contracts offered to them. However, it is important to acknowledge that this assumption may be considered restrictive in certain environments. For example, if an agent deviates by choosing low effort, thereby jeopardizing the prospects of other agents, and if other agents come to know of this deviation, a presumed altruistic value of $\alpha > 0$ *ex ante* may not apply *ex post*.[23] Likewise, it has been observed that in certain competitive environments, initial strangers who interact over time tend to develop adversarial attitudes, while cooperative environments foster friendship. This indicates that the type of contract offered can significantly influence preference linkages among agents in the long term. However, considering these phenomena involves *endogenous* preference interdependence, which falls outside the scope of this paper. Understanding this aspect is crucial for comprehending optimal design, but we defer it to future research.

References

Akerlof, George A and William T Dickens (1982) "The Economic Consequences of Cognitive Dissonance," *The American Economic Review*, 72 (3), 307–319.

Axelrod, Robert and William D Hamilton (1981) "The Evolution of Cooperation," *Science*, 211 (4489), 1390–1396.

Bandiera, Oriana, Iwan Barankay, and Imran Rasul (2005) "Social Preferences and the Response to Incentives: Evidence from Personnel Data," *The Quarterly Journal of Economics*, 120 (3), 917–962.

Banerjee, Abhijit V, Timothy Besley, and Timothy W Guinnane (1994) "Thy Neighbor's Keeper: The Design of a Credit Cooperative with Theory and a Test," *The Quarterly Journal of Economics*, 109 (2), 491–515.

Bergemann, Dirk and Stephen Morris (2009) "Robust Implementation in Direct Mechanisms," *The Review of Economic Studies*, 76 (4), 1175–1204.

Bergstrom, Theodore C (1999) "Systems of Benevolent Utility Functions," *Journal of Public Economic Theory*, 1 (1), 71–100.

Berman, Evan M, Jonathan P West, and Maurice N Richter, Jr (2002) "Workplace Relations: Friendship Patterns and Consequences (According to Managers)," *Public Administration Review*, 62 (2), 217–230.

[23] Anger and a sense of betrayal then enter into the analysis — at least in an extended dynamic model. For related discussions in a team setting, see Ray and Ueda (1996), Section 4.1., and in the specific context of microfinance, Ghatak and Guinnane (1999), Section 3.3.

Bernstein, Shai and Eyal Winter (2012) "Contracting with Heterogeneous Externalities," *American Economic Journal: Microeconomics*, 4 (2), 50–76.

Besley, Timothy and Stephen Coate (1995) "Group Lending, Repayment Incentives and Social Collateral," *Journal of Development Economics*, 46 (1), 1–18.

Camboni, Matteo and Michael Porcellacchia (2023) "Monitoring Team Members: Information Waste and Monitoring Team Members: Information Waste and the Self-Promotion Trap," *working paper*.

Che, Yeon-Koo and Seung-Weon Yoo (2001) "Optimal Incentives for Teams," *American Economic Review*, 91 (3), 525–541.

Cohen, Donald and Lawrence Prusak (2002) "In Good Company: How Social Capital Makes Organizations Work," *Harvard Business Review*, 80, 107–113.

DeMarzo, Peter M and Ron Kaniel (2023) "Contracting in Peer Networks," *The Journal of Finance*, 78 (5), 2725–2778.

Ely, Jeffrey C and Okan Yilankaya (2001) "Nash Equilibrium and the Evolution of Preferences," *Journal of Economic Theory*, 97 (2), 255–272.

Fehr, Ernst and Simon Gächter (2000) "Fairness and Retaliation: The Economics of Reciprocity," *Journal of Economic Perspectives*, 14 (3), 159–182.

Genicot, Garance and Debraj Ray (2006) "Contracts and Externalities: How Things Fall Apart," *Journal of Economic Theory*, 131 (1), 71–100.

Ghatak, Maitreesh (1999) "Group Lending, Local Information and Peer Selection," *Journal of Development Economics*, 60 (1), 27–50.

Ghatak, Maitreesh and Timothy W Guinnane (1999) "The Eonomics of Lending with Joint Liability: Theory and Practice," *Journal of Development Economics*, 60 (1), 195–228.

Green, Jerry R and Nancy L Stokey (1983) "A Comparison of Tournaments and Contracts," *Journal of Political Economy*, 91 (3), 349–364.

Halac, Marina, Ilan Kremer, and Eyal Winter (2020) "Raising Capital from Heterogeneous Investors," *American Economic Review*, 110 (3), 889–921.

_____ (2023) "Monitoring teams," *American Economic Journal: Microeconomics*.

Halac, Marina, Elliot Lipnowski, and Daniel Rappoport (2021) "Rank Uncertainty in Organizations," *American Economic Review*, 111 (3), 757–86.

Hamilton, Barton H, Jack A Nickerson, and Hideo Owan (2003) "Team Incentives and Worker Heterogeneity: An Empirical Analysis of the Impact of Teams on Productivity and Participation," *Journal of Political Economy*, 111 (3), 465–497.

Holmstrom, Bengt (1982) "Moral Hazard in Teams," *The Bell Journal of Economics*, 324–340.

Hori, Hajime and Sadao Kanaya (1989) "Utility Functionals with Nonpaternalistic Intergenerational Altruism," *Journal of Economic Theory*, 49 (2), 241–265.

Itoh, Hideshi (2004) "Moral Hazard and Other-Regarding Preferences," *The Japanese Economic Review*, 55, 18–45.

Jayaraman, Rajshri, Debraj Ray, and Francis De Véricourt (2016) "Anatomy of a Contract Change," *American Economic Review*, 106 (2), 316–358.

Kockesen, Levent, Efe A Ok, and Rajiv Sethi (1997) "Interdependent Preference Formation," *CV Starr WP*, 97–18.

Kreps, David M (1979) "A Representation Theorem for" Preference for Flexibility"," *Econometrica*, 565–577.

Lanzetta, John T and Basil G Englis (1989) "Expectations of Cooperation and Competition and Their Effects on Observers' Vicarious Emotional Responses," *Journal of Personality and Social Psychology*, 56 (4), 543.

Lazear, Edward P (1989) "Pay Equality and Industrial Politics," *Journal of Political Economy*, 97 (3), 561–580.

Lazear, Edward P and Sherwin Rosen (1981) "Rank-order Tournaments as Optimum Labor Contracts," *Journal of Political Economy*, 89 (5), 841–864.

Letina, Igor, Shuo Liu, and Nick Netzer (2020) "Delegating Performance Evaluation," *Theoretical Economics*, 15 (2), 477–509.

List, John (2009) "Social Preferences: Some Thoughts from the Field," *Annual Review of Economics*, 1, 563–583.

Luft, Joan (2016) "Cooperation and Competition Among Employees: Experimental Evidence on the Role of Management Control Systems," *Management Accounting Research*, 31, 75–85.

Ma, Ching-To (1988) "Unique Implementation of Incentive Contracts with Many Agents," *The Review of Economic Studies*, 55 (4), 555–572.

Ma, Ching-To, John Moore, and Stephen Turnbull (1988) "Stopping Agents from "Cheating"," *Journal of Economic Theory*, 46 (2), 355–372.

Marsh, Abigail A (2016) "Neural, Cognitive, and Evolutionary Foundations of Human Altruism," *Wiley Interdisciplinary Reviews: Cognitive Science*, 7 (1), 59–71.

Meyer, Margaret and Dilip Mookherjee (1987) "Incentives, Compensation, and Social Welfare," *The Review of Economic Studies*, 54 (2), 209–226.

Mookherjee, Dilip (1984) "Optimal Incentive Schemes with Many Agents," *The Review of Economic Studies*, 51 (3), 433–446.

Mookherjee, Dilip and Stefan Reichelstein (1992) "Dominant Strategy Implementation of Bayesian Incentive Compatible Allocation Rules," *Journal of Economic Theory*, 56 (2), 378–399.

Nalebuff, Barry J and Joseph E Stiglitz (1983) "Prizes and Incentives: Towards a General Theory of Compensation and Competition," *The Bell Journal of Economics*, 21–43.

Pearce, David G (2008) "Nonpaternalistic Sympathy and the Inefficiency of Consistent Intertemporal Plans," *Foundations in Microeconomic Theory*, 213.

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

Ray, Debraj ⓡ Rajiv Vohra (2020) "Games of Love and Hate," *Journal of Political Economy*, 128 (5), 1789–1825.

Robson, Arthur J (2017) "Group Selection: A Review Essay on Does Altruism Exist?" *Journal of Economic Literature*, 55 (4), 1570–1582.

Segal, Ilya (1999) "Contracting with Externalities," *The Quarterly Journal of Economics*, 114 (2), 337–388.

_____ (2003) "Coordination and Discrimination in Contracting with Externalities: Divide and Conquer?" *Journal of Economic Theory*, 113 (2), 147–181.

Sobel, Joel (2005) "Interdependent Preferences and Reciprocity," *Journal of Economic Literature*, 43 (2), 392–436.

Stiglitz, Joseph E (1990) "Peer Monitoring and Credit Markets," *World Bank Economic Review*, 4 (3), 351–366.

Trivers, Robert (2006) "Reciprocal Altruism: 30 Years Later," *Cooperation in Primates and Humans: Mechanisms and Evolution*, 67–83.

Vásquez, Jorge and Marek Weretka (2021) "Co-worker Altruism and Unemployment," *Games and Economic Behavior*, 130, 224–239.

Wantchekon, L (1994) "Tournaments and Optimal Contracts for Teams in Rural West Africa," *mimeo, Department of Economics, Northwestern University*.

Winter, Eyal (2004) "Incentives and Discrimination," *American Economic Review*, 94 (3), 764–773.

Zillman, Dolf and Joanne R Cantor (1977) "Affective Responses to the Emotions of a Protagonist," *Journal of Experimental Social Psychology*, 13 (2), 155–165.

APPENDIX

Proof of Proposition 5. *Part (a).* Recall that

$$S(\alpha) = (1+\alpha)\int_{\mathbf{Y}} f\left(\frac{\Psi(\mathbf{y})}{k}\right)\lambda(y_1)\lambda(y_2)\boldsymbol{d\mu^h} = (1+\alpha)\int_{-\infty}^{1}\underbrace{\left[\int_{-\alpha\lambda_2}^{1} f\left(\frac{\lambda_1+\alpha\lambda_2}{k}\right)\lambda_1 d\gamma^h(\lambda_1)\right]}_{\text{Weighting } W(\lambda_2)}\lambda_2 d\gamma^h(\lambda_2),$$

where we've simplified notation by changing variables to the likelihood values $\lambda_1 = \lambda(y_1)$ and $\lambda_2 = \lambda(y_2)$, noting that these ratios must have support contained in $(-\infty, 1]$, and writing γ^h for the induced distribution of λ under high effort.

Certainly, $S(0) = 0$ as argued in the main text. We claim that $S'(0) > 0$. Note that[24]

$$\begin{aligned} S'(\alpha) = {} & \frac{S(\alpha)}{1+\alpha} + \frac{1+\alpha}{k}\int_{-\infty}^{1}\left[\int_{-\alpha\lambda_2}^{1} f'\left(\frac{\lambda_1+\alpha\lambda_2}{k}\right)\lambda_1 d\gamma^h(\lambda_1)\right]\lambda_2^2 d\gamma^h(\lambda_2) \\ & - \frac{(1+\alpha)k'(\alpha)}{k^2}\int_{-\infty}^{1}\left[\int_{-\alpha\lambda_2}^{1} f'\left(\frac{\lambda_1+\alpha\lambda_2}{k}\right)\lambda_1^2 d\gamma^h(\lambda_1)\right]\lambda_2 d\gamma^h(\lambda_2) \end{aligned}$$

[24] In the expression that follows, the derivative with respect to α is registered in three locations. The first is at the endpoint of one of the integrals, but this derivative is readily seen to be zero. The second is with respect to the α inside the integral. The third, more implicit, is that k is itself a function of α. These second and third terms are seen in the expression for the derivative.

so that, evaluating this expression at $\alpha = 0$ and noting that $S(0) = 0$, we have:

$$\begin{aligned} S'(0) &= \frac{1}{k}\int_{-\infty}^{1}\left[\int_0^1 f'\left(\frac{\lambda_1}{k}\right))\lambda_1 d\gamma^h(\lambda_1)\right]\lambda_2^2 d\gamma^h(\lambda_2) - \frac{k'(0)}{k^2}\int_{-\infty}^{1}\left[\int_0^1 f'\left(\frac{\lambda_1}{k}\right)\lambda_1^2 d\gamma^h(\lambda_1)\right]\lambda_2 d\gamma^h(\lambda_2) \\ &= \frac{1}{k}\left[\int_0^1 f'\left(\frac{\lambda_1}{k}\right)\lambda_1 d\gamma^h(\lambda_1)\right]\left[\int_{-\infty}^{1}\lambda_2^2 d\gamma^h(\lambda_2)\right] - \frac{k'(0)}{k^2}\left[\int_0^1 f'\left(\frac{\lambda_1}{k}\right)\lambda_1^2 d\gamma^h(\lambda_1)\right]\left[\int_{-\infty}^{1}\lambda_2 d\gamma^h(\lambda_2)\right] \\ &> 0, \end{aligned}$$

where we use the likelihood property that $\int_{-\infty}^{1}\lambda_2 d\gamma^h(\lambda_2) = 0$.

Because $S(0) = 0$, the fact that $S'(0) > 0$ means that there is $\epsilon > 0$ such that $S(\alpha) < 0$ for $\alpha \in (\epsilon, 0)$ and $S(\alpha) > 0$ for $\alpha \in (0, \epsilon)$, which completes part (a) of the proof.

Part (b). For every λ_2, consider the weighting function above, defined as:

$$W(\lambda_2) \equiv \int_{-\alpha\lambda_2}^{1} \underbrace{f\left(\frac{\lambda_1 + \alpha\lambda_2}{k}\right)}_{\text{Weights for } \lambda_1} \lambda_1 d\gamma^h(\lambda_1)$$

This weighting function itself has weights within it as indicated just above. For any given value of λ_2, it is obvious that the weights are first flat at 0 and then increasing in λ_1, so using the fact that $\int \lambda_1 d\gamma^h(\lambda_1) = 0$, it should be clear that $W(\lambda_2) > 0$ for every λ_2.

Differentiation with respect to λ_2 tells us that

$$W'(\lambda_2) \equiv \frac{\alpha}{k}\int_{-\alpha\lambda_2}^{1} f'\left(\frac{\lambda_1 + \alpha\lambda_2}{k}\right)\lambda_1 d\gamma^h(\lambda_1), \tag{36}$$

where we use the fact that $f(0) = 0$. Now, by assumption, f is strictly convex so f' is strictly increasing in z, and therefore strictly increasing in λ_1 (for given λ_2) in (36) above. Using the fact that $\int \lambda_1 d\gamma^h(\lambda_1) = 0$, we must therefore conclude from (36) that $W'(\lambda_2) > 0$ for every λ_2.

Now we return to the formula for $S(\alpha)$, knowing that the weighting function is strictly increasing. Using again the fact that $\int \lambda_2 d\gamma^h(\lambda_2) = 0$, we must conclude that for $\alpha > 0$ we have $S(\alpha) > 0$ and for $\alpha < 0$ we have $S(\alpha) < 0$, as desired.

Part (c). Assume $\alpha > 0$ (the arguments for negative α are symmetric). Under our assumptions for this part and the fact that $\int \lambda d\gamma^h(\lambda) = 0$, the support of λ can be normalized without further loss of generality to $[-1, 1]$. Then $d\gamma^h(\lambda) = \frac{1}{2}$, and the domain of integration is $\Lambda \equiv [-1, 1] \times [-1, 1]$.

Let k be the normalizing constant in optimal contract under standard implementation.

$$\frac{4}{1+\alpha}S(\alpha) = \int_\Lambda f\left(\frac{\lambda_1 + \alpha\lambda_2}{k}\right)\lambda_1\lambda_2 d\lambda_1 d\lambda_2 = \int_\Lambda [f\left(\frac{\lambda_1 + \alpha\lambda_2}{k}\right) - f\left(\frac{\lambda_1}{k}\right)]\lambda_1\lambda_2 d\lambda_1 d\lambda_2,$$

where in the second equality, we utilize the fact that $\int_{[1,-1]} f\left(\frac{\lambda_1}{k}\right)\lambda_2 d\lambda_2 = 0$.

Partition Λ into four subsets A, B, C, D, depicted in Figure 7 in the main text.

Claim 1: The integral over D is zero, and over C it is strictly positive.

On D, where $\lambda_1 + \alpha\lambda_2 < 0$ and $\lambda_1 < 0$, $f\left(\frac{\lambda_1+\alpha\lambda_2}{k}\right) = 0$ and $f\left(\frac{\lambda_1}{k}\right) = 0$, and so $\int_D [f\left(\frac{\lambda_1+\alpha\lambda_2}{k}\right) - f\left(\frac{\lambda_1}{k}\right)]\lambda_1\lambda_2 d\lambda_1 d\lambda_2 = 0$. Next consider C. On this set, $\lambda_1 > 0$, leading to two possibilities. When $\lambda_2 > 0$, one has $f\left(\frac{\lambda_1+\alpha\lambda_2}{k}\right) - f\left(\frac{\lambda_1}{k}\right) > 0$ by monotonicity of f, while for $\lambda_2 < 0$, one has $f\left(\frac{\lambda_1+\alpha\lambda_2}{k}\right) - f\left(\frac{\lambda_1}{k}\right) < 0$. In either case, $[f\left(\frac{\lambda_1+\alpha\lambda_2}{k}\right) - f\left(\frac{\lambda_1}{k}\right)]\lambda_2 > 0$, and so the integral $\int_C [f\left(\frac{\lambda_1+\alpha\lambda_2}{k}\right) - f\left(\frac{\lambda_1}{k}\right)]\lambda_1\lambda_2 d\lambda_1 d\lambda_2$ is positive as well.

Claim 2: The absolute value of the negative integral over A is dominated by the positive value of the integral over B.

Observe that $\omega^A(\lambda_1) \equiv \int_{-\lambda_1/\alpha}^1 [f\left(\frac{\lambda_1+\alpha\lambda_2}{k}\right) - f\left(\frac{\lambda_1}{k}\right)]\lambda_2 d\lambda_2$ defined on region A is increasing in λ_1, and that

$$\int_A [f\left(\frac{\lambda_1 + \alpha\lambda_2}{k}\right) - f\left(\frac{\lambda_1}{k}\right)]\lambda_1\lambda_2 d\lambda_1 d\lambda_2 = \int_{-\alpha}^0 \omega^A(\lambda_1)\lambda_1 d\lambda_1.$$

Because $\int_{-\alpha}^0 [\lambda_1 + \frac{\alpha}{2}] d\lambda_1 = 0$ and $\omega^A(\lambda_1)$ is increasing, we have $\int_{-\alpha}^0 \omega^A(\lambda_1)[\lambda_1 + \frac{\alpha}{2}] d\lambda_1 \geq 0$, and so

$$\int_{-\alpha}^0 \omega^A(\lambda_1)\lambda_1 d\lambda_1 \geq -\frac{\alpha}{2}\int_{-\alpha}^0 .\omega^A(\lambda_1) d\lambda_1 \tag{37}$$

Similarly, $\omega^B(\lambda_1) \equiv \int_{-\lambda_1/\alpha}^0 [f\left(\frac{\lambda_1+\alpha\lambda_2}{k}\right) - f\left(\frac{\lambda_1}{k}\right)]\lambda_2 d\lambda_2$ defined on region B is increasing in λ_1, and $\int_B [f\left(\frac{\lambda_1+\alpha\lambda_2}{k}\right) - f\left(\frac{\lambda_1}{k}\right)]\lambda_1\lambda_2 d\lambda_1 d\lambda_2 = \int_0^\alpha \omega^B(\lambda_1)\lambda_1 d\lambda_1$. Because $\int_0^\alpha [\lambda_1 - \frac{\alpha}{2}] d\lambda_1 = 0$, it must be that $\int_0^\alpha \omega^B(\lambda_1)[\lambda_1 - \frac{\alpha}{2}] d\lambda_1 \geq 0$, and so

$$\int_0^\alpha \omega^B(\lambda_1)\lambda_1 d\lambda_1 \geq \frac{\alpha}{2}\int_0^\alpha \omega^B(\lambda_1) d\lambda_1 \tag{38}$$

Finally, observe that $\omega^A(\alpha + \lambda_1) = \omega^B(\lambda_1)$, so that $\int_{-\alpha}^0 \omega^A(\lambda_1) d\lambda_1 = \int_0^\alpha \omega^B(\lambda_1) d\lambda_1$. Combining this information with (37) and (38),

$$\int_{-\alpha}^0 \omega^A(\lambda_1)\lambda_1 d\lambda_1 + \int_0^\alpha \omega^B(\lambda_1)\lambda_1 d\lambda_1 \geq 0.$$

We can now combine Claims 1 and 2 to conclude that

$$S(\alpha) = \frac{1+\alpha}{4}\int_{A\cup B\cup C\cup D} [f\left(\frac{\lambda_1 + \alpha\lambda_2}{k}\right) - f\left(\frac{\lambda_1}{k}\right)]\lambda_1\lambda_2 d\lambda_1 d\lambda_2 > 0.$$

Robust Implementation With $n \geq 3$ Agents. If agent 1 chooses $e_1 = h$, her (renormalized) payoff with a set of shirkers $S \subseteq \{2, \ldots, n\}$ is

$$V_1^h = \left[\int_{\mathbf{Y}} u(m(\mathbf{y}))\prod_{s\in S}[1 - \lambda(y_s)]\boldsymbol{d\mu^h} - c\right] + \frac{\alpha}{1-\alpha(n-2)}\sum_{j\neq i}\left[\int_{\mathbf{Y}} u(m(\mathbf{y}^{[j]})\prod_{s\in S}[1-\lambda(y_s)]\boldsymbol{d\mu^h} - c|N-S|\right],$$

(λ_1, λ_2)-values	$(1,1)$	$(1,-1/4)$	$(-1/4,1)$	$(-1/4,-1/4)$
Probability	$1/25$	$4/25$	$4/25$	$16/25$
$\Psi = \lambda_1 + \alpha\lambda_2$	$3/2$	$7/8$	$1/4$	$-3/8$
$m = (\min\{\Psi, 0\}/k)^{1/\theta}$	1.569	0.862	0.214	0
$f\left(\frac{\lambda_1+\alpha\lambda_2}{k}\right)$	1.046	0.985	0.8572	0
$f\left(\frac{\lambda_1+\alpha\lambda_2}{k}\right) - f\left(\frac{\lambda_1}{k}\right)$	0.046	- 0.015	0.8572	0
$[f\left(\frac{\lambda_1+\alpha\lambda_2}{k}\right) - f\left(\frac{\lambda_1}{k}\right)]\lambda_1\lambda_2$	0.05	0.004	-0.21	0

Table 1. Information for the Example in Section 8.2.

where, as before, $\mathbf{y}^{[j]}$ is the "rotated vector" with y_j as its first entry. Likewise, if our agent chooses $e = \ell$, her expected payoff is given by

$$V_1^\ell = \left[\int_{\mathbf{Y}} u(m(\mathbf{y})) \prod_{s\in S}[1-\lambda(y_s)] d\mu_1^\ell \boldsymbol{d\mu}_{-1}^{\boldsymbol{h}}\right] + \frac{\alpha}{1-\alpha(n-2)} \sum_{j\neq i}\left[\int_{\mathbf{Y}} u(m(\mathbf{y}^{[j]})) \prod_{s\in S}[1-\lambda(y_s)] d\mu_1^\ell \boldsymbol{d\mu}_{-1}^{\boldsymbol{h}} - c|N-S|\right].$$

Remembering that the principal wishes to implement $e = h$, the incentive constraint is given by $V_i^h \geq V_i^\ell$, or, using the two equations above:

$$\left[\int_{\mathbf{Y}} u(m(\mathbf{y}))\lambda(y_1) \prod_{s\in S}[1-\lambda(y_s)] \boldsymbol{d\mu}^{\boldsymbol{h}}\right] + \frac{\alpha}{1-\alpha(n-2)} \sum_{j\neq i}\left[\int_{\mathbf{Y}} u(m(\mathbf{y}^{[j]}))\lambda(y_1) \prod_{s\in S}[1-\lambda(y_s)] \boldsymbol{d\mu}^{\boldsymbol{h}}\right] \geq c.$$

We now conduct a change of variables within the second integral, as we did before. For each $j \neq 1$, we "rotate" the entries in $\mathbf{y}^{[j]}$ so that j is replaced by 1, with all other indices rotated accordingly, including the index $i = 1$. As before, this will cause the index 1 to range over all the values $\{2, \ldots, n\}$ as different j's are replaced, while the set S will be replaced by an appropriately permuted set of indices; call it $S^{[j]}$. Remembering that $\mathbf{y}^{[1]}$ is just $\mathbf{y}$, we obtain an equivalent representation of the incentive constraint as

$$\int_{\mathbf{Y}} u(m(\mathbf{y}))\Phi(\mathbf{y}) \boldsymbol{d\mu}^{\boldsymbol{h}} \geq c, \tag{39}$$

where in an analogous manner to Ψ in (10), we've defined

$$\Phi(\mathbf{y}) \equiv \lambda(y_1) \prod_{k\in S}[1-\lambda(y_s)] + \frac{\alpha}{1-\alpha(n-2)} \sum_{j\neq 1} \lambda(y_j) \prod_{k\in S^{[j]}}[1-\lambda(y_s)]. \tag{40}$$

Cost-effective robust contracts must therefore minimize the expression

$$n \int_{\mathbf{Y}} m(\mathbf{y}) \boldsymbol{d\mu}^{\boldsymbol{h}} \tag{41}$$

copied from (11), subject to the set of robust incentive constraints (39) as we range over all sets $S \subseteq \{2, \ldots, n\}$. Note that when S equals the empty set, we obtain the old incentive constraint (9). Now there are more constraints in addition to (9), including a need to generate work incentives even when everyone else is shirking, captured by $S = \{2, \ldots, n\}$.

Details for the Example in Section 8.2. Note that under high effort, λ takes on two distinct values: $\lambda(0) = -\frac{1}{4}$ and $\lambda(1) = 1$, with probabilities $\frac{4}{5}$ and $\frac{1}{5}$ respectively. That yields four potential outcomes within the Λ-space of a two-agent team, as illustrated in Figure 10. Two

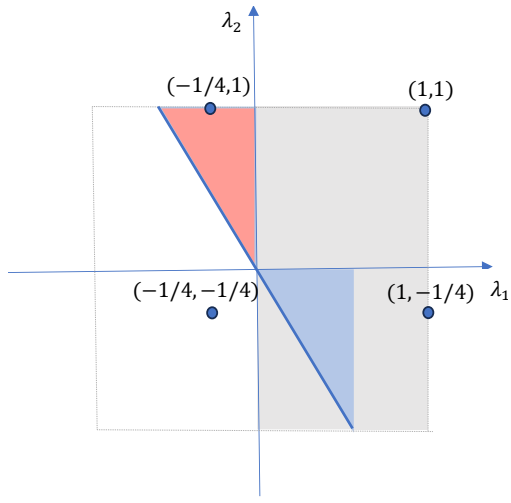

Figure 10. The four possible realizations of (λ_1, λ_2).

of these outcomes fall within set C, thus making positive contributions to the supermodularity function $S(\alpha)$, whereas the outcome in set A decreases its value (there is one more in D with zero contribution). Table 1 presents the probabilities, payments, payoffs, and the resulting contributions to $S(\alpha)$ for each of the four potential realizations. Although both outcomes in set C yield higher utility $f\left(\frac{\lambda_1+\alpha\lambda_2}{k}\right)$ compared to the one in set A, owing to the strict concavity of the realized utility in Ψ, their absolute values centered around $f\left(\frac{\lambda_1}{k}\right)$, are smaller. Integrating the term $[f\left(\frac{\lambda_1+\alpha\lambda_2}{k}\right) - f\left(\frac{\lambda_1}{k}\right)]\lambda_1\lambda_2$ over the three relevant realizations results in the negative value $S(\alpha) = -0.03$ when $\alpha = 0.5$.