**CES ifo** WORKING Papers

# Intentions Undercover – Hiding Intentions is Considered Unfair

Tim Friehe
Verena Utikal

**CESifo**
**Center for Economic Studies & Ifo Institute**

# Intentions Undercover – Hiding Intentions is Considered Unfair

## Abstract

Unfair intentions provoke negative reciprocity from others, making their concealment potentially beneficial. This paper explores whether people hide their unfair intentions from others and how hiding intentions is itself perceived in fairness terms. Our experimental data show a high frequency of cover-up attempts and that affected parties punish the concealment of intentions, establishing that people consider not only unkind intentions but also hiding intentions unfair. When choosing whether or not to hide intentions, subjects trade-off the lower expected punishment when the cover up of unfair intentions is successful against the higher expected punishment when cover up is unsuccessful. In an attempt to better understand fairness perceptions, we present a typology of punisher types and show that hiding unkind intentions is treated differently than unkind intentions, possibly establishing a behavioral category of its own.

JEL-Code: C900, D010, K420.

Keywords: intentions, reciprocity, fairness, avoidance, cover up, experiment.

*Tim Friehe*
*University of Bonn*
*Adenauerallee 24 - 42*
*Germany - 53113 Bonn*
*tim.friehe@uni-bonn.de*

*Verena Utikal*
*University of Erlangen-Nürnberg*
*Lange Gasse 20*
*Germany - 90403 Nürnberg*
*verena.utikal@fau.de*

## 1. Introduction

People care about the payoffs of others and this has important implications for behavior and well-being (e.g., Fehr and Schmidt 2006). For example, in Fehr and Schmidt (1999), people are worse off when others' payoffs exceed their own. However, such outcome-based theories have recently been criticized for their neglect of reciprocity (Charness and Rabin 2002, Falk et al. 2003). In this vein, Falk et al. (2008), among others, present evidence showing that theories that include roles for both outcomes and intentions in fairness perceptions have greater explanatory power. In models that incorporate intentions (e.g. Cox et al. 2007 or Falk and Fischbacher 2006), unfair intentions provoke negative reciprocity from others, making their concealment potentially beneficial.

The present contribution is the first to study the possibility of hiding one's intentions as well as how concealment is treated by parties who are affected by the cover-up. We are particularly interested in how covering-up intentions is perceived in fairness terms and for that reason allow affected parties to voice their feelings via a punishment option. In addition, we seek to establish a typology of subjects in order to shed light on the question of how many people care about outcomes, intentions, and/ or the concealment of intentions.

Intentions play a prominent role in legal codes. For a given outcome, the legal implications can vary widely according to the individual's intention. This can be seen, for example, in the practice of making punitive damages in civil cases contingent on the malicious intent of the tortfeasor. Similarly, with regard to the range of possible sanctions in criminal cases, it makes a huge difference whether a suspect is convicted of manslaughter, second-degree murder, or first-degree murder. Anticipating this decisive role of intentions, perpetrators go to great lengths to cast doubt on their malicious intentions when under scrutiny in order to limit the adverse legal implications resulting from the criminal act. This regularly culminates in another wrong, namely evidentiary misdeeds (e.g., concealing evidence), which are themselves punishable.[1] In other words, the legal system specifies sanctions for attempts to manipulate the legal decision-maker's information about any specifics related to the incident. Prior contributions to the literature on deterrence and legal proceedings have been criticized for focusing on evidence as something that investigators must uncover rather than something that violators may cover up (Sanchirico 2006), despite the fact that incentives for concealment activities are very important in practice (as they are relevant in almost every legal proceeding).

---

[1] For example, in the US, obstruction of justice, criminal contempt, and perjury are relevant categories that may be punished by fines or imprisonment (Sanchirico 2012).

To explore the actual use of cover-up activities and how offenders are treated by subjects who are affected by the cover-up, we rely on a laboratory experiment, as successful attempts at concealment are by definition difficult to monitor. In our one-shot, two-player experimental design, player A chooses between two probability distributions, while chance ultimately determines whether a given sum of points is distributed equally across both players or unequally (i.e., in favor of player A). One of the two probability distributions is heavily biased towards the allocation that favors player A; the other is skewed towards the equal allocation of points. The first (second) probability distribution thus represents an unkind (kind) procedural choice on the part of player A. Thus, in our setup, we model intention by the procedure choice. Next, contingent on the allocation, player A decides whether or not to conceal his or her procedural choice (i.e., to prevent player B from learning the initial choice of one of the two probability distributions). A given allocation may result from either a kind or an unkind procedural choice by player A, and player A can try to manipulate player B's information about this selection. Finally, player B can punish player A, where we differentiate punishment levels in various informational settings.

Turning to our results, we observe punishment for unkind intentions. Many subjects anticipate this and invest to hide such intentions. When concealment is successful, hiding intentions significantly reduces the level of punishment received. However, when the cover-up is not successful, players B impose a significantly higher level of punishment. In other words, player A's manipulation of information significantly increases punishment (with both the procedure and the allocation held constant); substantiating that hiding intentions is considered unfair and has some role to play with regard to fairness preferences. In our section on behavioral predictions, we show that this punishment may be traced back to the fact that successful concealment disallows affected parties to reciprocate according to the specifics of the case at hand.[2] The aspects of fairness dealt with in the prior literature are also important in our study. More specifically, we establish that punishment in our experiment depends on both the outcome (i.e., whether or not a fair allocation results) and on the first-mover's intentions (i.e., whether or not the unkind procedure was chosen).[3].

The present paper proposes a typology of subjects according to whether they are concerned about outcomes, intentions, and/or the hiding of intentions. We find that the majority of

---

[2] Due to the interaction being one-shot only, there is no role to play for deterrence of cover-up in our experimental design.

[3] Accordingly, our data is inconsistent with theories that relate fairness perceptions either solely to outcomes (e.g., Fehr and Schmidt 1999) or intentions (e.g., Dufwenberg and Kirchsteiger 2004). Instead, our evidence (like that presented by Falk et al. 2008) speaks in favor of understanding fairness as something influenced by both outcomes and intentions.

subjects display outcome-based as well as intention-based preferences. Interestingly, some subjects treat cover-up attempts differently than unkind intentions, allowing the conjecture that they represent a behavioral category of their own in our setting.

The remainder of the paper is organized as follows. The next section briefly discusses the related literature. Section 3 introduces the experimental design. Section 4 offers behavioral predictions. Section 5 presents the results, and Section 6 concludes.

## 2. Literature

The present research contributes to the discussion about fairness preferences. Early contributions to this line of research have emphasized that people dislike unfair allocations, that is, they focus on outcomes and may take steps to prevent advantageous or disadvantageous inequity (Bolton and Ockenfels 2000, Fehr and Schmidt 1999).[4] Charness and Rabin (2002) present a critique of this explanation, stressing the importance of efficiency and reciprocity concerns. Dufwenberg and Kirchsteiger (2004) and Sebald (2010) follow the lead of Rabin (1993) and discuss fairness with an emphasis on the importance of intentions. Falk and Fischbacher (2006) and Krawcyk (2011) present frameworks with a combined focus on both outcomes and intentions. We will derive our predictions on the basis of the tractable model developed by Cox et al. (2007). Falk et al. (2008) provide evidence in favor of the conjecture that both outcomes and intentions impact fairness, focusing on both positive and negative reciprocity and implementing a moonlighting game in which trust is efficiency-enhancing.[5] They compare the responses from a treatment in which the first-mover's action is determined by a random device to responses from a treatment in which the first mover makes the determination without stochastic influence, such that decisions in the latter treatment are only attributable to the intentions of the first-mover. In contrast, in our setting, the first-mover chooses between two probability distributions (i.e., procedures); thus, every allocation is possible, independent of the first-mover's choice. This feature allows us to include cover-up investment, that is, a payment made by the first-mover with the sole purpose of hiding his or her intentions from the second-mover.

Bolton et al. (2005) and Charness and Levine (2007) similarly consider the scenario in which an allocation is determined by the first-mover's choice and a move of nature. In the experimental setup used by Charness and Levine (2007), a worker's wage is co-determined by the employer's choice and luck, such that a given wage level may be the result of either a

---

[4] Trautmann (2009) extends the inequity aversion presented in Fehr and Schmidt (1999) to the case of uncertain payoffs, making inequity in expected payoffs decisive for social preferences.
[5] Falk et al. (2003) similarly provide evidence that cannot be dovetailed with an exclusive focus on outcomes.

generous employer and bad circumstances or a miserly employer and good circumstances. Importantly and in contrast to our setting, the second-mover has complete information (i.e., he or she knows the wage offered by the employer). The researchers show that –with the level of the effective wage held constant – workers repay a high wage with high effort and punish a low wage with low effort; that is, they exhibit behavior responsive to their employers' intentions. Bolton et al. (2005) study procedural fairness, finding that a fair procedure may substitute for a fair outcome and that randomness itself must be perceived as fair.

Motives apart from outcome and intentions can also make people punish others. For example, Utikal (2012) investigates the impact of confessions on punishment in a laboratory study. The results show that punishment after a confession is less likely than when randomly detected. Bartling et al. (2014) and Conrads and Irlenbusch (2013) determine that people can avoid punishment by remaining willfully ignorant about the possible negative consequences of their actions for others. Our paper focuses on punishment for people who have knowingly and intentionally chosen to inflict negative consequences on others and have tried to hide their intentions. We compare behavior in an experiment to theoretical predictions. Our paper is thereby also related to the theoretical literature on avoidance investments by offenders and their implications for optimal law enforcement (Malik 1990, Nussim and Tabbach 2009, Langlais 2008, Sanchirico 2006).

Our experimental design allows second-movers to punish first-movers in order to convincingly establish how second-movers perceive the hiding of intentions, thus including the possibility of peer punishment. Recently, Leibbrandt and López-Pérez (2012) have considered different motives for second-party and third-party punishment, arriving at the conclusion that inequity aversion and selfish preferences best explain their results. In contrast, our results clearly show that people focus significantly on intentions and possible attempts to hide intentions when determining punishment. To the best of our knowledge, our study is the first to address the costly opportunity to conceal one's intentions in order to influence the perception of others and how others respond to such behavior. In the spirit of Fehr and Schmidt (1999) who estimate the distribution of outcome-based preferences across the population, we present evidence on the distribution of outcome-based and intention-based preferences, and preferences against cover-up attempts.

## 3. Experimental Design

Our experimental setup is a one-shot game with two players and four stages (see Table 1). Both players receive an endowment of 10 points that can be invested during the game. In addition, a total of 100 points will be split between player A and player B. There are two possible allocations: either an equal split (50/50) or an unequal split favoring player A (80/20).

In **stage 1**, player A chooses between two allocation procedures. Procedure *kind* makes the equal split of the 100 points very likely, while procedure *unkind* makes the unequal split very likely. Specifically, choosing *kind* means selecting the lottery (0.1, 80/20; 0.9, 50/50), while *unkind* results in the lottery (0.95, 80/20; 0.05, 50/50).[6] In our instructions,[7] we refrain from using the terms "lottery", "(un-)kind", and "(un-)equal"; the lotteries are instead visualized as a left urn and a right urn with different compositions of red and black balls, where a red ball represents the unequal allocation.

|  | Players A and B receive an endowment of 10 points |
|---|---|
| **Stage 1** | Player A chooses between two lotteries<br><br>• *unkind* lottery: (0.95, 80/20; 0.05, 50/50)<br><br>• *kind* lottery (0.1, 80/20; 0.9, 50/50) |
| **Stage 2** | Nature determines the allocation, either 50/50 or 80/20 |
| **Stage 3** | Player A learns the allocation and chooses whether to invest his or her endowment in cover-up activities<br><br>**No cover-up:**<br>• Player B receives full information (**procedure, allocation, and cover-up decision**) with probability $a^{high}=0.8$<br>• Player B receives partial information (**allocation only**) with probability $1-a^{high}=0.2$<br><br>**Cover-up:**<br>• Player B receives full information (**procedure, allocation, and cover-up decision**) with probability $a^{low}=0.2$<br>• Player B receives partial information (**allocation only**) with probability $1-a^{low}=0.8$ |
| **Stage 4** | Player B receives full or partial information<br>Player B can punish player A<br>1 punishment point costs player B 1/6 points of his or her initial endowment |

**Table 1: Experimental setup (treatment CERT)**

In **stage 2**, the move of nature determines the allocation.

---

[6] The lotteries are asymmetric to make the treatment comparable to the treatment UNCERT, where we vary the likelihood of the unequal split in procedure *unkind* (holding constant the probability of the equal split in procedure *kind*).

[7] Translated instructions can be found in the appendix.

In **stage 3**, player A chooses whether to invest his or her endowment in cover-up activities or not. Such activities lower the probability that player B will learn about player A's intentions (i.e., player A's decision between *kind* and *unkind*). If player A does not invest in cover-up activities, player B receives full information about the procedure decision, the allocation, and the cover-up decision with probability $a^{high}=0.8$, and information about the allocation only with probability $1-a^{high}=0.2$. If player A invests in concealment, player B receives full information about the procedure decision, the allocation, and the cover-up decision with probability $a^{low}=0.2$, and information about the allocation only with probability $1-a^{low}=0.8$.

In **stage 4**, player B can punish player A. Deducting 1 point from player A's payoff costs player B 1/6 of the points of his or her endowment. Punishment is restricted such that it cannot yield negative payoffs for player A.[8]

Figure 1 presents the game tree, highlighting the possibility that player B may have to decide with imperfect information.
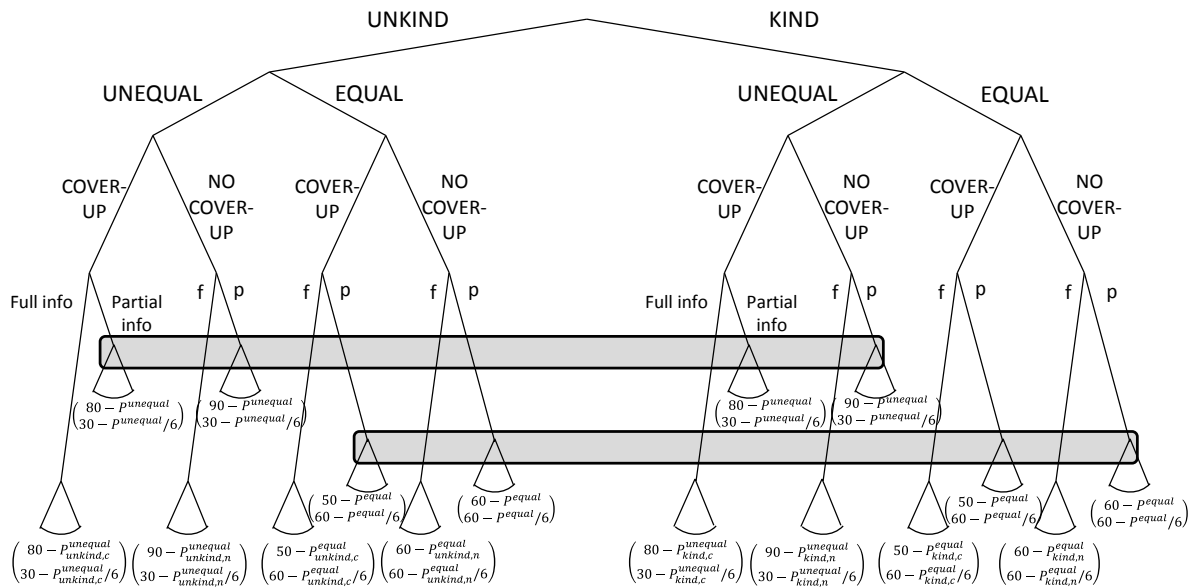


**Figure 1: Game tree**

**Treatments**

To control for the degree of procedural fairness and to check whether our chosen values drive behavior, we varied the likelihood of the unequal split in procedure *unkind* (holding constant the probability of the equal split in procedure *kind*). In treatment CERT, it is almost certain that the unequal allocation will be implemented when procedure *unkind* is selected. Choosing the *unkind* procedure in UNCERT is less unkind because the unequal split is less likely. More specifically:

---

[8] As a result, punishment is either restricted to [0,50] or to [0,60].

CERT: The likelihood of the unequal split in the *unkind* procedure is *p*=0.95.

UNCERT: The likelihood of the unequal split in the *unkind* procedure is *p*=0.5.

**Procedures**

We apply the strategy method, eliciting player A's cover-up decision for both the *unequal* and the *equal* allocation (Selten 1967). Player B decides on punishment levels for all 10 possible scenarios before being informed about the payoff-relevant history of events. The possible scenarios are as follows: Player B may observe only the allocation (*equal* or *unequal*); alternatively, player B observes player A's procedure choice (*kind* or *unkind*), the allocation, and player A's cover-up choice. The order of scenarios presented on the screen was randomized across players B.[9]

In order to better understand player A's motivation with regard to whether or not to hide intentions, we elicited the players' beliefs concerning player B's punishment in all scenarios. One belief specification was randomly determined for payoff. If the stated belief differed less than 5 points from the true value, subjects received 50 points.

We conducted 6 sessions with 24 subjects each in November 2013, for a total of 144 participants (42% male). All sessions were held at the *BonnEconLab*. The experiment took about 60 minutes and was conducted using z-Tree (Fischbacher, 2007). At the end of the experiment, one experimental point was exchanged for 0.15 € in compensation; the average income of participants was €9.90 ($13.26). We recruited participants using the online system ORSEE (Greiner 2004). Each subject sat at a randomly assigned PC terminal and was given a copy of the instructions. Control questions ensured that all participants understood the game. The experiment did not start until all subjects had answered all questions correctly. After the experiment, participants completed a questionnaire. Finally, participants were called one by one to the exit. They received their payment in cash outside the laboratory with sufficient time allowed between the exiting participants to ensure privacy with respect to the amount of money received.

---

[9] Brandts and Charness (2011) present evidence that results obtained for punishment using the strategy method are more likely than not to be lower bounds for the effect sizes achieved using the direct response method.

### 4. Behavioral Predictions

In this section, we explain the behavior expected in our experiment. We start with behavior when subjects care only about their own monetary consequences. Subsequently, we turn to the scenario of key interest, namely the case in which (at least some) participants have fairness preferences.

4.1 Self-interest prediction

When participants focus solely on their own monetary payoffs, no player B will sacrifice money to reduce the payoffs of player A at the last stage of our one-shot interaction. Without the threat of punishment, players A will have no reason to invest in concealing the history of events. Moreover, at the initial stage of the game, players A will always choose the *unkind* procedure independent of whether treatment CERT or UNCERT applies, as it promises a higher expected payoff.

4.2 Fairness predictions

It has repeatedly been documented that many people have fairness concerns (e.g., Fehr and Schmidt 2006). Accordingly, we expect that the predictions derived in this section will be of central interest in our analysis of the experimental data in Section 5.

Following the logic of backward induction, we first turn to player B's punishment choice in stage 4. It may be that player B observes only the allocation, *equal* or *unequal*; alternatively, player B might observe player A's choice in stage 1 (*kind* or *unkind*), the allocation, and player A's decision regarding cover-up investments. Central to our study is the fact that a given allocation (either *equal* or *unequal*) that player B faces when determining punishment can be the result of very disparate histories of events. Accounting for the endowment and the implemented allocation, effective point allocations may be either (90/30) or (60/60) when player A abstains from concealment and either (80/30) or (50/60) when player A invests in concealment.

According to the models proposed by Fehr and Schmidt (1999) and Bolton and Ockenfels (2000), individuals dislike distributional inequity and thus players B may punish to correct for any disadvantageous inequity between the payoffs of players A and B. Specifically, player B may punish player A when the unequal allocation has been implemented, as this implies an inequity that can be corrected by player B deducting points from player A. In contrast, no

punishment should be observed when the allocation *equal* results from the draw by nature. Importantly, individual well-being in these models is a function of final payoffs alone. This means that the punishment imposed by player B will differ between the possible allocations, but not according to the procedure chosen when the allocation is held constant. In other words, punishment subsequent to the observation that the unequal allocation has been implemented should be independent of whether player B knows that player A chose the *unkind* or the *kind* procedure. However, it could be argued that player A's decision to invest in cover-up affects relative payoffs and may thus be reflected in the punishment allotted to player A, where the reduced inequity due to cover-up costs would in principle make a smaller punishment acceptable from player B's standpoint.

In summary, our punishment predictions building on *outcome*-inequity models such as Fehr and Schmidt (1999) are as follows:

$H_{out}^1$: **Punishment contingent on allocation *unequal* will exceed punishment contingent on allocation *equal* (i.e., there will be outcome-based punishment).**

$H_{out}^2$: **Punishment will be independent of player A's choice between procedures (*unkind* and *kind*) for a given allocation (i.e., there will be no intention-based punishment).**

$H_{out}^3$: **Cover-up activities will decrease punishment.**

Theories that focus on outcomes have been criticized for their neglect of reciprocity (e.g., Charness and Rabin 2002, Falk et al. 2003). In the following paragraphs, we will derive our hypotheses by considering a very simple setup (building on Cox et al. 2007) in which the intentions of the other player influence individual decision-making by changing the *emotional state* of the decision-maker, impacting the marginal rate of substitution between one's own payoff and the other player's payoff for one's own utility.[10] Specifically, the individual utility function of player B is

$$u_B = m_B - c(P) + \theta * (m_A - P) \tag{1}$$

where $m_B$ is player B's gross income in the allocation drawn by nature, $m_A$ is the gross income of player A, $P$ is the number of points that player B deducts from player A at costs *c(P)*, and $\theta$ represents the emotional state which is influenced by reciprocity. We introduce the emotional state such that higher income for player A increases player B's utility only

---

[10] As emphasized by Cox et al. (2007), the model gains tractability relative to other approaches such as Falk and Fischbacher (2006) because it is a preference model, not an equilibrium model.

when $\theta > 0$. In this regard, we specifically assume that

$$\theta = \theta_0 + a(\Delta_{EmB}, \Delta_m) \tag{2}$$

In expression (2), $\theta_0$ is a parameter and may be equal to zero or be positive or negative. The function $a$ is increasing in both arguments.

The first argument represents the difference between player B's expected payoff resulting from player A's procedure choice and that implied by *normal* behavior by player A.[11] Our inclusion of this difference is consistent with the arguments presented in Falk and Fischbacher (2006) and Trautmann (2009), among others, in that people are concerned about the expected distribution of outcomes. Taking the procedure *kind* as normal behavior, we obtain $\Delta_{EmB} = 0$ when player A chooses *kind* and $\Delta_{EmB} = p\,20 + (1-p)50 - q\,20 - (1-q)50 < 0$ when player A chooses *unkind*. Obviously, player A's choice of the procedure *unkind* harms player B in expected terms. The harmfulness of the unkind procedure is greater in treatment CERT than in treatment UNCERT: In CERT (where *p=0.95* and *q=0.1*), it holds that $\Delta_{EmB} = \{0, -25.5\}$; in UNCERT (where *p=0.5* and *q=0.1*), it holds that $\Delta_{EmB} = \{0, -12\}$.

The second argument of the function $a$ is the difference in net payoffs between the players according to the allocation drawn by nature, that is, $\Delta_m = m_B - (m_A - k)$, where *k=10* represents player A's concealment costs. Our inclusion of this difference is consistent with the arguments presented in Krawcyk (2011), for example, in that the payoff-relevant allocation will influence the evaluation of the other player's behavior. In our experiment, $\Delta_m = \{10, 0, -50, -60\}$.

When player B decides upon punishment for a specific scenario in stage 4, he or she seeks to maximize the expression in (1) for a level of $\theta$ that is determined by choices at previous stages. Clearly, when $\theta$ is positive, player B altruistically benefits from seeing that player A has a higher payoff and will thus abstain from punishment. In contrast, when $\theta$ is negative, player B resents the fact that player A has a high gross payoff and may thus impose punishment on player A. Assuming a strictly convex punishment cost function for expositional ease, the first-order condition for the level of punishment is

---

[11] The inclusion of *normal* behavior and its definition is in line with Cox et al. (2007), but not critical for our argumentation.

$$-\theta = c'(P) \tag{3}$$

The comparative-statics analysis yields the prediction that punishment decreases with $\theta$ (i.e., the closer it approaches zero from the left, the smaller the punishment will be). Upon examination of expression (2), it becomes clear that procedure *unkind* and allocation *unequal* lower the level of the emotional state. Together with expression (3), this exerts upward pressure on the level of punishment. As a result, without specifying the relationship between the unfairness in procedure and the unfairness in outcomes with regard to their influence on the function a,[12] we obtain

$$P(\Delta_{EmB} < 0, \Delta_m^1) \geq P(\Delta_{EmB} < 0, \Delta_m^2) \tag{4}$$

$$P(\Delta_{EmB} < 0, \Delta_m^1) \geq P(\Delta_{EmB} = 0, \Delta_m^1) \tag{5}$$

$$P(\Delta_{EmB} < 0, \Delta_m^2) \geq P(\Delta_{EmB} = 0, \Delta_m^2) \tag{6}$$

$$P(\Delta_{EmB} = 0, \Delta_m^1) \geq P(\Delta_{EmB} = 0, \Delta_m^2) \tag{7}$$

where $\Delta_m^1 = \{-50, -60\}$ and $\Delta_m^2 = \{10, 0\}$. The first two inequalities state that punishment after the choice of the procedure *unkind* and the realization of the allocation *unequal* (weakly) exceeds the punishment that results when player A chooses procedure *unkind* but allocation *equal* results, or when player A opts for procedure *kind* but allocation *unequal* was drawn by nature. The punishments resulting for the latter scenarios in turn (weakly) exceed the punishment for the case with procedure *unkind* and allocation *equal* (as expressed in inequalities (6) and (7)). Since the first argument of the function *a* takes a much lower value when player A chooses procedure *unkind* in treatment CERT in comparison to treatment UNCERT, the level of punishment for the procedural choice of player A in treatment UNCERT is weakly lower than that in treatment CERT.

In summary, punishment predictions building on models incorporating *intentions* are as follows:

$H_{int}^1$: **There will be outcome-based punishment, that is, the level of punishment for player A will be weakly higher when allocation *unequal* results.**

---

[12] For example, Krawcyk (2011) uses the intuitive assumption that an adverse allocation is even worse when it comes about by means of an unfair procedure.

$H_{int}^2$: **There will be intention-based punishment, that is, the level of punishment for player A will be weakly higher when procedure _unkind_ is chosen. Intention-based punishment will be weakly higher in CERT than in UNCERT.**

Player B is also asked to assign punishment without knowledge of player A's procedure and cover-up decisions. Importantly, this scenario is possible in our design independent of player A's cover-up decision. In this case, player B knows about the outcome-based inequity in gross payoffs, but is not informed about the inequity in net payoffs (as there is no information about player A's cover-up decision) or about player A's intentions. As a result, player B must make use of some kind of expected level of the emotional state $\tilde{\theta}$ when determining the punishment level. This expected value may be conceived of as a convex combination of the four different emotional states possible for the given allocation ($\Delta_{EmB}$ may be either negative or equal to zero, and player A may have spent his or her endowment or not). Out of these four states, the one in which player A chooses _unkind_ and does not reduce the net payoff difference by spending his or her endowment on concealment is associated with the lowest level of the emotional state $\theta$, implying the highest level of punishment when player B can choose punishment contingent on the history of events.

Ignorance about player A's decisions is harmful to player B. This is a result of the inadequacy of punishment for each actual contingency. The actual difference in utility between a knowledgeable player B and an ignorant one using the true state (i.e., the level of $\theta$ that is actually mandated by the history of events) is:

$$m_B - c\left(P\big(\theta(\Delta_{EmB}, \Delta_m)\big)\right) + \theta(\Delta_{EmB}, \Delta_m)[m_A - P(\theta(\Delta_{EmB}, \Delta_m))] \qquad (8)$$
$$> m_B - c\left(P(\tilde{\theta})\right) + \theta(\Delta_{EmB}, \Delta_m)[m_A - P(\tilde{\theta})]$$

$$\Leftrightarrow c\left(P(\tilde{\theta})\right) - c\left(P\big(\theta(\Delta_{EmB}, \Delta_m)\big)\right) \qquad (9)$$
$$> -\theta(\Delta_{EmB}, \Delta_m)[P(\tilde{\theta}) - P\big(\theta(\Delta_{EmB}, \Delta_m)\big)]$$

where the last inequality follows from the optimality of the punishment level assigned when player B knows the history of events. The punishment level based on the expected emotional state may either be excessive (the additional punishment relative to the one based on accurate information is not cost-justified) or suboptimal (additional punishment relative to the one

based on accurate information is cost-justified but not imposed). Thus, cover-up activities are similar to player A's selection of procedure *unkind*, since they decrease the expected payoff of player B at the expense of the payoff of player A. As a result, player A's choice between cover-up and no cover-up should influence the emotional state of player B, such that a more general definition would be

$$\theta = \theta_0 + a(\Delta_{EmB}, \Delta_m, cu) \tag{10}$$

where the value of function $a$ is lower when there has been investment in concealment (i.e., when $cu = 1$), all else held constant. In other words, the actual concealing of intentions maybe considered as an additional way in which unfairness can emerge from player B's standpoint. Applying the same logic as above, cover-up investment decreases the level of $\theta$ and thereby calls for higher punishment when all else is held equal. The argumentation is, however, complicated by the fact that cover-up investment changes the comparison of net payoffs to the benefit of player B. This presumably dampens the increase in punishment without nullifying or reversing it. Consequently, we consider the following hypothesis in our empirical analysis:

$H_{int}^3$: **Cover-up activities will be punished.**

This concludes the description of our hypotheses for punishment levels in stage 4 of the game.

As a next step, we turn to player A's decision about cover-up activities in stage 3. Under theories focusing exclusively on distributional fairness, the only possible advantage gained from cover-up activities is the reduction in the inequity of payoffs resulting from the expenditure. This results because punishment will only respond to the outcome, not player A's choice in stage 1. In accordance with Fehr and Schmidt (1999) and Bolton and Ockenfels (2000), since the inequity-reducing motive for cover-up activities is absent when allocation *equal* applies, we predict that there will be no attempted cover-up in this case.

$H_{out}^4$: **Cover-up investments will be more likely for the *unequal* allocation.**

$H_{out}^5$: **Cover-up investments will not be correlated to the choice of procedure.**

Turning to theories incorporating both distributional fairness and intentions, the advantages of cover-up investments lie in their concealment of intentions and their effect on the comparison of net payoffs. It follows from our discussion of player B's punishment behavior that procedure *unkind* incites higher punishment in comparison to the scenario in which player B

does not know the procedural choice. As a result, player A may find that there is a concealment benefit from cover-up activities when he or she actually chose procedure *unkind*. In this scenario, player A can benefit from keeping player B in the dark about the history of events because he or she will receive a lower punishment. In treatment CERT, the procedure *unkind* may be considered particularly unfair. The pros and cons of cover-up investments depend on the realized allocation. The change in the comparison of net payoffs in player B's favor becomes relevant when the allocation *unequal* applies, as sinking 10 points into the cover-up investment reduces the difference in payoffs. In addition, the punishment decrease that results from hiding the history of events from player B will depend on the allocation. For example, when a player B will strongly suspect player A's unkind intentions simply by observing allocation *unequal*, this will moderate the benefit obtained from cover-up activities. Conversely, when a player B has the prior that almost no player A has opted for the procedure *unkind*, then concealment of malicious intentions when allocation *unequal* applies will offer great benefits. This argumentation may be summarized as follows:

$H_{int}^4$: **Cover-up investments will be more likely for the *unequal* allocation when the concealment benefit is not much weaker in this scenario.**

$H_{int}^5$: **Cover-up investments will be more likely when procedure *unkind* is chosen than when procedure *kind* is selected. This difference will be more pronounced in CERT than in UNCERT.**

Finally, we briefly turn to player A's decision about the procedure to be used in stage 1. A very inequity-averse player A will choose the procedure *kind* in both treatment CERT and treatment UNCERT. When inequity aversion is not as strong, player A's expectations about the punishment that will be imposed by player B become important. When player A believes that player B will be more lenient when procedure *unkind* promises the unfair allocation with only a 50% probability (i.e., in treatment UNCERT), this will entice more players A to opt for *unkind*. As argued above, such an expectation is consistent with the modeling inspired by Cox et al. (2007).

$H_{int}^6$: **Procedure *unkind* will be more likely in treatment UNCERT than in CERT.**

In a setting with pure outcome-inequity aversion, player A will choose *unkind* only when he or she expects that the punishment received in the unfair allocation will not dominate the payoff benefits. If that is the case, then the likelihood of *unkind* should be comparable across CERT and UNCERT since punishment does not rely on the treatment differences.

$H^6_{out}$**: Procedure *unkind* will be as likely in treatment UNCERT as in CERT.**

The argumentation above allows us to explore different aspects in order to establish whether the behavior of our subjects is more aligned with pure outcome-inequity aversion or with a theory emphasizing outcome and procedural fairness. Table 2 gives an overview of the behavior predicted by theories incorporating outcome only and theories incorporating both outcome and intentions. Both kinds of theory predict outcome-based punishment (summarized as Hypothesis $H^1$) and therefore outcome-based cover-up activities (Hypothesis $H^4$), whereas all other expectations differ. Most importantly, we develop very specific expectations about how the concealment of intentions will be evaluated and dealt with by those adversely affected. The hypotheses in bold are supported by our empirical analysis.

| | Theories incorporating outcome | Theories incorporating outcome and intentions |
|---|---|---|
| **Punishment** | unequal>equal ($H^1$) | |
| | unkind=kind ($H^2_{out}$) | **unkind>kind ($H^2_{int}$)** |
| | cover-up<no cover-up ($H^3_{out}$) | **cover-up>no cover-up ($H^3_{int}$)** |
| **Investment in cover-up** | unequal>equal ($H^4$) | |
| | unkind=kind ($H^5_{out}$) | **unkind>kind (CERT: unkind>>kind) ($H^5_{int}$)** |
| **Choice of unkind procedure** | UNCERT=CERT ($H^6_{out}$) | **UNCERT>CERT ($H^6_{int}$)** |

**Table 2: Hypotheses (Hypotheses supported by our results are in bold.)**

## 5. Results

The results section is split into four parts. Following the order of Table 2, we first address the punishment imposed by player B on player A. We then discuss player A's decision to invest in cover-up activities, followed by his or her procedure choice. The section concludes with a brief discussion of player A's payoff-maximizing behavior.

### Punishment

36 % of players B do not punish in any of the 10 punishment decisions. This is in line with our prediction for subjects who focus on own their monetary payoffs alone. 64% of players B punish in at least one of the different scenarios. Note that punishment does not significantly differ between treatments CERT and UNCERT,[13] Since punishment behavior is robust with respect to the chosen parameter values determining the lottery and thereby procedural fairness, we pool the data for the analysis.

We begin with the analysis of player B's punishment decision, followed by player A's beliefs on expected punishment. First, we turn to player B's punishment decisions when the history of events (procedure choice and investment in cover-up activities) was **not revealed** (grey bars in Figure 2). In this situation, player B is ignorant about player A's decisions; he or she only learns about the allocation. The average punishment contingent on allocation *unequal* is 12.54 points, which is significantly higher than the average punishment of 3.96 contingent on allocation *equal* (Wilcoxon signed-rank test, *p<0.01*). The punishment probability (grey bars in Figure 3) decreases from 0.51 to 0.21 when moving from *unequal* to *equal* (Chi$^2$ *p<0.01).* Both of these differences indicate outcome-based punishment. However, since the history was not revealed, punishment might also include punishment for unrevealed intentions or cover-up activities that player B expects given the allocation and the lack of information about player A's procedure and cover-up choices.

---

[13] Mann-Whitney, p>0.6 for all punishment conditions in CERT versus UNCERT
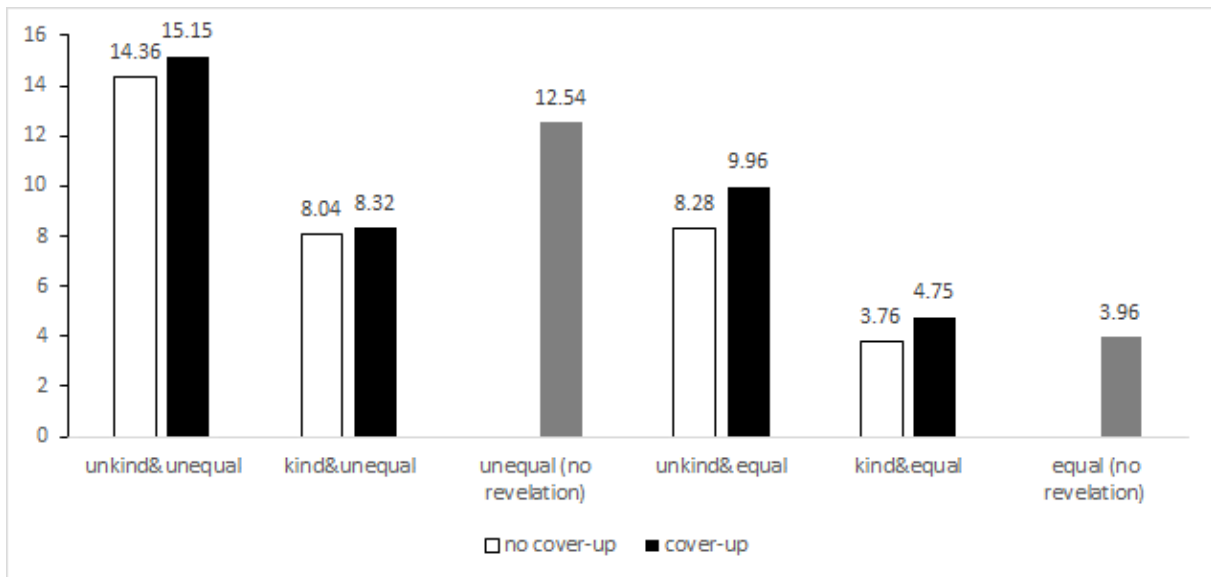
**Figure 2: Average punishment levels assigned to player A by player B, contingent on the eight situations in which the history of events was revealed (white/black bars) and the two situations in which the history of events was not revealed (grey bars), pooled data: CERT and UNCERT**
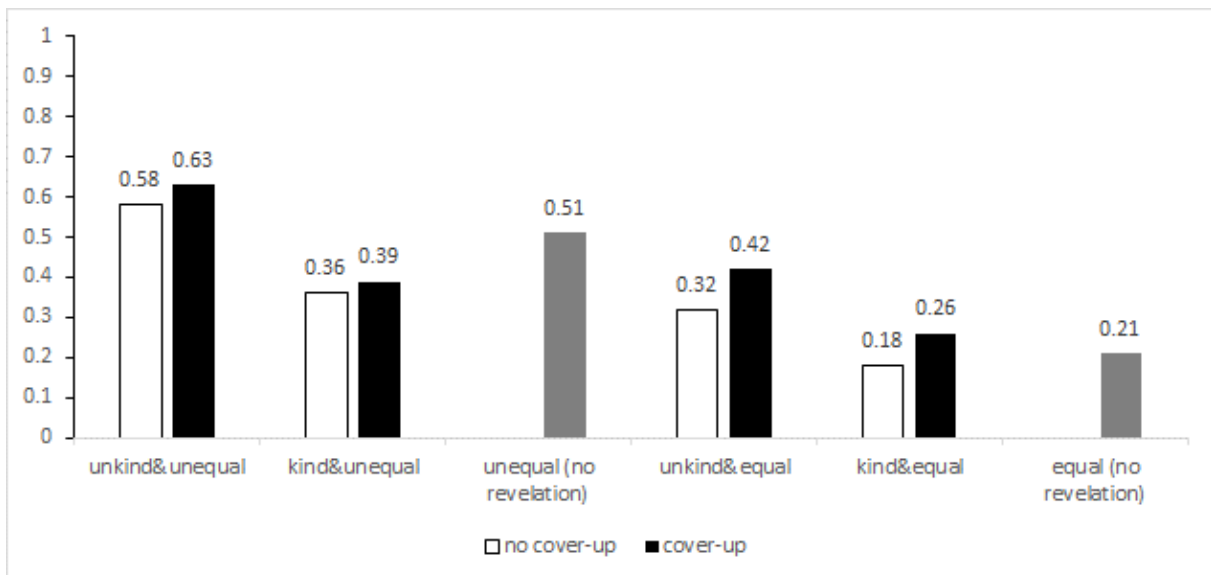


**Figure 3: Punishment probability, contingent on the eight situations in which the history of events was revealed (white/black bars) and the two situations in which the history of events was not revealed (grey bars), pooled data: CERT and UNCERT**

The average punishment when the history was not revealed differs significantly from that when the history was revealed (see grey bars vs. black/white bars in Figure2). We analyze punishment levels for all eight possible cases: player A has chosen the *kind* or *unkind* procedure, an *equal* or *unequal* allocation was implemented, and player A has invested or not invested in *cover-up* activities. Knowledge about the *unkind* procedure choice significantly increases average punishment and the punishment probability for both the *unequal* and *equal* outcomes. Knowledge about the *kind* procedure choice significantly decreases average punishment only when the *unequal* allocation was implemented, whereas the punishment probability contingent on the *kind* procedure choice is comparatively lower for both the

17

*unequal* and *equal* allocation. Table 3 presents the results of all possible comparisons in punishment levels (column 3). For the differences in probabilities, we obtain *p<0.01* for all possible comparisons (Chi$^2$).

**Result 1:** Successful concealment of *unkind* intentions decreases expected punishment.

**Result 2:** Successful concealment of *kind* intentions increases expected punishment.

These results are well in line with our reasoning in Section 4 about the influence of cover-up activities on player B's utility (as the level of punishment when the history is not revealed is off the mark in each actually relevant scenario). We argued that ignorant subjects will punish sub-optimally, under-punishing hidden negative intentions and over-punishing hidden positive ones, which is exactly what we find.

When the history of events **is revealed**, there are three different motives for punishment: Player B might punish the unequal allocation, unkind intentions, and/or cover-up activities. We find evidence for all three motives. We use average punishment levels to determine expected punishment (Figure 2) and analyze both the punishment probability (Figure 3) and conditional punishment[14] (Table 5) as possible drivers for expected punishment. First, there is outcome-based punishment: Average punishment is significantly higher when the *unequal* split of the 100 points was implemented in comparison to when the *equal* outcome resulted; this holds in all four possible situations (Wilcoxon signed-rank test: *p<0.01)*. Second, there is intention-based punishment: In all four possible situations, average punishment is significantly higher when the *unkind* procedure was chosen in comparison to when the *kind* procedure was selected (Wilcoxon signed-rank test: *p<0.01)*. Finally, there is punishment for cover-up activities: In three out of the four situations, average punishment is significantly higher after a *cover-up* attempt than after *no cover-up* attempt (see Table 4).

With respect to the punishment of cover-up activities, player B punishes the attempt to conceal *kind* as well as *unkind* intentions. However, the average punishment for the attempt to conceal *unkind* intentions is higher than when *kind* intentions were covered up (Wilcoxon signed-rank test: *p<0.01)*.

---

[14] Conditional punishment means the level of punishment conditional on those subjects who punished a positive amount in the comparable (less severe) situation. For example, to determine the conditional punishment of *unkind&unequal&cover* in comparison to *unkind&unequal&nocover*, we average punishment for *unkind&unequal&cover* of only those subjects who punished a positive amount in *unkind&unequal&nocover*.

| full revelation of history | no revelation of history | p-value for behavior | p-value for beliefs |
|---|---|---|---|
| unkind<br>unequal<br>cover-up | -<br>unequal<br>- | p<0.01 | p<0.01 |
| unkind<br>unequal<br>no cover-up | -<br>unequal<br>- | p=0.02 | p=0.02 |
| kind<br>unequal<br>cover-up | -<br>unequal<br>- | p=0.01 | p=0.5 |
| kind<br>unequal<br>no cover-up | -<br>unequal<br>- | p<0.01 | p<0.01 |
| unkind<br>equal<br>cover-up | -<br>equal<br>- | p<0.01 | p<0.01 |
| unkind<br>equal<br>no cover-up | -<br>equal<br>- | p=0.02 | p<0.01 |
| kind<br>equal<br>cover-up | -<br>equal<br>- | p=0.90 | p=0.03 |
| kind<br>equal<br>no cover-up | -<br>equal<br>- | p=0.77 | p=0.19 |

**Table 3: Significance levels of difference in average punishment between revelation of history and no revelation of history, results of Wilcoxon signed-rank test**

| | p-value for behavior | p-value for beliefs |
|---|---|---|
| **unkind unequal** | p=0.04 | p=0.16 |
| **unkind equal** | p=0.02 | p=0.31 |
| **kind unequal** | p=0.09 | p=0.01 |
| **kind equal** | p=0.21 | p=0.14 |

**Table 4: Significance levels of difference in average punishment between cover-up and no cover-up, results of Wilcoxon signed-rank test**

## Result 3

**(i)** There is outcome-based punishment.

**(ii)** There is intention-based punishment.

**(iii)** There is punishment for concealing intentions (i.e., hiding intentions is considered unfair).

Result 3 confirms hypotheses $H^1$, $H_{int}^2$, and $H_{int}^3$.

The mechanisms influencing average punishment vary across the three punishment motives. Unkind procedure choice, the unequal allocation, and cover-up activities significantly

increase the punishment probability for all possible comparisons (Chi$^2$: *p<0.01*). Outcome-based punishment has yet another driver, namely higher conditional punishment (Table 5). In contrast, punishment for unkind intentions and cover-up activities is driven by a higher punishment probability alone. Conditional punishment is basically not affected (Table 5). Unkind intentions and the attempt to hide them make subjects start punishing, but they do not cause subjects who had already punished in the respective kind intentions or no cover-up scenario to punish more.

| Outcome | unequal | equal | N | p-value |
|---|---|---|---|---|
| unkind&cover-up | 30.8 (23.2) | 23.9 (19.3) | 30 | <0.01 |
| kind&cover-up | 27.4 (23.7) | 18.0 (16.5) | 19 | <0.01 |
| unkind&no cover-up | 32.7 (24.4) | 25.9 (20.0) | 23 | <0.01 |
| kind&no cover-up | 28.0 (25.1) | 20.8 (19.23) | 13 | 0.04 |
| **Intentions** | **unkind** | **kind** | **N** | **p-value** |
| unequal&cover-up | 25.2 (22.6) | 21.4 (21.5) | 28 | 0.01 |
| equal&cover-up | 20.8 (20.5) | 18.0 (16.5) | 19 | 0.16 |
| unequal&no cover-up | 25.3 (23.9) | 22.3 (21.9) | 26 | 0.61 |
| equal&no cover-up | 22.2 (22.1) | 20.8 (19.2) | 13 | 0.91 |
| **Cover-up activities** | **Cover-up** | **No cover-up** | **N** | **p-value** |
| unkind&unequal | 25.0 (22.2) | 24.6 (21.8) | 42 | 0.21 |
| kind&unequal | 21.9 (22.4) | 22.3 (21.9) | 26 | 0.54 |
| unkind&equal | 26.4 (20.9) | 25.9 (20.0) | 23 | 0.91 |
| kind&equal | 19.3 (19.4) | 20.8 (19.2) | 13 | 0.19 |

**Table 5: Mean conditional punishment, standard errors in parentheses, p-values of Wilcoxon signed-rank test**

Next, we seek to derive a typology of punishment types. Based on individual average punishment levels across the eight situations after full revelation, players B can be classified according to their punishment type (see Figure 4). As noted above, 36% of players B never punish (presumably, because they are focused on their own monetary payoffs). The largest group, 42%, punishes in line with our intention-based hypotheses and displays both outcome-based and intention-based preferences and/ or a dislike of cover-up activities. In addition, 13% exhibit purely outcome-based preferences, whereas 8% exhibit either pure intention-

based preferences or pure aversion to concealment. Hence, punishment for cover-up activities does not necessarily accompany punishment for unkind intentions. Specifically, 24% punish unkind intentions but do not care about cover-up, whereas 8% punish cover-up activities but not unkind intentions.
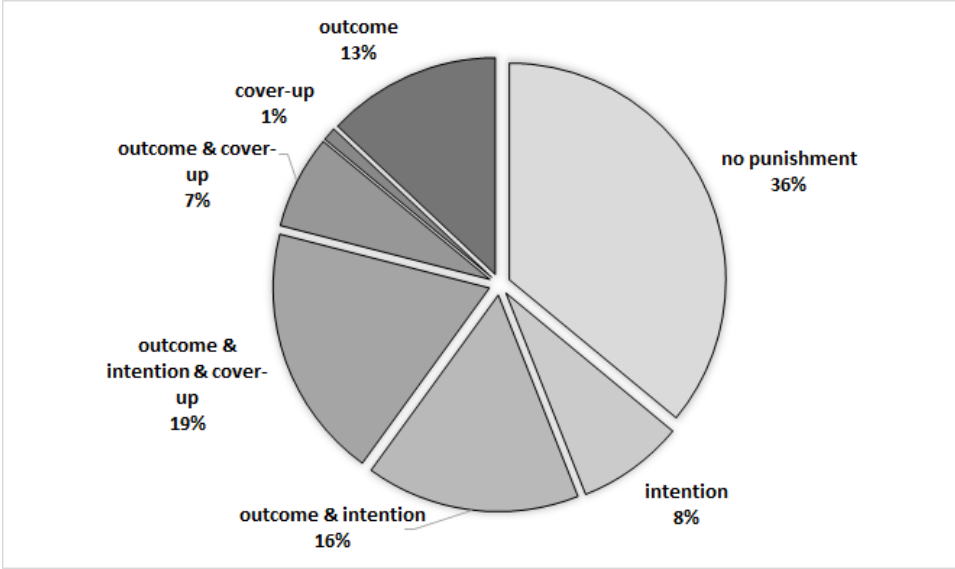


**Figure 4: Punishment types**

**Punishment beliefs**

Although players A on average clearly over-estimate expected punishment, Figure 5 shows that their beliefs about the effects are mainly in line with actual behavior. Players A expect outcome-based punishment after no and full revelation (Wilcoxon signed-rank test: *p<0.01*). Furthermore, players correctly expect unkind intentions to be punished more strongly (intention-based punishment, Wilcoxon signed-rank test: *p<0.1*). Successful cover-up attempts are correctly believed to decrease punishment when intentions were unkind and to increase punishment when intentions were kind (See Table 3, column 4). Although the effect is not significant for all comparisons, players A correctly predict higher punishment for failed cover-up attempts (See Table 4, column 3).



**Figure 5: Average punishment beliefs of player A, contingent on the eight situations in which the history of events was revealed (white/black bars) and the two situations in which the history of events was not revealed (grey bars), pooled data: CERT and UNCERT**

**Investment in cover-up activities**

We find that a substantial share of players A (32%) invest in cover-up activities for at least one of the two possible allocations (see Figure 6).[15] Cover-up activities are more likely for players with negative intentions (i.e., players who chose procedure *unkind*) when the *unequal* allocation was implemented in CERT (Chi$^2$, *p=0.052*).[16]

Note that some players also try to hide their kind intentions. One reason is that 23 % (32%) of players A falsely expect punishment to be lower after a successful cover-up attempt of kind intentions when the unequal (equal) allocation was implemented. Players A who chose the

---

[15] 24% invest in cover-up only when the *unequal* allocation is realized, while 8% invest for both allocations. No player A invests in cover-up activities only after the equal allocation was determined.
[16] The main effect is insignificant due to the behavior in treatment UNCERT. In UNCERT, it looks as if the effect were reversed but the difference is not significant (Chi$^2$, p=0.310).

kind procedure aiming for the equal allocation but ending up with the unequal allocation might use the cover-up option to reduce inequality investing their endowment. Also, some players might have a general preference for covering-up their history. However, our design does not allow us to evaluate the latter two explanations.

Cover-up activities are more likely after the *unequal* allocation is implemented (Chi$^2$, *p=0.015*). This behavior is consistent with predictions made by both theories ($H^4$). Cover-up is less likely in treatment UNCERT in comparison to treatment CERT when the *unkind* procedure is chosen and the *unequal* allocation is implemented (Chi$^2$, *p=0.057*). These findings are consistent with hypothesis $H^5_{int}$, but inconsistent with $H^5_{out}$. The latter fact may indicate that players A feel less compelled to hide their intentions when the procedure choice was not as detrimental to player B (when measured in terms of $\Delta_{EmB}$).

**Procedure choice**

We find that significantly more subjects choose the *unkind* procedure in UNCERT (72%) than in CERT (56%), (Chi$^2$: *p=0.024*). Our data thus support hypothesis $H^6_{int}$, while rejecting $H^6_{out}$. This may be interpreted as indicating that players A are sensitive to the harmfulness of their procedure choice in expected terms (abstracting from possible differences in the punishment levels).
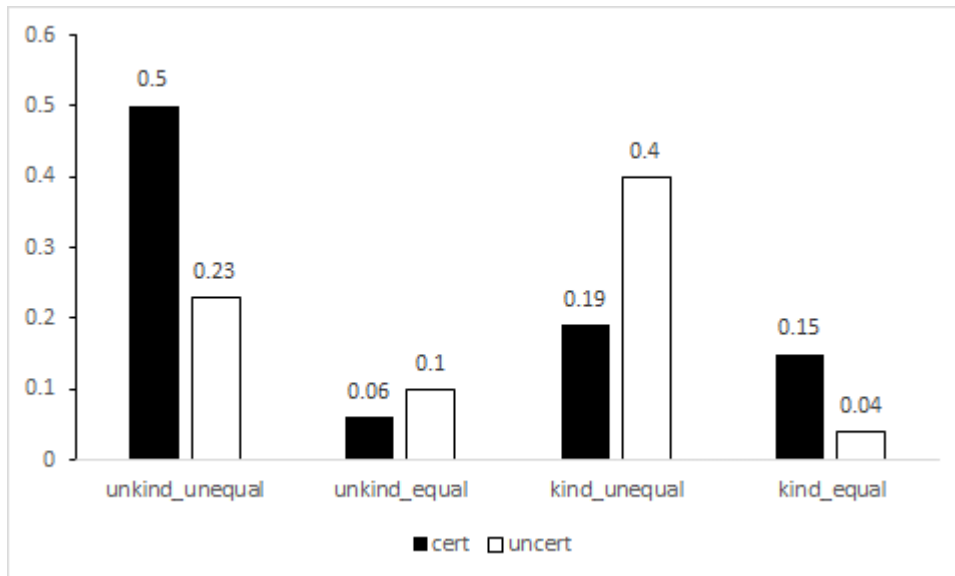


**Figure 6: Frequencies of cover-up activities**

**Payoff-maximizing behavior**

Finally, using all of our data, we discuss which behavior is payoff-maximizing. In our game hiding intentions does not pay off in expected terms although expected punishment falls.[17]

---

[17] Note that this result is due to our specific parameterization of the game (such as the cost of cover-up) and should not be generalized.

Choosing the *unkind* procedure yields 74.83 points without cover-up and 65.83 with cover-up (taking into account the endowment of 10 points). Choosing the *kind* procedure yields 58.31 points without cover-up and 47.95 points with cover-up. A payoff-maximizing player A should therefore opt for the *unkind* procedure and refrain from concealment. In our experiment, 42% percent of players A do so.

## 6. Discussion and Conclusion

Fairness considerations are important in many domains in life. They have real implications for behavior and well-being. To better understand how fairness considerations actually shape behavior, it is important to know what exactly determines fairness perceptions. Recent evidence suggests that people evaluate both outcomes and the intentions of others to arrive at a judgment about the fairness of specific events. In other words, in contrast to the emphasis of earlier contributions, intentions are of central importance in this context.

Taking into account the importance of intentions in peoples' evaluation of outcomes, our paper explores how hiding intentions enters fairness perceptions. To this end, we allow subjects to hide their intentions from others at a cost and those affected by the concealment of intentions to reciprocate by imposing punishment. We find that hiding intentions is itself a punishable act. In other words, people consider hiding intentions unfair. Nevertheless, many subjects still engage in cover-up activities in the hope that their attempts will successfully prevent the other party from learning about and punishing their intentions. In our experiment, a successful cover-up is indeed beneficial, in that a party who is in doubt about an actor's true intentions will moderate the level of punishment in comparison to the punishment assigned to actors with known unkind intentions. However, if concealment fails, the punishment will be even higher.

Our experimental results inform us about the use of an option that is available in numerous circumstances but difficult to observe in practice (i.e., cover-up of evidence). In addition, our findings show that the actual legal treatment of evidentiary foul play is actually aligned with peoples' fairness preferences.

Interestingly, when we consider a classification of the fairness preferences revealed by the punishment decisions of our players, we find that subjects concerned about unkind intentions are not necessarily bothered by concealment and that some participants who assign higher punishment for concealment do not allocate extra punishment to unkind intentions. This allows us the tentative conjecture that unkind intentions and the concealment of intentions can be considered as distinct categories.

## Acknowledgements

## References

Bartling, B., Engl, F., and R.A. Weber, 2014. Does willful ignorance deflect punishment? An experimental study. European Economic Review 70, 512-524.

Becker, G.S., 1968. Crime and punishment: an economic approach. Journal of Political Economy 76, 169-217.

Bolton, G.E., Brandts, J., and A. Ockenfels, 2005. Fair procedures: evidence from games involving lotteries. Economic Journal 115, 1054-1076.

Bolton, G.E., and A. Ockenfels, 2000. ERC: A theory of equity, reciprocity, and competition. American Economic Rebiew 90, 166-193.

Brandts, J., and G. Charness, 2011. The strategy versus the direct-response method: a first survey of experimental comparisons. Experimental Economics 14, 375-398.

Charness, G., and D. Levine, 2007. Intention and stochastic outcomes: an experimental study. Economic Journal 117, 1051-1072.

Charness, G., and M. Rabin, 2002. Understanding social preferences with simple tests. Quarterly Journal of Economics 117, 817-869.

Conrads, J., and B. Irlenbusch, 2013. Strategic ignorance in ultimatum bargaining. Journal of Economic Behavior and Organization 92, 104-115.

Cox, J.C., Friedman, D., and S. Gjerstad, 2007. A tractable model of reciprocity and fairness. Games and Economic Behavior 59, 17-45.

Dufwenberg, M., and G. Kirchsteiger, 2004. A theory of sequential reciprocity. Games and Economic Behavior 47, 268-298.

Falk, A., and U. Fischbacher, 2006. Testing theories of fairness – Intentions matter. Games and Economic Behavior 54, 293-315.

Falk, A., Fehr, E., and U. Fischbacher, 2003. On the nature of fair behavior. Economic Inquiry 41, 20-26.

Falk, A., Fehr, E., and U. Fischbacher, 2008. Testing theories of fairness – Intentions matter. Games and Economic Behavior 62, 287-303.

Fehr, E., and K.M. Schmidt, 1999. A theory of fairness, competition, and cooperation. Quarterly Journal of Economics 114, 817-868.

Fehr, E., and K.M. Schmidt, 1999. The economics of fairness, reciprocity and altruism - experimental evidence and new theories. In: Kolm, S.C., Ythier, J.M. (Eds.). Handbook on the Economics of Giving, Reciprocity and Altruism, Vol. 1, 615-691, Amsterdam: Elsevier.

Fischbacher, U., 2007. z-Tree: Zurich toolbox for readymade economic experiments. Experimental Economics 10, 171-178.

Greiner, B., 2004. The online recruitment system Orsee 2.0 - A guide for the organization of experiments in economics. University of Cologne, Department of Economics.

Krawcyk, M.W., 2011. A model of procedural and distributive fairness. Theory and Decision 70, 111-128.

Langlais, E., 2008. Detection avoidance and deterrence: some paradoxical arithmetic. Journal of Public Economic Theory 10, 371-382.

Leibbrandt, A., and R. López-Pérez, 2012. An exploration of third and second party punishment in ten simple games. Journal of Economic Behavior and Organization 84, 753-766.

Malik, A.S., 1990. Avoidance, screening, and optimum enforcement. Rand Journal of Economics 21, 341-353.

Nussim, J., and A. Tabbach, 2009. Deterrence and avoidance. International Review of Law and Economics 29, 314-323.

Rabin, M., 1993. Incorporating fairness into game theory and economics. American Economic Review 83, 1281-1302.

Sanchirico, C.W., 2006. Detection avoidance. New York University Law Review 81, 1331-1399.

Sanchirico, C.W., 2012. Detection avoidance and enforcement theory. In: Sanchirico, C.W. (Ed.) Procedural law and economics. Edward Elgar.

Sebald, A., 2010. Attribution and reciprocity. Games and Economic Behavior 68, 339-352.

Selten, R., 1967. Die Strategiemethode zur Erforschung des eingeschränkt rationalen Verhaltens im Rahmen eines Oligopolexperiments. In: Sauermann, H. (ed.) *Beiträge zur experimentellen Wirtschaftsforschung*, 136–168. Tübingen: Mohr.

Trautmann, S.T., 2009. A tractable model of process fairness under risk. Journal of Economic Psychology 30, 803-813.

Utikal, V., 2012. A fault confessed is half redressed—Confessions and punishment. Journal of Economic Behavior and Organization 81, 314-327.

**Appendix: Translated instructions for player A in treatment CERT**

**General Instructions**

Welcome to this economics experiment. By carefully reading the following instructions, you can earn money based on your decisions. It is therefore very important that you read these instructions carefully. If you have questions, please raise your hand. We will come to you and answer your question in private.

For the duration of the entire experiment, communication with other participants, the use of cell phones, and the use of other software on the computer is not allowed. Disregarding these rules will bar you from participating in the experiment and receiving payment. During the experiment, we will speak of payment in terms of points, not Euros. Your earnings will be calculated in points. The total number of points earned in the experiment will be converted into Euros at the end of the experiment using the exchange rate **1 point = 15 cents.** You will receive your earnings in cash right after the end of the experiment. On the following pages, we will explain the exact procedure of the experiment.

## Summary

In this experiment, there are two types of participants: participant 1 and participant 2. Participant 1 decides whether a ball is drawn from either the left or the right urn. The ball's color determines the participants' payoff. A black ball yields 50 points for both participants; a red ball yields 80 points for participant 1 and 20 points for participant 2. The probability of drawing a red ball is higher for the left urn. Participant 2 can deduct points from participant 1. A random mechanism determines whether participant 2 learns only the color of the ball, or additionally from which urn the ball was drawn. Participant 1 can decrease the likelihood of participant 2 learning the urn choice (left or right).

## Procedures in detail

In this experiment there are **two types of participants**: participant 1 and participant 2. **You are a participant 1.** You will not learn the identity of the participant 2 assigned to you during or after the experiment. Similarly, participant 2 will not learn your identity. Every pair of subjects consists of one participant 1 and one participant 2. Your payoffs depend on your decisions as well as on the decisions of participant 2.

You will receive an initial endowment of 10 points. Participant 2 will also receive an initial endowment of 10 points.

The experiment can be divided into 5 stages. In stages 1 and 2, only participant 1 makes a decision. In stage 3, only participant 2 makes a decision. In stage 4, both participant 1 and participant 2 make a decision. In stage 5, there are no more decisions, as a random mechanism determines the earnings.
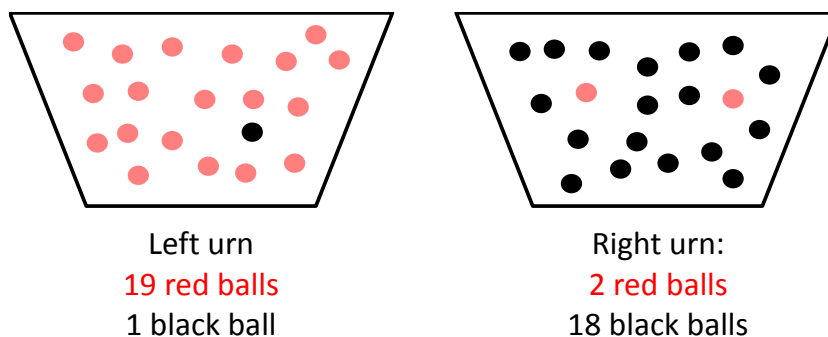


Figure 1: Left and right urns

Left urn
19 red balls
1 black ball

Right urn:
2 red balls
18 black balls

**Stage 1:**
First, you -- in the role of participant 1 -- decide from which urn the ball will be drawn in

stage 5. There are two different urns (See Figure 1). The left urn contains 20 balls, of which 19 are red and 1 is black. The right urn contains 20 balls, of which 2 are red and 18 are black. In other words, the probability of drawing a red ball from the left urn is 95%, and the probability of drawing a red ball from the right urn is 10%. Correspondingly, the probability of drawing a black ball from the left urn is 5%, and the probability of drawing a black ball from the right urn is 90%.

Your earnings depend on the color of the ball drawn in stage 5.

- A red ball yields 80 points for participant 1 and 20 points for participant 2.

- A black ball yields 50 points for both participants.

**Stage 2:**
In stage 5, there is a specific probability that participant 2 will learn about your urn choice. In stage 2, you can use your initial endowment of 10 points to reduce this revelation probability. If you spend your initial endowment, there is a 20% probability that participant 2 will learn your decisions. If you do **not** pay your initial endowment, there is an 80% probability that participant 2 will learn your decisions. In the event of revelation, participant 2 will be informed about both of your decisions, i.e., your urn choice in stage 1 and your decision of whether or not to use your initial endowment to decrease the revelation probability in stage 2. Otherwise, participant 2 will only learn whether a red or a black ball was drawn. As participant 1, you will see the following decision screen (Figure 2):
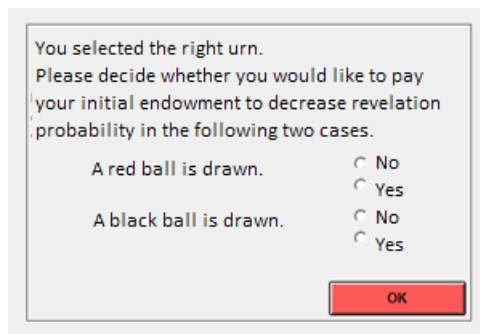


Figure 2: Screen of participant 1 when choosing
whether or not to decrease the revelation probability

**Stage 3:**
Participant 2 can deduct points from participant 1 by spending points from his or her initial endowment. When deducting one of your points, participant 2 pays 1/6 of a point.

Examples:
If zero points are deducted, participant 2 pays zero points.
If 24 points are deducted, participant 2 pays 4 points.
If 60 points are deducted, participant 2 pays 10 points.

Participant 2 cannot deduct more points than you have earned by the draw of the ball. Participant 2 uses the following decision screen (Figure 3):



Figure 3: Screen of participant 2 when choosing whether or not to deduct points

**Stage 4:**

In stage 4, all participants try to assess the behavior of the other participants in today's experiment. There are 15 assessments. One of these assessments will be randomly selected and will be relevant for your earnings. If your assessment is not more than 5 points above or below the actual value, you will receive 50 points.

**Stage 5:**

A random mechanism draws a ball from the urn selected by participant 1.

Next, a random mechanism determines whether participant 1's decisions are revealed to participant 2. This mechanism accounts for whether or not you have paid for a decrease in the revelation probability. If you paid for a decrease in the revelation probability, there is a 20% probability that participant 2 will learn your decisions. If you did **not** pay for a decrease in the revelation probability, there is an 80% probability that participant 2 will learn your decisions.

Finally, player 2's decision on point deduction will be implemented, and one of your expectations about the behavior of other players will be selected.

You will then learn your total earnings from the experiment. At the end of the experiment, all participants will receive their earnings in cash. If there are any questions, please raise your hand now. An experimenter will come to you and answer your questions.

Below, you will find some test questions. Please raise your hand when you have answered all the questions. The experiment will start when all participants have answered all the questions.

## Test questions

*Note: Correct answers presented here in parentheses.*

- A red ball yields

  _____ (80) for participant 1.

  _____ (20) for participant 2.

- A black ball yields

  _____ (50) for participant 1.

  _____ (50) for participant 2.

- What's the probability of drawing a red ball out of the left urn? _____ (95%)

- What's the probability of drawing a red ball out of the right urn? _____(10%)

- How many points does participant 1 have to spend to decrease the revelation probability? _____ (10)

- What's the revelation probability if participant 1 does NOT spend his or her initial endowment? _____ (80%)

- What's the revelation probability if participant 1 spends his or her initial endowment? _____(20%)

- How many points are deducted from participant 1 if participant 2 spends 3 points? _____ (18)