TI 2026-051/VII
Tinbergen Institute Discussion Paper

The Honesty Penalty: Hiding Mistakes in Teams

Randolph Sloof[1]
Julian Tait[1]
Jeroen van de Ven[1]

1 University of Amsterdam and Tinbergen Institute

The Honesty Penalty: Hiding Mistakes in Teams*

Randolph Sloof, Julian Tait & Jeroen van de Ven

University of Amsterdam and Tinbergen Institute

30th July 2026

Abstract

Organisations benefit from the timely disclosure of mistakes, yet workers often conceal their errors. We examine the role of blame avoidance, hypothesising that mistakes associated with a larger loss in potential earnings increase blame and discourage disclosure. This creates a paradox: mistakes are concealed precisely when concealment is most costly to the organisation. We formalise this argument in a model and test it in a vignette study and two incentivised laboratory experiments. The vignette study supports the proposed mechanism, showing that blame increases with the loss in potential earnings. In the laboratory experiments, larger losses do not reduce disclosure, but we do document an important honesty penalty in team settings: the desire to avoid blame induces free-riding by leaving the disclosure of mistakes up to their team members. This pattern emerges both when new teams are formed after each interaction and when teams stay together.

JEL codes: D91, M54
Keywords: organisational failures, blame, experiment, mistakes, lying

*All studies were pre-registered on AsPredicted (194 423, 249 370, 258 655). Ethical approval was granted by the University of Amsterdam Economics and Business Ethics Committee (EB-9525, EB-18649). We thank the Amsterdam Center for Behavioural Change (ACBC) for funding the project.

1 Introduction

In organisations, the quality of decisions depends critically on what information is shared. This is especially important when something has gone wrong, because timely disclosure of errors can enable corrections and mitigate damage. Yet for various reasons, mistakes are frequently deliberately concealed. In the US construction industry alone, over a quarter of workers report having failed to disclose an injury, citing fear of negative consequences as a primary reason (Moore et al., 2013). With over 1,000 workers killed in the US construction industry alone each year (Bureau of Labor Statistics, 2024) and the total cost of workplace injuries and illnesses across all industries estimated at least $250 billion annually (Leigh, 2011), each unreported incident represents a missed opportunity to identify hazards and prevent the serious accidents that follow.[1] A large literature on error management emphasises the benefits of encouraging the disclosure of mistakes (van Dyck et al., 2005; Frese and Keith, 2015), but comparatively little is known about the causal conditions under which people choose to conceal their mistakes.

In this study, we examine avoidance of blame as a key driver of concealment, and investigate factors that can affect the level of blame and the resulting concealment of mistakes. In particular, we test whether blame increases with the magnitude of potential earnings losses associated with a mistake. Based on theoretical work (Battigalli and Dufwenberg, 2007), we hypothesise that more costly errors elicit greater blame, which in turn reduces the willingness to disclose mistakes and prevents managers from taking actions in order to minimise the damage. If so, concealment is most likely precisely when disclosure is most valuable. Somewhat paradoxically, the implication would be that organisations are more likely to persist in an unpromising project when the prospective losses from admitting failures are large.

We report evidence from a vignette study and two laboratory experiments in which we examine the impact of the magnitude of the loss in potential earnings resulting from a mistake on blame and disclosure behaviour. In all studies, participants are confronted with a situation in which a manager of a team must decide between two projects. The default project has the highest upside potential, but in case mistakes are made the alternative project yields higher earnings. Disclosing a mistake to the manager enables the manager to choose effectively between projects, thereby minimising damage. Managers are given the option to punish workers for mistakes.

We exogenously vary the magnitude of the maximum payoff from the default project, and predict that blame increases in the gap between those potential payoffs and the actually realised payoffs. We test if this affects punishments inflicted by managers on workers who disclose making a mistake, and how this affects mistake disclosure. Of course, punishing a worker for disclosing a mistake creates

[1] A telling example of organisational failure due to lack of information transmission about safety problems mentioned by Garicano and Rayo (2016) is the large 2005 explosion at the BP Texas City refinery, causing many deaths and billions of dollars of damage. From interviews, the review panel investigating the case documented that many workers did not report incidents and safety-related concerns for fear of repercussions: "...many workers, particularly at Texas City, Toledo, and Whiting, believe that some incidents, near misses, or other safety concerns do not get reported for a variety of reasons, including fear of being blamed or retaliated against...The interviews and survey data also reveal that at all five refineries, significant doubt existed about whether minor process-related incidents, accidents, and near misses were reported." (BP U.S. Refineries Independent Safety Review Panel, 2007, p. 79).

perverse incentives for workers, stimulating them to hide mistakes. In a team context, it will induce free-riding on others' disclosures. In an ongoing relationship, it can therefore be better for managers to avoid punishing for mistakes to encourage disclosure. To examine whether managers take such reputation effects into consideration, we additionally exogenously vary whether participants stay together in the same team or new teams are formed in each round.

In the vignette study (without real workers), we find support for our primary hypothesis that the level of blame is affected by the loss in potential earnings. Managers express a willingness to punish mistakes, but punish workers who disclose their mistakes less than workers who do not disclose a mistake. Even so, disclosing mistakes becomes relatively less attractive with a larger loss in potential earnings, in line with our hypothesis. In the two laboratory experiments (with real workers and real incentives), a larger loss in potential earnings does not translate in more punishments or lower disclosure rates. We also find no effect of staying in the same team.

Interestingly, mistakes often remain hidden from managers. Workers fail to disclose their mistakes in 34 percent of cases in Experiment 1 and 43 percent in Experiment 2. Although managers tend to reward teams collectively for honesty, disclosure is not encouraged at the individual level. In the first lab experiment we even document an "honesty penalty": compared with concealing a mistake, disclosing it results in a larger punishment for the reporting team member. This creates a free-rider problem, as workers benefit from leaving disclosure to their team members. Overall, however, punishments are relatively small, and not enough to deter disclosure on their own, suggesting that image concerns also play a role. The low level of punishment may indicate that managers recognise its potentially adverse effect on the disclosure of mistakes, but do not fully account for the free-riding incentives and image concerns.

Our results have important implications for organisational approaches to mistakes. We show that concealment persists even when the stakes are low and punishments are small and disclosure is beneficial. In team settings where free-riding on others' disclosures is possible, even a small honesty penalty can substantially reduce reporting. This suggests that the mere possibility of punishment may suffice to deter disclosure, and organisations cannot rely on managerial discretion alone but must commit to institutional mechanisms that protect those who report errors. The desire to encourage the avoidance of future mistakes must be carefully balanced against the chilling effect of punishing disclosures.

Related Literature. Our paper relates to several strands of the literature making several contributions. First, we provide causal evidence of error reporting. Existing work underlines the importance of an error-management culture that encourages disclosing errors. This work relies on observational data or field experiments (Edmondson, 1996; Grant and Larsen, 2007) and unincentivised vignette studies (van Dyck et al., 2005; Gronewold et al., 2013; Gold et al., 2014). Identifying causal factors that promote disclosure of mistakes is particularly challenging, as mistakes often go undetected for some time, and the choice to disclose or conceal mistakes is an endogenous response to the environment. This is even more complex in a team setting, where, even if a mistake is detected, it is often hard to determine where it originated. We add to this literature by providing causal evidence using incentives. The only other experimental work we are aware of in this context, is the study by Baader et al. (2024). Their focus is on different comparative statics, manipulating whether or not mistakes can be reported anonymously and whether or not reporting a mistake is of instrumental value to managers in making choices. Unlike in our setting, workers cannot tell if they are responsible for their mistake or if the mistake was caused by factors outside of their control. In their experiment, managers can punish workers by firing them. Their results similarly show that managers punish workers who disclose a mistake. Our study is also related to the theoretical work of Levitt and Snyder (1997), who analyse how optimal contracts can elicit early warning from an agent with private information about a project's prospects. We instead study the disclosure of realised mistakes in teams, abstracting from effort incentives and focusing on blame, discretionary punishment, and the reporting externality between workers.

Second, we isolate and test a new mechanism for why people conceal mistakes. Specifically, we conjecture that people get blamed for making mistakes, and the more so when the mistake results in a larger loss of potential earnings. Concealing mistakes is then a way to avoid getting blamed. Gurdal et al. (2013) show that workers are held accountable (get blamed) for unfavourable outcomes even if the worker had no control over the event. We build on their work by examining whether blame is related to performance rather than a random event, and by examining how this affects disclosure of mistakes. As Gurdal et al. (2013), and in line with the theoretical work by Battigalli and Dufwenberg (2007), we hypothesise that blame depends on outcomes that could have been obtained if the worker had not made a mistake, i.e., the loss in potential earnings. Our treatment variation in potential earnings is an explicit test of this idea.

Third, we causally test the effect a manager's reputation has on the willingness to punish and disclose mistakes. In a one-shot relationship, punishments can serve to express anger. In a repeated relationship, managers face a more complicated trade-off: punishments can be used to discipline workers to exert effort, but at the same time punishments also discourage workers from disclosing mistakes. This fundamentally differs from other situations such as the provision of public goods, where the option to punish typically improves outcomes (Fehr and Gächter, 2000). In public good settings it is the free-riders who get punished, discouraging free-riding. This differs from our setting, where punishing *encourages* free-riding behaviour if it is targeted to workers who disclose mistakes. By varying whether or not participants stay in the same group for several rounds, we are able to determine whether managers use punishments to express anger and how they take the above strategic considerations into account.

Finally, we contribute to the literature on lying and deception, in particular in the context of preserving one's reputation (for instrumental purposes or social image reasons). Several studies show that in some situations it can be beneficial to lie about an event in order to preserve a good image (see Gneezy et al. (2018); Abeler et al. (2019); Dufwenberg and Dufwenberg (2018); Sorrapera et al. (2019) for theoretical contributions and Sorrapera et al. (2019) and Barron (2019) for experimental evidence). In all these studies people lie about random events that are unrelated to ability.

We examine a related but distinct form of deception, in which deception relates to performance. Suvorov et al. (2024) show that people are willing to withhold valuable information about their performance from group members if the information hurts their image. The focus of that study is how such selective information sharing affects confidence in the ability of their group. Related work also includes studies on confessing behaviour. Utikal (2012) studies confessing to mistakes, while Galeotti and Rose (2024) study confessions to lies. Both studies investigate the impact of possible punishments on the willingness to report. A key difference with our setting is that in both those studies participants confess to a deliberate choice rather than an unintentional mistake. In Utikal (2012), participants gain by making a mistake, and in Galeotti and Rose (2024) participants gain by lying. Thus, in contrast to our study, confessions are not informative about a participant's ability but about their past choices.

The remainder of the paper is structured as follows. Section 2 establishes the theoretical framework, which we use as a basis for our pre-registered hypotheses summarised in Section 3. Section 4 introduces a vignette study to test blame directly, before Section 5 and 6 directly tests our theoretical predictions in two laboratory experiments. Section 7 extends the main results with exploratory analyses, and Section 8 closes with a discussion.

2 Theoretical Framework

We model a team production setting consisting of a risk-neutral manager and two risk-neutral workers.

The workers are responsible for completing individual tasks. Workers can make mistakes while working on their task. We abstract away from moral hazard and simply assume that a mistake happens with a fixed probability m. Each worker privately learns whether they made a mistake. If they made a mistake, they have the option to disclose this to the manager. Each worker independently sends a report $r_i \in \{0, 1\}$ to the manager. The worker can either disclose a mistake ($r_i = 1$) or claim to have made no mistake ($r_i = 0$). For simplicity, we assume that workers can only disclose a mistake when they actually made a mistake.

The manager is responsible for selecting a project, and has the choice between two projects: A and B. The payoff structure is summarised in Table 1. Project A is the more ambitious alternative. It is risky but has a relatively high upside, yielding a payoff of π^S to the manager and each worker if it succeeds and $\pi^F < \pi^S$ if it fails. The success of the project depends on the workers' performance and on an exogenous shock. Specifically, if both workers complete their task without mistakes, the project succeeds with probability q and fails with probability $1 - q$. The project always fails if one or both workers make a mistake. Project B is the safer backup option, and always yields π^B to the manager and each worker, independent of the workers' performance on their tasks. To make things interesting, we assume that $q\pi^S + (1 - q)\pi^F > \pi^B > \pi^F$. This implies that when no mistakes are made, the manager prefers project A over B. Project B always yields higher payoffs when one or both workers made a mistake.

Table 1: Project payoffs per team member, by team performance

Team performance	Project A (ambitious)	Project B (backup)
No mistakes	$\{q : \pi^S \,;\, 1-q : \pi^F\}$	π^B
Mistake by one or both workers	π^F	π^B

After selecting a project, the payoffs of that project are realised. If the manager selected project A, the team members learn whether the project succeeded or failed. Note that the existence of the exogenous shock means that a failure is a noisy signal about the worker's performance, and does not immediately imply that a mistake was made by one or both workers. This provides workers with a cover in case they did not disclose a mistake while the project failed.

Team members benefit from mistakes being disclosed, as it allows the manager to take a corrective action and turn to project B. This will increase each team member's payoff by the amount $\pi^B - \pi^F$. In practice, workers may be hesitant to disclose mistakes for reputational concerns. These reputational concerns can be instrumental (e.g., affecting the chance of being promoted or getting fired) or of an intrinsic nature (social image concerns). In the current model, we focus on instrumental reasons for hiding mistakes, by allowing managers to punish workers. After observing the outcome of the selected project and the reports by the two workers, the manager can punish the workers. The punishment amount is deducted from the worker's payoff and need not be the same for each worker. Punishing is costless to the manager.

To summarise, the timeline is:

1. Each worker works on their tasks and privately learns whether or not they individually made a mistake

2. Each worker independently decides whether to disclose any mistake to the manager

3. The manager selects a project

4. Project payoffs are realised and the manager decides on the punishment amount

2.1 Blame and Punishment

Managers may punish workers for at least two reasons. First, punishments can serve as an incentive for agents to put in effort. Second, punishments can be driven by emotional responses such as anger and frustration. Because we abstract away from the effort decision, we focus on emotional responses as the motivation to punish.

We assume that the amount of punishment for worker i given reports r_i and r_j is determined by:

$$p_i(r_i, r_j) = w_b b_i(r_i, r_j)(\pi^S - \pi^F) - w_h h_i(r_i, r_j)\Delta\pi, \tag{1}$$

The punishment function has two main elements: a *blame* part and an *honesty* part, with respective weights $w_b > 0$ and $w_h > 0$. We discuss each of these elements in turn.

The blame part arises from the manager's anger or frustration, and is proportional to the loss in potential earnings, $\pi^S - \pi^F$, as well as to the extent to which a worker can be held accountable for this loss, denoted by blame function b_i. A variety of theories exists describing how accountability is assigned. We model accountability as a probability-weighted average of full responsibility (when worker i alone made a mistake) and shared responsibility (when both workers made a mistake), based on the manager's posterior beliefs. Let $\hat{m}_{ab}(r_i, r_j) = \Pr(m_i = a, m_j = b \mid r_i, r_j)$ for $a, b \in \{0, 1\}$ denote the manager's posterior about the workers' mistake profile given the report profile (r_i, r_j). When no confusion arises, we suppress the arguments and write $\hat{m}_{ab}$. Thus, for instance, $\hat{m}_{10}$ is the posterior that worker i made a mistake while worker j did not.[2] Worker i is then blamed as:

$$b_i(r_i, r_j) = \hat{m}_{10} + \tfrac{1}{2}\hat{m}_{11}. \tag{2}$$

The blame function $b_i(\cdot)$ has several desirable properties. First, when workers are equally likely to have made a mistake, they get an equal share of the blame. Second, not disclosing a mistake does not eliminate blame altogether. The manager is sophisticated enough to understand that a worker who did not disclose a mistake may still have made one. This is factored in through the posteriors $\hat{m}_{ab}$. Third, reporting a mistake increases blame given any report by the other worker, i.e., $b_i(1, r_j) > b_i(0, r_j)$ for any r_j. If only one worker disclosed a mistake, that worker will get the larger share of the blame, but the other worker will still receive some of the blame. Finally, the blame shares of the two workers add up to one if a mistake is reported, i.e. if the manager knows for sure that a mistake has been made within the team.

The honesty part rewards disclosing a mistake proportional to $\Delta\pi$, which is the gain in payoffs from disclosing a mistake. If the manager's project choice is unaffected by disclosing a mistake, $\Delta\pi = 0$. If managers switch to the backup project after a mistake has been disclosed, then part of the loss can be recovered and $\Delta\pi = \pi^B - \pi^F$. A worker receives full credit for the gain in payoffs if that worker is the only one disclosing a mistake, and half the credit if both workers disclosed a mistake.

$$h_i = r_i(1 - \frac{1}{2}r_j). \tag{3}$$

Note that also the honesty shares of the two workers add up to one if at least one of them reports a mistake (and equal zero otherwise).

Our specification of the blame component in punishment function (1) is conceptually guided by the formal notion of anger-from-blaming behavior (ABB) in Battigalli et al. (2019). Applied to our

[2] The posterior beliefs depend on the workers' reporting strategies, the mistake rate m, the exogenous failure rate q, and the project outcome. For ease of presentation we drop these arguments in the notation.

setting, ABB implies that managers become frustrated when realised payoffs fall short of what a successful project would have paid. In (1), this frustration is captured by term $(\pi^S - \pi^F)$. Managers take this frustration on their workers only insofar as the lower payoffs can be attributed to workers' mistakes. Under ABB, workers' intentions are thus irrelevant. In turn, our specification of relative blame across workers via blame function $b_i(\cdot)$ follows the expected degree of responsibility measure in Halpern (2016, Chapter 6). Workers are responsible for project A yielding less than expected only to the extent that they made a mistake and, if both did so, responsibility is shared. Finally, the honesty component of the punishment function applies the same logic to workers' reporting behaviour.

The resulting specification of the blame and honesty components implies several intuitive properties of the overall punishment function $p_i(r_i, r_j)$. First, reporting a mistake increases the blame and punishment for that worker, but lowers the punishment for the co-worker: $p_i(0,1) \leq p_i(0,0)$. Disclosing a mistake therefore has a positive externality on the teammate. Second, the total team punishment $p_T = p_i(r_i, r_j) + p_j(r_j, r_i)$ is increasing in the likelihood that a (team) mistake was made. Third, total team punishment p_T is the same irrespective of whether one or both workers disclose a mistake (i.e. $p_i(0,1) + p_j(1,0) = 2p_i(1,1)$), even though the punishment is distributed differently over the workers in those two cases. This follows from both the blame and the honesty shares adding up to one if at least one mistake is reported.

2.2 Equilibrium

In this section, we derive the equilibrium strategies of the workers (in terms of reporting) and manager (in terms of project choice). Our focus is on symmetric equilibria, which is reasonable given that we do not allow workers to communicate directly with each other.

Project choice. We start by examining the manager's project choice. Let $y^*_{ab} \in \{A, B\}$ be the manager's optimal project choice when $r_i = a$ and $r_j = b$. When one of the workers discloses a mistake, the manager always selects project B as it yields a higher payoff with certainty: $y^*_{10} = y^*_{01} = y^*_{11} = B$. If none of the workers discloses a mistake, the manager forms beliefs based on the reporting strategies of workers. Suppose that, conditional on having made a mistake, each worker is believed to disclose the mistake with probability r. Then conditional on no mistakes being disclosed, the probability that no worker made a mistake is given by $\hat{m}_{00}(0,0) = \left(\frac{1-m}{1-mr}\right)^2$. The manager then prefers project A when the reporting rate exceeds a threshold:

$$r \geq \hat{r} := \frac{1}{m}\left[1 - (1-m)\sqrt{q\frac{\pi^S - \pi^F}{\pi^B - \pi^F}}\right]. \tag{4}$$

Intuitively, when workers are believed to disclose mistakes with a high probability (high r), the absence of a report is a strong indicator that no mistakes are made, and project A is relatively attractive. By contrast, if workers tend to hide mistakes, then it remains likely that project A will fail, making project B the more attractive choice.

Collecting results and assuming the manager selects A if indifferent:

$$y_{ij}^* = \begin{cases} A & \text{if } (i,j) = (0,0) \text{ and } r \geq \hat{r}, \\ B & \text{otherwise.} \end{cases} \tag{M}$$

Disclosure decision. We next turn to the worker's decision to hide or disclose a mistake, given the strategies of the manager and the other worker. Let $\pi^{y_{00}}$ be the project payoffs when no mistakes are disclosed and the manager selects project $y_{00} \in \{A, B\}$. To ease notation, we write y for y_{00} in what follows. If worker i believes that worker j discloses a mistake with probability r, then reporting increases the worker's expected payoffs by:

$$\Delta U_i(r, y) = mr[p_i(0,1) - p_i(1,1)] + (1 - mr)[\pi^B - \pi^{y_{00}} + p_i(0,0) - p_i(1,0)]. \tag{W}$$

With probability mr, the other worker is expected to make and disclose a mistake. Disclosing a mistake then affects the worker's punishment ($p_i(1,1)$ versus $p_i(0,1)$) but has no effect on the manager's project selection; the manager will select project B anyway. With the remainder probability $1 - mr$, the other worker does not disclose a mistake. In that case, disclosing a mistake affects punishments ($p_i(1,0)$ versus $p_i(0,0)$), and is potentially pivotal for project selection, resulting in a change in payoffs equal to $\pi^B - \pi^{y_{00}}$.

The expressions (M) and (W) jointly govern the possible equilibria of the game. Following Sobel (2013), we distinguish between equilibria in which communication is influential and equilibria in which communication is not influential. Applied to our setting, communication is said to be influential if the manager's equilibrium action depends on the worker's report. From (M), this is equivalent to saying that the worker's equilibrium reporting rate is at least $\hat{r}$.

Our main interest is in the comparative statics with respect to the project payoffs. If communication is not influential, it is easy to see that workers never report in equilibrium: reporting would increase their punishment without any possible benefits of a better project choice. In that case, comparative statis are trivial. In the following, we therefore focus on equilibria in which disclosure is influential.[3]

Proposition 1. *In any equilibrium with influential communication, the disclosure rate r^* is (weakly) decreasing in the payoffs of a successful project (π^S).*

This result is intuitive. As π^S increases, managers punish mistakes more harshly, lowering the worker's willingness to disclose mistakes. Note that, while the proposition is formulated in terms of π^S, it can be equivalently formulated in terms of π^F with π^F working in the opposite direction provided w_b is large enough. An interesting, somewhat paradoxically implication is that

[3]The two types of equilibria can coexist. A full characterisation of the equilibrium set is provided in Appendix C.

the equilibrium disclosure rate can be lowest when the team has most to gain from disclosing mistakes: when π^F is low, the disclosure rate can be low while $\pi^B - \pi^F$ is high.

2.3 Discussion

Our model is based on a single interaction between the workers and manager, taking the weights w_b and w_h as given, arguing that they stem from emotions rather than deliberate choices to incentivise workers. In practice, the relationship is often ongoing, and managers will also take incentives into consideration. Given that the mistake rate is exogenous in our theoretical framework, there is indeed no role for punishments in terms of disciplining workers to exert effort. However, in an ongoing relationship, managers will benefit from mistakes being disclosed. Rather than explicitly modelling this, we take a reduced form approach and note that the equilibrium disclosure rate is weakly increasing in the weight given to honesty (w_h). Thus, managers can encourage workers to disclose mistakes by rewarding workers for their honesty, i.e., by setting w_h higher. This resembles the principal-agent model of early warning in Levitt and Snyder (1997). When an agent's private bad-news signal induces project cancellation, the optimal contract pays the reporting agent more than an agent whose project continues and subsequently produces a bad outcome. In our setting, analogous incentives arise when the weight placed on honesty, w_h, is sufficiently large relative to the weight placed on blame, w_b. Note that changing honesty weights will not change the comparative statics with respect to π^S, unless the manager sets a higher weight to honesty when the payoffs of success are higher.[4]

Although our theoretical framework is based on the specific punishment function reflected in (1), it allows for a variety of punishment structures via the weights w_b and w_h. It is insightful to rewrite expression (W) as follows:

$$\Delta U_i(r,y) = (1-mr)[\pi^B - \pi^{y_{00}}] - \{(1-mr)[p_i(1,0) - p_i(0,0)] + mr[p_i(1,1) - p_i(0,1)]\} \qquad (5)$$

The term within curly brackets can be defined as the *disclosure* or *honesty penalty*. It equals the expected increase in the punishment received from the manager if a mistake is revealed rather than kept hidden. Although labelled a penalty, it can be negative—and hence constitute a reward—when w_h is sufficiently large relative to w_b. A sufficient condition for it being a true penalty is that $p_i(0,0) < p_i(1,0)$.[5] This happens when honesty weight w_h is not too large. As noted above, reporting a mistake also has an external effect on the expected punishment received by the teammate. The size of this external effect immediately follows from flipping the arguments in the p_i functions in the above expression (in curly brackets) for the disclosure penalty. Adding this external effect

[4] One may wonder why the manager would not set w_h arbitrarily high. In practice, rewarding workers may be costly to the manager. Furthermore, rewarding honesty too much would make it attractive for workers to falsely report mistakes or even deliberately make them.

[5] In that case we have from $p_i(0,1) \leq p_i(0,0)$ necessarily that $p_i(0,1) \leq p_i(0,0) < p_i(1,0)$. In turn, it follows from $p_i(0,1) + p_i(1,0) = 2p_i(1,1)$ that $p_i(1,1) > p_i(0,1)$.

to the disclosure penalty yields the total change in team punishment; the expected team benefit of disclosing a mistake is thus $2(1-mr)[p_i(0,0)-p_i(1,1)]$. The latter can go either way; it is positive and thus a true (team) *benefit* only if weight w_h is non-negligible. Overall, in case $p_i(1,1) < p_i(0,0) < p_i(1,0)$, honesty is beneficial at the team level, but harmful at the individual worker level. Such a situation theoretically occurs if the honesty weight w_h is moderate relative to the blame weight w_b.[6]

3 Hypotheses

The model serves as a basis for the following pre-registered hypotheses. First, when π^S is larger, a mistake causes a larger loss in potential earnings, making mistakes more costly. We expect that a larger payoff π^S therefore results in more blame and a higher punishment.

Hypothesis 1. *Workers are punished more after reporting a mistake when the stakes are higher.*

Relatedly, we expect that a larger payoff π^S discourages workers from disclosing mistakes. This follows directly from Proposition 1 if communication is influential, which we will verify is the case.

Hypothesis 2. *Conditional on making a mistake, workers are less likely to disclose a mistake when the stakes are higher.*

The next two hypotheses concern team formation. We informally discussed the effect that repeated interaction has on the weight given to honesty (w_h), arguing that w_h will be higher in an ongoing relationship compared to one-shot interactions. Rewarding the disclosure of mistakes becomes more attractive if the relationship is ongoing. By building up a reputation for not punishing mistakes, managers can create an improved 'error-management culture'. This is expected to result in a higher disclosure rate.

Hypothesis 3. *Managers punish mistakes less in ongoing relationships compared to one-shot interactions.*

Hypothesis 4. *Workers disclose mistakes more often in ongoing relationships compared to one-shot interactions.*

4 Measuring Blame: Vignette Study

We start by describing the results from a vignette study. We designed this vignette to explore manager's attitudes towards workers making and disclosing mistakes. More details and instructions can be found in Appendix D.[7]

[6]In particular, this happens when $w_b(1-\hat{m}_T)\frac{(\pi^S-\pi^F)}{(\pi^B-\pi^F)} < w_h < w_b\frac{(\pi^S-\pi^F)}{(\pi^B-\pi^F)}$, where $\hat{m}_T$ denotes the manager's posterior belief that the team made a mistake in case project A is chosen and fails.

[7]We ran this vignette study between running Experiments 1 and 2, but find it more instructive to present it first.

4.1 Setup and procedures

Participants were asked to imagine being the manager of a small factory, where a mechanic is responsible for machine maintenance. On days where the machine runs smoothly, profits are $10,000 (treatment HIGH) or $600 (treatment LOW). However, machine breakdowns can occur, either because of random factors beyond the mechanic's control, or as a result of an avoidable servicing mistake made by the mechanic. If the mechanic makes a mistake they can choose to disclose it or hide it. If they hide it, the machine breaks down and profits drop to zero. In contrast, if the mechanic discloses their mistake, corrective action can be taken resulting in profits of $400. This scenario captures key aspects of our experimental design (see later), but focuses on a single worker rather than a team and eliminates the project choice by the manager, to simplify the exposition.

We asked participants to evaluate the mechanic along several dimensions, in each of the three possible outcomes (successful production, a disclosed mistake paired with corrective action, and machine breakdown without a disclosed a mistake). Our primary outcome is the size of a bonus awarded to the mechanic (between $0 and $2,000). We also asked participants to evaluate the mechanic's overall performance (on a 9-point Likert scale), whether they would re-hire the mechanic (yes or no), and, if a machine breakdown occurred, the extent to which they blame the mechanic for this (on a 9-point Likert scale).

Our main hypothesis for the vignette study (equivalent to Hypothesis 1 in the previous section) is that the levels of blame and punishment increase in the loss in potential earnings when there is a machine breakdown. In other words, we expect more blame and more punishment in treatment HIGH compared to treatment LOW.

We recruited 243 participants via Prolific. The treatment variation was the maximum profits in case of no mistake ($600 or $10,000), using a between-subject design. The order in which the three scenarios were presented was randomised between participants, with an approximately equal share seeing each scenario first. The study was pre-registered.

4.2 Results

As hypothesised, we find that participants blame the mechanic more when the loss in potential earnings is higher (Appendix Fig. B.1). After disclosing a mistake, the mean blame increases from 4.6 (treatment LOW) to 5.2 (treatment HIGH), and the difference is significant (rank sum test, $p = .028$, one-sided test).[8]

This also translates into higher punishment. Participants were told that $1,000 is the average performance bonus for a mechanic at this firm. Thus, any bonus below $1,000 can be interpreted as a punishment. The average bonus awarded after disclosing a mistake is $874 in treatment LOW, and decreases to $674 in treatment HIGH. This difference of $200 is significant (rank-sum test, $p < .001$). Thus, in line with our hypothesis, the punishment for disclosing a mistake is larger when

[8] Though not pre-registered, we also find an (non-significant) increase in blame when the project failed but the mechanic did not disclose a mistake (from 5.6 in treatment LOW to 5.9 in HIGH, rank-sum test, $p = .235$, one-sided test).

the loss in potential earnings is higher. Our pre-registered test takes into account that bonuses may also differ across treatments when there is no breakdown. Indeed, when no breakdown occurs, workers receive \$48 more bonus when the stakes are higher; from \$1,185 in LOW to \$1,233 in HIGH. The DID estimate of \$-248 (=-200-48) is significantly different from zero (rank-sum test, $p < .001$); that is, the punishment for disclosing is \$248 larger in treatment HIGH.[9] We see similar patterns in terms of overall evaluations and re-hire decisions, though in those cases differences are small and not significant.[10]

From the perspective of the worker, the most relevant outcome variable is how a disclosed mistake affects the bonus after a mistake has been made.[11] We already saw that when a mechanic discloses a mistake, they are being "punished", with a mean bonus of 874 in treatment LOW and 674 in treatment HIGH. Ex ante, it is not evident whether hiding a mistake will result in a higher or lower punishment. In the absence of a disclosure, there is uncertainty about the source of a machine breakdown. A disclosed mistake is conclusive evidence that the mechanic made a mistake, and should therefore increase the level of blame. On the other hand, disclosing allows for corrective action to be taken, which should decrease blame. This latter effect dominates in our setting, resulting in bonuses that are lowest after a breakdown without a disclosed mistake (\$408 in treatment LOW and \$377 in treatment HIGH).[12]

In both treatments, hiding a mistake results in a lower bonus compared to disclosing a mistake. In monetary terms, it would therefore never be profitable for a worker to hide a mistake. However, workers may have other motives to hide mistakes besides monetary reasons, such as image concerns. They will then trade off monetary losses against image concerns. Since disclosing a mistake becomes relatively less attractive in monetary terms when the loss in potential earnings increases (larger punishment), we would expect to see less disclosures in treatment HIGH, where the monetary cost of hiding a mistake is relatively lower.

Result 1. *(Vignette study). With larger losses in potential earnings, people are blamed more for making a mistake and punished harsher.*

The results from the vignette study also suggest that the honesty weight w_h might well be moderate relative to the blame weight w_b. On the one hand, it appears sufficiently low for honesty per se not to trump making mistakes in the determination of punishment; in both treatments the bonus is significantly lower after honestly reporting a mistake than after conclusive evidence (i.e. no breakdown) that no mistake has been made.[13] At the same time, w_h is sufficiently high as to make

[9] Table A.1 reports the corresponding OLS estimates, which match the tests reported here.

[10] For overall performance evaluations, the DID estimate is .263 with $p = .272$, for the re-hiring decision the DID estimate is .070 with $p = .093$.

[11] We did not pre-register this part of the analysis, but we did pre-register such an analysis for the two lab experiments. For comparison, we also report it for the vignette study.

[12] In the vignette we thus find an honesty *reward* equal to 466 ($= 874 - 408$) in LOW and a lower reward equal to 297 ($= 674 - 377$) in HIGH.

[13] Applying the theoretical framework of Section 2 to a single worker setting, this translates into $w_h < w_b \frac{(\pi^S - \pi^F)}{(\pi^B - \pi^F)}$ to hold.

it beneficial to reveal the mistakes actually made.[14] Translated to the team setting, this implies that it may actually be quite possible that honesty hurts individually but helps the team. That is, the positive externality on the teammate of disclosing one's own mistake may well outweigh the individual harm.

4.3 Discussion

The main takeaway is that the vignette study supports our main hypothesis: people are blamed more for making a mistake when there are larger losses in potential earnings, and this translates into higher punishments. From the perspective of the worker, this would mean that disclosing a mistake becomes less attractive when there are larger losses in potential earnings.

The vignette study also has some potential limitations. It does not have actual workers and choices are not incentivised. Moreover, the vignette does not consider a team setting and thus does not allow for free-riding and external effects in disclosing mistakes. We now present the design of a lab experiment which addresses these limitations.

5 Experiment 1

5.1 Experimental Setup and Procedures

In the experiment, teams of one manager and two workers are formed. Participants are randomly assigned the role of manager or worker, and keep this role throughout the experiment.

Workers first complete a task. The task consists of a series of simple math problems and variants on the Stroop task under a time constraint. In the Stroop task, workers are asked to name the font colour of a word while the word itself may spell a different colour (e.g., the word "gray" printed in blue ink). These tasks are not difficult to complete within the time limit, but are prone to mistakes when insufficient attention is paid to them. A more detailed description of the task and instructions are provided in Appendix E.

After finishing the task, workers learn privately whether or not they made a mistake (where a time-out also counts as a mistake). If so, they can keep this private (hide the mistake) or disclose it to the team (both the manager and the other worker). The manager then chooses between an *Ambitious Project* or a *Backup Project*. The Ambitious Project pays €10 (treatment LOW) or €20 (treatment HIGH) to each team member if it is successful and €2 if it fails. It fails if either a mistake is made by one or both workers, or with some exogenous chance $1 - q = .2$ even if no mistakes are made. The Backup project always pays €4.

Once the project payoff has been realised, the manager can choose to punish a worker by subtracting money from their payoff. The total team punishment cannot exceed €2.00 and individual punishments must be in multiples of €0.25. The manager's own payoff is not affected by the chosen punishment level.

[14] In the theoretical framework this translates into $w_b(1 - \hat{m}_{T=1})\frac{(\pi^S - \pi^F)}{(\pi^B - \pi^F)} < w_h$, where $\hat{m}_{T=1}$ is now the manager's posterior belief that the single worker made a mistake in case a breakdown is observed.

At the end of each round, all team members learn which project was chosen, the resulting project payoff, and the punishment given to each worker, so a worker observes both their own punishment and that of the other worker. Mistakes themselves are never revealed: neither the manager nor the other worker learns whether a mistake was made unless it was reported, so a hidden mistake is indistinguishable from no mistake at all.

5.1.1 Treatments

We implemented two treatment variations. The first treatment variation is the payoff of the ambitious project when it is successful. The payoff could be low (€10, treatment LOW) or high (€20, treatment HIGH).

The second treatment variation is whether or not teams stay together. In treatment PARTNERS, participants stay in the same team for all rounds in a block. At the start of a new block, they are assigned to a completely new team. In treatment STRANGERS, new teams are formed at the start of each round, randomly matching managers to workers. Although it is possible that they were paired with the same participants as in a previous round, they could not know this, as their identities were never revealed.

5.1.2 Procedures

The experiment was programmed in oTree (Chen et al., 2016). Participants played two blocks of 10 rounds each. A participant was always assigned to either treatment PARTNERS or treatment STRANGERS (between-subjects), but always did one block in treatment LOW and one block in treatment HIGH (within-subjects). Across sessions, we balanced the order in which they saw the low or high payoffs.

In total, 264 participants were recruited over 17 sessions. Within a session, we formed independent matching groups (always consisting of 12 participants in the strangers treatment, and between 9-12 participants in the partners treatment). Teams were always formed with participants from the same matching group. Each session lasted one hour and participants on average earned €17. 56 percent of our participants are female, and the mean age is 20.3. Treatments are largely balanced in terms of these demographics.[15]

5.1.3 Design features

Task selection. The ideal task for the purpose of our setting is one that is perceived as easy (so that managers will blame workers for making mistakes) and at the same time results in frequent mistakes (to generate enough data). We believe the selected set of tasks accomplishes this. The Stroop task in particular appears easy, but in practice people find it hard to avoid making mistakes because

[15] We cannot reject that the share of females is the same across treatments (Fisher Exact test, $p = .511$. The mean age is almost identical across treatments (varying between 20.0 and 20.6) but a Kruskal-Wallis test rejects equality ($p = .059$). Our results are robust against controlling for gender and age.

that requires suppressing the urge to give the intuitive response. To further increase the perception that it is an easy task, we showed the questions with the correct solutions to the managers.

5.2 Results

Before reporting treatment effects, we first present results aggregated over all treatments. To be conservative, we take the matching group as the independent unit of observation for statistical testing.

On average, a worker makes a mistake in 51 percent of the rounds. At the team level, a mistake is made by at least one of the two workers in 74 percent of the rounds. The mistake rates are comparable across treatments.

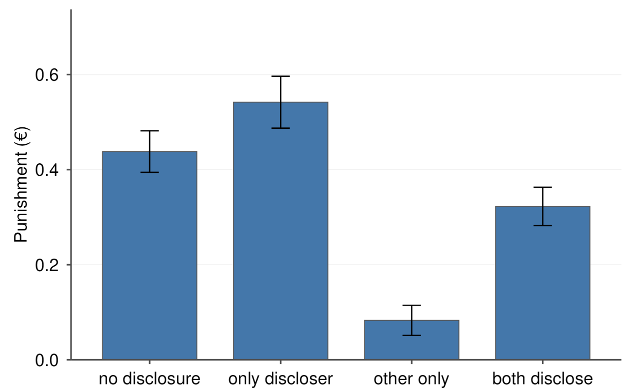
Workers often do not disclose their mistake to the manager. On average, they hide their mistake 34 percent of the time. This has consequences for the payoffs of each team member, because communication is influential: Managers almost always select the ambitious project when no mistakes are disclosed, and select the fallback project when a mistake is disclosed (around 95 percent of the time in each case). Since a mistake always results in a failure of the ambitious project, hiding the mistake results in lower earnings. In teams where a mistake was made but not disclosed, mean earnings are €2.17 for each member of the team (excluding any punishments). This is almost 50 percent lower than earnings of teams where a mistake is disclosed (€3.88).

Managers' punishments vary systematically with realised outcomes and workers' disclosure behaviour. As expected, punishment is minimal when the ambitious project is selected and the high payoff is realised. In that case, the mean total team punishment is €0.16, relative to a maximum of €2.00. When the ambitious project fails and no worker disclosed a mistake, the mean total team punishment increases to €0.84. If a mistake is disclosed and the manager selects the fallback project, the mean team punishment is €0.60.

Two features of the manager's punishment strategy are noteworthy. First, conditional on failure of the ambitious project, total team punishment is lower when one or both workers disclose a mistake. This pattern is not obvious ex ante: a disclosure confirms that a mistake was made, whereas a failed project without disclosure leaves open the possibility that the project failed due to random factors outside the worker's control. The lower punishment following disclosure is consistent with managers valuing honesty, as it enables timely corrective action. For total team punishment, it appears not to make a difference if one or both workers disclose a mistake, in line with our theoretical framework. Second, disclosures do not eliminate blame. While total team punishment is lower, when only one worker discloses, that worker bears a disproportionate share of the total team punishment (roughly 85 percent).

At the team level, disclosure is rewarded (Appendix Fig. B.2), whereas at the individual level there is a penalty for being honest. Of course, a worker knows at the time of disclosure whether they personally made a mistake, but not whether the other worker will disclose one. The relevant comparisons are therefore conditional on the worker's own mistake status and on whether the other worker discloses. Figure 1 shows the mean punishment across the different possibilities: neither

Figure 1: Punishment amounts by disclosure of mistakes (Experiment 1)



Notes: Mean punishment received by a worker, by own and other worker's disclosure of a mistake, conditional on the plotted worker having made a mistake. Means are computed per matching group. Error bars indicate +/- 1 s.e.

of the workers discloses a mistake ('no disclosure', yielding punishment $p(0,0)$), only the other worker discloses ('other only', $p(0,1)$), they are the only one disclosing ('only discloser', $p(1,0)$), and both workers disclose ('both disclose', $p(1,1)$). The lowest punishment is received when only the other worker discloses a mistake (€0.09); this is roughly the same punishment as after a successful ambitious project, suggesting that one's teammate disclosing a mistake completely absolves a worker of any suspicion of having made a mistake themselves. The highest punishment is received when the worker is the only one disclosing (€0.54). The very structure of punishments thus seem to foster free-riding in disclosing. On average, disclosing a mistake rather than hiding it increased a worker's punishment – and thus constitutes an honesty penalty – by €0.18 (computed at the matching-group level, the unit of our statistical tests; sign test, $p = .023$, two-sided).

Workers must trade-off the honesty penalty against the benefits of enabling the manager to take a corrective action. Despite the penalty for disclosing a mistake, hiding a mistake is net costly to a worker: the benefits of enabling a corrective action (€2) outweigh the penalty for being honest.

We summarise our findings so far as follows:

Result 2. *Workers often hide their mistakes even if this results in a monetary loss to each team member including themselves.*

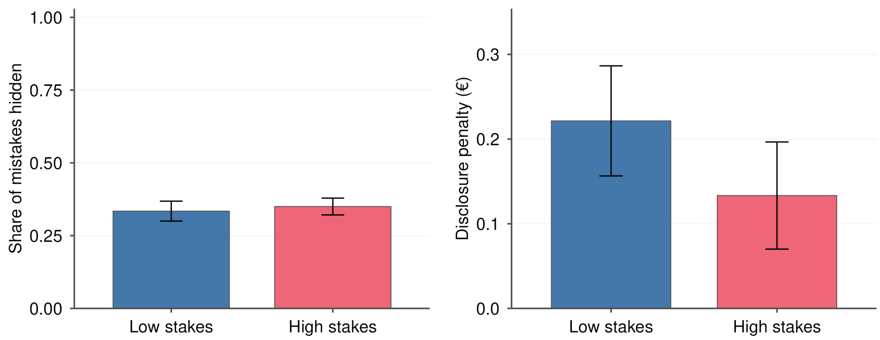
5.2.1 Treatment Effects

We next report our main treatment comparisons as pre-registered.

Our first two hypotheses are that managers punish workers more for disclosing a mistake and workers (therefore) hide mistakes more often when the loss in potential earnings is larger. We do not find evidence of this, as shown in Figure 2. Workers hide mistakes 33 percent of the time in treatment LOW and 35 percent of the time in treatment HIGH. This difference is not significant (paired Wilcoxon signed-rank test, $p = .342$, one-sided). Consistent with this behaviour, punishments

Figure 2: Effect of stakes on mistake hiding rates and disclosure penalty (Experiment 1)

(a) Hiding mistakes (b) Disclosure penalty



Notes: The sample consists of workers who made a mistake. LOW/HIGH denotes the ambitious project paying €10/€20 if successful. Panel (b) is the mean punishment after disclosing a mistake minus after hiding one. Means are computed per matching group. Error bars indicate +/- 1 s.e.

also do not differ much across treatments. In treatment LOW, a worker can expect €0.22 more punishment after disclosing a mistake compared to hiding the mistake – that is, there is a disclosure penalty of €0.22. In treatment HIGH, the disclosure penalty is on average €0.13. Thus, if anything, the disclosure penalty is lower in treatment HIGH, though the difference is not significant (paired-sign test, $p = .419$, one-sided). We thus find no support for hypotheses H1 and H2:

Result 3. *Workers do not hide their mistakes more often and are not punished more for disclosing mistakes when the loss in potential earnings is higher.*

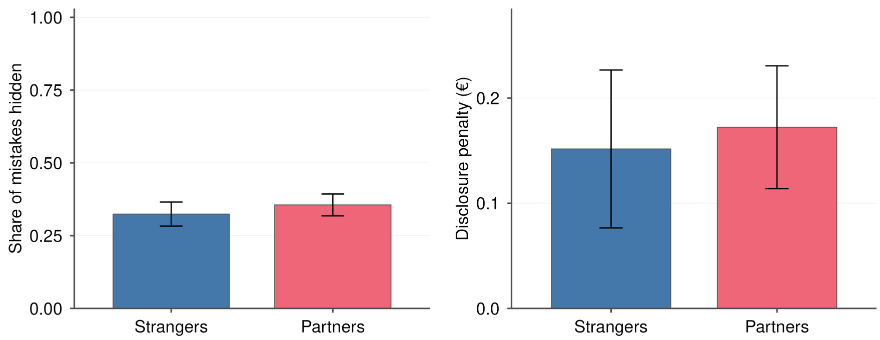
We also hypothesised that hiding mistakes and disclosure penalties would be less prevalent when teams stay together in all rounds. Here as well we find no treatment differences, as shown in Figure 3. Mistakes are hidden 32 percent of the time in treatment STRANGERS compared to 36 in treatment PARTNERS, and the difference is not significant (rank-sum test, $p = .705$, one-sided). The disclosure penalties are also nearly identical. Disclosing a mistake in treatment STRANGERS does not result in higher expected punishment than in PARTNERS (.24 against .24, rank-sum test, $p = .511$, one-sided).

Result 4. *Workers do not hide their mistakes less often and are not punished less for disclosing mistakes when they stay in the same team.*

Appendix Table A.2 confirms these results with OLS regressions. These null results hold when adding demographic controls and when excluding the first 5 rounds of each block. In the block-1 between-subject specification (column (7)), the disclosure penalty is significantly smaller under high stakes – the opposite of the hypothesised direction – a pattern that appears in neither the full sample nor the group-level tests.

Figure 3: Effect of team-formation on mistake hiding rates and disclosure penalty (Experiment 1)

(a) Hiding mistakes (b) Disclosure penalty



Notes: The sample consists of workers who made a mistake. In treatment STRANGERS new teams are formed each round; in treatment PARTNERS the team stays together for all rounds of a block. Panel (b) is the mean punishment after disclosing a mistake minus after hiding one. Means are computed per matching group. Error bars indicate +/- 1 s.e.

5.3 Discussion

The low punishments should not deter workers from disclosing mistakes, so why do workers frequently hide mistakes? One possibility is that they are trying to free-ride on disclosures by the other worker. This would, however, require quite extreme beliefs about other workers' mistake and disclosure rates.[16] Unincentivised survey responses are inconsistent with this explanation. At least at the aggregate level, beliefs are fairly accurate: workers expect others to make a mistake in 65 percent of the rounds (instead of the actual 51) and disclose a mistake 57 percent of the time (instead of the actual 63). Alternatively, workers may care about their social image, and hide their mistakes wishing to look capable to team members. Even in anonymous laboratory settings such social image concerns have been shown to affect behaviour (Andreoni and Petrie, 2004; Andreoni and Bernheim, 2009; Friedrichsen et al., 2018, Bolton et al., 2021).

More puzzling is that, unlike in the vignette study, the amount of punishing is low and does not vary across treatments. Possibly, the scope to punish was too limited, creating a ceiling effect. A substantial fraction (roughly 70 percent) of managers used the maximal punishment of €2.00 at least once, and may have wished to punish harsher. Even when managers did not use the maximum punishment, their punishment may have been affected by the range of allowed punishments, if they interpreted this as a signal of what is appropriate (Tversky and Kahneman, 1981; Ockenfels and Werner, 2014). Another limiting factor could be the relatively small difference in potential earnings across treatments (10 or 20), which is a lot smaller than the treatment difference in the vignette study (600 versus 10,000). We address these potential limitations in a follow-up experiment.

[16] Just to illustrate, based on equation (W) and an average disclosure penalty of €0.14 over all treatments, this would require $mr \geq 0.93$.

6 Experiment 2

To test whether the absence of a treatment effect in Experiment 1 is driven by a ceiling or framing effect of punishment possibilities, or by a too narrow spread in potential earnings, we ran an additional lab experiment.

6.1 Experimental Design and Procedures

The setup is identical to that of Experiment 1 with two main changes. First, we increased the maximum team-punishment from €2.00 to €12.00. Second, we changed the payoffs in case of a successful ambitious project to €8.00 (treatment LOW) and €40.00 (treatment HIGH), thus increasing the spread. Besides these main changes, we also simplified the setup by setting the payoff in case of failure to zero and of the backup project to €2.00. To gain more insights into the mechanisms, we also added some questions to elicit beliefs.

We recruited 144 new participants. As in Experiment 1, we varied the maximum payoffs within participants, balancing the order across sessions. We only implemented the strangers treatment in this experiment, with random re-matching of workers and managers after every round. Each participant was in a matching group of 12. We collected 6 independent matching groups in each treatment order (LOW first and HIGH first).

The mean age of participants is 20 years and 44 percent of participants are female. These demographics again appear balanced across treatment orders.[17]

6.2 Results

The mistake rate is 50 percent, similar to that in Experiment 1. Workers who made a mistake hide this from their manager around 43 percent of the time. This is higher than the 34 percent in Experiment 1, but the difference is not significant (rank sum test, $p = .106$, two-sided). Possibly workers fear more punishments overall with the higher cap on punishments. However, also now we find no treatment differences in hiding rates; workers hide their mistakes 42 percent of the time in treatment LOW and 44 percent of the time in HIGH.

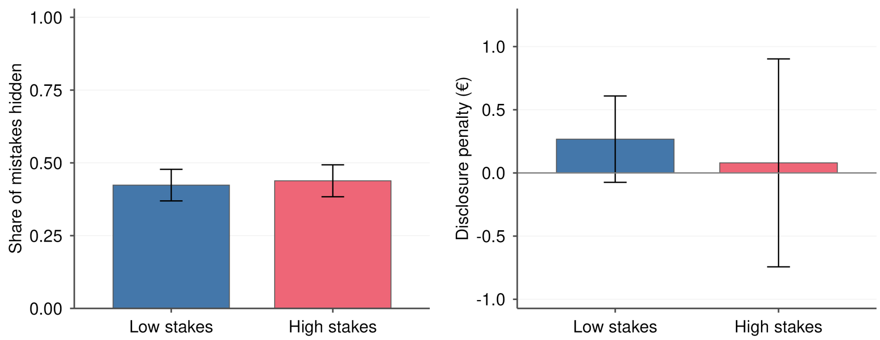
The punishment pattern is qualitatively similar to that of Experiment 1 (see also Appendix Fig. B.3). A worker is punished the most when they are the only discloser (€1.99) and lowest when the other worker is the only discloser (€0.50). As in Experiment 1, the punishment is higher after no disclosures (and project failure) than when both workers disclose: at the team level, disclosure is again rewarded (Appendix Fig. B.2). On average, disclosing a mistake rather than hiding it increased a worker's punishment by €0.17, though the disclosure penalty is not statistically significant (sign test over the 12 matching groups, $p = .774$, two-sided).

Consistent with the absence of a treatment difference in disclosure rates, we also find no differences in the disclosure penalty across treatments. The disclosure penalty in treatment LOW is €.27, compared with €.08 in treatment HIGH, though neither differs from 0 (Wilcoxon signed-rank

[17] For the share of females a Fisher Exact test yields $p = .402$, for age a Kruskal Wallis test gives $p = .912$

Figure 4: Effect of stakes on mistake hiding rates and disclosure penalty (Experiment 2)

(a) Hiding mistakes (b) Disclosure penalty



Notes: The sample consists of workers who made a mistake. LOW/HIGH denotes the ambitious project paying €8/€40 if successful. Panel (b) is the mean punishment after disclosing a mistake minus after hiding one. Means are computed per matching group. Error bars indicate +/- 1 s.e.

test, $p = .569$ and $p = .733$ respectively, two-sided). As in Experiment 1 we find that, if anything, the disclosure penalty is higher in treatment LOW, though this is not significant (paired sign-test, $p = .387$, one-sided).

These small increases in punishments after disclosing a mistake are a reflection of low punishment rates in general. Mean punishments remain low throughout, despite the possibility to punish more. When a mistake was disclosed and the fallback project selected, the maximum punishment was inflicted only 7 percent of the time. Even when no mistakes were disclosed and the ambitious project failed, the maximum punishment was inflicted only 23 percent of the time, and far more often managers did not punish at all (50 percent of the time).

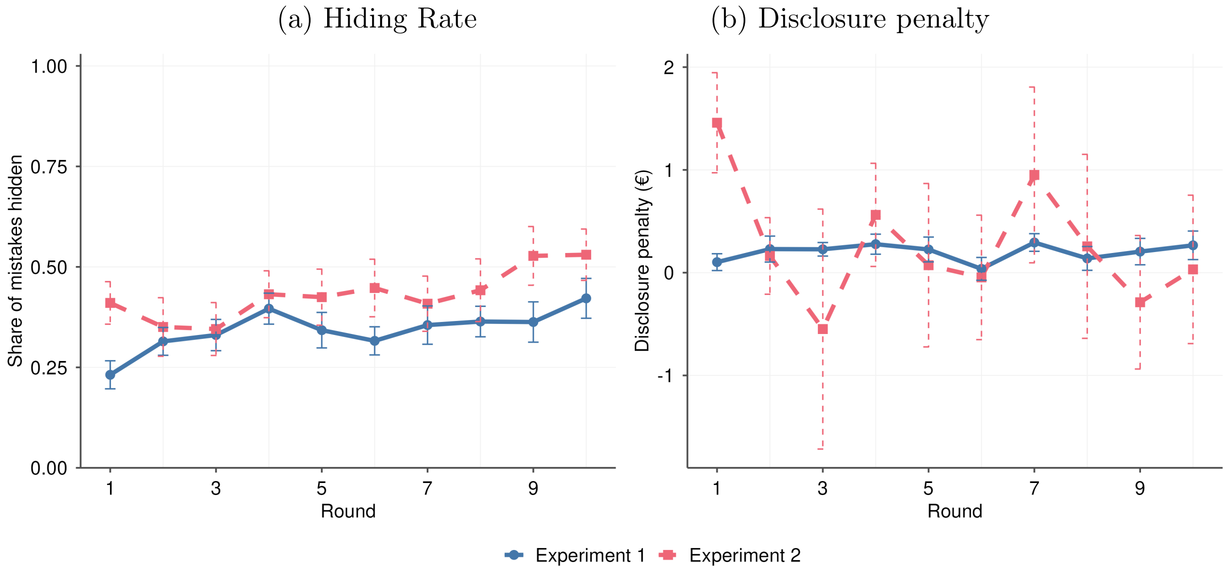
Thus, despite the higher maximum possible punishments and the much larger spread in maximum earnings, we again find no treatment differences in terms of disclosure rates and disclosure penalties, as shown in Figure 4. Like before, these null results hold when adding demographic controls, when restricting to block 1 only, and when excluding the first 5 rounds of each block to allow for learning (see Appendix Table A.3). In the concluding section, we offer some explanations for the absence of treatment effects.

7 Exploratory Analyses

In this section we present some exploratory analyses and heterogeneous effects. Given that we do not find treatment effects of stakes and matching, our focus is on aggregate results.

Learning. Workers' performance in the real effort task initially increases with experience. In the first round, 80 percent of workers make a mistake. This drops to roughly 45 percent in round five, after

Figure 5: Hiding rates and disclosure penalties over time (Experiments 1 and 2)



Notes: Means by round within a block, conditional on a mistake, pooled over blocks. Means are computed per matching group. Error bars indicate +/- 1 s.e.

which it remains stable. More interesting is whether subjects learn to hide or disclose mistakes, and how the punishment evolves over time. Panel (a) in Figure 5 shows the hiding rate over the rounds. In both experiments, we observe a positive time trend. On average,the hiding rate goes up from roughly 30 percent at the start to 45 percent by the last round. Thus, the problem of hiding mistakes does not improve but worsens over time. Panel (b) shows the disclosure penalty, conditional on having made a mistake. In Experiment 1, the disclosure penalty is stable over the rounds. In Experiment 2, it fluctuates across rounds, with no apparent time trend after the initial drop in round 1.

Beliefs. The decision to disclose a mistake or not depends not only on the expected punishment, but also on beliefs about the other worker's mistake rate and disclosure strategy. A worker can free-ride on the other worker if the other worker is believed to disclose mistakes. We can use beliefs elicited after each block in Experiment 2 to examine whether workers behave in accordance with free-riding behaviour. It turns out to be the opposite: workers who believe that others disclose more are themselves more likely to disclose (Pearson correlation of .57 across workers; $p < .001$). This could point to fairness considerations, but is also consistent with ex post rationalisations: workers may justify hiding mistakes by reporting that they expect others to do so as well.

Overall, most reported beliefs are fairly accurate. Workers hold somewhat pessimistic beliefs about their peers' performance and disclosure behaviour. On average, they believe that mistakes occur 66 percent of the time, instead of the observed mistake rate of 50 percent, and believe that mistakes are hidden 52 percent of the time, compared to the actual hiding rate of 43 percent. If a manager observed a single worker disclosing a mistake, they estimate that the other worker is hiding a mistake 35 percent of the time, which is close to the true rate of 29 percent. By contrast, managers' beliefs that a worker is hiding a mistake when the ambitious project fails without any disclosed mistakes – namely 41 percent of the time – underestimate the observed rate of 50 percent. Moreover, although beliefs are on average fairly accurate, they vary considerably across individuals (Appendix Fig. B.4).

Gender Differences. Turning to gender differences, we find differences in punishment behaviour by managers, but not in performance or hiding rates. The performance levels are comparable. The mean mistake rates are 48 percent for males and 53 percent for females, and the difference is not significant (rank-sum test, $p = .218$, two-sided). Hiding rates are also similar. In Experiment 1, mistakes are hidden 31 percent by both males and females. In Experiment 2, females are slightly less likely to hide mistakes (38 versus 43 percent, rank-sum test, $p = .423$, two-sided). However, we do find a gender difference in punishment in Experiment 2. Male managers on average punish more harshly than female managers (€2.14 versus €0.87, $p < .001$), but do not differ in their disclosure penalty ($p = .880$). While the literature on gender differences in punishment is mixed, these findings are consistent with evidence that women's social preferences are generally more context-sensitive and more reciprocal than men's (Croson and Gneezy, 2009). In our setting, where punishment is costless but directly reduces a worker's payoff, this greater reciprocity may manifest as leniency, while male managers are more inclined to punish on principle (Eckel and Grossman, 1996).

Error-management styles. Looking at how managers punish depending on how mistakes are disclosed we classify managers into 3 types.[18] There are *never-punishers*, who do not punish regardless of the project outcome or disclosure of mistakes (16% across both experiments). The largest share are *disclosure-punishers*, who encourage hiding mistakes by punishing a worker who discloses a mistake more than they punish workers after a failed project without disclosures (49%). A substantial fraction are *suspicious-silence-punishers*, who use punishment to encourage truthful disclosure by punishing most heavily when neither worker discloses a mistake but the ambitious project fails (35%). The distribution of managers is not different when they are rematched each round or remain with the same team. Fig. B.5 visualises the distinct punishment profiles.

We can also assess whether different manager strategies impact team outcomes. Appendix Table A.4 compares the mean hiding rate, payoff, and mistake frequency at the team level across manager types. As managers in treatment STRANGERS cannot directly influence worker expectations beyond the matching group *average*, we expect – and indeed find – no significant differences between types (Panels (b) and (c)). For managers in treatment PARTNERS, managers can plausibly use repeated interaction with the same workers to shape the error-culture in their team to improve team-level outcomes. However, even in partnered teams we do not find significant differences across manager types (Panel (a)).

[18] We exclude those managers who never saw a round with no-disclosure and project failure from the manager-type analysis as we cannot classify their punishment profile. This excludes three managers in Experiment 1 and two managers in Experiment 2.

8 Discussion

We hypothesised that workers would be blamed, and consequently punished, for making mistakes, and that this punishment would increase with the potential earnings loss caused by the mistake. Across three studies, the evidence for this stakes-based blame mechanism is mixed. The vignette study supports the mechanism: larger potential losses increase blame and reduce bonuses after a disclosed mistake. In the incentivised laboratory experiments, however, stakes do not significantly affect either punishment or disclosure, even when the payoff difference and punishment range are substantially increased in Experiment 2. We also find no evidence that repeated interaction within stable teams reduces punishment or increases reporting. Thus, the comparative-statics predictions that motivate the experimental treatments are not supported in the incentivised team environment.

This divergence between the vignette and the laboratory experiments is informative. The vignette isolates blame in a simplified one-worker setting without real incentives, strategic interaction, or free-riding on another worker's disclosure. In that setting, respondents acting as manager react to the size of the foregone payoff in the predicted direction. The laboratory experiments add real incentives and a two-worker team environment in which disclosure has both informational benefits and distributional consequences. In this richer setting, managers do punish mistakes, but punishment does not scale systematically with the stakes. One possible interpretation is that managers distinguish unintentional mistakes from deliberate shirking or free-riding: higher losses make the outcome worse, but not necessarily the worker more culpable. Because our design does not separately manipulate intentions, this interpretation should be treated as suggestive rather than conclusive.

Although the laboratory experiments are designed primarily to test comparative statics rather than to estimate the absolute prevalence of concealment, the reporting patterns are still insightful. Disclosure is influential: managers typically use disclosed information to switch to the safer project. Moreover, punishments are generally modest relative to the material benefits of corrective action. The fact that reporting nevertheless remains incomplete across treatments suggests that even limited individual costs of disclosure can matter in team settings.

The team setting helps explain why disclosure may remain fragile. Disclosure benefits all team members by allowing the manager to switch to the safer project, but the reputational or monetary cost of disclosure falls primarily on the worker who comes forward. The team environment therefore creates a reporting externality. A worker who discloses her own mistake provides information that benefits the entire team, because it allows the manager to switch to the safer project. Yet the private cost of disclosure is concentrated on the reporting worker. A worker who has made a mistake may therefore benefit if the teammate reports his own mistake, while avoiding the additional exposure associated with reporting herself. This differs from standard public-good punishment settings, where punishment typically targets free-riders and can support cooperation (Fehr and Gächter, 2000). In our setting, punishment can fall on the source of the informational benefit, thereby discouraging the behaviour needed for efficient decision making.

The punishment patterns are consistent with this interpretation. At the team level, managers do not punish disclosed mistakes particularly harshly. In both experiments, total team punishment is lower after a disclosed mistake than after project failure with undisclosed mistakes. Yet the distribution of punishment within the team matters. When only one worker discloses, that worker receives a disproportionate share of the punishment. In Experiment 1, this produces an honesty penalty at the individual level. In Experiment 2, the same pattern is weaker, but disclosure still does not receive a robust individual reward. Thus, even when managers appear to recognise the value of disclosure at the team level, their punishment decisions may still make reporting less attractive for the individual worker. This pattern contrasts with the optimal-contract analysis of (Levitt and Snyder, 1997), in which early warning is supported by paying an agent who reports bad prospects more than an agent whose project continues and subsequently fails. Our experiment does not provide managers with such an ex ante commitment device: punishment remains discretionary, and the reporting worker receives no robust individual reward.

The absence of an effect of stable teams also qualifies the role of reputational incentives. We expected managers in ongoing relationships to have stronger incentives to avoid punishing disclosure, because a reputation for leniency could encourage future reporting. We do not find support for this prediction: workers in stable teams do not disclose more, and managers do not punish less, than in randomly rematched teams. This result should be interpreted as evidence about the repeated-interaction structure implemented here. It shows that merely keeping the same team together is not sufficient, in our environment, to generate a more favourable reporting culture. It does not imply that reputation, leadership, or explicit reporting norms would be ineffective in richer organisational settings.

Taken together, the results qualify a simple blame-avoidance account of mistake concealment. The vignette study shows that larger potential losses can increase blame and punishment, as predicted. In the incentivised team experiments, however, increasing the stakes does not lead to more punishment or less disclosure, and stable teams do not generate more reporting than randomly rematched teams. The more robust pattern is that disclosure has asymmetric consequences within teams. A disclosed mistake helps the manager take corrective action and thereby benefits the team, but the worker who reports her own mistake may bear a disproportionate share of the subsequent punishment. The problem is therefore not only whether mistakes are punished, but also how the costs of disclosure are allocated across team members. This helps explain why voluntary reporting may remain incomplete even when disclosure is payoff-relevant and punishments are limited.

The results also point to directions for future work. First, our design abstracts from costly effort. This helps isolate disclosure incentives, but it also separates the punishment of mistakes from the disciplining role that punishment may have when workers can reduce the probability of errors. Second, disclosure is always potentially useful for managerial decision making in our experiments. This is central to our setting, because reporting can both expose the worker to blame and create value by allowing the manager to respond to the mistake. Baader et al. (2024) provide related evidence on this trade-off by varying anonymity and the instrumental value of reporting. Our results complement this by studying a setting in which workers know whether they personally made a mistake and in which punishment can be allocated across individual team members. Future work could further examine how observability, instrumental value, effort choice, and explicit protection for reporters jointly affect both punishment and reporting in teams.

References

Abeler, J., Nosenzo, D., and Raymond, C. (2019). Preferences for Truth-Telling. *Econometrica*, 87(4):1115–1153.

Andreoni, J. and Bernheim, B. D. (2009). Social Image and the 50-50 Norm: A Theoretical and Experimental Analysis of Audience Effects. *Econometrica*, 77(5):1607–1636.

Andreoni, J. and Petrie, R. (2004). Public goods experiments without confidentiality: A glimpse into fund-raising. *Journal of Public Economics*, 88(7-8):1605–1623.

Baader, M., Bowen, S., Hochleitner, A., and Mills, R. (2024). To Hide or not to Hide? How Fear and Futility Affect the Decision to Report a Mistake. *SSRN Electronic Journal*.

Barron, K. (2019). Lying to appear honest. Technical report, WZB Discussion Paper.

Battigalli, P. and Dufwenberg, M. (2007). Guilt in Games. *American Economic Review*, 97(2):170–176.

Battigalli, P., Dufwenberg, M., and Smith, A. (2019). Frustration, aggression, and anger in leader-follower games. *Games and Economic Behavior*, 117:15–39.

Bolton, G., Dimant, E., and Schmidt, U. (2021). Observability and social image: On the robustness and fragility of reciprocity. *Journal of Economic Behavior & Organization*, 191:946–964.

BP U.S. Refineries Independent Safety Review Panel (2007). The report of the BP u.s. refineries independent safety review panel. Technical report, BP p.l.c. Panel chaired by James A. Baker, III.

Bureau of Labor Statistics (2024). Census of fatal occupational injuries summary, 2023. Technical report, U.S. Department of Labor. Accessed February 2025.

Chen, D. L., Schonger, M., and Wickens, C. (2016). otree: An open-source platform for laboratory, online, and field experiments. *Journal of Behavioral and Experimental Finance*, 9(C):88–97.

Croson, R. and Gneezy, U. (2009). Gender differences in preferences. *Journal of Economic Literature*, 47(2):448–474.

Dufwenberg, M. and Dufwenberg, M. A. (2018). Lies in disguise–a theoretical analysis of cheating. *Journal of Economic Theory*, 175:248–264.

Eckel, C. C. and Grossman, P. J. (1996). The relative price of fairness: Gender differences in a punishment game. *Journal of Economic Behavior & Organization*, 30(2):143–158.

Edmondson, A. C. (1996). Learning from Mistakes is Easier Said Than Done: Group and Organizational Influences on the Detection and Correction of Human Error. *The Journal of Applied Behavioral Science*, 32(1):5–28.

Fehr, E. and Gächter, S. (2000). Cooperation and Punishment in Public Goods Experiments. *American Economic Review*, 90(4):980–994.

Frese, M. and Keith, N. (2015). Action Errors, Error Management, and Learning in Organizations. *Annual Review of Psychology*, 66(1):661–687.

Friedrichsen, J., König, T., and Schmacker, R. (2018). Social image concerns and welfare take-up. *Journal of Public Economics*, 168:174–192.

Galeotti, F. and Rose, J. (2024). Costly confessions. Working paper.

Garicano, L. and Rayo, L. (2016). Why Organizations Fail: Models and Cases. *Journal of Economic Literature*, 54(1):137–192.

Gneezy, U., Kajackaite, A., and Sobel, J. (2018). Lying Aversion and the Size of the Lie. *American Economic Review*, 108(2):419–453.

Gold, A., Gronewold, U., and Salterio, S. E. (2014). Error Management in Audit Firms: Error Climate, Type, and Originator. *The Accounting Review*, 89(1):303–330.

Grant, M. J. C. and Larsen, G. Y. (2007). Effect of an Anonymous Reporting System on Near-miss and Harmful Medical Error Reporting in a Pediatric Intensive Care Unit. *Journal of Nursing Care Quality*, 22(3):213.

Gronewold, U., Gold, A., and Salterio, S. E. (2013). Reporting Self-Made Errors: The Impact of Organizational Error-Management Climate and Error Type. *Journal of Business Ethics*, 117(1):189–208.

Gurdal, M. Y., Miller, J. B., and Rustichini, A. (2013). Why Blame? *Journal of Political Economy*, 121(6):1205–1247.

Halpern, J. (2016). *Actual Causality: Chapter 6*. MIT Press.

Leigh, J. P. (2011). Economic burden of occupational injury and illness in the United States. *The Milbank Quarterly*, 89(4):728–772.

Levitt, S. D. and Snyder, C. M. (1997). Is no news bad news? Information Transmission and the role of "Early Warning" in the principal-agent model. *The RAND Journal of Economics*, 28(4):641–661.

Moore, J., Cigularov, K., Sampson, J., Rosecrance, J., and Chen, P. (2013). Construction Workers' Reasons for Not Reporting Work-Related Injuries: An Exploratory Study. *International journal of occupational safety and ergonomics : JOSE*, 19:97–105.

Ockenfels, A. and Werner, P. (2014). Scale manipulation in dictator games. *Journal of Economic Behavior & Organization*, 97:138–142.

Sobel, J. (2013). Giving and Receiving Advice. In Acemoglu, D., Arellano, M., and Dekel, E., editors, *Advances in Economics and Econometrics*, pages 305–341. Cambridge University Press, 1 edition.

Sorrapera, I., Suvorov, A., van de Ven, J., and Villeval, M. C. (2019). Doing bad to look good: Negative consequences of image concerns on pro-socal behavior. *Revue Economique*, 70(6):pp–945.

Suvorov, A., van de Ven, J., and Villeval, M. C. (2024). Selective information sharing and group delusion.

Tversky, A. and Kahneman, D. (1981). The Framing of Decisions and the Psychology of Choice. *Science*, 211(4481):453–458.

Utikal, V. (2012). A fault confessed is half redressed—Confessions and punishment. *Journal of Economic Behavior & Organization*, 81(1):314–327.

van Dyck, C., Frese, M., Baer, M., and Sonnentag, S. (2005). Organizational Error Management Culture and Its Impact on Performance: A Two-Study Replication. *Journal of Applied Psychology*, 90(6):1228–1240.

A Additional Tables

Table A.1: Robustness Vignette: OLS regression results

	Dependent variable:			
	Blame	Evaluate	Rehire	Bonus
	(1)	(2)	(3)	(4)
High stakes	0.61*	−0.17	0.01	48.36
	(0.33)	(0.17)	(0.01)	(49.17)
Disclose		−1.86***	−0.14***	−310.57***
		(0.18)	(0.04)	(36.48)
High×Disclose		−0.26	−0.07	−249.01***
		(0.28)	(0.05)	(56.08)
Constant	4.59***	8.00***	0.98***	1,184.55***
	(0.22)	(0.11)	(0.01)	(31.97)
Observations	243	486	486	486

Notes: High stakes denotes profits of \$10,000 rather than \$600 when the machine runs smoothly. Blame is not elicited in the Success scenario. Blame and Evaluate are 1–9 scales, Rehire is a dummy, and Bonus is in dollars. Clustered standard errors at the respondent level in parentheses. Significance levels: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.

Table A.2: Robustness Experiment 1: mistake hiding and punishment (parametric)

	Hiding (prob.)				**Punishment (€)**			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Sample	Full	Full	Block 1	Final 5 rounds	Full	Full	Block 1	Final 5 rounds
High Stakes	0.008	0.012	-0.028	-0.012	0.032	0.038	0.143*	0.023
	(0.026)	(0.026)	(0.058)	(0.040)	(0.049)	(0.052)	(0.074)	(0.075)
Partners	0.026	0.023	-0.010	0.058	-0.034	-0.028	-0.030	0.051
	(0.053)	(0.055)	(0.059)	(0.060)	(0.073)	(0.072)	(0.074)	(0.098)
Disclose × High Stakes					-0.106	-0.116	-0.284**	-0.077
					(0.083)	(0.083)	(0.107)	(0.112)
Disclose × Partners					0.027	0.031	0.025	-0.099
					(0.087)	(0.088)	(0.107)	(0.110)
Disclose					0.184**	0.188**	0.286***	0.211**
					(0.077)	(0.078)	(0.098)	(0.086)
High before Low		-0.037		0.002		-0.005		0.021
		(0.054)		(0.060)		(0.083)		(0.094)
Male		-0.013	-0.022	0.011		-0.011	-0.039	-0.016
		(0.041)	(0.045)	(0.049)		(0.071)	(0.068)	(0.091)
Age		-0.0008	-0.009	0.006		-0.010***	-0.011***	-0.010**
		(0.010)	(0.012)	(0.011)		(0.003)	(0.003)	(0.003)
Constant	0.326***	0.365	0.538**	0.218	0.308***	0.507***	0.458***	0.516***
	(0.044)	(0.224)	(0.256)	(0.243)	(0.048)	(0.108)	(0.100)	(0.133)
Observations	1,803	1,803	1,007	810	1,803	1,803	1,007	810
R^2	0.001	0.003	0.003	0.004	0.016	0.022	0.042	0.018

Notes: OLS regressions. The unit of observation is a worker-round in which that worker made a mistake. Samples vary by column as indicated; Final 5 rounds excludes the first 5 rounds of each block. Columns 2–4 include the age and gender of the worker, columns 6–8 those of the manager. High before Low indicates participants facing high-stakes rounds before low-stakes rounds. Clustered standard errors at the group level in parentheses. Significance levels: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.

Table A.3: Robustness Experiment 2: mistake hiding and punishment (parametric)

	Hiding (prob.)				Punishment (€)			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Sample	Full	Full	Block 1	Final 5 rounds	Full	Full	Block 1	Final 5 rounds
Stakes High	0.000	0.022	-0.111	0.004	0.147	0.227	-0.991*	0.073
	(0.047)	(0.047)	(0.091)	(0.051)	(0.190)	(0.202)	(0.496)	(0.239)
Stakes High × Disclose					0.160	0.260	0.618	0.215
					(0.378)	(0.317)	(0.649)	(0.432)
Disclose					0.044	0.141	-0.077	-0.163
					(0.343)	(0.290)	(0.401)	(0.248)
High before Low		-0.131		-0.120		-1.013		-1.192*
		(0.092)		(0.120)		(0.625)		(0.647)
Male		0.006	0.005	0.034		1.405**	1.332**	1.745**
		(0.092)	(0.084)	(0.101)		(0.586)	(0.529)	(0.633)
Age		0.007	0.001	0.037***		0.006	-0.045	0.016
		(0.007)	(0.008)	(0.010)		(0.126)	(0.152)	(0.163)
Constant	0.433***	0.339*	0.431**	-0.243	1.411***	0.922	2.032	0.868
	(0.052)	(0.175)	(0.185)	(0.264)	(0.343)	(2.649)	(3.199)	(3.313)
Observations	954	954	557	422	954	954	557	422
R^2	0.000	0.019	0.013	0.065	0.003	0.095	0.077	0.147

Notes: OLS regressions. The unit of observation is a worker-round in which that worker made a mistake. Samples vary by column as indicated; Final 5 rounds excludes the first 5 rounds of each block. Columns 2–4 include the age and gender of the worker, columns 6–8 those of the manager. High before Low indicates participants facing high-stakes rounds before low-stakes rounds. Clustered standard errors at the group level in parentheses. Significance levels: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.

Table A.4: Mean outcomes by manager type, by experiment and matching

(a) Experiment 1: Partners

Outcome	Never punishers	Disclosure punishers	Suspicious-silence punishers	p-value
Mean Punishment	0.00	0.40	0.27	0.003
Mean Expected Payoff	5.42	5.16	6.16	0.290
Mean Hiding Rate	0.331	0.364	0.327	0.704
Share project A (reports=0)	0.814	0.908	0.992	0.177
Share of rounds with team mistake	0.738	0.796	0.705	0.345
Share (n)	10.0% (4)	65.0% (26)	25.0% (10)	

(b) Experiment 1: Strangers

Outcome	Never punishers	Disclosure punishers	Suspicious-silence punishers	p-value
Mean Punishment	0.00	0.52	0.30	<0.001
Mean Expected Payoff	5.54	5.48	5.91	0.359
Mean Hiding Rate	0.263	0.377	0.347	0.162
Share project A (reports=0)	1.000	0.936	0.942	0.151
Share of rounds with team mistake	0.756	0.752	0.700	0.402
Share (n)	17.8% (8)	44.4% (20)	37.8% (17)	

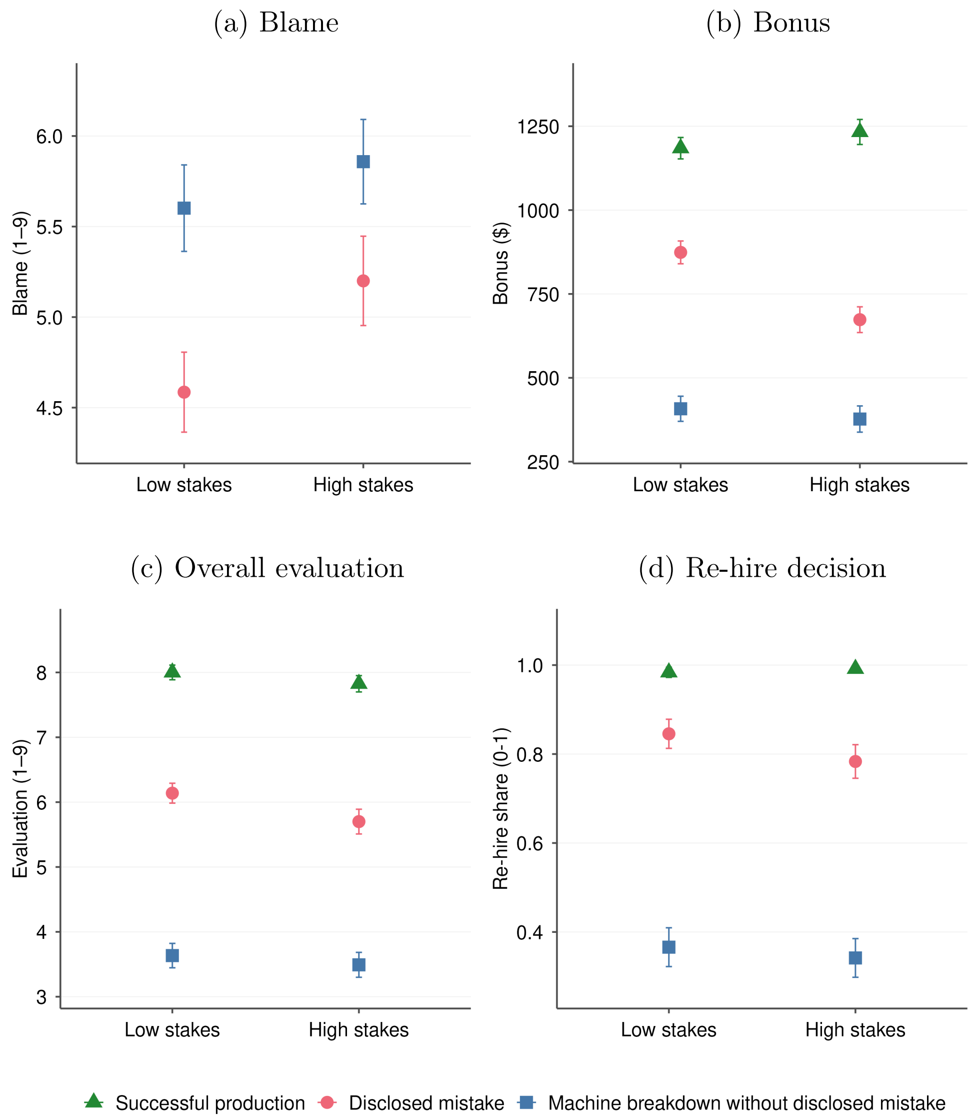
(c) Experiment 2 (Strangers only)

Outcome	Never punishers	Disclosure punishers	Suspicious-silence punishers	p-value
Mean Punishment	0.00	2.07	1.57	<0.001
Mean Expected Payoff	4.87	6.01	6.22	0.767
Mean Hiding Rate	0.404	0.437	0.469	0.641
Share project A (reports=0)	0.940	0.933	0.895	0.521
Share of rounds with team mistake	0.789	0.733	0.724	0.565
Share (n)	19.6% (9)	39.1% (18)	41.3% (19)	

Notes: The unit of observation is a manager; entries are means of each manager's average across rounds. Mean Punishment is the average punishment per worker in euro over rounds with a team mistake or a failed ambitious project; the cap is €2 in Experiment 1 and €12 in Experiment 2. Mean Expected Payoff is the per-round payoff per team member in euro, using the expected value when the ambitious project was chosen without a team mistake. Partners = matched teams; Strangers = unmatched teams. P-values are from Kruskal–Wallis tests across the three types. See *Error-management styles* in Section 7 for a description of manager types.

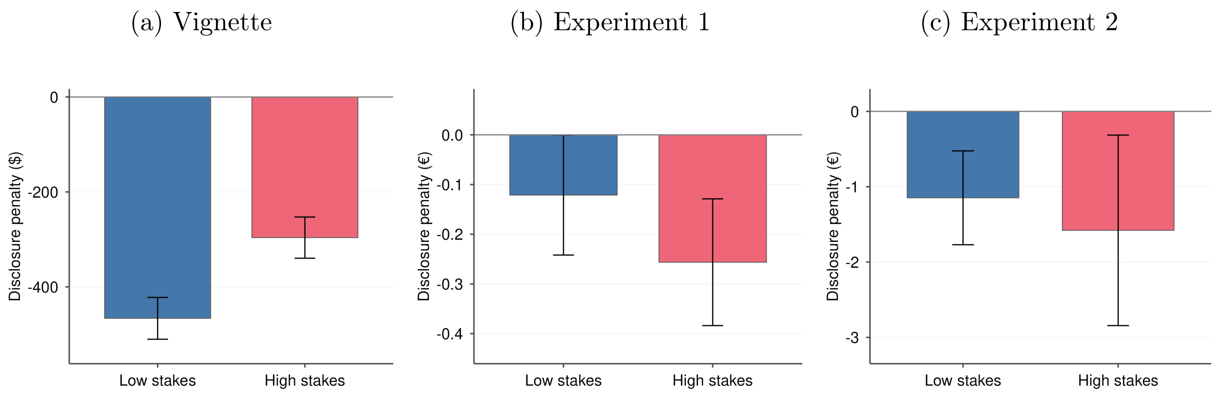
B Additional Figures

Figure B.1: Manager evaluations of worker performance by stakes and project outcome (vignette study)



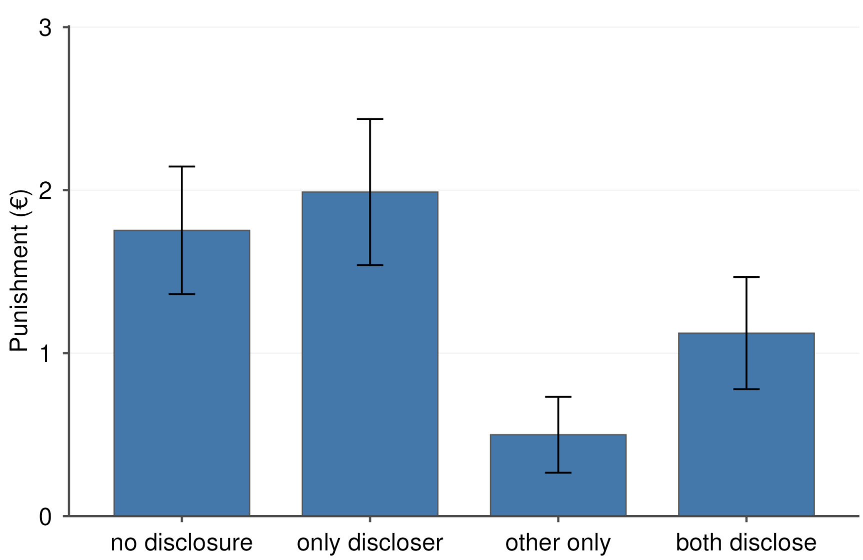
Notes: Means across the 243 vignette respondents, who each evaluated the mechanic in the three scenarios shown in the legend. LOW/HIGH stakes denotes profits of \$600/\$10,000 when the machine runs smoothly, varied between subjects. Blame is not elicited after successful production. Respondents were told that \$1,000 is the average bonus. Error bars indicate +/- 1 s.e.

Figure B.2: Disclosure penalty at the team level



Notes: The disclosure penalty is the change in punishment when a mistake is disclosed rather than hidden; negative values mean disclosure is rewarded. Panel (a) covers the vignette's single mechanic and is the bonus after a breakdown with no disclosed mistake minus the bonus after a disclosed mistake, differenced within participant. Panels (b) and (c) are the total team punishment when at least one worker discloses minus when neither does, in rounds with at least one team mistake, differenced within matching group. LOW/HIGH stakes are \$600/\$10,000 in (a), €10/€20 in (b) and €8/€40 in (c). Error bars indicate +/- 1 s.e.

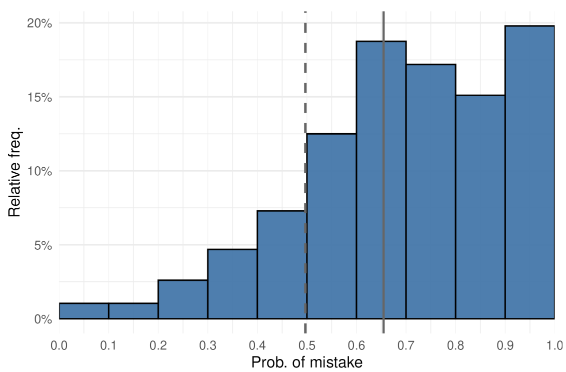
Figure B.3: Punishment amounts by disclosure of mistakes (Experiment 2)



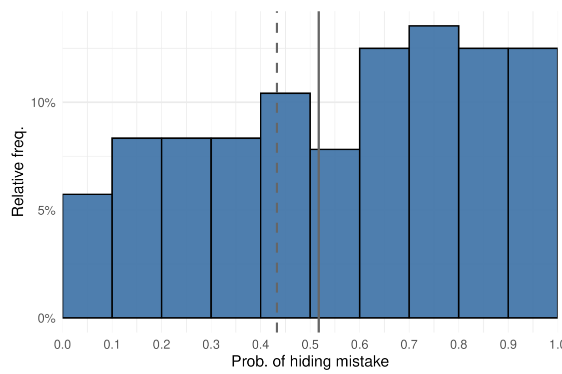
Notes: Mean punishment received by a worker, by own and other worker's disclosure of a mistake, conditional on the plotted worker having made a mistake. The punishment cap is €12, against €2 in Experiment 1. Means are computed per matching group. Error bars indicate +/- 1 s.e.

Figure B.4: Histogram of beliefs (Experiment 2)

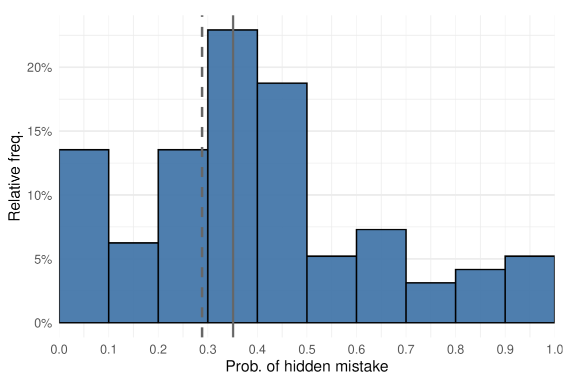
(a) Worker mistake rate



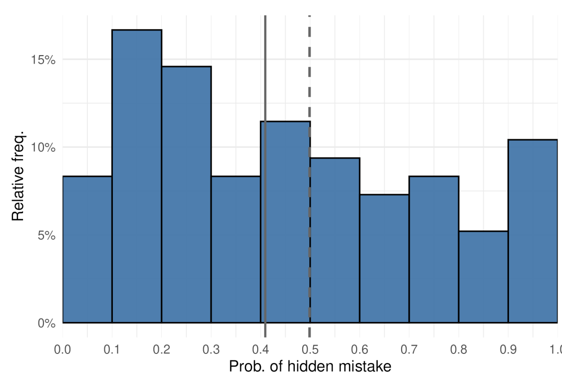
(b) Worker hiding rate



(c) Worker hiding rate
(only *other* worker disclosed a mistake)

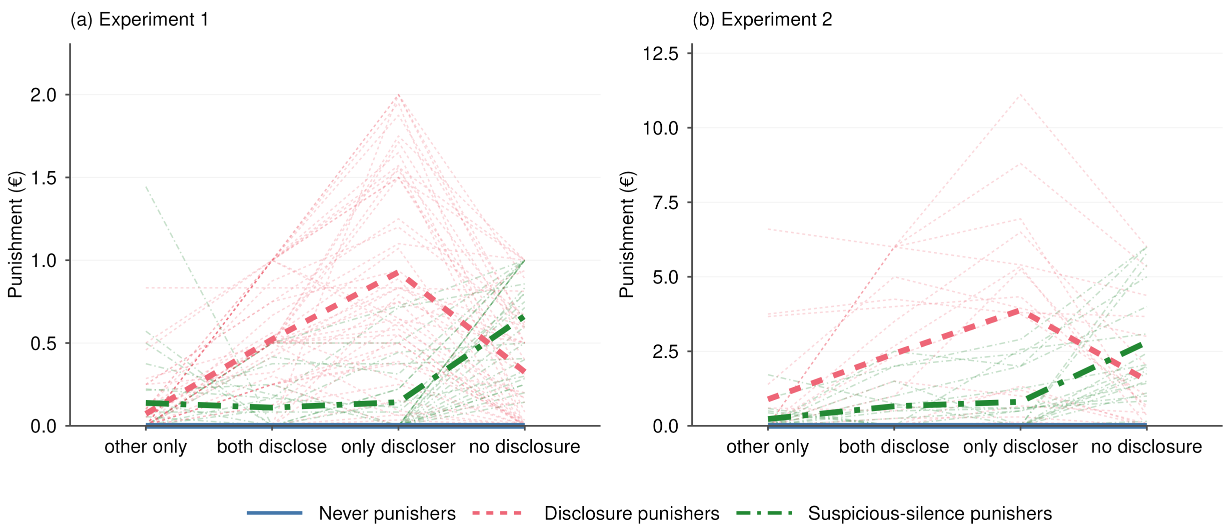


(d) Worker hiding rate
(no disclosures and project failure)



Notes: This figure plots the distribution of incentivised beliefs elicited in Experiment 2 at the end of each block. Panels B.4a and B.4b show worker beliefs about the rate at which other workers make a mistake and about the rate at which they hide a mistake they have made. Panels B.4c and B.4d show manager beliefs about the probability that a worker is hiding a mistake, conditional on the other worker disclosing one in panel B.4c, and conditional on the ambitious project failing when neither worker disclosed a mistake in panel B.4d. Bars give relative frequencies across participants, pooling the two blocks. In each panel the grey solid line marks the mean belief and the grey dashed line marks the corresponding rate observed in the data.

Figure B.5: Punishment profiles by manager type



Notes: Faint lines are individual managers; thick lines are the mean profile of each type. Punishment is measured per individual worker: when exactly one worker disclosed, the punishment of the disclosing worker ('only discloser') and of the non-disclosing worker ('other only') are shown separately; when both or neither disclosed, the average punishment per worker is shown. Only rounds in which the ambitious project failed or a mistake was made are included. This profile therefore conditions on a team mistake (any round in which the ambitious project failed or any team member made a mistake), in contrast to Figures 1 and B.3, where punishment is conditioned on the specific worker under consideration having personally made a mistake. The punishment cap is €2 in Experiment 1 and €12 in Experiment 2. See *Error-management styles* in Section 7 for a description of manager types.

C Proofs

Since disclosing a mistake is only relevant after making a mistake, we focus on the situation where at least one worker made a mistake, and so project A will necessarily fail if selected. Assume worker i made a mistake and let r the belief of worker i that worker j discloses a mistake. We start by examining properties of $\Delta U_i(r, y)$. The following lemma will prove helpful.

Lemma C.1. *$\Delta U_i(r, y)$ is a decreasing function in r for any $y \in \{A, B\}$.*

Proof. We need to distinguish two cases, depending on which project the manager selects in the absence of disclosed mistakes by any of the workers.

Case I: $y = A$; the manager selects project A when no worker disclosed a mistake. We have:

$$\Delta U_i(r, A) = mr[p_i(0,1) - p_i(1,1)] + (1 - mr)[\pi^B - \pi^F + p_i(0,0) - p_i(1,0)]. \tag{6}$$

The four possible punishments are then as follows. If none of the workers reported a mistake and the project fails, the manager updates the probability that the failure is due to a mistake to:

$$\hat{m}_T(0,0) = 1 - \frac{(1-m)^2(1-q)}{a(r)}, \tag{7}$$

where $a(r) := m^2(1-r)^2 + 2m(1-m)(1-r) + (1-m)^2(1-q)$. This is the probability that at least one worker made a mistake in the absence of any reports. Each worker shares equally in the blame. The punishment for worker i becomes:

$$p_i(0,0) = w_b \tfrac{1}{2} \hat{m}_T(0,0)(\pi^S - \pi^F).$$

In all other cases, at least one worker disclosed a mistake, and the manager is certain that a mistake has been made. If worker i is the only worker disclosing a mistake, then the manager's posterior that worker j also made a mistake is:

$$\hat{m}_{11}(1,0) = \frac{m(1-r)}{1-mr}. \tag{8}$$

The weighted share of worker i in the blame is then:

$$b_i(1,0) = \hat{m}_{10}(1,0) + \tfrac{1}{2}\hat{m}_{11}(1,0) = 1 - \tfrac{1}{2}\hat{m}_{11}(1,0) = 1 - \frac{m(1-r)}{2(1-mr)}. \tag{9}$$

Since worker i is the only one disclosing a mistake, that worker gets the full credits for being honest, and the total punishment for worker i becomes:

$$p_i(1,0) = w_b \left[1 - \frac{m(1-r)}{2(1-mr)}\right](\pi^S - \pi^F) - w_h(\pi^B - \pi^F).$$

If worker j is the only one disclosing, the total punishment for worker i becomes:

$$p_i(0,1) = w_b \left[\frac{m(1-r)}{2(1-mr)}\right](\pi^S - \pi^F),$$

as the worker gets no credits for reporting. If both workers report, they share the blame and credits for reporting equally. The total punishment for worker i becomes:

$$p_i(1,1) = w_b \tfrac{1}{2}(\pi^S - \pi^F) - w_h \tfrac{1}{2}(\pi^B - \pi^F).$$

Plugging all punishments in $\Delta_i(r)$ gives, after some rewriting:

$$\begin{aligned}\Delta U_i(r,A) &= mr\left[\frac{m-1}{2(1-mr)} w_b(\pi^S - \pi^F) + w_h \tfrac{1}{2}(\pi^B - \pi^F)\right] \\ &+ (1-mr)\left[w_b \tfrac{1}{2}\left(\frac{m-1}{1-mr} - \frac{(1-m)^2(1-q)}{a(r)}\right)(\pi^S - \pi^F) + (1+w_h)(\pi^B - \pi^F)\right].\end{aligned} \tag{10}$$

The blame term in the first brackets is decreasing in r, since $m < 1$ and $mr/(1-mr)$ is increasing in r. The blame term in the second brackets is decreasing in r whenever $-(1-mr)/a(r)$ is decreasing in r. Taking the derivative of this fraction, the sign of the numerator can be shown to be equal to the sign of $-m[(mr-1)^2 + q(1-m)^2]$, which is strictly negative. Collecting the honesty terms across the brackets gives $1 + w_h - mr(1 + \frac{1}{2}w_h)$ which is also decreasing in r. Thus, $\Delta U_i(r,A)$ is a strictly decreasing function in r.

Case *II*: $y = B$; the manager selects project B when no worker disclosed a mistake. We have :

$$\Delta U_i(r,B) = mr[p_i(0,1) - p_i(1,1)] + (1-mr)[p_i(0,0) - p_i(1,0)]. \tag{11}$$

In this case, the worker never gets credit for disclosing a mistake. The punishments can be determined by setting $w_h = 0$ in the punishment expressions above, and by setting $q = 0$ in $a(r)$ (since in this case the manager will never observe the outcome of project A, they cannot distinguish between a failed or successful project A). It is then easy to show that $\Delta U_i(r,B)$ is decreasing in r. ■

We next characterise the equilibrium set in symmetric strategies. Let $\tilde{r}(y)$ be such that $\Delta U_i(\tilde{r}, y) = 0$, i.e., $\tilde{r}(y)$ is the reporting rate of worker j that makes worker i indifferent between hiding or disclosing a mistake, when the manager selects project y in the absence of disclosed mistakes.

Lemma C.2 (Equilibrium characterisation)**.** *The equilibrium set in symmetric strategies is characterised as follows. If $\tilde{r}(A) > 0 \geq \hat{r}$, there exists a unique equilibrium outcome in which workers disclose mistakes with positive probability ($r^* > 0$) and managers select project A in the absence of disclosed mistakes and B otherwise. If $\tilde{r} \geq \hat{r} > 0$, both a positive equilibrium ($r^* > 0$) and a never-disclosing equilibrium ($r^* = 0$) exist. In all remaining cases, the equilibrium outcome is unique and workers never disclose ($r^* = 0$), and managers always select project B.*

Proof. Let r be the expected disclosure rate of workers. For any $r < \hat{r}$, the expected gain of disclosing a mistake is given by $\Delta U_i(r,B)$. Since $p_i(1,r_j) > p_i(0,r_j)$, it follows that $\Delta U_i(r,B) < 0$.

Hence, if worker i expects worker j to never disclose ($r_j = 0$), then worker i prefers not to disclose. If $0 < \hat{r}$, the manager strictly prefers project B in the absence of disclosed mistakes. Thus, $r^* = 0$ constitutes an equilibrium and it exists whenever $\hat{r} > 0$.

Consider next $r \geq \hat{r}$. The expected gain of disclosing a mistake is then given by $\Delta U_i(r, A)$. If $\tilde{r}(A) < \hat{r}$, then by Lemma C.1, $\Delta U_i(r, A) < 0$ for any $r \geq \hat{r}$. Hence, worker i is not willing to mix, and $r^* > 0$ cannot be an equilibrium. By contrast, if $1 > \tilde{r}(A) \geq \hat{r}$ and $\tilde{r}(A) \geq 0$, an equilibrium exists with $r^* > 0$. To see this, take $r^* = \tilde{r}(A)$. This implies $\Delta U_i(r^*, A) = 0$, so that worker i is willing to mix and disclose with probability r^*. Since $r^* \geq \hat{r}$, the manager prefers project A in the absence of disclosed mistakes. Hence, this constitutes an equilibrium. If $\tilde{r}(A) \geq 1$, then each worker prefers to disclose a mistake if the other worker always discloses, and so always disclosing is an equilibrium. If an equilibrium exists with $r^* > 0$, then r^* is uniquely determined: By Lemma C.1, the function $\Delta U_i(r, A)$ is strictly decreasing in r so there is at most one value of r solving $\Delta U_i(r, A) = 0$ for $r \geq \hat{r}$. ■

Proof of Proposition 1. In any equilibrium with $0 < r^* < 1$, workers must be indifferent, implying $\Delta U_i(r^*, A) = 0$. By Lemma C.1, $\Delta U_i(r, A)$ is strictly decreasing in r, so $\Delta U_i(r, A) = 0$ has at most one solution. Moreover, $\Delta U_i(r, A)$ is decreasing in π^S, so an increase in π^S shifts $\Delta U_i(r, A)$ downward. It follows that the solution r^* weakly decreases in π^S.

D Instructions & Details - Vignette Study

Recruitment. We recruited 243 participants from the US via Prolific in November 2025. We restricted recruitment to those having a 95% approval rating on the platform.
Treatment randomisation. The treatments HIGH and LOW are randomised between subject. The order the scenarios are presented in are randomised between participant, with an approximately equal share seeing each of the scenarios first.

Instructions

[INTRO SCREEN]

Please read the following scenario carefully and answer the questions. **[STORY SCREEN]**

Imagine you are the supervisor at a small drinks factory. The factory's profits depend on a bottling machine. On a normal day, if the machine runs without problems, the factory earns $10,000/$600 in profit.

Sometimes, however, the machine breaks down and production stops. This can happen for two reasons:

- Normal wear and tear: About 2 out of every 10 days, a part wears out and needs to be replaced.

- A servicing mistake: The mechanic might accidentally use the wrong type of oil, since two bottles look very similar.

If the wrong oil is used, the mechanic will notice the mistake quickly, when putting the bottle back in its place. The mechanic then has a choice:

- Report the mistake to you: You can call a technician to fix the problem. Production continues at a reduced level, and the factory still makes $400 in profit.

- Hide the mistake from you: You cannot tell whether a later breakdown is due to normal wear and tear or a mistake. In this case, production stops completely, and the factory makes no profit.

You are asked to evaluate the mechanic based on whether they report an error and the profits that result. You will see three possible cases (in random order):

- The mechanic does not report an error, and the machine runs smoothly.

- The mechanic does not report an error, but the machine later breaks down.

- The mechanic reports an error, allowing you to authorise the cheaper fix before production starts.

In each case, your feedback will help decide the mechanic's bonus and chances for promotion.

[VIGNETTE SCREENS]

Success:

Scenario 1: No errors were reported, and the machine runs smoothly. Profits are XXX.

Report:

Scenario 2: An error was reported. A technician was hired to fix the machine and profits are $400.

No Report:

Scenario 3: No errors were reported but the machine broke down. There are no profits.

[BELOW EACH VIGNETTE STORY - IN RANDOM ORDER]

Please evaluate the mechanic.

Evaluation: How would you evaluate the mechanic's performance? (1-very bad to 9–very good)

[if error] Blame: To what extent do you blame the mechanic for the error? (1-not at all to 9–very much)

Bonus: Per the factory's terms of employment each employee receives a monthly bonus. For mechanics the average performance bonus is $1,000. Based on their performance, how much bonus would you recommend they receive? (slider with integers $0 to $2,000 & starting position at $1,000.)

Re-hire: Would you re-hire the mechanic? (Yes or No)

E Instructions - Lab Experiment

The instructions below are those shown to participants in Experiment 1, lightly edited for print: treatment-conditional content is shown jointly, while participants saw only the values and text for their own treatment. Experiment 2 used identical instructions up to the parameter changes described in Section 6, a €12.00 base wage paid each round, and one paid round instead of two. Experiment 2 additionally elicited incentivised beliefs at the end of each block: participants stated, for 10 randomly selected rounds, in how many a worker made (or reported) a mistake; one belief question per block was randomly selected and paid €4.00 if the stated count exactly matched the realised count. Participants read the role-specific manager/worker instructions (Sections E.1 and E.2) and then needed to answer all of the comprehension questions (Section E.3) before continuing. Each *Part* is shown on screens sequentially.

E.1 Manager Instructions

Part 1: General Information

Please do not communicate with other participants. If you have any questions, please raise your hand and an experimenter will come to you to answer your question in private.

You will receive a guaranteed show-up fee of **€5.00**. On top of this, you can earn more money. This experiment has **2 blocks**. Each block has **10 rounds**. At the end of the experiment, 2 of the 20 rounds of the experiment will be randomly selected to determine your payment.

We will now explain the instructions for the first block. Instructions for the other blocks will be provided at the start of each block.

Part 2: Roles and Tasks

You will be part of a team of three, consisting of one manager and two workers. **You are a manager**. You will keep this role throughout the entire experiment.

Each round each worker starts by working on simple tasks. We will first describe the tasks, and then explain how you can earn money. You do not do the tasks.

The Tasks Workers will get 3 types of tasks:

1. a simple addition task where they must sum 2 numbers between 1 and 10;

2. a simple multiplication task where they must multiply 2 numbers between 1 and 10;

3. a simple colour identification task.

In the colour identification task they see a coloured word. Each word spells out a colour. Like in the picture below, the font colour can be different from what the word spells out.

The task is to click the coloured box on the bottom that corresponds to either the **spelled-out colour** (what the word means), or the **font colour** (the colour the word is written in).

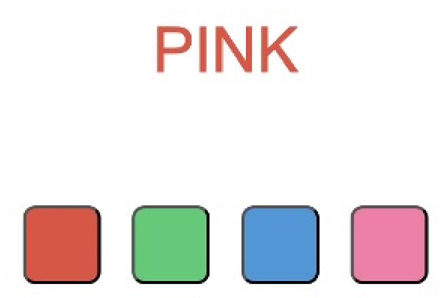


Figure E.1: Example of the colour identification task (manager instructions).

For example, in the screenprint above: If the task is to click the **font colour, they should click the red box**. If the task is to click the **spelled-out colour, they should click the pink box**. Clicking the wrong coloured box is a mistake.

In each task they see two words sequentially. Before each pair of words they will see an instruction screen, as in the example screen below. The first instruction refers to the first word in the sequence, and the second instruction refers to the second word in the sequence.

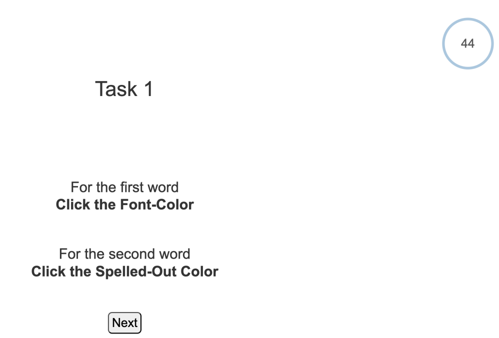


Figure E.2: Example of the instruction screen for a pair of words (manager instructions).

The colour options are displayed on the screen as coloured boxes (blue, red, green, orange, purple, yellow, black, and gray).

There will be **20 tasks** per round. There is a countdown that varies from round to round. It is between 60 to 75 seconds. **They should complete all 20 tasks before the countdown reaches 0**. The countdown is displayed in the top right corner. It starts when they start the first task and they are cut off when the countdown reaches 0. **Any tasks they do not complete in the time count as a mistake.**

Workers get one round of 20 tasks to practice after the instructions.

You will see the solutions to the tasks in each round. You do not see their answers.

Part 3: Payoffs, Projects, and Punishment

We next explain how you can earn money.

After the workers have each worked on their tasks, you choose between two projects: an **Ambitious Project** or a **Backup Project**.

Ambitious Project This project gives the highest possible earnings if it is successful. However, it gives the lowest payoff if it fails. Its success depends on whether mistakes were made and chance. Success is determined by the following two cases:

- **Case 1:** One or both workers made a mistake on any of their tasks that round. In this case, the project fails and the team loses the option to get a high payoff. The Ambitious Project then always pays **€2.00** to each team member.

- **Case 2:** None of the workers made a mistake on their tasks. In this case, the Ambitious Project is likely to be successful and result in a high payoff. Without mistakes the project is successful with 80 percent chance and the **payoff is then €10.00 (LOW) / €20.00 (HIGH) for each team member**. With 20 percent chance it still fails and the payoff is €2.00 for each team member.

Backup Project The payoff of this project is fixed and always **€4.00** for each team member.

Before you decide between the two projects, each worker will **privately learn whether or not they made a mistake** on any of their tasks. Only they learn this.

They can choose to report their mistake to you and the other worker or keep it hidden. If they make a mistake and do not report it to you and you select the Ambitious Project, the payoff will be €2.00, but you will not be able to tell whether they made a mistake or whether the low payoff occurred by chance.

After you select the project, the payoffs are determined and shown to each team member. **You then have the option to punish** one or both of the workers by lowering their payoff. You can choose how to hand out the €2.00 punishment. You can:

- not lower any worker's payoff;

- lower the payoff of exactly one of the workers by up to **€2.00**;

- lower the payoffs of both workers, where both punishments together cannot exceed **€2.00**.

Your payoff is not affected by handing out punishment.

Your payoff per round is the project payoff. This means you can earn up to €10.00 (LOW) / €20.00 (HIGH) per round.

Part 4: Timing and Matching

To summarise, the timing per round is as follows:

1. Each worker works on their tasks and privately learns whether or not they individually made a mistake.

2. Each worker decides whether or not to report their mistake to the manager.

3. The manager selects a project.

4. Payoffs are revealed and the manager decides on any punishment.

The above describes the procedure of one round. In total, there will be 10 rounds in this block.

[Treatment PARTNERS:] You remain with the **same team for all 10 rounds** this block. For Block 2 you will be paired with a **completely new team with two new workers** and remain with this team for the 10 rounds in Block 2.

[Treatment STRANGERS:] At the start of each round, you will be **randomly paired with two workers**.

E.2 Worker Instructions

Part 1: General Information

Please do not communicate with other participants. If you have any questions, please raise your hand and an experimenter will come to you to answer your question in private.

You will receive a guaranteed show-up fee of **€5.00**. On top of this, you can earn more money. This experiment has **2 blocks**. Each block has **10 rounds**. At the end of the experiment, 2 of the 20 rounds of the experiment will be randomly selected to determine your payment.

We will now explain the instructions for the first block. Instructions for the other blocks will be provided at the start of each block.

Part 2: Roles and Task

You will be part of a team of three, consisting of one manager and two workers. **You are a worker**. You will keep this role throughout the entire experiment.

Each round each worker starts by working on a simple task. We will first describe the task, and then explain how you can earn money from working on the task. Managers do not do the tasks.

The Task You will get 3 types of tasks:

1. a simple addition task where you must sum 2 numbers between 1 and 10;

2. a simple multiplication task where you must multiply 2 numbers between 1 and 10;

3. a simple colour identification task.

In the colour identification task you see a coloured word. Each word spells out a colour. Like in the picture below, the font colour can be different from what the word spells out.

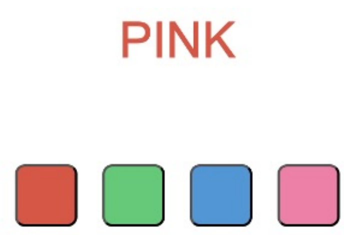


Figure E.3: Example of the colour identification task (worker instructions).

The task is to click the coloured box on the bottom that corresponds to either the **spelled-out colour** (what the word means), or the **font colour** (the colour the word is written in).

For example, in the screenprint above: If the task is to click the **font colour, you should click the red box**. If the task is to click the **spelled-out colour, you should click the pink box**. Clicking the wrong coloured box is a mistake.

In each task there are two sequential coloured words. Before each pair of words you will see an instruction screen, as in the example screen below. The first instruction refers to the first word in the sequence, and the second instruction refers to the second word in the sequence.

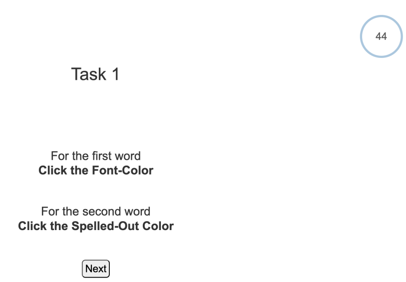


Figure E.4: Example of the instruction screen for a pair of words (worker instructions).

The colour options are displayed on the screen as coloured boxes (blue, red, green, orange, purple, yellow, black, and gray).

There will be **20 tasks** per round. **You should complete all 20 tasks before the count-down reaches 0**. The time limit varies by round between 60 to 75 seconds. It is displayed at the start of each round in the top right corner of the screen. It starts when you press Start on the first task and you are cut off when the countdown reaches 0. **Any tasks you do not complete in the time count as a mistake.** If you refresh the page this also counts as a mistake.

You get one round of 20 practice tasks after the instructions.

In each round the managers see the solutions to the tasks. They do not see your answers.

Part 3: Payoffs, Projects, and Punishment

We now explain how you can earn money.

After the workers have each worked on their tasks, the manager chooses between two projects: an **Ambitious Project** and a **Backup Project**.

Ambitious Project This project gives the highest possible earnings if it is successful. However, it gives the lowest payoff if it fails. Its success depends on whether mistakes were made and chance. Success is determined by the following two cases:

- **Case 1:** One or both workers made a mistake on their tasks. In this case, the project fails and the team loses the option to get a high payoff. The Ambitious Project then always pays **€2.00** to each team member.

- **Case 2:** None of the workers made a mistake on their tasks. In this case, the Ambitious Project is likely to be successful and result in a high payoff. Without mistakes the project is successful with 80 percent chance and the **payoff is then €10.00 (LOW) / €20.00 (HIGH) for each team member**. With 20 percent chance it still fails and the payoff is €2.00 for each team member.

Backup Project The payoff of this project is fixed and always **€4.00** for each team member.

Before the manager decides between the two projects, you will **privately learn whether or not you made a mistake** on any of the tasks. Only you will know this. The other worker will also privately learn if they made a mistake.

You can choose to report your mistake to the manager and the other worker or keep it hidden. If you make a mistake and do not report it to the manager and the manager selects the Ambitious Project, the payoff will be €2.00, but the manager will not be able to tell whether you made a mistake or whether the low payoff occurred by chance.

After the manager selects the project, the payoffs are determined and shown to each team member. The manager then has the option to **punish** one or both of the workers by lowering their payoff. The manager can choose between the following options:

- Not lower any worker's payoff;

- Lower the payoff of exactly one of the workers by up to **€2.00**;

- Lower the payoffs of both workers, where both punishments together cannot exceed **€2.00**.

The manager's payoff is not affected by handing out punishment.

Your payoff per round is the project payoff minus any punishment. This means you can earn up to €10.00 (LOW) / €20.00 (HIGH) per round.

Part 4: Timing and Matching

To summarise, the timing per round is as follows:

1. Each worker works on their tasks and privately learns whether or not they individually made a mistake.
2. Each worker decides whether or not to report their mistake to the manager.
3. The manager selects a project.
4. Payoffs are revealed and the manager decides on any punishment.

The above describes the procedure of one round. In total, there will be 10 rounds in this block.

[Treatment PARTNERS:] You remain with the **same team for all 10 rounds** this block. For the second block you will be paired with a **completely new team with a new worker and new manager**.

[Treatment STRANGERS:] At the start of each round, you will be **randomly paired with a worker and manager**.

E.3 Comprehension Questions

The questions and answer options below are those shown in Experiment 1; they were identical for all participants and blocks, with only the answer graded as correct for Question 3 depending on the participant's first block (€10.00 in LOW, €20.00 in HIGH).

Answer the questions below before continuing.

1. What is the largest punishment the manager can hand out to one worker at the end of each round?

◯ €2.00

◯ €0.50

◯ €5.00

2. How much does the Backup Project pay?

◯ €4.00

◯ €1.00

◯ €0.50

3. What is the maximum amount that the Ambitious project pays to each team member when none of the workers made a mistake?

◯ €2.00

◯ €10.00

◯ €20.00

4. What is the maximum amount that the Ambitious project pays to each team member when one or both workers made a mistake?

- ◯ €2.00
- ◯ €10.00
- ◯ €20.00

5. True or false: If one or more workers report a mistake, then the Backup project is guaranteed to pay more than the Ambitious project.

- ◯ False
- ◯ True

6. If a worker made a mistake:

- ◯ The worker can choose whether or not to report this to the team
- ◯ This will always be reported to the team

7. If a worker made one or more mistakes:

- ◯ The Ambitious project always fails
- ◯ There is still a chance that the Ambitious project can be successful

8. If none of the workers made any mistakes:

- ◯ The Ambitious project will always be successful
- ◯ There is still a chance that the Ambitious project can fail