Sharing the Fame but Taking the Blame: When Declaring a Single Person Responsible Solves the Reputation Free Rider Problem

Xinyu Li* Wendelin Schnedler†

January 23, 2023

Abstract

Teams are formed because input from different people is needed. According to received wisdom, providing incentives to team members by declaring them all responsible fails because real responsibility for team output 'diffuses'. But why? And why and when does *formally* declaring one member 'responsible' mean that this member can be attributed *real* responsibility?

We offer a model that answers these questions. We identify when declaring a team jointly responsible results in reputation free-riding. We show that formally assigning responsibility to one person can overcome this problem but only if all other team members are protected from being sanctioned.

JEL: M54, D23, D86, L23, K12, K13

Keywords: reputation free riding, collective punishment, social capital, formal and real responsibility

*PBL Netherlands Environmental Assessment Agency, xinyu.li@pbl.nl

†Faculty of Economics and Business Administration, Paderborn University, wendelin.schnedler@upb.de

We like to thank Sylvain Chassang, Florian Engl, Urs Fischbacher, Andreas Diekmann, Bentley MacLeod, Gerd Mühlheusser, Nora Szech, Björn Bartling, Simeon Schudy, Georg Weizsäcker, Claus-Jochen Haake, Eberhard Feess, Florian Kerzenmacher and seminar participants at the University of Groningen and from the Organisational Economics Group of the VfS for their valuable comments.

> Great leaders give credit to others and accept the blame themselves.
>
> John Wooden

> And a lean, silent figure slowly fades into the gathering darkness, aware at last that in this world, with great power there must also come great responsibility!
>
> from Amazing Fantasy #15 Spider-Man

Introduction

For organizations and firms, motivating teams is difficult because members can free ride on each other's efforts. Management scholars emphasize two factors that help teams to overcome these difficulties. Firstly, team members employ their 'social capital' and help each other. For example, Katzenbach and Smith (1995) conclude in their widely quoted work on teams that a 'high degree of personal commitment to one another differentiates people on high performing teams from people on other teams'. Secondly, only one member should be given responsibility for the output because 'assigning responsibility to teams of people can mean that no one takes responsibility for anything' (Wilson, 1999, p. 182)[1]—this member should then take all blame in case of failure but share all fame in case of success.[2]

This conventional wisdom conflicts with traditional incentive theory. In Holmström's seminal principal-agent model (1982), an outsider to the team,

[1] Schwaber and Sutherland (2012) insist that 'a person not a committee' should 'remain accountable' and Melissa Valentine claims that performance in flash teams is higher if they have a 'directly responsible individual' ('Reinventing the Way we Work', June 27, 2016 by Edmond L. Andrews, accessed January 9, 2021).

[2] Tobias Fredberg's devotes a Harvard Business Review Article to 'Why Good Leaders Pass the Credit and Take the Blame'. Simon Sinek's claims at Live2Lead 'when things go right, you have to give away all the credit and when things go wrong you have to take all the responsibility', (video min. 3:51). Both were accessed on May 27, 2022.

the principal, solves a free rider problem among its members, the agents, by sanctioning the whole team whenever a pre-specified target is not met (see Holmström's Theorem 2). This suggests that declaring a team jointly responsible helps to overcome free riding. So, in what sense does joint responsibility create a free rider problem? And why is 'social capital' and selecting one individual to 'take all blame but share all fame' crucial to overcome this problem?

This paper is (to our knowledge) the first to offer an incentive theoretical model that answers these questions. First, we need to explain why Holmström's solution does not work and free riding arises although the principal commands sufficient resources to provide incentives to all agents. For this, we introduce into Holmström's model that the principal is unwilling (or unable) to sanction members who are not responsible for failure; very much in line with the fundamental right in various legal traditions that no person should be 'punished for an offence that he or she has not committed' (See, for example, Article 33 of the Geneva Convention, August 12, 1949).[3] Following legal practice, we regard an agent to be really responsible if he can be 'reasonably expected' to contribute to team success but has caused failure.[4] (We formalize these 'reasonable expectations' using the concept of Perfect Bayesian Equilibrium and follow Lewis (1974) who considers a choice to be causal if an outcome would not have occurred 'but for' that choice—see our Definitions 1 and 2.)

An implicit assumption for Holmström's solution to work is that team members can coordinate their actions. We render this assumption explicit by assuming that agents decide sequentially under perfect knowledge. The principal is fully aware that agents are capable of coordinating their choices.

[3]Not punishing 'innocent' members can also be derived from fairness preferences if the principal cares about intent (Chassang and Zehnder, 2016).

[4]In U.S. tort law, negligence is 'based upon a failure to comply with the duty of care of a reasonable person, which failure is the actual cause and proximate cause of damages. That is, but for the tortfeasor's act or omission, the damages to the plaintiff would not have been incurred, and the damages were a reasonably foreseeable consequence of the tortious conduct'— See Wikipedia article on U.S. tort law, accessed 17th of July 2020.

She does not, however, know who did what within the team and infers this from incentives (by ruling out that agents play strictly dominated strategies).

Given these assumptions, a principal who cares enough about responsibility is unwilling to collectively punish (Proposition 1). The reason is the following. Since agents can perfectly coordinate, only one agent is really responsible (Lemma 2). The principal, however, does not know which. In the words of de Voltaire (1747), the principal prefers 'sparing the guilty to condemning the innocent'; just as a majority of subjects in a large field experiment by Cappelen et al. (2018).

Being unable to punish collectively, a principal who cares about real responsibility, or more succinctly, a caring principal, cannot employ Holmström's solution to the free riding problem. Since all members' contributions are required for team success but providing them all with incentives is impossible, willingness to help each other (social capital) becomes necessary to counter free riding (Proposition 2)—just as suggested by management scholars.

Social capital, however, is not sufficient for overcoming free riding. In a committed team that is jointly declared responsible, responsibility diffuses: assigning real responsibility to any member becomes impossible, a caring principal is then limited by her unwillingness to use incentives rather than her lack of resources. As a result, agents ride free (Theorem 1).

A priori, it is not clear why formally assigning responsibility to one agent solves this free rider problem. How can the principal be sure that an agent will *later* cause failure when declaring this agent responsible *before* the team starts working? Declaring one agent responsible neither shifts the balance of power between him and other agents nor does it alter the production technology. It also does not change who can observe what about agents' contributions and therefore does not affect the scope for formal or informal agreements. Finally, free riding occurs here although agents face no coordination problem. So, the declaration does not help with coordination, either.

Nevertheless, the management wisdom about declaring one member responsible can be rationalized for committed teams. The crucial condition is that no other agent can be sanctioned—irrespective of what he has done (Theorem 2). As long as the bonus from another agent may be withdrawn, the declaration is cheap talk. The formally responsible agent can 'pass the blame' by claiming that he thought the other agent will be sanctioned. This, however, is no longer possible if only the formally declared agent can be sanctioned. Then, real follows from formal responsibility and even a caring principal sanctions in good conscience.

Interestingly, having to pay bonuses even in the case of failure does not restrict a caring principal; she would have paid anyway (Proposition 3). Instead, it enables her to overcome reputational free riding by focusing blame on the formally responsible agent. Since contributions of all team members are required for success, the same agent also has to share all fame (Corollary 4); echoing the notion that one person should 'take all blame but share all fame.'

The management wisdom breaks down when only one agent has sufficient social capital. With the greater power of this agent to elicit others' contributions directly comes real responsibility—irrespective of who has been formally declared responsible (Corollary 5).

1 Model

We build on the canonical team incentive model by Holmström (1982) in which a principal P (she) employs agents (he), $i \in N = \{1, \ldots, n\}$ with $n \geq 2$, who then produce some joint output. We extend this model in two ways. Firstly, agent i may be able to draw on the support of others to a degree $\gamma_i \in \mathbb{R}_{\geq 0}$, where $\gamma_i = 0$, brings us back to the original model. Secondly, the principal suffers from sanctioning an agent who is not really responsible, where $\kappa \in \mathbb{R}_{\geq 0}$, describes the degree to which the principal cares. Holmström's assumption that the principal does not care is represented by $\kappa = 0$.

Strategies

The principal P (she) initially proposes a project consisting of an outcome target $\hat{y}$, a bonus $\hat{b}_i$ promised to each agent i, and a declaration for each agent i whether he is formally responsible, $\hat{r}_i = 1$, or not, $\hat{r}_i = 0$.

For reference, we call the game that begins after the principal's proposal and which is characterized by $(\hat{y}, \widehat{\mathbf{b}}, \widehat{\mathbf{r}})$ *team game,* where $\hat{\mathbf{b}} = (\hat{b}_1, \ldots, \hat{b}_n)$ and $\hat{\mathbf{r}} = (\hat{r}_1, \ldots, \hat{r}_n)$. In this game, any agent i can leave the project at any time before his contribution, in which case he incurs arbitrarily small costs of $\epsilon > 0$. (This is our way of resolving indifference with respect to participation.)

In order to avoid any coordination problems within the team that might hamper success, we assume that agents decide sequentially whether to take charge. As long as no one has taken charge yet, agent k can take charge, contribute $c_k > 0$ and request contributions $\hat{c}_l \in \mathbb{R}_{\geq 0}$ from all other agents $l \in N \backslash \{k\}$, who then decide on their actual contributions $c_l \geq 0$ in some pre-determined sequence. Alternatively, agent k can decide not to take charge.

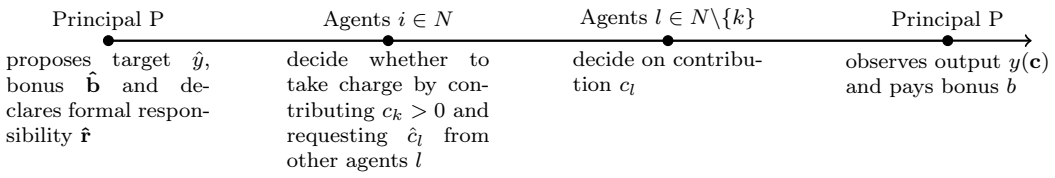
For notational completeness, all requests are initially set to zero, $\hat{c}_i = 0$ for all i and remain at this level unless they are altered by an agent who takes charge. If no agent has taken charge, agents decide in some pre-determined sequence how much to contribute c_i. Members can observe the current state of the project, i.e., agents have full information on all contributions made before them. Importantly, the order in which agents have the opportunity to take charge is unknown to the principal. Being ignorant, she considers every sequence equally likely. We use the notation $\hat{\mathbf{c}} = (\hat{c}_1, \ldots, \hat{c}_n)$, for the requests to contribute and $\mathbf{c} = (c_1, \ldots, c_n)$ for the actual contributions.

The sequential nature of contributions reflects that members of a team typically have ample means of communication and need not fear that their contributions are wasted—unlike in the context of the 'volunteers dilemma', where players move simultaneously and 'free ride' because of this fear (See, e.g., Diekmann, 1985; Harrington, 2001; Krueger and Massey, 2009; Sliwka,

2006). If anything, the ability to coordinate and take initiative simplifies the problem faced by agents. Nevertheless, we will later observe free riding.

Joint team output y is a function that is continuous and increases in the contributions of all agents, $y : \mathbb{R}^n_{\geq 0} \to \mathbb{R}_{\geq 0}$.. The reason for the team to exist is that contributions from all are necessary to produce output: $y(c) = 0$ if $c_i = 0$ for some agent i. The specific functional form does not matter as long as the production is concave enough such that $y(c) - \sum_i c_i$ has a unique interior maximizer $c^{\text{FB}} = (c_1^{\text{FB}}, \ldots, c_n^{\text{FB}})$, with $c_i^{\text{FB}} > 0$.[5] This maximizer later turns out to be the first-best contribution (see Lemma 4 in the Appendix), hence the name. The team's output generates a benefit for an organization or firm that is represented by the principal. This output cannot be traded on a market, which is why the team was created by the organization and the output was not bought in.

Figure 1: Sequence of Moves



After output y has been produced, the principal can decide on the actual bonuses $\mathbf{b} = (b_1, \ldots, b_n)$— see Figure 1. Hidden action models customarily assume that the principal can credibly commit to pay the promised bonus $\hat{b}_i$ in case of success. Put differently, institutions (courts, self-enforcing agreements, etc) ensure $b_i \geq \hat{b}_i$ if $y \geq \hat{y}$ and the principal can only take away the bonus in case of failure. After failure, institutions may differ in whether the agent has to be warned that his bonus is at stake or not. We capture this by distinguishing between institutions that do not restrict the principal after failure ($\omega = 0$) and those which require that an agent is formally declared responsible, $\hat{r}_i = 1$, for his bonus to be taken away ($\omega = 1$). This distinction later turns out to be

[5] One example, for $n = 2$ would be $y(c) \equiv \sqrt[4]{c_1 \cdot c_2}$.

crucial for our main result.

Summarizing, the principal has to pay the promised bonus $\hat{b}_i$ in case of success but also in case of failure to an agent i who is not formally responsible ($\hat{r}_i = 0$) within an institution that protects his bonus ($\omega = 1$):

$$b_i \geq \hat{b}_i \cdot \begin{cases} 1 & \text{if } y \geq \hat{y} \\ (1 - \hat{r}_i) \cdot \omega & \text{if } y < \hat{y} \end{cases} \tag{1}$$

The restrictions imposed on the principal are lower bounds on the bonus. She is, of course, free to pay more if she wants to.

Payoffs of Agents

Contributing c_i to the team outcome is costly for agents. Apart from being motivated by a bonus, members may contribute because they have been asked to help. They may do so because of peer pressure (Kandel and Lazear, 1992), gift exchange, or team norms, which might be sustained by repeated interactions (Kandori, 1992; Itoh, 1992, 1993; Che and Yoo, 2001). Since we are not interested in the origins of this willingness but its consequences, we simply assume that team member i feels 'pain' or some other cost if he does not answer a request $\hat{c}_i$ of a team colleague k who has taken charge. Agent i's 'pain' is larger the higher k's ability to elicit help, γ_k, and the higher k's commitment as measured by his contribution c_k. For simplicity, let agent i's costs from not meeting the request $\hat{c}_i$ amount to $\gamma_k c_k$. In Holmström's model, agents cannot elicit contributions from other team members; this can be captured here by setting $\gamma_k = 0$.

We also refer to agent k's 'ability to elicit help', γ_k, as his 'social capital'. Our attempt here is not to contribute to the debate on the many ways in which social capital can be defined (See, for example Robison et al., 2002). Instead, we use the term to describe an aspect typically associated with social capital, namely, the ability to obtain help from others.

Together with the enjoyment of getting a bonus b_i and the costs of contributing c_i, agent i's utility becomes:

$$u_i(b_i, c_i, \hat{c}_i) = b_i - c_i - \begin{cases} \gamma_k c_k & \text{for } c_i < \hat{c}_i \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

The last cost term drops out for an agent k who has taken charge because he has not received any request, $\hat{c}_k = 0$, so that $c_k \geq \hat{c}_k$. If no agent has taken charge, we have $\hat{c}_i = 0$ for all i, and the cost term is also eliminated.

Causality and real responsibility

Modern legislation and jurisdiction emphasize the importance of individual responsibility for punishments. The Geneva Convention, for example, states that no person should be "punished for an offence that he or she has not committed"[6] In line with this view, we want to allow for the possibility that the principal cares about who is actually responsible for the production outcome when deciding whether and how much to sanction agents. By 'actually responsible', we mean that the agent has caused the outcome in the sense that his choices made a difference to the outcome (Lewis, 1974).[7] We thus need to predict (on and off the equilibrium path) what would have happened had player $i \in \{P\} \cup N$ acted differently at some point that was actually reached in the game. Since this prediction also depends on other players' behavior and since the principal does not know $\hat{c}$ or c, we use the Perfect Bayesian Equilibrium (PBE).

We have to deal with the problem that the player may not fully determine the outcome, i.e., multiple PBE may follow a player's decision. If, for example,

[6] Article 33 on Individual responsibility, collective penalties, pillage, reprisals of Convention (IV) relative to the Protection of Civilian Persons in Time of War. Geneva, 12 August 1949.

[7] Lewis writes (p. 557,1974): "We think of a cause as something that makes a difference, and the difference it makes must be a difference from what would have happened without it."

one team member does not call the others to arrange a meeting, another member might still do so. Some choices, however, prevent certain sets of outcomes Y. Losing the key for the meeting room, for example, renders it impossible for the meeting to take place in this room, which may mean that the team cannot be successful. The player who loses the key in this example causes the outcome 'failure' because his choice is not to be expected to result in the outcome 'success'.

Definition 1 (Causality)**.** *Let outcome Y be a set of outputs $y \in Y \subset \mathbb{R}_{\geq 0}$. Player i then causes outcome Y by making a choice at some decision node such that no PBE consistent with this choice results in a different outcome $y \in \mathbb{R}_{\geq 0} \backslash Y$.*

Experimental evidence suggests that causing an outcome matters to whether subjects are considered responsible or regard themselves to be responsible. Bartling et al. (2015) find that pivotal voters, who by definition rule out one policy, are assigned more responsibility for passing policies that are unfair to other experimental subjects. Falk et al. (2020) show that believing to be pivotal is related to the willingness of killing a mouse. Feess et al. (2020) observe that pivotality affects whether subjects vote for taking money from a charity. Lübbecke and Schnedler (2020) provide evidence that individuals are willing to pay for having 'authored' an attractive outcome in the sense that they personally have excluded failure.

Just because a player could have made a difference to the outcome, the player is not necessarily attributed responsibility for it. For example, Olympic skier Noah Hoffman has argued that sponsors should do more to boycott the winter Olympics in Beijing because individual athletes cannot be expected to risk their career.[8] If causing success is prohibitively costly for a player, she can hardly be held responsible. We formalize this idea consistently with how we predict future behavior. A player can only be 'expected' to produce success

[8] In the BBC program 'business daily' (2nd of February 2022, min 7.50).

if doing so is an equilibrium strategy.[9]

Definition 2 (Real responsibility)**.** *Let $\hat{y}$ be some output target. Player i is really responsible for the outcome, $Y = \{y \geq \hat{y}\}$ or $Y = \{y < \hat{y}\}$, whenever she causes this outcome by a choice at a decision node at which some PBE results in success, $Y = \{y \geq \hat{y}\}$.*

Let us denote with $\sigma_i \in \{0, 1\}$ whether player i is really responsible for success, $\sigma_i = 1$, or not $\sigma_i = 0$. In case of failure, $Y = \{y < \hat{y}\}$, we use $\phi_i \in \{0, 1\}$ to capture whether player i is really responsible for it, $\phi_i = 1$, or not $\phi_i = 0$. With this notation, a player may be responsible in case of success, $\sigma_i = 1$, while she is not responsible in case of failure, $\phi_i = 0$. In principle, it is thus possible that player i has a 'claim to fame', $\sigma_i = 1$, but cannot be blamed, $\phi_i = 0$.

Payoffs for the principal

We are now in the position to write down the principal's utility. The principal benefits from output y and has costs from providing bonuses b_i to the agents. As in Holmström (1982), she may not care about real responsibility, which is represented here by $\kappa = 0$. She may, however, also suffer $\kappa > 0$ from wrongly 'punishing' agent i, where $\mathbb{1}_{[0,\hat{b}_i)}(b_i) = 1$ indicates that the principal at least partially withheld the promised bonus, $b_i < \hat{b}_i$ and $\mathbb{1}_{[0,\hat{b}_i)}(b_i) = 0$ that she did not. Sanctioning agent i feels wrong to the principal either if the target has been reached and agent i is responsible, $\sigma_i = 1$, or the target has not been reached but agent i is not responsible, $\phi_i = 0$.

Taken together, we get the following utility function for the principal:

$$u_P(y, \hat{y}, b, \hat{b}) = y - \sum_{i \in N} b_i - \kappa \sum_{i \in N} \mathbb{1}_{[0,\hat{b}_i)}(b_i) \left(\sigma_i + 1 - \phi_i\right). \tag{3}$$

With this utility function, a principal who cares enough about real responsibility

[9] For a responsibility criterion that considers how far a player is from being pivotal, see Engl (2018).

later turns out to be unwilling to sanction an innocent agent—see Lemma 1.[10]

If the principal does not know whether an agent i is really responsible for failure, she forms beliefs, $P(\phi_i = 1)$, which can be interpreted as i's reputation. In line with standard textbook practice (Gibbons, 1997, p. 237), the principal believes that deviations do not come from agents with a strictly dominant strategy. In our context, this will imply that an agent who clearly loses out from causing failure is not attributed real responsibility for it.

Utility functions of the principal and the agents are common knowledge. In particular, everyone knows the social capital γ_j of agent j and how much the principal cares about real responsiblity κ.

2 Analysis

This section has two aims. First, we want to shed light on why and when Holmström's proposed incentive scheme does not work and a new free rider problem emerges although the principal has sufficient resources to reward and punish agents. Second, we want to understand why and when assigning responsibility to one agent can overcome this problem.

2.1 Holmström's bonus scheme

In an ideal (first-best) world, where principal and agents could commit to payments and contributions, agents contribute c^{FB} and produce output $y^{\text{FB}} = y(c^{\text{FB}})$—see Lemma 4 in the appendix. In a world without such commitment and in which the principal does not condition bonus payments on output, no output is produced in equilibrium—see Lemma 6 in the appendix. The reason is the fundamental externality at the heart of every incentive problem: the contributions benefit the principal[11] but the respective costs are incurred by

[10] Perhaps more realistically, the principal's pain may increase in the relative size of the punishment $\frac{\hat{b}_i - b_i}{\hat{b}_i}$. This, however, would complicate notation without affecting our results.

[11] Recall that the benefit y is generated within the organization and cannot be traded. Otherwise, it might be possible for one agent to 'buy the benefit', commit to contribute and

the agents. The externality has to be internalized at least partially for some output to be produced, where full internalization would generate the first-best output.

Holmström (1982) famously suggested to solve the incentive problem by conditioning bonuses on joint output. In our setting, Holmström's bonus scheme can be represented as follows. The principal sets the first-best output as a target $\hat{y} = y^{\text{FB}}$, promises a bonus that compensates for contributions $\hat{b}_i = c_i^{\text{FB}}$, and then pays out the bonus to all agents whenever the target is met $y \geq \hat{y}$ and nothing to any agent, otherwise.

In order for this scheme to be feasible if the institution does not allow for punishing agents who are not formally responsible ($\omega = 1$), the principal has to declare the team jointly responsible—see Lemma 7 in the Appendix. Declaring all agents responsible, which is seen as the root of a free rider problem by management scholars, is actually a necessary condition to implement Holmström's solution to this problem.

This solution works in our model if the principal does not care about real responsibility ($\kappa = 0$)—see Corollary 6 in the Appendix, which directly follows from Holmström's Theorem 2. If we want to explain, why assigning formal responsibility to all agents leads to free riding, we thus need to consider a principal for whom real responsibility matters.

2.2 Limits of collective punishment

A principal who is interested in real responsibility ($\kappa > 0$) faces a dilemma. Paying a promised bonus $\hat{b}_i$ is costly but so is withholding it from an 'innocent' agent. If the principal cares enough about real responsibility or is pretty convinced that an agent is innocent, i.e., not responsible for failure, she will pay the bonus even after failure.

Lemma 1. *A principal with $\kappa > 0$ does not withdraw the bonus of team*

get the other agents to contribute.

member i after failure, $y < \hat{y}$, if she deems it unlikely that i is really responsible for this failure: $b_i \geq \hat{b}_i \Leftrightarrow P(\phi_i = 1) \leq 1 - \frac{\hat{b}_i}{\kappa}$.[12]

Proof. Recall the principal's utility:

$$u_P(y, \hat{y}, b_i, \hat{b}_i) = y - \sum_i b_i - \kappa \sum_i \mathbb{1}_{[0,\hat{b}_i)}(b_i) \left(\sigma_i + 1 - \phi_i\right).$$

In case of failure, we have $\sigma_i = 0$ and the principal pays agent i the lowest bonus meeting the promise, $b_i = \hat{b}_i$, if and only if $y - \hat{b}_i \geq y - \kappa(1 - \phi_i)$, or equivalently, $\hat{b}_i \leq \kappa(1 - \phi_i)$. If the principal does not know ϕ_i, she uses her beliefs to compute the expected utility and we find that paying the bonus does not reduce the principal's utility if and only if $\hat{b}_i \leq \kappa(1 - P(\phi_i = 1))$. Solving for $P(\phi_i = 1)$ yields the above inequality. □

From the lemma, we can conclude that a principal who cares about responsibility is only willing to employ Holmström's bonus scheme if she is sufficiently sure that all agents are really responsible for failure. On the other hand, achieving the target is prohibitively costly for other members after one player has caused failure. They are thus not really responsible, but the first player is.

Lemma 2 (Unique responsible agent)**.** *Suppose $y^* \geq \hat{y}$ is supported by some PBE. Then, a single player is really responsible for failure, $y < \hat{y}$: $\sum_i \phi_i = 1$.*

Proof. The proof works by contradiction. Suppose several players are really responsible: $\sum_i \phi_i > 1$, say, for example, player i and j. By the definition of real responsibility, Definition 2, these players must have caused failure. Without loss of generality, let player i be the first to have caused $y < \hat{y} \leq y^*$. By the definition of causality, Definition 1, there is hence no PBE consistent with i's choice that results in success. This means that success, $y^* \geq \hat{y}$ can no longer be achieved at the moment, where player j is causing failure and player j cannot be really responsible for failure by Definition 2. This in turn

[12] If the principal knows ϕ_i with certainty, the inequality still describes when the principal pays the bonus with the degenerate probability distribution $P(\phi_i = 1) \in \{0, 1\}$.

contradicts the assumption that player i and j both hold real responsibility for failure. □

Given that only one agent can be really responsible, a principal who sufficiently 'cares about real responsibility' finds it hard to collectively punish.

Proposition 1 (No collective punishment)**.** *In the team game* $(\hat{y}, \widehat{\mathbf{b}}, \widehat{\mathbf{r}})$*, a principal does not withdraw the bonuses of all team members after failure,* $y < \hat{y}$*, if she cares enough about responsibility:* $\kappa > \frac{\sum_i \hat{b}_i}{n-1}$.

Proof. P is only willing to withdraw the bonus from all agents, if she makes no loss from doing so. This is the case whenever $\kappa \cdot (1 - \phi_i) \leq \hat{b}_i$ for all i—see Proof of Lemma 1. Adding the inequalities for all agents, we get $\kappa \cdot \sum_i (1 - \phi_i) \leq \sum_i \hat{b}_i$ or $\kappa(n - \sum_i \phi_i) \leq \sum_i \hat{b}_i$. Using that by Lemma 2, $\sum_i \phi_i = 1$, we get the following necessary condition for collective punishment:

$$\kappa(n-1) \leq \sum_i \hat{b}_i. \tag{4}$$

□

For large values of κ, the principal thus adheres to the maxim that 'sparing the guilty' is better than 'to condemn the innocent' (de Voltaire, 1747) or that 'it is better that ten guilty persons escape, than that one innocent suffers' (Blackstone, 1765-1770, p. 358).

The proposition offers a possible explanation why some companies pay bonuses even after failure.[13] An immediate consequence of the proposition is that a sufficiently caring principal is unwilling to employ Holmström's scheme.

Corollary 1. *In team game* $(\hat{y}, \widehat{\mathbf{b}}, \widehat{\mathbf{r}})$*, a principal who cares sufficiently about real responsibility,* $\kappa(n-1) > \sum_i \hat{b}_i$*, will not employ Holmström's scheme.*

[13]For an example, see Anthony Mason's report on 'Why Failing Companies Still Pay Bonuses', March 18, 2009, accessed on 30th of June, 2022.

Proof. The proof follows immediately from observing that Holmström's scheme requires collective punishment, which a sufficiently caring principal is unwilling to engage in by Proposition 1. □

While a sufficiently caring principal cannot rely on Holmström's scheme, there may be other ways to implement the first-best. The next section shows that any such way requires social capital but also that social capital is not sufficient.

2.3 Reputation free riding and diffusion of responsibility

Since contributions of all agents are required for success and collective punishment is not viable for a principal who cares about real responsibility, one agent has to take charge and asks others to contribute.

Lemma 3 (One agent needs to take charge)**.** *In team game* $(\hat{y}, \widehat{\mathbf{b}}, \widehat{\mathbf{r}})$*, with a principal who cares sufficiently about real responsibility,* $\kappa > \frac{\sum_i \hat{b}_i}{n-1}$*, a positive output* $\hat{y} > 0$ *can only be achieved in equilibrium if some agent* k *takes charge and requests help* $\hat{c}_l > 0$ *from all other agents* l*.*

Proof. The proof works by contradiction. Suppose no agent takes charge. Then, $\hat{c}_i = 0$ for all i and the last term in the agents' utility (2) drops out and any arbitrary agent i is only willing to contribute if he is sanctioned in case of failure, $b_i = 0$ whenever $y < \hat{y}$. By assumption, all agents need to contribute for success: $y > 0 \Rightarrow c_i > 0$ for all i, so all agents need to be sanctioned in case of failure. Using that $\kappa > \frac{\sum_i \hat{b}_i}{n-1}$ in Proposition 1, the principal is unwilling to collectively punish, the required incentives cannot be provided, no agent contributes, and $y = 0$. □

Asking for help alone, does not mean that it is forthcoming. We want to characterize when an agent has enough social capital for getting in the contributions $\tilde{\mathbf{c}}$ that are needed to produce the desired output, $y(\tilde{c}) \geq \hat{y}$.

Definition 3. *Suppose success can be produced using contributions* $\tilde{\mathbf{c}}$*:* $y(\tilde{c}) \geq \hat{y}$*. Then, agent* i*'s social capital suffices to produce success* $\hat{y}$ *using contributions* $\tilde{\mathbf{c}}$ *whenever* $\gamma_i > \frac{\tilde{c}_j}{\tilde{c}_i}$ *for all* j*.*

If none of the agents has sufficient capital to implement success using whatever contributions, no request for contributions will be met. Then, none of the agents can ensure success and hence become really responsible for causing failure. In the case of lacking social capital, a principal who cares about real responsibility thus faces a free rider problem.

Proposition 2 (Free Rider Problem in Absence of Social Capital)**.** *Consider a team game* $(\hat{y}, \widehat{\mathbf{b}}, \widehat{\mathbf{r}})$*, with a principal who cares sufficiently about real responsibility,* $\kappa > \frac{\sum_i \hat{b}_i}{n-1}$*. If no agent* i *has sufficient social capital to produce success* $y(\tilde{c}) \geq \hat{y} > 0$ *for any contributions* $\tilde{c}$*, real responsibility diffuses* $\phi_i = 0$ *and no positive output can be produced.*

Proof. By Lemma 3, some agent k needs to take charge for any positive output $\hat{y} > 0$ to be a achieved in the team game. By Lemma 5 in the Appendix, a necessary condition for agent l to contribute positively after k's request is that $\hat{c}_l < \gamma_k c_k$. Since for every $\tilde{c}$ with $y(\tilde{c}) > \hat{y}$ and any agent k, $\tilde{c}_l > \gamma_k c_k$, no request $\hat{c}_l$ for a contribution $\tilde{c}_l$ that would ensure success is ever met and $c_l = 0$ for all l, again by Lemma 5. Accordingly, success cannot be sustained in any equilibrium, which implies by Definition 2 that no agent is really responsible. □

The proposition shows us that social capital is necessary to avoid free riding. After dealing with the case of no social capital, we now turn to the polar case of a 'high degree of personal commitment to one another' by Katzenbach and Smith (1995).

Definition 4 (Committed Team)**.** *Suppose success can be produced using contributions* $\tilde{\mathbf{c}}$*:* $y(\tilde{c}) \geq \hat{y}$*. Then, a team is committed to success* $\hat{y}$ *if for all agents* $i \in N$ *social capital suffices to implement success using contributions* $\tilde{\mathbf{c}}$*.*

In a committed team, any agent may generate success. This, however, does not mean that any agent wants to generate success. If agents are jointly responsible for team output, they can free-ride on each others' reputation in such a way that real responsibility diffuses entirely. The intuition is the following. The first agent with the chance to take charge is not yet causing failure because another agent may take charge later. This argument holds for all but the last agent. Therefore, failure can only be caused by the last agent. In order to be really responsible, this last agent must also have an incentive to take charge. His bonus must hinge on output. Since the principal cares about real responsibility, she is unwilling to withdraw the bonus from anyone but this agent. As an outsider to the team, however, she does not know who this agent is. As a result, all agents will keep their bonus. This in turn means that no agent can be expected to contribute. They free ride. Moreover, their reputation does not even suffer because none of them is expected to contribute and hence really responsible.

Theorem 1 (Reputation Free Riding and the Diffusion of Responsibility)**.** *Consider a team game* $(\hat{y}, \widehat{\mathbf{b}}, \widehat{\mathbf{r}})$ *with a positive target* $\hat{y} > 0$*, and a team that is committed to success* $\hat{y}$ *and jointly declared responsible (*$\hat{r}_i = 1$ *for all* i*). Then, a sufficiently caring principal will not consider any agent really responsible for failing to implement* $\hat{y}$*, agents slack, and* $\hat{y}$ *cannot be produced:*

$$\kappa > \frac{1}{1-\frac{1}{n}} \cdot \max_i \hat{b}_i \quad \Rightarrow \quad \phi_i^* = 0, c_i^* = 0, \text{ for all } i \text{ and } y^* = 0 < \hat{y}.$$

Proof. The proof works by contradiction. Suppose that an equilibrium with $\mathbf{c}^*$ such that $y^* = y(\mathbf{c}^*) \geq \hat{y} > 0$ exists. Then, $\kappa > \frac{1}{1-\frac{1}{n}} \cdot \max_i \hat{b}_i$ implies:

$$n\kappa > n\frac{1}{1-\frac{1}{n}} \cdot \max_i \hat{b}_i = \frac{1}{1-\frac{1}{n}} \cdot n \cdot \max_i \hat{b}_i \geq \frac{1}{1-\frac{1}{n}} \sum_i \hat{b}_i, \tag{5}$$

where the last inequality follows from $\max_i \hat{b}_i \geq \hat{b}_i$ for all i. Dividing (5) by n yields: $\kappa > \frac{1}{n}\frac{1}{1-\frac{1}{n}} \sum_i \hat{b}_i = \frac{\sum_i \hat{b}_i}{n-1}$, which directly implies that one agent l has to

take charge by Lemma 3 in the Appendix.

Observe that for a positive production level, all agents' participation constraints must be met in equilibrium, $c_k^* \leq \hat{b}_k$ for $k \in N$. Moreover, since the team is committed to success $\hat{y}$, $\gamma_k > \frac{c_l^*}{c_k^*}$ for all k and l. By Lemma 5 in the appendix, any agent k can thus request $\hat{c}_l = c_l^*$ from agents $l \in N \backslash \{k\}$ and this request is met favorably if $c_k = c_k^*$: $c_l = \hat{c}_l = c_l^*$. Any agent k can thus ensure $y(\tilde{\mathbf{c}}) \geq \hat{y} > 0$. Finally, any agent k can be sanctioned in case of failure because $\hat{r}_k = 1$ for all k :

$$b_k \geq 0 \text{ if } y < \hat{y}. \tag{6}$$

Suppose that in equilibrium positive output $y^* = y(\mathbf{c}^*) > \hat{y} > 0$ is produced because agent k contributes c_k^* and takes charge and requests contributions $\hat{c}_l = c_l^*$, which are then forthcoming. Then, there is also an equilibrium in which the last agent with the opportunity to take charge $\underline{k}$, elicits $\hat{c}_l = c_l^*$ and contributes $c_{\underline{k}}^*$. This, however, implies that no other agent but agent $\underline{k}$, can cause failure by not taking charge; if any other agent $k \neq \underline{k}$, does not take charge, there still is an equilibrium in which the target is reached because the last agent $k = \underline{k}$ takes charge.

If positive output can be produced, responsibility for failure can thus be attributed to agent $\underline{k}$ for not appropriately taking charge, i.e., contributing $c_{\underline{k}} < c_{\underline{k}}^*$ or not requesting $\hat{c}_l = c_l^*$, or to any other agent l for not meeting the request, i.e., contributing $c_l < \hat{c}_l^*$. Meeting the request, however, is a strictly dominant strategy by Lemma 5. Since the principal does not believe that agents deviate from a strictly dominant strategy, she can only attribute real responsibility to the last agent with the opportunity to take charge. Since the principal cannot identify agent $\underline{k}$, and since all agents are equally likely to be $\underline{k}$, the probability that some pre-determined agent i whose bonus is withdrawn in case of failure is agent $\underline{k}$ is $\frac{1}{n}$. Since agent $\underline{k}$ is only really responsible if a respective equilibrium can be constructed, the probability that

an arbitrary agent i is really responsible is restricted: $P(\phi_i = 1) \leq \frac{1}{n}$. This implies $\frac{1}{1-\frac{1}{n}}\hat{b}_i \geq \frac{\hat{b}_i}{1-P(\phi_i=1)}$ and since $\kappa > \frac{1}{1-\frac{1}{n}}\hat{b}_i$, it follows that $\kappa > \frac{1}{1-P(\phi_i=1)}\hat{b}_i$ and by Lemma 1, the principal is unwilling to sanction any agent i. As a result, no agent has an incentive to ask for help or contribute, which would be necessary for any positive output by Lemma 6 in the appendix.

Assuming the existence of an equilibrium that implements $y^* = \hat{y} > 0$ has thus led to $y^* = 0 < \hat{y}$, which is a contradiction. □

The theorem reflects the management wisdom that joint formal responsibility leads to free riding.

When agents are committed, the reputation free riding problem emerges because multiple agents may successfully take charge and would do so if they are sanctioned in case of failure. For the argument it was important that the principal can take away the bonus of any agent after failure. The reason why the principal can take away the bonus is immaterial. In the theorem, this was possible because she declared all agents responsible. The same, however, also happens if the organisation does not protect agents' bonuses $\omega = 0$, which leads to the following corollary.

Corollary 2. *Consider a team game $(\hat{y}, \widehat{\mathbf{b}}, \widehat{\mathbf{r}})$ with a positive target $\hat{y} > 0$, and a team that is committed enough to generate success using some contributions $\tilde{\mathbf{c}}$ in which any member can be sanctioned ($\omega = 0$). Then, a sufficiently caring principal will not consider any agent really responsible for failing to implement $\hat{y}$, agents slack, and $\hat{y}$ cannot be produced:*

$$\kappa > \frac{1}{1-\frac{1}{n}} \max_i \hat{b}_i : \phi_i = 0, c_i = 0, \text{ for all } i \text{ and } y^* = 0 < \hat{y}.$$

Proof. The proof is analogous to that of Theorem 1, where inequality (6) follows from constraint (1) on the principal's bonus payments because $\omega = 0$ rather than $r_i = 1$. □

The preceding sequence of results are, to our knowledge, the first to formally

identify conditions for the claim by social psychologists that punishment for individual members of failing groups is 'non-existent': agents must be unable to rely on each other (Proposition 2) or be very committed to each other and either jointly responsible for output (Theorem 1) or vulnerable to losing out in case of failure (Corollary 2).

In the context of the volunteer's dilemma, free riding problems have been modeled as the result of coordination failure (Diekmann, 1985; Harrington, 2001; Krueger and Massey, 2009; Sliwka, 2006). This may suggest that the free riding problems in this section can also be viewed as coordination failure between agents. In our game, however, where agents have perfect information about each others' decisions and move sequentially, coordination problems can be ruled out as reasons for free riding.

This leaves us with a puzzle. If declaring one team member responsible does not solve a coordination problem how can it prevent free riding? The declaration does not shift any decision rights, does not alter the production technology, does not give the respective agent more power, etc. At best, the declaration prevents the caring principal from doing something that she is not keen to do anyway: taking away the bonus in case of failure from agents who are not formally responsible. This subtle shift is indeed key for overcoming free riding.

2.4 When formal translates into real responsibility

If only one agent is assigned formal responsibility and the other agents cannot be sanctioned, it becomes impossible for the formally responsible agent to free ride on the others' reputation and 'pass on' real responsibility. Other agents can simply not be expected to take charge because they have no incentive to do so; their bonus is guaranteed. This guarantee is crucial for the formal assignment to work. If all agents may lose their bonus, we are back in the situation where the formal assignment is cheap talk.

Theorem 2 (Preventing reputation free riding)**.** *Consider a team game* $(\hat{y}, \widehat{\mathbf{b}}, \widehat{\mathbf{r}})$ *with a positive target* $\hat{y} > 0$ *such that for some adequately compensated contributions* $\tilde{\mathbf{c}}$ *with* $\hat{b}_i \geq \tilde{c}_i$ *the target can be met,* $y(\tilde{\mathbf{c}}) > \hat{y}$*, a team committed to success* $\hat{y}$*, and only one agent* k *being formally responsible,* $r_k = 1$ *and* $r_l = 0$ *for all* $l \neq k$*, and a principal who sufficiently cares about real responsibility:* $\kappa > \frac{1}{1-\frac{1}{n}} \max_i \hat{b}_i$*. Then, success can be implemented if and only if agents who are not formally responsible cannot be sanctioned,* $\omega = 1$*.*

Proof. First examine the case $\omega = 1$ and consider the following PBE candidate. Agent k with $\hat{r}_k = 1$ takes charge by contributing $c_k^* = \tilde{c}_k$ and asking all agents l to contribute $\hat{c}_l^* = \tilde{c}_l$. Agents l contribute $c_l^* = \hat{c}_l$ if $c_k = \tilde{c}_k$ and nothing otherwise $c_l^* = 0$. The principal believes that failure is caused by agent k only, $P(\phi_k = 1) = 1$ and $P(\phi_l = 1) = 0$, always pays any agent $l \neq k$ the promised bonus, $b_l^* = \hat{b}_l$ while agent k only gets the bonus in case of success $b_k^* = \hat{b}_k$, $y \geq \hat{y}$ and $b_k^* = 0$ otherwise.

Given $P(\phi_l = 1) = 0$ for all $l \neq k$, paying the bonus to agent l is optimal for the principal. On the other hand, the principal maximizes her utility by withdrawing agent k's bonus if and only if there is failure because $P(\phi_k = 1) = 1$.

Since agent k loses $\hat{b}_k \geq \tilde{c}_k$ when not requesting $\hat{c}_l = \tilde{c}_l$ or not contributing $\tilde{c}_k$, agent k has no reason to deviate from this behavior. Agent l's optimal response then follows from Lemma 5 in the Appendix.

Finally, we need to check whether real responsibility is correctly attributed and matches behavior. Due to $\omega = 1$, $b_l = \hat{b}_l$ for all agents l, agent l only finds it optimal to contribute $c_l > 0$ as response to a request by agent k. Consequently, l can only be really responsible for failure if his contribution is below the request $c_l < \hat{c}_l$. Meeting the request, however, is a strictly dominant strategy by Lemma 5. Agent l thus has no reason to deviate and the principal attributes $P(\phi_l = 1) = 0$. Agent k is really responsible for failure if he deviates from requesting contributions $\hat{c}_l = \tilde{c}_l$ or contributing $\tilde{c}_k$ which in the PBE

lead to $y(\tilde{c}) \geq \hat{y}$. This cannot be ruled out because such a deviation would be profitable if, for example, the principal also pays $b_k = \hat{b}_k$ in case of failure. Hence, the principal's attribution of $P(\phi_k = 1) = 1$ is consistent.

For the case where $\omega = 0$, Corollary 2 directly implies that a positive target $\hat{y} > 0$ cannot be implemented. □

The theorem provides a formal justification for the management wisdom that declaring only one team member responsible can solve a free rider problem. Moreover, it identifies that this solution only works if institutions protect any member who is not formally responsible from suffering in case of failure. Under this condition and with sufficient commitment, even the first-best outcome can be implemented—see Corollary 7 in the Appendix.

We have seen that protecting agents who are not formally responsible by restricting the principal is crucial for implementing success. Typically restrictions are imposed by institutions on players to curb their opportunistic behavior. This is not the case here. The restriction does not actually prevent the caring principal from sanctioning the agent. Indeed, a caring principal has no interest in sanctioning any agent even if formal declarations do not come with any protection —recall Corollary 2.

Corollary 3 (Role of protection)**.** *Consider a team game* $(\hat{y}, \widehat{\mathbf{b}}, \widehat{\mathbf{r}})$ *with a positive target* $\hat{y} > 0$ *and a team committed to success. Then, a sufficiently caring principal pays at least as much without protection* ($\omega = 0$) *as with protection* ($\omega = 1$)*:* $b_i^1 \leq b_i^0$*, where* b_i^ω *denotes the benefit paid to agent* i *with and without protection.*

Proof. In absence of protection, $\omega = 0$, the principal assigns no real responsibility to any agent by Corollary 2 and pays the promised bonus to any agent $b_i^0 = \hat{b}_i$. Following directly from her utility function, the principal never pays more than the promised bonus $b_i \leq \hat{b}_i$. This implies that the bonus when agents are protected is at most $\hat{b}_i : b_i^1 \leq \hat{b}_i$. Taken together, we get $b_i^1 \leq \hat{b}_i = b_i^0$. The

inequality is even strict with respect to the formally responsible agent k in case of failure because $b_k^1 = 0 < b_k^0 = \hat{b}_k$. □

While protection does not restrict the principal in equilibrium, it ensures that the formally responsible agent has to take all the blame in case of failure. Failure hinges only on one specific agent's unwillingness to elicit help but success is only possible if all agents contributed. Put differently, any agent can prevent success by not contributing and is thus causing success. In this sense, the formally responsible agent has to share the fame in case of success, while taking the blame in case of failure.

Corollary 4 (Sharing fame but not blame)**.** *In any team game* $(\hat{y}, \widehat{\mathbf{b}}, \widehat{\mathbf{r}})$ *with a positive target* $\hat{y} > 0$ *and a committed team in which success* $y \geq \hat{y} > 0$ *can be supported by some PBE, real responsibility for success,* σ_i*, is attributed to the whole team, while failure,* ϕ_i *belongs to the formally responsible agent:*

$$P(\sigma_i = 1) = 1 \text{ and } P(\phi_i = 1) = \hat{r}_i \text{ for all } i.$$

Proof. Let us first deal with the attribution of fame. The proof works by contradiction. Assume agent i has not caused success $y \geq \hat{y} > 0$, which implies that agent i cannot alter the outcome from $y \geq \hat{y}$ to $y < \hat{y}$ by changing his contribution c_i. From $y(c) = y \geq \hat{y} > 0$, we get $c_i > 0$. Consider a change by agent i from $c_i > 0$ to $c_i' = 0$. Output then becomes $y = 0$, i.e., success is no longer possible. This contradicts that agent i cannot alter the outcome. Hence, agent i has caused the outcome. Since success can be reached in equilibrium, he is also really responsible for it.

Now consider the attribution of blame. Theorem 2 tells us that the only way to support success is by assigning formal responsibility to one agent k. From the proof of this theorem, it is clear that this agent is also really responsible: $\phi_k = 1$. Lemma 2 then implies $\phi_l = 0$ for any other agent l.

□

The corollary offers a precise interpretation in which sense 'good' leaders should 'take all the blame, while sharing all fame' and uncovers conditions under which this 'wisdom' holds. In particular, all agents' contributions must be required for success, the principal must care about real responsibility, and sanctions must be limited to those who are formally responsible.

2.5 Social capital and responsibility

Earlier we have seen that free riding can arise if no agent (Proposition 2) or all agents have sufficient capital (Theorem 1). This section deals with the intermediate cases.

If exactly one agent has enough capital to generate success, this means that only this agent can bring about success and is really responsible.

Corollary 5 (With greater power comes real responsibility)**.** *Consider the team game* $(\hat{y}, \widehat{\mathbf{b}}, \widehat{\mathbf{r}})$ *with a positive target* $\hat{y} > 0$ *in which only agent* k*'s social capital suffices to produce success and this agent can be sanctioned in case of failure (e.g., because he is formally responsible,* $r_k = 1$*, or formal responsibility is irrelevant,* $\omega = 0$*). Then, agent* k *is really responsible for failure and success can be implemented.*

Proof. There can be no equilibrium supporting success in which any agent $l \neq k$ whose social capital does not suffice brings about success; this agent can simply not elicit the necessary contributions from the others. If agent k's bonus depends on success, k has an incentive to contribute and elicit help from the other agents l such that sufficient output for success is produced. Any agent l will contribute because k has sufficient social capital—see Lemma 5. Due to the strategic dominance of contribution for all other agents l, the principal only attributes real responsibility for failure (off-equilibrium) to agent k. Given this belief, the principal optimally sanctions k in case of failure by withdrawing the bonus, which is possible either because $r_k = 1$ or $\omega = 0$. □

If power is interpreted as the ability to elicit the coorporation of others, this corollary provides a formal underpinning for the notion that power implies real responsibility. Only a 'powerful' agent can make a difference and avert failure and consequently be attributed real responsibility in case of failure.

If two (or more) agents command sufficient social capital, the situation is very similar to that in which all have sufficient capital: responsibility diffuses (Corollary 8) and the principal can only prevent free riding by declaring one agent formally responsible if all others cannot be sanctioned (Corollary 9).

3 Contribution to the literature and conclusion

Our analysis points to the limits of collective punishment as a way to counter free riding. Once the principal is sufficiently concerned about real responsibility, commitment within the team starts to matter (as suggested by management wisdom). Without social capital, free riding is unavoidable. If just one person has sufficient power to elicit contributions and this person can be sanctioned, this person is really responsible and no free riding occurs. The common advice to declare one agent responsible only works if several members are capable of obtaining support from others and if the declaration protects anyone who is not declared responsible from punishment. All this is summarized in Table 1.

We found that the free rider problem re-emerges in jointly responsible teams because responsibility diffuses (Theorem 1). The term 'diffusion of responsibility' can be traced back to Darley and Latané (1968), who associate it with the well-documented 'bystander effect': the larger a group, the less pronounced is 'helping'—see Latané and Nida (1981) and Fischer et al. (2011) for an overview. Closely related, economic experiments find that subjects act more morally questionably and behave less fairly in groups.[14] Adding to the literature, our paper offers an explicit definition of responsibility and proves

[14]For examples, see Charness (2000), Fershtman and Gneezy (2001), Dana et al. (2007), Bartling and Fischbacher (2011), Coffman (2011), Oexl and Grossman (2013), Falk and Szech (2013), and Behnk et al. (2017).

principal cares sufficiently about responsibility	agents with sufficient capital: none	one	more than one
no	free riding can be prevented using collective punishment (Holmström's Theorem 2)		
yes	free riding cannot be prevented (Proposition 3)	if agent can be sanctioned, he is really responsible no free riding occurs (Corollary 7)	free riding can be prevented by declaring one member responsible if and only if all others cannot be sanctioned (Theorem 2 and Corollary 9)

Table 1: Free riding and counter measures depending on whether the principal cares about real responsibility and agents' ability to elicit contributions (social capital).

that real responsibility diffuses when formal responsibility is assigned to several people.

Some contributions take diffusion of responsibility in groups as a given and examine how it can be used to encourage a certain behavior, for example, risk taking—see Borland (1992) or Milgrom and Roberts (1992, p. 431). In contrast, we derive diffusion of responsibility and its behavioral effects from fundamental principles.

Our paper also highlights that diffusion of responsibility is not necessarily the result of a coordination problem. In the classical bystander problem, one member of a group can engage in a costly action to prevent some bad outcome. The bystander effect can then be explained as a failure to coordinate on who that person should be (Diekmann, 1985; Harrington, 2001; Krueger and Massey, 2009). If this person is determined, for example, on the basis of lower costs (Diekmann, 1993) or by offering promotion to this person (Sliwka, 2006), this enables coordination and eliminates the bystander effect.[15] Our finding holds even though our model is stripped of any coordination failures by

[15]There is some experimental evidence that such devices help (Przepiorka and Diekmann, 2013; Diekmann and Przepiorka, 2015).

having players move sequentially. This shows that responsibility may diffuse even if people have the means to coordinate their contributions. For example, an after-sales agent may call, chat with, or email her colleague in order to ensure that a complaint has not already been dealt with before reaching out to a client. According to our findings, responsibility may nevertheless diffuse, suggesting that regarding diffused responsibility as a coordination problem is perhaps too narrow. Indeed, experimental subjects have been observed to act more morally questionably in groups—even when there were no coordination problems (Falk et al., 2020; Feess et al., 2020).

Another contribution of our paper is that it analyzes 'diffused responsibility' in the context of divided labor. While psychologists believe that divided labor plays a crucial role (Bandura et al., 1975; Bandura, 1999), the above mentioned explanations for diffused responsibility build on models in which contributions are additive, so that dividing labor has no benefits in comparison to one person doing the job. According to the principle of unity of responsibility by Milgrom and Roberts (1992, p. 410) all tasks should then be given to one person. In our model, this is not possible because inputs from all team members are needed in order for the team to perform. Such complementarities are typically the reason for forming a team in the first place[16] and why one might consider declaring everyone responsible. We thus complement the literature by examining diffusion of responsibility in an arguably natural setting. We also add to the existing literature in economics on why individuals may act more immorally in groups[17] by examining the role of formal and real assignment of responsibility for behavior.

Our paper also offers a new type of free rider problem that does not hinge on limited resources to incentivize performance. Following Holmström (1982),

[16]It may even be key in what seems traditional 'bystander' scenarios. Bregman (2020) vividly describes how the cooperation of four bystanders is required to save a mother and a toddler from drowning in a canal in Amsterdam.

[17]See, e.g., Huck and Konrad (2005), Lindbeck et al. (1999), Dufwenberg and Patel (2017) or Rothenhäusler et al. (2018).

a lower performance of teams has been attributed to the fact that team output can only be shared once. Breaking this constraint, for example with the help of a principal outside the team (Holmström, 1982)[18] or an 'innocent' team member (Rahman and Obara, 2010), can solve the problem. Other solutions include random punishments (Rasmusen, 1987), stochastic output (Legros and Matsushima, 1991; Legros and Matthews, 1993), allowing members to send messages either explicitly (Miller, 1997) or implicitly through the level of output (Strausz, 1999) or increasing reciprocity and hence the scope for informal contracting within the group (Dur and Sol, 2010).

When resources to reward or punish are limited, announcing one subgroup to be particularly scrutinized can help. For instance considering only one employee for promotion (Sliwka, 2006) or cracking down on an arbitrary subgroup of potential criminals (Eeckhout et al., 2010) provides very steep incentives for the respective person or group, while other people or groups' incentives are weakened. In the aggregate, this shift of attention may be beneficial, e.g., yield more effort or less crime.

At the heart of all these incentive problems are exogenously limited resources to provide incentives, e.g., output can only be shared once, only one vacancy is available as reward, the police cannot monitor everyone, etc. The principal here does not face any such exogenous limitations. Instead, she is restricted by her desire to limit sanctions to 'guilty members' and the fact that *fully* blaming *all* members leads to a logical inconsistency: if one member is already given full actual responsibility for having caused the failure, another member cannot have been causal (Lemma 2). Complementing the existing literature, we thus offer a rationale for a reputation free rider problem. This problem is consistent with the observation by social psychologists that people have to fear 'fewer negative social consequences' when being part of a group (Guerin,

[18]Holmström (1982) argues that capitalist firms are superior to cooperatives because the capitalist takes over the role of principal, which is challenged by MacLeod (1984, 1987, 1988) who shows that once members interact repeatedly, collectives can be more efficient.

1999, 2003). This new free rider problem is suited to describe situations like group work in classes. The teacher could easily give a bad grade to the whole group but might shy away from inflicting harm on students who did their job. Students will anticipate this and free ride.

Our results also point to the limits of mechanisms to ensure the provision of public goods. In numerous public good experiments (See Chaudhuri, 2011, for an overview) group members have an interest in joint contributions. As a result, they are willing to opt into (Kosfeld et al., 2009) or vote for (Dal Bo et al., 2010; Markussen et al., 2014) institutions that ensure contributions using the threat of sanctioning those who do not contribute. The principal in our model can be seen as such an institution. We show that this institution is of little help when it cannot identify who failed to contribute and is unwilling to sanction possibly innocent members (Theorem 1).

We offer a theoretical underpinning for the claim by management scholars that social capital is important for team performance (Leana and Pil, 2006; Clopton, 2011). Experimental economists have shown that social ties matter for team performance (Gächter et al., 2019; De Paola et al., 2019) and interact with group incentives (Delfgaauw et al., 2021). In our theory, social capital helps to translate formal into real responsibility and then enables the principal to incentivize the team. Glaeser et al. (2002) examine the trade-offs that lead to the formation of social capital and Itoh (1991) studies how incentives impact on agents' decision to help each other. In contrast, we study how the ability to elicit help affects the provision of incentives. By focusing on eliciting help as a very specific common aspect of most definitions of social capital since the seminal works by Putnam (1995) and Bourdieu (1986), we avoid the danger identified by Portes (1998) that the term 'social capital' has become 'vacuous' by its many different uses.

Our theory provides a specific formal reading of a general notion that has been around at least since the French revolution[19] and been attributed to

[19]On May 8th, 1783, the French Convention Nationale reminded the representatives of

various people from Voltaire, Roosevelt, and Churchill to Spider Man: namely that great power comes with great responsibility[20]—see Corollary 5.

We also show that a rather small organizational change, like a declaration of responsibility, can have substantial effects on performance. Grossman and Hart (1986) find that being the residual claimant should align with residual rights of control. Holmstrom (1980), Radner (1992, 1993) and many more examine where decision rights should be allocated within a firm to increase performance—for an excellent overview see Mookherjee (2006) and for a recent field experiment Bandiera et al. (2021). Relatedly, Prendergast (1995) and Aghion and Tirole (1997) look at when decision rights are transferred and how this can be achieved (e.g. by not informing oneself). Baliga and Sjöström (1998) analyze the transferal of decision rights in a situation that is particularly close to ours. They show that having a hierarchy in a team (by making the first or the second mover the grand contractor who subcontracts with the other player) leads to smaller rents than contracting with both individually. While all these contributions focus on delegation, performance here increases without any transfer of decision rights; all agents engage in the same decisions—irrespective of the formal responsibility assignment. Gains do not come from squeezed rents, either. Incentives are shifted by the assignment because the credibility to punish in case of failure changes.

In addition, we provide a new argument why being 'weak' can turn out to be beneficial. Grout (1984) finds that investors are less afraid of being expropriated when the trade union is weak and hence more willing to invest, which ultimately benefits the union. In contrast, the weakness here concerns the principal's inability to sanction agents ($\omega = 1$), which closes the wiggle room for shifting real responsibility. This ultimately benefits the principal in the sense that she can overcome free riding (See Theorem 2).

the people of the responsibilities that come with their power: 'Ils [les representants] doivent envisager qu'une grande responsabilité est la suite inséparable d'un grand pouvoir'

[20]See entry 'With Great Power Comes Great Responsibility' on quoteinvestigator.com from the 23rd of July 2015, accessed on the 18th of January 2021.

Finally, we complement the existing rationale that laws are necessary to curb unwanted behavior or shift threat points. In law and economics, the literature focuses on how liability rules affect behavior, both, theoretically (for an excellent overview, see Schweizer, 2015) as well as empirically (for examples, see Currie and MacLeod, 2008; Carvell et al., 2012). In one variation of our model ($\omega = 1$), rules restrict the principal to pay agents who are not formally responsible. A principal who cares about real responsibility, however, would on his own accord pay them. Accordingly, there is neither a need to curb unwanted opportunistic behavior nor a shift in threat point. Still, the restriction is crucial. It ensures that a non-responsible agent cannot lose the reward and hence has a strictly dominant strategy. This enables the principal to incentivize the formally responsible agent and produce a positive outcome (Theorem 2).

In summary, we are the first to explain why real responsibility diffuses when the whole team is declared responsible, why declaring one agent responsible can overcome this problem, and when this is the case. We offer a rationale why fame is shared among team members while blame rests with the formally responsible member. Finally, we present a first rationale why power entails real responsibility.

References

Aghion, Philippe and Jean Tirole, "Formal and Real Authority in Organizations," *Journal of Political Economy*, 1997, *105* (1), 1–29.

Baliga, Sandeep and Tomas Sjöström, "Decentralization and Collusion," *Journal of Economic Theory*, 1998, *83*, 196–232.

Bandiera, Oriana, Michael Carlos Best, Adnan Qadir Khan, and Andrea Prat, "The allocation of authority in organizations: A field experiment with bureaucrats," *The Quarterly Journal of Economics*, 2021, *136* (4), 2195–2242.

Bandura, Albert, "Moral disengagement in the perpetration of inhumanities," *Personality and Social Psychology Review*, 1999, *3* (3), 193–209.

_ , Bill Underwood, and Michael E Fromson, "Disinhibition of aggression through diffusion of responsibility and dehumanization of victims," *Journal of Research in Personality*, 1975, *9* (4), 253–269.

Bartling, Björn and Urs Fischbacher, "Shifting the blame: On delegation and responsibility," *The Review of Economic Studies*, 2011, *79* (1), 67–87.

_ , _ , and Simeon Schudy, "Pivotality and responsibility attribution in sequential voting," *Journal of Public Economics*, 2015, *128*, 133–139.

Behnk, Sascha, Li Hao, and Ernesto Reuben, "Partners in crime: Diffusion of responsibility in antisocial behaviors," Discussion paper 11031, IZA 2017.

Blackstone, William, *Commentaries on the Laws of England*, Vol. 4, Clarendon Press, 1765-1770.

Borland, Jeff, "Career concerns: Incentives and endogenous learning in labour markets," *Journal of Economic Surveys*, 1992, *6* (3), 251–270.

Bourdieu, Pierre, "The forms of capital," in John G Richardson, ed., *Handbook of Theory and Research for the Sociology of Education*, New York, Greenwood, 1986.

Bregman, Rutger, *Humankind: A hopeful history*, Bloomsbury Publishing, 2020.

Cappelen, Alexander W, Cornelius Cappelen, and Bertil Tungodden, "Second-best fairness under limited information: The trade-off between false positives and false negatives," Technical Report 18 2018.

Carvell, Daniel, Janet Currie, and W Bentley MacLeod, "Accidental death and the rule of joint and several liability," *The Rand Journal of Economics*, 2012, *43* (1), 51–77.

Charness, Gary, "Responsibility and effort in an experimental labor market," *Journal of Economic Behavior & Organization*, 2000, *42* (3), 375–384.

Chassang, Sylvain and Christian Zehnder, "Rewards and punishments: Informal contracting through social preferences," *Theoretical Economics*, 2016, *11* (3), 1145–1179.

Chaudhuri, Ananish, "Sustaining cooperation in laboratory public goods experiments: a selective survey of the literature," *Experimental Economics*, 2011, *14* (1), 47–83.

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

Clopton, Aaron W, "Social capital and team performance," *Team Performance Management: An International Journal*, 2011.

Coffman, Lucas C, "Intermediation reduces punishment (and reward)," *American Economic Journal: Microeconomics*, 2011, *3* (4), 77–106.

Currie, Janet and W Bentley MacLeod, "First do no harm? Tort reform and birth outcomes," *The Quarterly Journal of Economics*, 2008, *123* (2), 795–830.

Dal Bo, Pedro, Andrew Foster, and Louis Putterman, "Institutions and Behavior: Experimental Evidence on the Effects of Democracy," *The American Economic Review*, 2010, *100* (5), 2205–2229.

Dana, Jason, Roberto A Weber, and Jason Xi Kuang, "Exploiting moral wiggle room: experiments demonstrating an illusory preference for fairness," *Economic Theory*, 2007, *33* (1), 67–80.

Darley, John M and Bibb Latané, "Bystander intervention in emergencies: Diffusion of responsibility.," *Journal of Personality and Social Psychology*, 1968, *8* (4p1), 377.

De Paola, Maria, Francesca Gioia, and Vincenzo Scoppa, "Free-riding and knowledge spillovers in teams: The role of social ties," *European Economic Review*, 2019, *112*, 74–90.

de Voltaire, Jean Francois Marie Arout, *Zadig ou la Destinée* 1747.

Delfgaauw, Josse, Robert Dur, Oke Onemu, and Joeri Sol, "Team incentives, social cohesion, and performance: A natural field experiment," *Management Science*, 2021.

Diekmann, Andreas, "Volunteer's dilemma," *Journal of Conflict Resolution*, 1985, *29* (4), 605–610.

_ , "Cooperation in an asymmetric Volunteer's dilemma Game: Theory and Evidence," *International Journal of Game Theory*, 1993, *22*, 75–85.

_ **and Wojtek Przepiorka**, "Punitive preferences, monetary incentives and tacit coordination in the punishment of defectors promote cooperation in humans," *Scientific Reports*, 2015, *5*, 10321.

Dufwenberg, Martin and Amrish Patel, "Reciprocity networks and the participation problem," *Games and Economic Behavior*, 2017, *101*, 260–272.

Dur, Robert and Joeri Sol, "Social interaction, co-worker altruism, and incentives," *Games and Economic Behavior*, 2010, *69* (2), 293–301.

Eeckhout, Jan, Nicola Persico, and Petra E Todd, "A theory of optimal random crackdowns," *American Economic Review*, 2010, *100* (3), 1104–35.

Engl, Florian, "A theory of causal responsibility attribution," *Available at SSRN 2932769*, 2018.

Falk, Armin and Nora Szech, "Morals and markets," *Science*, 2013, *340* (6133), 707–711.

_ , Thomas Neuber, and Nora Szech, "Diffusion of being pivotal and immoral outcomes," *The Review of Economic Studies*, 2020.

Feess, Eberhard, Florian Kerzenmacher, and Gerd Muehlheusser, "Moral Transgressions by Groups: What Drives Individual Voting Behavior?," Discussion Paper 13383, Institute of Labor Economics (IZA) June 2020.

Fershtman, Chaim and Uri Gneezy, "Strategic delegation: An experiment," *RAND Journal of Economics*, 2001, pp. 352–368.

Fischer, Peter, Joachim I Krueger, Tobias Greitemeyer, Claudia Vogrincic, Andreas Kastenmüller, Dieter Frey, Moritz Heene, Magdalena Wicher, and Martina Kainbacher, "The bystander-effect: a meta-analytic review on bystander intervention in dangerous and non-dangerous emergencies.," *Psychological Bulletin*, 2011, *137* (4), 517.

Gächter, Simon, Chris Starmer, and Fabio Tufano, "The surprising capacity of the company you keep: revealing group cohesion as a powerful factor of team production," Technical Report, CeDEx Discussion Paper Series 2019.

Gibbons, Robert, "An Introduction to Applicable Game Theory," *Journal of Economic Perspectives*, Winter 1997, *11* (1), 127–149.

Glaeser, Edward L, David Laibson, and Bruce Sacerdote, "An economic approach to social capital," *The Economic Journal*, 2002, *112* (483), F437–F458.

Grossman, Sanford J. and Oliver D. Hart, "The Costs and Benefits of Ownership: A Theory of Vertical and Lateral Integration," *Journal of Political Economy*, 1986, *94*, 691–719.

Grout, Paul, "Investment and Wages in the Absence of Binding Contracts: A Nash Bargaining Approach," *Econometrica*, March 1984, *53* (2), 449–460.

Guerin, Bernard, "Social behaviors as determined by different arrangements of social consequences: Social loafing, social facilitation, deindividuation, and a modified social loafing," *The Psychological Record*, 1999, *49* (4), 565–577.

_ , "Social behaviors as determined by different arrangements of social consequences: Diffusion of responsibility effects with competition," *The Journal of Social Psychology*, 2003, *143* (3), 313–329.

Harrington, Joseph E, "A simple game-theoretic explanation for the relationship between group size and helping," *Journal of Mathematical Psychology*, 2001, *45* (2), 389–392.

Holmstrom, Bengt, "On the theory of delegation," Technical Report, Discussion Paper 1980.

Holmström, Bengt, "Moral Hazard in Teams," *Bell Journal of Economics*, 1982, *13* (2), 324–340.

Huck, Steffen and Kai A Konrad, "Moral cost, commitment, and committee size," *Journal of Institutional and Theoretical Economics (JITE)/Zeitschrift für die gesamte Staatswissenschaft*, 2005, pp. 575–588.

Itoh, Hideshi, "Incentives to Help in Multi-Agent Situations," *Econometrica*, May 1991, *59* (3), 611–636.

_ , "Cooperation in Hierarchical Organizations: An Incentive Perspective," *Journal of Law, Economics and Organization*, April 1992, *8* (2), 321–345.

_ , "Coalitions, incentives, and risk sharing," *Journal of Economic Theory*, 1993, *60* (2), 410–427.

Kandel, Eugene and Edward Lazear, "Peer Pressure and Partnerships," *Journal of Political Economy*, 1992, *100* (4), 801–817.

Kandori, Michihiro, "Social Norms and Community Enforcement," *Review of Economic Studies*, 1992, *59*, 63–80.

Katzenbach, Jon R and Douglas K Smith, *The wisdom of teams: Creating the high-performance organization*, Harvard Business Review Press, 1995.

Kosfeld, Michael, Akira Okada, and Arno Riedl, "Institution formation in public goods games," *American Economic Review*, 2009, *99* (4), 1335–55.

Krueger, Joachim I and Adam L Massey, "A rational reconstruction of misbehavior," *Social Cognition*, 2009, *27* (5), 786–812.

Latané, Bibb and Steve Nida, "Ten years of research on group size and helping.," *Psychological bulletin*, 1981, *89* (2), 308.

Leana, Carrie R and Frits K Pil, "Social capital and organizational performance: Evidence from urban public schools," *Organization Science*, 2006, *17* (3), 353–366.

Legros, Patrick and Hitoshi Matsushima, "Efficiency in partnerships," *Journal of Economic Theory*, December 1991, *55* (2), 296–322.

_ **and Steven A Matthews**, "Efficient and nearly-efficient partnerships," *The Review of Economic Studies*, 1993, *60* (3), 599–611.

Lewis, David, "Causation," *The Journal of Philosophy*, 1974, *70* (17), 556–567.

Lindbeck, Assar, Sten Nyberg, and Jörgen W. Weibull, "Social Norms and Economic Incentives in The Welfare State," *Quarterly Journal of Economics*, 1999, *114* (1), 1–35.

Lübbecke, Silvia and Wendelin Schnedler, "Don't patronize me! An experiment on preferences for authorship," *Journal of Economics and Management Strategy*, 2020, pp. 420–438.

MacLeod, Bentley, "Behavior and the Organization of the Firm," *Journal of Comparative Economics*, 1987, *11* (2), 207–220.

MacLeod, W Bentley, "A Theory of Cooperative Teams," resreport 8441, CORE Discussion Paper 1984.

MacLeod, W. Bentley, "Equity, efficiency, and incentives in cooperative teams," *Advances in the Economic Analysis of Participatory and Labor Managed Firms*, 1988, *3* (798), 5–23.

Markussen, Thomas, Louis Putterman, and Jean-Robert Tyran, "Self-organization for collective action: An experimental study of voting on sanction regimes," *The Review of Economic Studies*, 2014, pp. 301–324.

Milgrom, Paul and John Roberts, *Economics, Organization and Management*, New Jersey: Prentice Hall, 1992.

Miller, Nolan H, "Efficiency in partnerships with joint monitoring," *Journal of Economic Theory*, 1997, *77* (2), 285–299.

Mookherjee, Dilip, "Decentralization, Hierarchies, and Incentives: A Mechanism Design Perspective," *Journal of Economic Literature*, 2006, *44* (2), 367–390.

Oexl, Regine and Zachary J Grossman, "Shifting the blame to a powerless intermediary," *Experimental Economics*, 2013, *16* (3), 306–312.

Portes, Alejandro, "Social capital: its origins and applications in modern sociology," *Annual Review of Sociology*, 1998, *24* (1), 1–25.

Prendergast, Canice J, "A theory of responsibility in organizations," *Journal of Labor Economics*, 1995, *13* (3), 387–400.

Przepiorka, Wojtek and Andreas Diekmann, "Individual heterogeneity and costly punishment: a volunteer's dilemma," *Proceedings of the Royal Society B: Biological Sciences*, 2013, *280* (1759), 20130247.

Putnam, Robert D, "Bowling alone: America's declining social capital," *Journal of Democracy*, 1995, *1* (6), 65–78.

Radner, Roy, "Hierarchy: The Economics of Managing," *Journal of Economic Literature*, September 1992, *XXX*, 1382–1415.

_ , "The Organization of Decentralized Information Processing," *Econometricy*, September 1993, *51* (5), 1109–1146.

Rahman, David and Ichiro Obara, "Mediated partnerships," *Econometrica*, 2010, *78* (1), 285–308.

Rasmusen, Eric, "Moral hazard in risk-averse teams," *The RAND Journal of Economics*, 1987, pp. 428–435.

Robison, Lindon J, A Allan Schmid, and Marcelo E Siles, "Is social capital really capital?," *Review of social economy*, 2002, *60* (1), 1–21.

Rothenhäusler, Dominik, Nikolaus Schweizer, and Nora Szech, "Guilt in voting and public good games," *European Economic Review*, 2018, *101*, 664–681.

Schwaber, Ken and Jeff Sutherland, *Software in 30 days: how agile managers beat the odds, delight their customers, and leave competitors in the dust*, John Wiley & Sons, 2012.

Schweizer, Urs, *Spieltheorie und Schuldrecht*, Mohr Siebeck, 2015.

Sliwka, Dirk, "On the notion of responsibility in organizations," *Journal of Law, Economics, and Organization*, 2006, *22* (2), 523–547.

Strausz, Roland, "Efficiency in Sequential Partnerships," *Journal of Economic Theory*, 1999, *85* (1), 140 – 156.

Wilson, S.F., *Analyzing Requirements and Defining Solution Architectures: MCSD Training Kit : for Exam 70-100* Dv-McSd Training Kit, Microsoft Press, 1999.

Appendix

Lemma 4 (First-best)**.** *The contributions c_i^{FB} that maximize $y(c) - \sum_i c_i$ are also the first-best contributions that would result if contracts could be written about c.*

Proof. For finding the first-best, maximize the principal's utility under the side constraint that the agents are not worse off:

$$\max_{c,b} y(c) - \sum_{i \in N} b_i - \kappa \sum_{i \in N} \mathbb{1}_{[0,\hat{b}_i)}(b_i)\,(\sigma_i + 1 - \phi_i) \tag{7}$$

$$\text{s.t. } b_i - c_i - \left\{ \begin{array}{rl} \gamma_j c_j & \text{for } c_i < \hat{c}_i \\ 0 & \text{otherwise.} \end{array} \right\} \geq -\epsilon \tag{8}$$

Observe that $\mathbb{1}_{[0,\hat{b}_i)}(b_i)\,(\sigma_i + 1 - \phi_i) > 0$ negatively affects the principal without generating gains. By setting the first-best promise to the first-best bonus $\hat{b}_i^{\text{FB}} = b_i^{\text{FB}}$, these costs drop out and the purely negative effects can be avoided.

Requesting help $\hat{c}_i > 0$ may impose a cost of $\gamma_j c_j$ on agent i without generating any benefit to anyone. If agents do not request anything from each other $\hat{c}_i^{\text{FB}} = 0$, these costs can be eliminated.

Using the choices of $\hat{b}_i^{\text{FB}} = b_i^{\text{FB}}$ and $\hat{c}_i^{\text{FB}} = 0$ in the maximization problem, we get:

$$\max_{c,b,\hat{c},\hat{b}} y(c) - \sum_{i \in N} b_i \text{ s.t. } b_i - c_i \geq -\epsilon \tag{9}$$

This problem is independent of γ_i and κ. Taking ϵ to zero in the constraint, we get $b_i^{\text{FB}} = c_i^{\text{FB}}$ and plugging this in the main condition results in

$$\max_{c} y(c) - \sum_{i} c_i,$$

which leads to the maximizers c_i^{FB}. □

Lemma 5 (Optimal response to request)**.** *Suppose the principal does not condition the bonus for agent l on output y, agent k requested $\hat{c}_l > 0$ from l and contributed c_k. Then, the dominant strategy for agent l is*

$$c_l^* = \begin{cases} \hat{c}_l & \text{if } \hat{c}_l \leq \gamma_k c_k \text{ and } \hat{c}_l \leq \hat{b}_l. \\ 0 & \text{otherwise.} \end{cases} \tag{10}$$

If $\hat{c}_l < \gamma_k c_k$, this strategy is strictly dominant.

Proof. When it is l's turn to contribute, there are three possible cases. First, the agent can leave the relationship and incur the small loss of $\epsilon > 0$. Alternatively, he can stay and receive the bonus of $\hat{b}_l$. If he stays, he has the choice between ignoring the request, contributing $c_l < \hat{c}_l$, and getting $u_l(b_l, c_l) = b_l - c_l - \gamma_l c_l$ with $c_l \geq 0$. Finally, he can meet the request and get $u_l(b_l, c_l) = b_l - c_l$ with $c_l \geq \hat{c}_l$. In all cases, utility is maximized by contributing the lowest possible level. This means $c_l = 0$ in whenever the agent leaves the relationship or does not meet the request, $c_l < \hat{c}_l$, and $c_l = \hat{c}_l$, otherwise. The agent thus contributes $\hat{c}_l$ if and only if $\hat{c}_l \leq \hat{b}_l$ and $\hat{c}_l \leq \gamma_k c_k$, where $\hat{c}_l < \gamma_k c_k$ means that l is strictly better off from meeting the request. □

Lemma 6 (Necessity of intervention)**.** *If the principal does not condition the bonus for some agent k on output y, agents do not contribute: $b_k \equiv b(y)$ constant in y for all k: $\Rightarrow c_l^* = 0$ for all l.*

Proof. By Lemma 5, agent l only contributes $c_l > 0$ when some agent k requests some $\hat{c}_l > 0$ with $\hat{c}_l \leq \min(\gamma_k c_k, \hat{b}_k)$. Being in charge, agent k, however, has no incentive to contribute because her benefit $\hat{b}_k$ is independent of her contribution. She hence chooses $c_k = 0$, which then implies $c_l = 0$. □

Lemma 7 (Need to declare all formally responsible)**.** *When members who are not formally responsible cannot be sanctioned ($\omega = 1$), declaring the whole team jointly responsible is a necessary condition, $\hat{r}_i = 1$ for all $i \in N$, for implementing Holmström's incentive scheme.*

Proof. Consider the case of failure, $y < \hat{y}$. Using in equation (1) that $\omega = 1$ and $y < \hat{y}$, we get $b_i \geq \hat{b}_i(1 - \hat{r}_i)$. Since Holmström's scheme requires $b_i = 0$ for all $i \in N$ in case of failure, we must have $\hat{r}_i = 1$ for all $i \in N$. □

Corollary 6 (to Holmström's Theorem 2)**.** *If the principal does not care about who is really responsible ($\kappa = 0$), first-best contributions can be achieved in a PBE using Holmström's bonus scheme.*

Proof. Consider the following equilibrium candidate. The principal sets $\hat{y} = y^{\text{FB}}$, $\hat{b}_i = c_i^{\text{FB}}$, and $\hat{r}_i = 1$ and pays the bonus $b_i = \hat{b}_i$ whenever $y \geq \hat{y}$ and $b_i = 0$, otherwise. Agents contribute first-best levels c_i^{FB} as long as the promised bonus covers at least their costs $\hat{b}_i \geq c_i^{\text{FB}}$ but leave the relationship, otherwise. Principal's beliefs about agents behavior may be arbitrary; due to $\kappa = 0$ they do not affect behavior.

This behavior constitutes a PBE. The principal cannot further reduce bonus payments (or promises) without agents leaving the relationship and has no interest in increasing the payments because this lowers her benefit. Agents do not lose out by contributing because contribution costs c_i^{FB} are exactly set off by the bonus $\hat{b}_i = c_i^{\text{FB}}$. If agents had to contribute more than c_i^{FB}, they would lose more than $\epsilon > 0$ by staying in the relationship and leave it.

Agents might send requests as part of the equilibrium but these do not affect the outcome. While requests $\hat{c}_i > c_i^{\text{FB}}$ will not be met, requests $\hat{c}_i \leq c_i^{\text{FB}}$ impose no constraint. Taking ϵ to zero, principal's earnings approach the first-best rent: $y^{\text{FB}} - \sum_i c_i^{\text{FB}}$. □

Corollary 7 (Implementing the first-best)**.** *Suppose the team is committed to bring about success $y(\mathbf{c}^{FB})$ using c^{FB}, then the first-best can be implemented however much the principal cares about real responsibility whenever only formally responsible agents can be sanctioned $\omega = 1$.*

Proof. Set $\hat{y} = y^{\text{FB}}, \hat{b}_i = c_i^{\text{FB}}$, for all i in Theorem 2. □

Corollary 8 (to Theorem 1)**.** *Consider a team game* $(\hat{y}, \widehat{\mathbf{b}}, \widehat{\mathbf{r}})$ *with a positive target* $\hat{y} > 0$*, and a group* G *consisting of* m *of the* n *team members that is committed to success* $\hat{y}$ *and jointly declared responsible (*$\hat{r}_i = 1$ *for all* $i \in G$*).*[21] *Then, a sufficiently caring principal will not consider any agent really responsible for failing to implement* $\hat{y}$*, agents slack, and* $\hat{y}$ *cannot be produced:*

$$\kappa > \frac{1}{1-\frac{1}{m}} \cdot \max_i \hat{b}_i : \phi_i^* = 0, c_i^* = 0, \text{ for all } i \text{ and } y^* = 0 < \hat{y}.$$

Proof. The proof follows that of Theorem 1. The condition $\kappa > \frac{1}{1-\frac{1}{m}} \cdot \max_i \hat{b}_i$ implies that $\kappa > \frac{1}{1-\frac{1}{n}} \cdot \max_i \hat{b}_i$ and by Lemma 3 one agent needs to take charge. Then, all agents in G can produce success by requesting $\hat{c}_l = c_l^*$ and contributing appropriately. Only the last agent $\underline{k} \in G$ with the opportunity to take charge can cause failure by not taking charge. Since all agents in G are equally likely to be $\underline{\text{k}}$, $P(\phi_i = 1) \leq \frac{1}{m}$ for $i \in G$ (and zero for all others). Again Lemma 1 can be evoked to show that the principal does not take away any promised bonus, which then implies that no one can be expected to contribute not even the agents from G. □

Corollary 9 (to Theorem 2)**.** *Consider a team game* $(\hat{y}, \widehat{\mathbf{b}}, \widehat{\mathbf{r}})$ *with a positive target* $\hat{y} > 0$ *such that for some adequately compensated contributions* $\tilde{\mathbf{c}}$ *with* $\hat{b}_i \geq \tilde{c}_i$ *the target can be met,* $y(\tilde{\mathbf{c}}) > \hat{y}$*, and a group* G *consisting of* m *of the* n *team members that is committed to success* $\hat{y}$ *and only one agent* $k \in G$ *being declared formally responsible,* $r_k = 1$ *and* $r_l = 0$ *for all* $l \neq k$*.*[22] *Then, success can be implemented irrespective of how much the principal cares about real responsibility if and only if agents who are not formally responsible cannot be sanctioned,* $\omega = 1$*.*

Proof. The proof follows from that of Theorem 2. The declared agent k takes charge, the principal thinks this agent is responsible for failure and this is consistent because all other agents have a strictly dominant strategy, namely to meet any request by k. □

[21] The definition of a group being committed to success can naturally be derived from a team being committed to success where G replaces N.

[22] The definition of a group being committed to success can naturally be derived from a team being committed to success where G replaces N.