I Z A Institute
of Labor Economics
Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 17478

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

Xinyu Li
Wendelin Schnedler

NOVEMBER 2024

I Z A Institute
of Labor Economics

Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 17478

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

Xinyu Li
PBL Netherlands Environmental Assessment Agency

Wendelin Schnedler
Paderborn University and IZA

NOVEMBER 2024

ISSN: 2365-9793

IZA – Institute of Labor Economics

Schaumburg-Lippe-Straße 5–9
53113 Bonn, Germany

Phone: +49-228-3894-0
Email: publications@iza.org

www.iza.org

ABSTRACT

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

Teams are formed because input from different people is needed. Providing incentives to team members, however, can be diffcult. According to received wisdom, declaring all members 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 jointly declaring a team responsible results in reputation free-riding. We show that declaring one person responsible can overcome this problem but only if all other team members are protected from being sanctioned.

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

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

Corresponding author:
Wendelin Schnedler
Faculty of Economics
University of Paderborn
Warburger Straße 100
33098 Paderborn
Germany
E-mail: wendelin.schnedler@upb.de

* We like to thank Björn Bartling, Sylvain Chassang, Anastasia Danilova, Robert Dur, Florian Engl, Eberhard Feess, Urs Fischbacher, Stephane Gonzalez, Andreas Diekmann, Matthias Fahn, Claus-Jochen Haake, Jenny Kragl, Florian Kerzenmacher, Matthias Kräkel, Bentley MacLeod, Gerd Mühlheusser, Dirk Sliwka, Simeon Schudy, Nora Szech, Georg Weizsäcker, and seminar participants at the University of Groningen, the Meeting of the OrgEcon Group of the VfS, and COPE 2023 for their valuable comments. All errors remain our own.

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

Introduction

Management scholars emphasize several factors that help teams with overcoming free riding. Firstly, team members must be willing to help each other. Katzenbach and Smith (1995), for example, claim 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). Finally, if this member decides to lead the team, she has to 'share all the fame and take all the blame'.[1]

Traditional incentive theory offers no justification for this management wisdom. Itoh (1991; 1992; 1993) famously examined how incentives shape helping behavior but assumes rather than explains why help in teams may be desirable. In Holmström's seminal principal-agent model (1982), the principal as an outsider to the team can solve the free rider problem. In sharp contrast to the above management wisdom, this solution involves 'blaming' the whole team: all members suffer whenever a pre-specified target is not met (see Holmström's Theorem 2). So in what sense does giving a group joint responsibility create rather than solve a free rider problem? And why is a commitment to 'one another' crucial for overcoming this problem?

For answering these questions, consider a project that only succeeds if

[1] 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.

two or more diligent people put in sufficient effort. The principal forms a team and formally gives responsibility to one or many members. Members can coordinate contributions and ask for help, which will come forth if they are committed to each other. Members' reason to contribute is that they prefer not to be blamed but backed by the principal, for example, when it comes to pay rises, promotions or being assigned to interesting projects.

We later tie this description to the principal-agent model by Holmström (1982). For rationalizing the management wisdom, we have to deviate from this model. We do so by supposing that the principal cares about who is really responsible. More specifically, she only wants to blame members whom she expected to produce success (in the sense of Perfect Bayesian Nash equilibrium) but who then caused failure (in the sense of Hume (1748) and Lewis (1974)). This preference could reflect that the principal does not like punishing someone for an offense they have not committed[2] or that she fears unjust punishment undermines morale. In line with the standard principal-agent framework but complicating matters, the principal can only observe the outcome of the project; she has to infer who did what from members' incentives (by ruling out that agents play strictly dominated strategies).

In this situation, team members' commitment to 'one another' is crucial for success (Proposition 2). The reason is the following. We can show that at most one member can be really responsible for failure (Lemma 2). Not knowing who this member is but caring about real responsibility, the principal rather 'spares the guilty than condemns the innocent' (de Voltaire, 1747) and refrains from blaming groups (Proposition 1).[3] As a result, the principal can (at best) provide incentives to one team member. Success, however, requires the input of at least one other member. Consequently, commitment of team members to help 'one another' is necessary for success.

[2]This notion is, for example, enshrined in international martial law (See Article 33 of the Geneva Convention, August 12, 1949).

[3]This behavior is in line with a majority of subjects in a large field experiment by Cappelen et al. (2018)

Commitment to one another within the team, however, is not enough for success. Even if every member can ensure success by eliciting help from the others, one member has to actually step forward to do so. Consider a team that is jointly declared responsible and fails. Then, the principal cannot identify who should have stepped forward, cannot assign real responsibility for failure to any member and hence cannot credibly blame anyone. Members anticipate this, free ride on each others' reputation and neither elicit help nor contribute. Although members are perfectly capable of coordinating their contributions and are formally declared to be responsible, no member is really responsible (Theorem 1).

According to management wisdom, declaring one agent responsible can prevent the diffusion of responsibility. But how can the formal assignment *in advance* affect real responsibility *later*? The formal assignment neither shifts the balance of power between the 'responsible' member and other members nor does it alter the production technology. It also does not change who can observe whose contributions and therefore does not affect the scope for formal or informal agreements. Since members have no trouble coordinating, the declaration does not help with coordination, either.

Still, the management wisdom about declaring one member responsible can be rationalized (Theorem 2). Suppose the principal can only blame members who are formally declared responsible. Then, declaring one (and only one) member responsible means that only this member can be blamed. Accordingly, this member is the only one who can have an incentive to elicit contributions. This means that this member becomes really responsible for failure. Knowing this, the principal 's threat to blame this member becomes credible, the formally declared member then literally turns into a 'leader': she moves first and elicits help, the other members follow and the team succeeds.[4]

[4]A recent series of field experiments highlights how leaders affect productivity by creating a supportive environment (Haeckl and Rege, forthcoming; Castro et al., 2022), regularly communicating with subordinates (Manthei et al., 2023), recognizing their work (Bradler et al., 2016), or being charismatic (Antonakis et al., 2022). Czura et al. (forthcoming) provide

This argument only works because the principal cannot blame members that are not formally declared responsible. If she could blame anyone, the formal declaration becomes cheap talk and similar to the case, where several members are jointly responsible, real responsibility diffuses. Responsibility declaration thus only works if the leader 'takes all the blame'. At the same time, the leader has to 'share fame' with anyone essential for success (Corollary 4).

If the team has a natural leader (in the sense that only one member has the power to elicit effort from the other members), then formal declarations of responsibility are pointless. In line with a notion that has been around at least since the French revolution (Comittee of Public Safety, 1793, p. 72), real responsibility then follows from the greater power of this member (Corollary 5).

1 Contribution to the literature

A first contribution of our paper is that it offers a novel and arguably relevant reason for why effort provision in teams may break down. A shortfall in effort provision has been associated with limited resources to reward or punish—see Holmström (1982) and various articles following his pioneering contribution (Rasmusen, 1987; Legros and Matsushima, 1991; Legros and Matthews, 1993; Miller, 1997; Strausz, 1999; Sliwka, 2006; Dur and Sol, 2010; Eeckhout et al., 2010). Here, we exclude this explanation by assuming that resources to provide incentives are unrestricted. Output may also be low because it is noisy and agents require insurance which in turn means weaker incentives and ultimately lower effort (Holmström, 1982). Here, output is deterministic. Finally, members may face a coordination problem (Diekmann, 1985; Harrington, 2001; Krueger and Massey, 2009). Then, singling out a single person (Diekmann, 1993; Sliwka, 2006), or treating a group of ex-ante identical agents differently (Winter, 2004) may help them to coordinate better. Here, we suppose that team members

evidence that a transformational leadership style improves well-being and performance in times of crisis.

move sequentially and are fully informed about previous choices to eliminate any such coordination issues. If the whole team is declared responsible, effort provision breaks down, although traditional reasons for lower output are absent from our model. This reputation free riding is distinct from extant forms of free riding. It can explain the shortfall of team output even if the principal can easily motivate agents by withdrawing her backing, while agents have fairly good control over the outcome and no problems with coordinating their contributions.

A second contribution is that we provide an explanation for why declaring multiple agents responsible is problematic.[5] The reason is that responsibility 'diffuses' (Theorem 1). Social psychologist use the term to explain why people are less pro-active in groups (Darley and Latané, 1968), associate it with divided labor (Bandura et al., 1975; Bandura, 1999), and trace it back to individuals having to fear 'fewer negative social consequences' in groups (Guerin, 1999, 2003). In economics, Milgrom and Roberts (1992, p. 431) argue that diffusion of responsibility can explain why teams may make more risky decisions than individuals. Economic studies of liability rules,[6] ownership (Grossman and Hart, 1986; Grout, 1984) and organizational design (Milgrom and Roberts, 1992, p. 410) typically assume that full incentives can only be provided to one agent and then go on to study, who this agent should be. Contributing to this literature, we neither take the diffusion of responsibility nor the need to focus on one agent as a given. Instead, we derive them from first principles. We find that members of collectives indeed have less to fear (Proposition 1), that responsibility diffuses entirely out of the team rather than being shared among

[5]Plenty of experimental evidence suggests that experimental subjects in groups act more morally questionably and behave less fairly (Charness, 2000; Fershtman and Gneezy, 2001; Dana et al., 2007; Bartling and Fischbacher, 2011; Coffman, 2011; Oexl and Grossman, 2013; Falk and Szech, 2013; Behnk et al., 2017), even if they face no coordination problems Falk et al. (2020); Feess et al. (2020). In social psychology, the tendency of groups to help less is a major topic—see the reviews by Latané and Nida (1981) and Fischer et al. (2011)

[6]For an excellent overview, see Schweizer (2015), for empirical evidence Currie and MacLeod (2008); Carvell et al. (2012)

team members (Theorem 1),[7] and that declaring only one agent responsible can prevent the diffusion (Theorem 2).

As a third contribution, we identify when declaring a single agent responsible works, i.e., when formal translates into real responsibility, so that the principal can commit to blame. Principals who formally delegate a decision[8] must commit themselves not to interfere with the agent's choice in order to really transfer authority, which can be achieved by being busy (Prendergast, 1995) or staying ignorant (Aghion and Tirole, 1997). Here, in the case of committing to blame, the crucial condition is that members who are not formally responsible cannot be blamed (Theorem 1). This could explain why in many firms and organizations only formally responsible people are held accountable.

The key assumption driving our results is that the principal has qualms about blaming an agent who is not really responsible for failure. The paper thus contributes to a broader literature that analyzes contracting in teamwork with non-standard preferences (Bartling, 2011; von Siemens, 2011, 2013; Bierbrauer and Netzer, 2016)—only that we focus on the preferences of the principal and not the agent. Perhaps closest to our paper is that of Chassang and Zehnder (2016) who show that a principal with ex-ante fairness preferences later updates beliefs about behavior and has qualms about punishing. Since their model only features one active player, it can, however, not be used to study the central questions of our paper on how declaring and later attributing responsibility to group members affects incentives and performance.

In summary, we contribute to the literature by proposing a novel free rider problem, by rationalizing why real responsibility 'disappears' when a group is formally responsible, and by identifying the importance of immunity of those not declared responsible for formal declarations to result in real responsibility.

[7]For other explanations why people behave differently in groups, see, e.g., Huck and Konrad (2005), Lindbeck et al. (1999), Dufwenberg and Patel (2017) or Rothenhäusler et al. (2018).

[8]See, Baliga and Sjöström (1998) for an example close to ours, Mookherjee (2006) for an excellent overview and Bandiera et al. (2021) for a recent field experiment.

2 Model

For bringing out the difference between reputation free riding and the free riding in the canonical team incentive model by Holmström (1982), we use his model as a starting point. A principal P (she) employs agents (he), $i \in N = \{1, \ldots, n\}$ with $n \geq 2$, who then produce some joint output.

Principal's proposal and agents' participation

The principal P initially proposes a project consisting of an outcome target $\hat{y}$ for the team, a bonus $\hat{b}_i$ promised to each agent i, and a declaration whether this agent is formally responsible, $\hat{r}_i = 1$, or not, $\hat{r}_i = 0$. In line with our leading example, $\hat{b}_i$ can also be interpreted as the promised backing that the principal is willing to give to agent i, for example, when the agent's yearly bonus is determined or he is considered for promotion. We summarize the promised backing and declarations: $\hat{\mathbf{b}} = (\hat{b}_1, \ldots, \hat{b}_n)$ and $\hat{\mathbf{r}} = (\hat{r}_1, \ldots, \hat{r}_n)$.

Taking charge, contributions, and output

Holmström (1982) implicitly assumes that agents coordinate their contributions. We render this assumption explicit with the simplest possible structure. We assume that initially one agent k is drawn randomly, where all agents have the same chance to be drawn. This might be the team member who first thinks about the project or first finds time to work on it. This agent may then 'take charge', positively contribute to the team $c_k > 0$ and ask all other agents $l \in N \backslash \{k\}$ to contribute $\hat{c}_l \in \mathbb{R}_{\geq 0}$. If agent k does not 'take charge', some other agent is randomly selected (where all remaining agents are equally likely to be drawn). This agent is then given the opportunity to contribute and ask for contributions. This process is repeated until finally one agent takes charge or all agents had at least the chance to take charge. If some agent k took charge, we collect his requests in a vector $\hat{\mathbf{c}} = (\hat{c}_1, \ldots, \hat{c}_n)$, where $\hat{c}_k = 0$ reflects

that nothing is requested of the agent k in charge. Likewise, $\hat{\mathbf{c}} = (0, \ldots, 0)$ represents the requests if no one has been asked because no one took charge.

Next, all agents who have not yet contributed sequentially decide on their contributions $c_i \in \mathbb{R}_{\geq 0}$, resulting in the vector $\mathbf{c} = (c_1, \ldots, c_n)$.

Imposing a sequential structure with perfect information, we ensure that no agent needs to fear that his contribution is wasted due to communication failure—unlike in the context of the 'volunteers dilemma', where players move simultaneously and may 'free ride' because of this fear (See, e.g., Diekmann, 1985; Harrington, 2001; Krueger and Massey, 2009; Sliwka, 2006).

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 more than one agent are necessary to produce output. We reflect this by assuming that $y(c) = 0$ unless $c_i > 0$ and $c_j > 0$ for at least two distinct agents $i, j \in N$, with $i \neq j$. The specific functional form of the production function does not matter as long as it 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 $c_i^{\text{FB}} > 0$.[9] This maximizer later turns out to be the first-best contribution (see Lemma 4 in the appendix), hence the name. For eliminating any insurance motives, output is deterministic. The essence of all results, however, is preserved even with stochastic output.[10]

Principal's scope for incentives

Alchian and Demsetz (1972) argue that one characteristic of teams is that outsiders (like the principal) do not know what happens within the team. At the same time, the principal believes the team to be capable to overcome coordination problems (otherwise there would be no point in providing incentives). We reflect both by assuming that the principal neither observes the

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

[10] Only Theorem 2 and Corollary 4 are affected. For the principal to still provide incentives when output is stochastic, output must not be too volatile and the principal must be willing to accept occasionally and accidentally blaming an agent who is not really responsible.

communication between agents, $\hat{\mathbf{c}}$, nor their contributions $\mathbf{c}$, or the sequences of moves (the realizations of the random draws) but knows the structure of the *team game* that starts after her proposal $(\hat{y}, \widehat{\mathbf{b}}, \widehat{\mathbf{r}})$.

The output y cannot be verified by third parties and hence not be purchased on a market. This is why the organization created the team. The principal can observe whether team output meets a pre-specified outcome target $\hat{y}$ and then decide on $\mathbf{b} = (b_1, \ldots, b_n)$, i.e. on how much to back agents or which bonus to give— see Figure 1.

In case of success, the principal cannot give agent i less backing b_i than she promised: $b_i \geq \hat{b}_i$ if $y \geq \hat{y}$. This assumption could be seen as a reduced form of a relational contract between the principal and the agent, where future interactions discipline the principal to reward the agent.

In case of failure, the principal may also be restricted in her choice of b_i. Consider an organisation, where P can only blame agents that are formally declared responsible ($\hat{r}_i = 1$). Let's distinguish this case ($\omega = 1$) from the more traditional case in which P is unrestricted ($\omega = 0$), i.e., she can blame for failure whoever she wants (or equivalently, she can withhold the bonus from anyone).

Summarizing these restrictions on the principal's choice of b_i, we get:

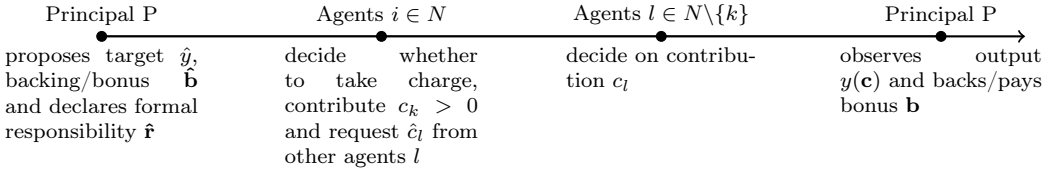
$$b_i \geq \hat{b}_i \cdot \begin{cases} 1 & \text{if } y \geq \hat{y} \\ (1 - \hat{r}_i) \cdot \omega & \text{if } y < \hat{y}. \end{cases} \tag{1}$$

These restrictions impose a lower bound on the principal's backing (or bonus). She is, of course, free to provide more backing (or higher bonus payments) if she wants to.

Payoffs to Agents

Contributing c_i is costly for agents. Following up on the observation that members 'help each other' in teams, we allow for agents to be willing to

Figure 1: Sequence of Moves (Agents may leave at any time)



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 the 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 $\gamma_k = 0$.

Together with the enjoyment from being backed b_i and the costs of contributing c_i, agent i's utility becomes:

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

An agent i who has not received a request does not feel 'pain' from failing to meet the request: if $\hat{c}_i = 0$, we get $c_i \geq \hat{c}_i$, and the condition for suffering pain, $c_i < \hat{c}_i$, is never met. For resolving indifference between joining or leaving the project, we assume that an agent who leaves before contributing, incurs an arbitrarily small costs of $\epsilon > 0$ and forgoes the backing.

Real responsibility

A key aspect of our model is that the principal might care about who is actually responsible. Before we can incorporate this aspect into the principal's utility function, we need to define what we mean by 'actual responsibility'.

A necessary condition for a person to be responsible for failure is that the person has caused failure[11] in the sense that her choices made a difference to the outcome (Hume, 1748; Lewis, 1974).[12] In order to know what happens after a choice, we predict the outcome following that choice by player $i \in \{P\} \cup N$ using the Perfect Bayesian Equilibrium (PBE) concept.[13] We have to deal with the problem that a specific choice, say x, may result in an outcome, say Y, as well as the opposite of this outcome, Y^C, because multiple PBEs may follow x. We can thus only say that x caused Y if all PBEs consistent with this choice result in Y.

Definition 1 (Causality)**.** *Let outcome Y be a set of outputs $y \in Y \subset \mathbb{R}_{\geq 0}$ and the complementary set $Y^C := \mathbb{R}_{\geq 0} \backslash Y$ be the counterfactual outcome. Player i's choice of x causes outcome Y if*

- *this choice leads to an outcome Y, i.e., all PBEs result in some $y \in Y$.*
- *another choice x^c may lead to the counterfactual outcome Y^C, i.e., some PBE involving x^c result in $y \in Y^C$.*

Although inspired by well-known fundamental ideas in the philosophy of science, this definition deviates from those typically proposed (Lewis, 1974;

[11] The link between responsibility and causality is backed by various experiments. 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.

[12] 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."

[13] The ordinary Nash equilibrium concept does not suffice because the principal does not know $\hat{c}$ or c and the realized sequence of moves in the team game.

Halpern and Pearl, 2005) because we do not allow for arbitrary future behavior following a choice but restrict ourselves to behavior that can be expected in equilibrium —similar to Engl (2018).[14] This avoids a host of problems that causality definitions based on counterfactuals typically run into (see Collins et al., 2004).

While causality is crucial for responsibility, a person might cause failure but not be responsible for it. Consider the example of Linda and Robert who need to submit a report on a specific day. In the morning of that day, Linda has prepared everything for Robert because she has an important meeting with a client in the afternoon. Robert has ample time to finalize and submit the report. Although Robert anticipates that the meeting is so important for Linda that she will not sacrifice it for the submission, he leaves the office before Linda's meeting starts. By leaving, he causes the submission to fail because Linda is not going to cancel her meeting. If Linda attends her meeting as planned, however, she also causes failure because this means that the report is not finalized. Although both cause failure, only Robert is really responsible. Since cancelling the meeting is prohibitively costly for Linda, no one expects her to make up for Robert's slack. Robert, on the other hand, can be expected to finalize the report. The example shows that in addition to causing failure, an agent must also be expected to produce success to be responsible for failure.

Definition 2 (Real responsibility)**.** *Let $\hat{y}$ be some output target. Player i is really responsible for failure, $\phi_i = 1$ rather than $\phi_i = 0$, whenever i's choice of x causes failure and some x^c leading to success is consistent with a PBE.*

Payoffs for the principal

Equipped with a notion of real responsibility, we are now in the position to write down the principal's utility. The principal benefits from output y and

[14] Unlike him, we do not establish gradual causality on the basis of how far a player must deviate to cause a different outcome.

has costs from backing 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 withheld some of the promised backing, $b_i < \hat{b}_i$ and $\mathbb{1}_{[0,\hat{b}_i)}(b_i) = 0$ that she did not. Sanctioning agent i feels wrong to the principal if agent i is not responsible for failure, $\phi_i = 0$.

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

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

With this utility function, a principal who cares enough about real responsibility is unwilling to sanction an innocent agent—see Lemma 1, later.[15]

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 ability to elicit help γ_j of agent $j = 1, \ldots, n$ and how much the principal cares about real responsibility κ.

3 Analysis

Our analysis has two aims. First, we want to shed light on why giving responsibility to a whole team is problematic. Second, we want to understand why and when declaring one agent responsible can overcome this problem.

The analysis proceeds in five steps. Section 3.1 and Section 3.2 set the scene

[15] 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.

by linking to Holmström's model and showing that a principal who sufficiently cares is unwilling to collectively punish (Proposition 1) and hence cannot use his scheme (Corollary 1). Section 3.3 rationalizes why team members must help each other for the team to be successful (Proposition 2) and why declaring several members responsible is problematic (Theorem 1) . With Section 3.4 we get to the heart of the analysis. We identify that declaring one (and just one) member formally responsible only translates to real responsibility and solves the problem under specific conditions: it must be impossible to sanction people who are not formally responsible (Theorem 2). Up to this point the analysis will only have dealt with the case that none or many members can elicit sufficient support to achieve the team target. For completeness, Section 3.5 looks at the missing case in which only one member of the team has this ability and finds that this member is always really responsible—irrespective of formal declarations (Corollary 5).

3.1 The externality problem and Holmström's solution

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 b_i 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 but the respective costs are incurred by the agents. Notice that each agent may well be able to produce the output by taking charge and getting the others to contribute. Still, they are not willing. The externality has to be internalized at least partially for some output to be produced.[16]

[16]Recall that the benefit y is generated within the organization and cannot be traded. Otherwise, free riding might be avoided by 'selling the shop' to one agent who then elicits contributions of the others.

Holmström (1982) famously suggested to solve the incentive problem by sanctioning all agents in case of failure. In our setting, Holmström's scheme can be represented as follows. The principal sets the first-best output as a target $\hat{y} = y^{\text{FB}}$, promises a bonus or backing 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 an institution that requires a formal declaration before agents can be punished ($\omega = 1$), a principal who wants to implement Holmström's solution has to declare all team members responsible whose contributions are required (Lemma 7 in the appendix). Declaring all agents responsible, which is seen as the root of a free rider problem by management scholars, is thus necessary for overcoming this problem using Holmström's solution.

This solution works in our model if the principal does not care about real responsibility ($\kappa = 0$)—see Corollary 6 in the appendix. 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.

3.2 Limits of collective punishment

A principal who is interested in real responsibility ($\kappa > 0$) faces a dilemma. Backing the agent $\hat{b}_i$ is costly but so is blaming an 'innocent' agent. If the principal cares enough about real responsibility or is relatively certain that an agent is not responsible for failure, she will back the agent even after failure.

Lemma 1. *After failure, $y < \hat{y}$, the principal does not withdraw $\hat{b}_i$ from agent i if she deems it unlikely that i is really responsible for this failure:*[17]

$$b_i \geq \hat{b}_i \Leftrightarrow P(\phi_i = 1) \leq 1 - \frac{\hat{b}_i}{\kappa}.$$

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

Proof. Recall the principal's utility:

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

The principal always provides the lowest possible backing. If she wants to meet the promise to agent i, this is $b_i = \hat{b}_i$. In case of failure, she then keeps to the promise, 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 fulfilling the promise 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. In other words, they are not really responsible.

Lemma 2 (at most one responsible player)**.** *Suppose $y^* \geq \hat{y}$ is supported by some PBE. Then, at most one player is really responsible for failure, $y < \hat{y}$:* $\sum_i \phi_i \leq 1$.

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 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 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 impossible to punish collectively.

Proposition 1 (No collective punishment)**.** *In the team game* $(\hat{y}, \widehat{\mathbf{b}}, \widehat{\mathbf{r}})$*, a principal does not withdraw the backing/bonuses of more than one agent after failure,* $y < \hat{y}$*, if she cares enough about responsibility,* $\kappa > 2 \cdot \max_i \hat{b}_i$*.*

Proof. P is only willing to withdraw the backing from any agent i from a group a group $\tilde{N} \subseteq N$ of size $\tilde{n} > 1$ if the gains from not backing each agent $i \in \tilde{N}$ more than outweigh the loss from hurting an agent who is not really responsible $\hat{b}_i \geq \kappa \cdot (1 - \phi_i)$ for all $i \in \tilde{N}$—see Proof of Lemma 1. Adding the inequalities for all $\tilde{n} \in \tilde{N}$ agents, we get $\kappa \cdot \sum_{i \in \tilde{N}} (1 - \phi_i) \leq \sum_{i \in \tilde{N}} \hat{b}_i$ or $\kappa(\tilde{n} - \sum_i \phi_i) \leq \sum_{i \in \tilde{N}} \hat{b}_i$. By Lemma 2, $\sum_{i \in N} \phi_i \leq 1$, so that $\sum_{i \in \tilde{N}} \phi_i \leq 1$ and we get the following necessary condition for punishing group $\tilde{N}$:

$$\kappa(\tilde{n} - 1) \leq \sum_{i \in \tilde{N}} \hat{b}_i. \tag{4}$$

From $\kappa > 2 \cdot \max_i \hat{b}_i$ and $\max_{i \in N} \hat{b}_i \geq \frac{1}{\tilde{n}} \sum_{i \in \tilde{N}} \hat{b}_i$ for any set $\tilde{N} \subseteq N$, it follows that $\kappa > \frac{2}{\tilde{n}} \sum_{i \in \tilde{N}} \hat{b}_i$. Using that $\frac{2}{\tilde{n}} \geq \frac{1}{\tilde{n}-1}$ for $\tilde{n} > 1$, we get $\kappa > \frac{1}{\tilde{n}-1} \sum_{i \in \tilde{N}} \hat{b}_i$. The necessary condition (4) can thus not be met and the principal never punishes more than one agent. □

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 provides justification for why 'it is better that ten guilty persons escape, than that one innocent suffers' (Blackstone, 1765-1770, p. 358). A direct 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,* $\kappa > 2 \cdot \max_i \hat{b}_i$ *will not employ Holmström's scheme.*

Proof. The proof follows immediately from observing that Holmström's scheme requires punishing more than one agent, which a sufficiently caring principal is unwilling to do by Proposition 1. □

While a sufficiently caring principal cannot rely on Holmström's scheme, there may be other ways to avoid free riding. The next section shows that any such way requires that agents are willing to help each other but that this willingness is not sufficient.

3.3 Reputation free riding and diffusion of responsibility

Since contributions of multiple 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 ask others to contribute.

Lemma 3 (One agent needs to take charge)**.** *In the team game* $(\hat{y}, \widehat{\mathbf{b}}, \widehat{\mathbf{r}})$*, with a principal who cares sufficiently about real responsibility,* $\kappa > 2 \cdot \max_i \hat{b}_i$*, 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 some other agent* l*.*

Proof. The proof works by contradiction. Suppose no agent takes charge. Then, $\hat{c}_i = 0$ for all i, 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, at least two agents i and j need to contribute for positive output to be produced: $y > 0 \Rightarrow c_i > 0$ and $c_j > 0$ for some $i \neq j$. A positive output thus requires that these two agents are sanctioned in case of failure. Since $\kappa > 2 \cdot \max_i \hat{b}_i$, the principal is unwilling to collectively punish more than one agent by Proposition 1. Accordingly, positive output can only be achieved in equilibrium if at least one agent takes charge. □

While asking for help is an important step, this help might not be forthcoming if the member who asks 'has no credit' with his team mates.

Definition 3. *Team member i has sufficient power γ_i to elicit contributions $\widetilde{\mathbf{c}}$ whenever $\gamma_i \tilde{c}_i > \tilde{c}_j$ for all members $j \neq i$.*

Members might not be willing to contribute enough irrespective of who is asking them. Then, none of the agents can ensure success and hence be really responsible for causing failure, and a principal who cares about real responsibility cannot provide incentives.

Proposition 2 (Free Rider Problem When Agents are Unable to Elicit Help)**.** *Consider a team with a positive target $\hat{y} > 0$ that has been promised $\widehat{\mathbf{b}}$ by a principal who cares sufficiently about real responsibility, $\kappa > 2 \cdot \max_i \hat{b}_i$. If no agent has sufficient power to elicit the smallest contributions $\widetilde{\mathbf{c}}$ that could result in participation, $\widehat{\mathbf{b}} \geq \widetilde{\mathbf{c}}$,[18] as well as success, $y(\widetilde{\mathbf{c}}) = \hat{y}$, real responsibility diffuses, $\phi_i = 0$ for all i and success cannot be produced, $y^* < \hat{y}$.*

Proof. In order to produce success with minimal contributions $y(\tilde{c}) = \hat{y}$ agents must be willing to participate. Since they get at most $\hat{b}_i$, this must be sufficient to compensate for their costs $\tilde{c}_i$, or more succinctly $\widehat{\mathbf{b}} \geq \widetilde{\mathbf{c}}$. Using $\kappa > 2 \cdot \max_i \hat{b}_i$ in Lemma 3, some agent k needs to take charge for any positive output $\hat{y} > 0$ to be generated in the team game. Fix an arbitrary agent k who takes charge. Then, the lack of sufficient power of this agent implies that for every $\tilde{c}$ that might result in participation, $\widehat{\mathbf{b}} \geq \widetilde{\mathbf{c}}$, as well success, $y(\tilde{c}) > \hat{y}$, there is at least one agent l with $\tilde{c}_l > \gamma_k \tilde{c}_k$. Fix an arbitrary $\tilde{c}$ and distinguish between agents who contribute $N = \{i | \tilde{c}_i > 0\}$ and those who do not $N^C = \{i | \tilde{c}_i = 0\}$. For all agents who do not contribute, $l \in N^C$, $\tilde{c}_l \leq \gamma_k \tilde{c}_k$ follows from $\tilde{c}_l = 0$. Since $\tilde{c}_l \leq \gamma_k \tilde{c}_k$ for all $l \in N^C$, $\tilde{c}_l > \gamma_k \tilde{c}_k$ for some $l \in N$ who positively contributed. Lemma 5 (see appendix) then implies that agent l does not contribute, i.e., $c_l^* = 0$—even if the request $\hat{c}_l$ is the lowest that still ensures success and participation $\hat{c}_l = \tilde{c}_l$. Since output is continuous and increasing in the contribution of l, $\hat{y} = y(\tilde{c}) > y(c^*) = y^*$ and success cannot be reached. By Definition 2, no agent is then really responsible for failure. □

[18] Here, this notation describes the component-wise comparison, i.e., $\hat{b}_i \geq c_i$ for all i.

The proposition shows us that being able to elicit help is necessary to avoid free riding. After dealing with the case of no willingness to help, 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* $\hat{y}$ *(using* $\tilde{\mathbf{c}}$*) if all agents* $i \in N$ *are able to elicit contributions* $\tilde{\mathbf{c}} : \gamma_i \tilde{c}_i > \tilde{c}_j$ *for all* $j \neq i$*.*

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 once more. 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 with the opportunity to take charge. In order to be really responsible, this last agent must also have an incentive to take charge. His backing must hinge on output. Since the principal cares about real responsibility, she is unwilling to withdraw the backing from anyone but this agent. As an outsider to the team, however, she does not know who this agent is and will back all agents. This in turn means that all agents ride free and indeed cannot be expected to contribute. Consequently, not even the last agent is really responsible.

Theorem 1 (Reputation Free Riding and the Diffusion of Responsibility)**.** *Consider a team with a positive target* $\hat{y} > 0$ *that has been promised backing/bonuses* $\widehat{\mathbf{b}}$*, 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:*

$$\kappa > 2 \cdot \max_i \hat{b}_i \quad \Rightarrow \quad \phi_i^* = 0, c_i^* = 0, \text{ for all } i \text{ and } y^* < \hat{y}.$$

Proof. The proof works by contradiction. Suppose that an equilibrium with $\mathbf{c}^*$ such that $y^* = y(\mathbf{c}^*) \geq \hat{y} > 0$ exists. Using $\kappa > 2 \cdot \max_i \hat{b}_i$ in Lemma 3, one agent l has to take charge. Observe that for a positive production level, all agents' participation constraints must be met in equilibrium, so $c_k^* \leq \hat{b}_k$ 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, any agent k can request $\hat{c}_l = c_l^*$ from agents $l \in N \backslash \{k\}$ and this request is met favorably if $c_k = c_k^*$: $c_l = \hat{c}_l = c_l^*$ by Lemma 5 in the appendix. Any agent k can thus ensure $y(\mathbf{c}^*) \geq \hat{y} > 0$. Using that $\hat{r}_k = 1$ for all k in Equation (1) implies that the principal may withhold the promised backing from any agent k:

$$b_k \geq 0 \text{ if } y < \hat{y}. \tag{5}$$

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, say agent $\underline{k}$, contributes $c_{\underline{k}}^*$ and elicits $\hat{c}_l = c_l^*$. 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 either 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., not participating or contributing $c_l < \hat{c}_l$. In equilibrium, agents must participate, i.e., $b_l \geq c_l^*$. This, however, implies that agent l loses at least $b_l - c_l^* + \epsilon > 0$ if he does not participate. By Lemma 5, participating but not contributing $\hat{c}_l$ also results in losses. Put differently, participating and meeting the request is a strictly dominant strategy for agent l. 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 backing 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: $P(\phi_i = 1) \leq \frac{1}{n}$ and $\kappa\left(1 - P(\phi_i = 1)\right) \geq \kappa\left(1 - \frac{1}{n}\right)$. Using that the principal cares enough, we have $\kappa\left(1 - \frac{1}{n}\right) > 2 \cdot \max_i \hat{b}_i \cdot \frac{n-1}{n} \geq \max_i \hat{b}_i$, where the last inequlity follows from $2(n-1) \geq n$ for $n \geq 2$. All in all, we get $\kappa\left(1 - P(\phi_i = 1)\right) > \max_i \hat{b}_i \geq \hat{b}_i$ and by Lemma 1, the principal is then 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 backing of any agent after failure. The reason why the principal can take away the backing 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, $\omega = 0$, which leads to the following corollary.

Corollary 2. *Consider a team with a positive target $\hat{y} > 0$ that has been promised backing/bonuses $\widehat{\mathbf{b}}$ and is committed to success operating in an institution where no declaration is necessary to sanction members ($\omega = 0$). 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:

$$\kappa > 2 \cdot \max_i \hat{b}_i \quad \Rightarrow \quad \phi_i^* = 0, c_i^* = 0, \text{ for all } i \text{ and } y^* < \hat{y}.$$

Proof. The proof is analogous to that of Theorem 1, where inequality (5) 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), 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 backing in case of failure from agents who are not formally responsible. This subtle shift is indeed key for overcoming free riding.

3.4 When formal translates into real responsibility

For seeing how declaring one agent responsible might help, consider an institution in which only formally responsible agents can lose the principal's backing. If the principal then assigns responsibility to only one agent, the backing for all other agents is guaranteed and they cannot be expected to take charge and indeed will not take charge. They are, however, happy to contribute as response to a reasonable request, for example, by the formally responsible agent. This agent then cannot 'pass on' the real responsibility for failure to some other agent. He has to ensure success himself by taking charge, contributing and requesting contributions. The guaranteed backing is crucial for the formal assignment to work. If the principal can withdraw the backing from several agents, the declaration becomes cheap talk, real responsibility diffuses and success cannot be implemented for the same reasons as in Theorem 1.

Theorem 2 (Preventing reputation free riding)**.** *Consider a team with a positive target $\hat{y} > 0$ that is committed to success using $\tilde{\mathbf{c}}$, has been promised $\widehat{\mathbf{b}}$, which would adequately compensate for contributions $\hat{b}_i \geq \tilde{c}_i$, and a principal who sufficiently cares about real responsibility, $\kappa > 2 \cdot \max_i \hat{b}_i$. Suppose the principal declares a single agent k formally responsible, $r_k = 1$ and $r_l = 0$ for all $l \neq k$. Then, the formally responsible agent k will take charge, elicit contributions and ensure that success is implemented but 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 backs any agent $l \neq k$, $b_l^* = \hat{b}_l$, while agent k only gets the backing in case of success $b_k^* = \hat{b}_k$, $y \geq \hat{y}$ and $b_k^* = 0$ otherwise.

We will now check whether this candidate is actually an equilibrium, starting with the principal's decision given her beliefs. Given $P(\phi_l = 1) = 0$ for all $l \neq k$, backing agent l is optimal for the principal. On the other hand, the principal maximizes her utility by withdrawing agent k's backing 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 he does not participate or his contribution is below the request $c_l < \hat{c}_l$. Participating and 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.

Restricting the principal's ability to blame is thus 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 with a positive target $\hat{y} > 0$ that is committed to success using $\tilde{\mathbf{c}}$, has been promised $\widehat{\mathbf{b}}$, which would adequately compensate for contributions $\hat{b}_i \geq \tilde{c}_i$. Then, a sufficiently caring principal, $\kappa > 2 \cdot \max_i \hat{b}_i$, 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 keeps the promise to any agent $b_i^0 = \hat{b}_i$. Following directly from her utility function, the principal never backs more than promised: $b_i \leq \hat{b}_i$. This implies that the backing 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 several agents contribute. Put differently, any 'essential' agent can sabotage 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)**.** *Consider a team with a positive target $\hat{y} > 0$ that is committed to success using $\tilde{\mathbf{c}}$, has been promised $\widehat{\mathbf{b}}$, which would adequately compensate for contributions $\hat{b}_i \geq \tilde{c}_i$, and a principal*

who sufficiently cares about real responsibility, $\kappa > 2 \cdot \max_i \hat{b}_i$*, and who only declared agent* k *formally responsible,* $r_k = 1$ *and* $r_l = 0$ *for all* $l \neq k$ *in an organisation with* $\omega = 1$*. Then, success,* $\sigma_i = 1$*, rather than failure,* $\sigma_i = 0$*, is attributed to all agents* $E \subseteq N$ *whose contributions are essential for production,* $E = \{i | c_i = 0 \Rightarrow y = 0\}$*, while failure belongs alone to the formally responsible agent:*

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

Proof. Let us first deal with the attribution of fame. From $y(c) = y \geq \hat{y} > 0$, we get that all agents $i \in E$ must have contributed $c_i > 0$. Consider a change by some $i \in E$ from $c_i > 0$ to $c_i' = 0$. Output then becomes $y = 0$, i.e., success is no longer possible. Agent i has thus caused the outcome. Further, output can only be produced because in equilibrium, the formally responsible agent k has asked the other agents to contribute. Not asking an agent $i \in E$, i.e., $\hat{c}_i = 0$, would have resulted in $c_i = 0$. In other words, k has also caused success and is really responsible. 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 more precise interpretation in which sense 'good' leaders should 'take all the blame, while sharing all fame'. In particular, fame has to be shared with all agents who are deemed essential. Moreover, the 'wisdom' requires that the principal cares about real responsibility, and sanctions are limited to those who are formally responsible.

3.5 Ability to elicit help and responsibility

Earlier we have seen that free riding can arise if no agent (Proposition 2) or all agents can elicit help (Theorem 1). Now suppose that only one agent can elicit help.

Corollary 5 (With greater power comes real responsibility)**.** *Consider a team with a positive target $\hat{y} > 0$ that has been promised backing/bonuses $\widehat{\mathbf{b}}$, which suffice to compensate for contributions $\hat{b}_i \geq \tilde{c}_i$, in which agent k can be sanctioned in case of failure ($r_k = 1$ or $\omega = 0$) and is the only agent with sufficient power to elicit contributions $\tilde{c}$ required for success $y(\tilde{c}) \geq \hat{y}$. Then, agent k is really responsible for failure $\phi_k = 1$ and success can be implemented.*

Proof. Any agent $l \neq k$ cannot elicit the necessary contributions from the others required for success. There cannot be an equilibrium in which they take charge. If agent k's backing 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 is able to elicit contributions—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 backing, which is possible either because $r_k = 1$ or $\omega = 0$. □

If power is interpreted as the ability to elicit the cooperation 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 are sufficiently able to elicit help, 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).

Taken together, this can be interpreted as follows. A team with a clear leader, i.e., only one agent who can elicit support, does not require a formal declaration of responsibility. Adding another leader-type to the team, i.e., a second agent who can elicit support, leads to a free-rider problem—very much in line with the saying that too many cooks spoil the broth.[19] Then, declaring one of them formally responsible helps overcoming the problem.

4 Discussion and Implications

Let us summarize our main results and what they imply in terms of free riding. A supervisor who does not particularly care about real responsibility, can employ collective punishment to counter free riding—see top line in Table 1. Our rationale for the management wisdom on commitment in teams and taking blame thus crucially relies on managers being sufficiently concerned about real responsibility. This naturally leads to the question why and when managers may be worried about harming innocent employees.

The first and most direct reason is that managers genuinely dislike punishing someone who has not done anything wrong. Such a dislike permeates most societies and organizations. This ranges from parents being unwilling to punish their child for acts they have not committed (Shavell, 2023) to the presumption of innocence until proven guilty, which is a corner stone of most legal systems and which only makes sense together with the legal principle that people cannot be punished for others' actions or omissions ('Nulla poena since culpa').[20] Philosopher McCloskey (1963, 1965) appeals to a 'common moral consciousness' to which punishment of innocent people or 'persons not responsible for their actions' is unacceptable. Since managers are also human beings, their decisions

[19] For a very different take on this saying in an incentive context see Ratto and Schnedler (2008).

[20] Article 33 of the Geneva Convention (August 12, 1949) affirms this principle and the European Court of Justice frequently refers to it (see, for example, the court's ruling in case Case C-210 / 00). The German Supreme Court recognizes the principle as 'constitutional' (Wolff, 1999, p. 56).

are likely to be affected by this moral consciousness. Another reason why supervisor might not want to punish innocent subordinates is that employees may feel unfairly treated and reciprocate. Management scholars have long been emphasizing the importance of justice for the well-functioning of organizations (Greenberg, 1990). This includes specifically the positive effects of being able to give a causal account (Bies and Shapiro, 1987; Bies, 1987). Indeed, being able to accept responsibility for substandard performance seems key for employees to respond well to negative feedback (Ilgen and Davis, 2000). In addition, managers find it hard to communicate bad news to subordinates, for example, when giving negative feedback (Larson Jr, 1984; Smith et al., 2000; Villanova et al., 1993). Not supporting the employee is a decision that ultimately has to be communicated. This communication is easier if the feedback appears 'rational, legitimate, objective, and reasonable,' the so-called evidence tactic (Brown et al., 2016). Anticipating the need to communicate their decision, managers may shy away from a withdrawal of support if this cannot be linked to evidence of under-performance. Finally, a manager may not want to discipline an innocent subordinate to avoid formal consequences. Codes of conduct typically clearly require the documentation of misconduct of an employee before disciplinary action against this employee can be taken. The subordinate may appeal or even threaten with legal action, both of which come with costs to the supervisor. While there certainly are managers without qualms, all this suggests that many managers will find it hard to withdraw support unless there is sufficient evidence that the subordinate has caused poor performance. Such managers cannot use collective punishments to overcome free riding problems.

Being unable to counter free riding with collective punishment also has implications for the provision of public goods. In numerous public good experiments (See Chaudhuri, 2011, for an overview) group members benefit from the contributions of others. 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. According to our analysis, such institutions become pointless if they cannot identify who failed to contribute and are unwilling to sanction possibly innocent members (Theorem 1). Like the principal in our model, they would then shy away from sanctioning in case of failure. Group members would foresee this, stop contributing and the public good could not be provided.

principal cares sufficiently about responsibility	agents able to elicit sufficient contributions: none	one	more than one
no	collective punishment can prevent free riding (Holmström's Theorem 2)		
yes	free riding cannot be prevented (Proposition 3) *Commitment is key.*	if agent can be sanctioned, he is really responsible no free riding occurs (Corollary 7) *With power comes responsibility.*	free riding can be prevented by declaring one member responsible if and only if all others cannot be sanctioned (Theorem 2 and Corollary 9) *Taking blame, sharing fame.*

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

Given the novel free rider problem, commitment within the team starts to matter (as suggested by management wisdom)—see bottom line in Figure 1. Without commitment, free riding is unavoidable—see bottom left corner. The common advice to declare one agent responsible works whenever several members are capable of obtaining support from others and anyone who is not declared responsible can also not be punishment. Finally, if just one person has sufficient power to elicit contributions and can be sanctioned, the unique position of power of this person implies real responsibility.

The implication for management is that the advice of declaring one group member responsible only works in a suitable context: team members must be sufficiently willing to help each other and those who are not declared

responsible must be sure not to be negatively affected in case of failure.

References

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

Alchian, Armen A. and Harold Demsetz, "Production, Information Costs, and Economic Organization," *The American Economic Review*, December 1972, *62* (5), 777–795.

Antonakis, John, Giovanna d'Adda, Roberto A Weber, and Christian Zehnder, ""Just words? Just speeches?" On the economic value of charismatic leadership," *Management Science*, 2022, *68* (9), 6355–6381.

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, "Relative performance or team evaluation? Optimal contracts for other-regarding agents," *Journal of Economic Behavior & Organization*, 2011, *79* (3), 183–193.

_ **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.

Bierbrauer, Felix and Nick Netzer, "Mechanism design and intentions," *Journal of Economic Theory*, 2016, *163*, 557–603.

Bies, RJ, "The predicament of injustice: The management of moral outrage," *Research in organizational behavior/JAI Press*, 1987.

Bies, Robert J. and Debra. L. Shapiro, "Interactional Fairness Judgments: The influence of Causal Accounts," *Social Justice Research*, 1987, *1*, 199–218.

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

Bradler, Christiane, Robert Dur, Susanne Neckermann, and Arjan Non, "Employee recognition and performance: A field experiment," *Management Science*, 2016, *62* (11), 3085–3099.

Brown, Michelle, Carol T. Kulik, and Victoria Lim, "Managerial tactics for communicating negative performance feedback," *Personnel Review*, 2016, *45* (5), 969–987.

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.

Castro, Silvia, Florian Englmaier, and Maria Guadalupe, "Fostering psychological safety in teams: Evidence from an rct," *Available at SSRN 4141538*, 2022.

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.

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

Collins, John, Ned Hall, and Laurie Ann Paul, "Counterfactuals and Causation: History, Problems and Prospects," in "Causation and Counterfactuals," MIT Press, 2004.

Comittee of Public Safety, "Plan de Travail, de Surveillance et de Correspondance," in "Collection Générale des Décrets Rendus par la Convention Nationale," Paris: Baudouin, May 1793.

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.

Czura, Kristina, Florian Englmaier, Hoa Ho, and Lisa Spantig, "Employee performance and mental well-being: The mitigating effects of transformational leadership during crisis," *Management Science*, forthcoming.

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

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.

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.

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

Greenberg, Jerald, "Organizational justice: Yesterday, today, and tomorrow," *Journal of management*, 1990, *16* (2), 399–432.

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.

Haeckl, Simone and Mari Rege, "Effects of Supportive Leadership Behaviors on Employee Satisfaction, Engagement, and Performance: An Experimental Field Investigation," *Management Science*, forthcoming.

Halpern, Joseph Y and Judea Pearl, "Causes and explanations: A structural-model approach. Part I: Causes," *The British journal for the philosophy of science*, 2005.

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.

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.

Hume, David, *An Enquiry Concerning Human Understanding*, LaSalle, Illinois: Open Court Press, 1748. Reprinted in 1958.

Ilgen, Daniel and Cori Davis, "Bearing bad news: Reactions to negative performance feedback," *Applied Psychology*, 2000, *49* (3), 550–565.

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.

Jr, James R Larson, "The performance feedback process: A preliminary model," *Organizational behavior and human performance*, 1984, *33* (1), 42–76.

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.

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.

Manthei, Kathrin, Dirk Sliwka, and Timo Vogelsang, "Talking about performance or paying for it? A field experiment on performance reviews and incentives," *Management Science*, 2023, *69* (4), 2198–2216.

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.

McCloskey, Henry John, "A note on Utilitarian Punishment," *Mind*, October 1963, *72* (288), p. 599.

_ , "A non-utilitarian approach to punishment," *Inquiry*, 1965, *8* (1-4), 249–263.

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.

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

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

Ratto, Marisa and Wendelin Schnedler, "Too Few Cooks Spoil the Broth: Division of Labour and Directed Production," Technical Report 468, Department of Economics, University of Heidelberg 2008.

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

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

Shavell, Steven, "On the Law of the Household: The Principles Used by Parents in Disciplining Their Children," Technical Report, National Bureau of Economic Research 2023.

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

Smith, Wanda J, K Vernard Harrington, and Jeffery D Houghton, "Predictors of performance appraisal discomfort: A preliminary examination," *Public Personnel Management*, 2000, *29* (1), 21–32.

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

Villanova, Peter, H John Bernardin, Sue A Dahmus, and Randi L Sims, "Rater leniency and performance appraisal discomfort," *Educational and psychological measurement*, 1993, *53* (3), 789–799.

von Siemens, Ferdinand A, "Heterogeneous social preferences, screening, and employment contracts," *Oxford Economic Papers*, 2011, *63* (3), 499–522.

_ , "Intention-based reciprocity and the hidden costs of control," *Journal of Economic Behavior & Organization*, 2013, *92*, 55–65.

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

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

Wolff, Heinrich A., "Der Grundsatz 'nulla poena sine culpa' als Verfassungsrechtssatz," *Archiv des öffentlichen Rechts*, 1999, *124* (1), 55–86.

Appendix

Lemma 4 (First-best contributions)**.** *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:

$$\max_{c,b} y(c) - \sum_{i \in N} b_i - \kappa \sum_{i \in N} \mathbb{1}_{[0,\hat{b}_i)}(b_i)\,(1 - \phi_i) \tag{6}$$

$$\text{s.t. } b_i - c_i - \left\{ \begin{array}{rl} \gamma_j c_j & \text{for } c_i < \hat{c}_i \\ 0 & \text{otherwise.} \end{array} \right\} \geq -\epsilon \tag{7}$$

Observe that the costs from not meeting the promise, $\mathbb{1}_{[0,\hat{b}_i)}(b_i)\,(1 - \phi_i) > 0$, can be avoided without doing harm to the agent by setting the first-best promise to the first-best backing / bonus $\hat{b}_i^{\text{FB}} = b_i^{\text{FB}}$.

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 likewise 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:

$$\max_{c,b,\hat{c},\hat{b}} y(c) - \sum_{i \in N} b_i \text{ s.t. } b_i - c_i \geq -\epsilon \tag{8}$$

This problem is now 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 the new objective function:

$$\max_c y(c) - \sum_i c_i,$$

which is maximized by c_i^{FB}. □

Consider an agent l whose backing does not depend on output. This agent may still contribute as response to a request by agent k. The next lemma establishes when contributing is optimal for this agent.

Lemma 5 (Optimal response to request)**.** *Suppose the principal does not condition the backing / 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*

$$c_l^* = \begin{cases} \hat{c}_l & \text{if } \hat{c}_l \leq \gamma_k c_k \text{ and } \hat{c}_l \leq \hat{b}_l. \\ 0 & \text{otherwise.} \end{cases} \tag{9}$$

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 backing / bonus of $\hat{b}_l$. If he stays, he has the choice between ignoring the request, contributing $c_l < \hat{c}_l$, and getting $u_l(b_l, c_l) = b_l - c_l - \gamma_k c_k$ 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 agent l contributes nothing, $c_l = 0$, whenever the agent leaves the relationship or stays but does not meet the request, $c_l < \hat{c}_l$, and agent l contributes the requested level $c_l = \hat{c}_l$, otherwise. Plugging in these contributions in agent l's utility, we get that the utility from contributing the requested input $\hat{c}_l$ is at least as larger as that of the other two options, $b_l - c_l \geq \max(b_l - \gamma_k c_k, -\epsilon)$ for all ϵ, if and only if $\hat{c}_l \leq \hat{b}_l$ and $\hat{c}_l \leq \gamma_k c_k$. In case that $\hat{c}_l \leq \hat{b}_l$ and $\hat{c}_l < \gamma_k c_k$, agent l is strictly better off from meeting the request for all ϵ. □

Lemma 6 (Necessity of intervention)**.** *If the principal does not condition the backing / bonus for some agent k on output y, agents do not contribute: $b_k \equiv b(y)$ constant in y for all k: $c_l^* = 0$ 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 $\hat{c}_l \leq \min(\gamma_k c_k, \hat{b}_k)$. Take an arbitrary agent k who is in charge an requested $\hat{c}_l > 0$. Notice that agent k has not incentive to contribute because her benefit $\hat{b}_k$ is independent of her contribution. She hence chooses $c_k = 0$, which then implies $c_l = 0$ for all l. □

Lemma 7 (Need to declare all formally responsible)**.** *When members who are not formally responsible cannot be sanctioned ($\omega = 1$), implementing Holmström's incentive scheme requires that all all members whose contributions are requi 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 $b_i \geq \hat{b}_i(1 - \hat{r}_i)$. Holmström's scheme requires a promised bonus, $\hat{b}_i > 0$, that is withdrawn after failure, i.e., $b_i = 0$ for all $i \in N$ if $y < \hat{y}$. This in turn implies that $\hat{b}_i(1 - \hat{r}_i) = 0$, which is equivalent to $\hat{r}_i = 1$ for all $i \in N$ because $\hat{b}_i > 0$. □

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 $\hat{y} = y^{\mathrm{FB}}$, $\hat{b}_i = c_i^{\mathrm{FB}}$, 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 $\hat{b}_i \geq c_i^{\mathrm{FB}}$ 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 $\hat{b}_i = c_i^{\mathrm{FB}}$. 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 $\hat{c}_i > c_i^{\mathrm{FB}}$ will not be met, requests $\hat{c}_i \leq c_i^{\mathrm{FB}}$ impose no constraint. Taking ϵ to zero, principal's earnings approach the first-best rent: $y^{\mathrm{FB}} - \sum_i c_i^{\mathrm{FB}}$. □

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 $G \subseteq N$ consisting of $m \geq 2$ 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:*

$$\kappa > 2 \cdot \max_i \hat{b}_i : \phi_i^* = 0, c_i^* = 0, \text{ for all } i \text{ and } y^* = 0 < \hat{y}.$$

Proof. The proof follows that of Theorem 1. From $\kappa > 2 \cdot \max_i \hat{b}_i$ and Lemma 3, it follows that 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). Using this, we get: $\kappa\,(1 - P(\phi_i = 1)) \geq \kappa\left(1 - \frac{1}{m}\right) > 2 \cdot \max_i \hat{b}_i \cdot \frac{m-1}{m} \geq \max_i \hat{b}_i$ because $2(m-1) \geq m \Leftrightarrow m \geq 2$. Using $\kappa\,(1 - P(\phi_i = 1)) > \max_i \hat{b}_i$ in Lemma 1, we get 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

[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.

consistent because all other agents have a strictly dominant strategy, namely to meet any request by k. □