

Three Essays on

**Moral Self-Regulation of Honesty and
Impression Management**

Der Wirtschaftswissenschaftlichen Fakultät
der Universität Passau eingereichte

DISSERTATION

zur Erlangung des Grades eines
doctor rerum politicarum (Dr. rer. pol.)

vorgelegt von

Diplom-Volkswirt Volker Nagel

Passau, November 2014

Tag der Disputation:

07.05.2015

Promotionskommission:

Prof. Dr. Johann Graf Lambsdorff (Erstprüfer)

Prof. Dr. Stefan Bauernschuster (Zweitprüfer)

Prof. Dr. Michael Grimm

Prof. Dr. Carolin Häussler

Prof. Dr. Markus Diller

Summary

Human behavior is influenced by numerous different determinants. Behavioral research has tried to track down these influences and identify the factors that drive behavior in different situations. Starting with the homo oeconomicus, research has come up with multiple models explaining human behavior. Because the neoclassical view that human beings are utility maximizing economic agents was met with increasing critique over the last decade, new models were developed. With novel research techniques a vast amount of evidence contradicting the homo oeconomicus was generated. Those techniques included empirical as well as experimental data and forced researchers to update and extend the old models to fit reality. In consequence models such as the *satisficer*¹, the *homo reciprocans*², or the *homo heuristicus*³ have emerged. Deviations from utility maximization can be explained by "satisficing" – behavior which does not aim to achieve the absolute best outcome but rather a sufficient one, where costs and benefits find a satisfying balance. Another observation is that reciprocity determines behavior as well. Experiments have shown how participants were willing to forego money in order to return a favor to others (positive reciprocity) or to take revenge (negative reciprocity). Research also identified that humans deviate from rational behavior because they often rely on simple rules of thumb (heuristics) as guidance for everyday behavior.

Additional to these explanations of human behavior, another strand of research has looked at individual decisions and how these are embedded in a succession of decisions. This research states that one decision may have repercussions on a subsequent decision and therefore may influence future behavior. When one decision follows another, especially inconsistencies are puzzling: how can it be explained that a person first states to not be sexist, but afterwards engages in sexist behavior? How is it possible that someone who describes himself as charitable refuses to give money to a charity only minutes later? Such inconsistencies can be explained by moral self-regulation.

This thesis is situated in this field of research. In study 1 an introduction to the research on moral self-regulation is provided alongside with an explanation of the two manifestations of moral self-regulation: moral licensing and moral cleansing. At the core of the first study is an experiment which was designed to identify moral licensing and cleansing in the domain of honesty. The experiment merges relevant studies from social psychology and experimental economics. It assesses the question if moral self-regulation exists within the domain of honesty – or more precisely, if the truth and lies are told in such a way as to balance each other out. After manipulating participants' moral balances (either positively or negatively), rates of truth-telling are compared to a neutral baseline scenario. Since neither moral licensing nor moral cleansing is observed, the results provide no support to the initial hypothesis that moral self-regulation exists within the domain of honesty.

Study 2 builds on these results and discusses possible reasons for the absence of moral self-regulation. The research on moral hypocrisy and self-concept maintenance are presented and discussed as possible explanations. In order to shed more light on participants' behavior, a coding procedure is presented that was used on the dataset from study 1. This approach makes it possible to quantify participants' handwritten stories that resulted from the moral manipulation in study 1 and gain more insights on how truth-telling and lying affect the moral balance. By analyzing (dis)honesty on a more

¹ e.g. Schwartz, B.; Ward, A.; Monterosso, J.; Lyubomirsky, S.; White, K. and Lehman, D. R. (2002): Maximizing versus satisficing: Happiness is a matter of choice, *Journal of Personality and Social Psychology* 83(5), 1178-1197

² Falk, A.; Dohmen, T.; Huffman, D. and Sunde, U. (2009): *Homo Reciprocans: Survey Evidence on Behavioural Outcomes*, *The Economic Journal* 119(536), 592-612

³ Gigerenzer, G. and Brighton, H. (2009): *Homo Heuristicus: Why Biased Minds Make Better Inferences*, *Topics in Cognitive Science*, 1, 107-143

detailed level, results show that participants tend to act consistent to what they revealed about themselves in their stories.

Study 3 links together aspects of moral self-regulation, moral hypocrisy and impression management. The "looting game" is presented which lets participants loot money from a charity box being subject to altruistic punishment from observers. For their punishment decision observers are provided with a history of participants' past actions. This design allows to assess how misconduct, punishment and the creation of a favorable impression interact and ultimately impact profits. The results indicate that moral cleansing – and not the desire to trick observers – is the reason for manipulation. Participants who loot money from the charity box do not expect to receive less punishment, rather they simply want to present a more favorable picture of themselves. On the other hand, observers fully account for the possibility of manipulation and tend to disregard a manipulated history. The looting game therefore brings the hypothesis into question that impressions are managed and manipulated to increase profits.

This thesis adds to the current research by identifying a domain where moral self-regulation fails (study 1), offering explanations for this observation and identifying consistent behavior with regard to truth-telling and lying (study 2). Finally, it presents a novel experimental design which challenges profit-maximization as a motive for impression management and offers moral cleansing as an alternative explanation for such behavior (study 3).

Licensed to Lie?


Moral Self-Regulation of Truth-Telling and Lying

Volker Nagel 

Abstract

To find evidence for the existence of moral licensing or moral cleansing in the domain of truth-telling and lying I constructed an experiment that brings together moral self-regulation and Gneezy's (2005) deception game. Participants' moral balances were manipulated by two tasks that either increased or decreased moral credentials as well as moral credits. Afterwards participants engaged in a deception game. With this approach I quantify how changes in the moral balance affect truth-telling and lying. Compared to a baseline treatment where credentials and credits remained unchanged, I find no significant differences in truth-telling and lying. These experimental results suggest that same domain moral self-regulation does not exist within the domain of honesty.

Keywords: moral licensing, moral cleansing, moral credentials, moral credits, dishonesty, lying, deception game

 Volker Nagel is a doctoral scholar at the University of Passau, Germany

1. Introduction

Do people compensate good actions with bad ones and vice versa in order to balance their morality? This question plays an important role in human interactions and in the scientific quest to analyze and understand human behavior. For example for members of the Catholic Church the Sacrament of Penance, commonly known as confession, is a well known and wide spread way of absolving ones mortal sins. The penitent tells his sins to a priest and is subsequently granted absolution. Of course such behavior roots in the religious belief that not absolving mortal sins will condemn a person to hell. Nevertheless, it is an example of moral balancing behavior that seeks to compensate bad actions with good ones. This need exists for environmentalism as well. There are many websites that offer to offset a person's carbon dioxide emissions by donating money to a charity⁴. These sites let you easily calculate the amount of greenhouse gas emitted by a flight or a car drive and display the required donation to offset this exact amount. Here, environmentally bad behavior can be balanced with just a few clicks and the payment of the calculated fee. These examples show the need of human beings to equalize good and bad actions through balancing behavior. In social psychology such behavior is called *moral self-regulation*. Research on this topic (e.g. Monin and Miller 2001; Khan and Dhar 2005; Effron et al. 2009; Sachdeva et al. 2009; Mazar and Zhong 2010; Merritt et al. 2010) has produced many new insights on human behavior during the last decade. Various experiments gathered evidence on how and where moral self-regulation is at work. These have shown that moral self-regulation works through two main mechanisms: *moral licensing* (feeling free to act bad) and *moral cleansing* (feeling the need to do something good). The religious act of confession would be an example of moral cleansing: a person is cleansed of his previously bad behavior. Donating to a charity on the other hand can subsequently lead to moral licensing: by giving money, a person has done something good and in turn feels less or no discomfort if he engages in behavior that is otherwise considered wrong. These mechanisms have been proven to exist in many different domains such as political incorrectness (Monin and Miller 2001), selfishness (Sachdeva et al. 2009) or dieting (Fishbach and Dhar 2005). Yet, little research has been carried out on moral licensing and cleansing in the domain of honesty so far. This paper aims at filling this gap and seeks to enrich our understanding of moral self-regulation. This article discusses if the forces of moral self-regulation work for truth-telling and lying as well and asks the two main questions: (1) Does past honest behavior endow people with a positive self-worth in such a way that they subsequently feel free to lie (moral licensing)? (2) Do people feel the need to cleanse themselves of their lies by telling the truth more often (moral cleansing)?

This paper is structured as follows: an overview of the relevant literature is provided in section 2. The experimental design is presented in section 3 and the hypotheses in section 4. I present the main results in section 5 and additional results in section 6. Section 7 concludes.

2. Related literature

Research on the dynamics of moral self-regulation describes two ways in which licensing can occur. One model is called moral credentials, the other moral credits. In short, the first can be thought of colorful glasses one puts on to make things look different. This model argues that previously conducted good acts cast a different light on subsequent bad acts that would on their own be considered a moral transgression. Moral credentials are acquired by prefacing a dubious action with evidence that moderates this same action. An example is if one tells a racist joke but states upfront that his best friend belongs to the race that he is making fun of (Zhong et al. 2009, p. 83). The second

⁴ See for example <http://www.myclimate.org>

model resembles the idea of a bank account for moral currency. The moral credits model states that doing something good adds currency to one's moral balance while doing something bad deducts from the balance. Unlike moral credentials, moral credits do not change the way a transgression is assessed. When transgressing, one is fully aware of this fact but still feels free to do so because of his positive moral balance. A typical example of moral credits is the practice of purchasing the carbon offsets mentioned earlier. By purchasing carbon offsets a person tries to compensate for an environmentally harmful activity, e.g. taking a transcontinental flight (Miller and Effron 2010, p. 125). Both models and the related literature are shortly discussed in the subsequent two sections.⁵

Moral Credentials

The idea of moral credentials was established by Monin and Miller (2001). They performed an experiment where in a first step participants were asked to indicate if they agree or disagree to blatant sexist statements about women (e.g. "Most women are not really smart.") and in a second step had to take a hypothetical hiring decision. There, participants had to decide if they would hire a man or a woman for a job in the building industry that was described as being obviously best suited for a man. The initial statements were formulated with the purpose to result in a high rate of disagreement. A second treatment, where the word "most" was substituted by "some" (e.g. "Some women are not really smart."), resulted in significantly less disagreement among participants. By dividing participants into two groups with two different rates of disagreement about politically incorrect statements it was possible to show that the first group had acquired moral credentials by demonstrating that they are not sexist (Monin and Miller 2001, p. 35). These credentials were visible because the first group preferred a man in the hiring decision, while the second group did not as they feared appearing sexist by such a decision. By disagreeing to the blatant sexist statements the first group had prefaced the subsequent task with evidence of being a non-sexist person and thus felt free to engage in a decision that could look sexist on its own. Identical results were acquired when the experiment was repeated in a racism version (Monin and Miller 2001, study 2). Having established credentials as non-racist individuals, participants favored a White person over an African-American one.

Effron et al. (2009) do support these results with a similar experiment, showing that endorsing US-President Obama made people more willing to choose a white job applicant over a black one and thus potentially show political incorrect behavior. Again, this is an example of moral credentials at work. Showing support for Obama provides participants with the credentials as a non-racist person. This frees them to later take a hiring decision in a hypothetical situation which favors a White over a Black candidate and could appear racist on its own.

Moral Credits

Other research states that licensing takes place because of behavior that aims at balancing good with bad actions. Labels for such behavior are moral balancing (Nisan 1991), moral self-regulation (Sachdeva et al. 2009) or compensatory moral action (Jordan et al. 2011). Good deeds are assumed to generate some sort of currency (cp. Hollander 1958) that may later be spent to engage in bad deeds with impunity. Instead of changing the meaning of a transgression (what moral credentials do) the moral credits model assumes that participants are fully aware of their transgression but nevertheless engage in it as they have sufficient moral currency to afford such behavior.

For example, Sachdeva et al. (2009) ran an experiment to identify balancing behavior due to moral credits. Participants had to write stories about themselves with certain positive (e.g. caring, generous, fair) or negative (e.g. selfish, mean, greedy) trait words. Afterwards they decided how much to give to a charity. Results indicate that when in the positive trait word treatment, participants decreased their

⁵ For an additional overview of moral licensing and related topics see Miller and Effron (2010)

giving to a charity significantly. A self-relevant story filled with positive words licensed participants to behave selfishly afterwards. These participants were reminded of all their good behavior and thus their positive moral balance. This made them realize that they had accumulated moral credits in the past and provided them with a reason to abstain from donating money. Sachdeva et al. (2009) demonstrate evidence for moral cleansing as well. After having written a story about oneself with negative trait words, charitable giving increased compared to a control treatment with neutral words. This shows that participants felt the need to do something good (by donating money) after being reminded of their lack of moral credits.

Jordan et al. (2011) used a similar experimental approach. Participants were asked to recall "a time when they helped other people" (moral condition) or "used others to get something they wanted" (immoral condition). In a subsequent math task participants had to solve 15 math problems that consisted of adding or subtracting ten two-digit numbers each. Here, it was possible to cheat by design of the experiment. If participants did not press the space bar within the first 3.5 seconds of each math problem an answer appeared on the screen that varied +/-1 from the correct one. This made it possible to identify participants who did not hit the space bar in time and let the answer appear on screen, as well as participants who furthermore used this wrong answer. Not only did participants in the moral condition allow the answer to appear more often on screen, they furthermore used it significantly more often than in the immoral condition. This shows how participants increase cheating after being reminded about their positive balance of moral credits.

Brañas-Garza et al. (2013) provide further evidence for moral licensing and cleansing. In their experiment participants played 16 dictator games in sequence. Their results show that a consistent pattern of moral self-regulation emerges: participants equalize their donations systematically around their mean donation. A round with a higher donation is subsequently balanced with a lower donation. Since in every round players are matched with a new partner, this behavior cannot be explained by the desire to compensate the receiver but is entirely due to the balancing of the moral self.

Licensing effects emerge in other domains as well, for example in consumer choice, as Khan and Dhar (2005) show. Within their experiment, participants were asked to imagine and furthermore give hypothetical reasons for having done one of two community services that were described in detail. This simple task made participants appear charitable and boosted their moral self-worth. In a second task they were more likely to choose a luxury item over a necessity item than the control group.

In my experiment I extend the research on moral self-regulation to the domain of honesty. There is no conclusive evidence for licensing and cleansing in this special domain yet. Evidence exists that lying can be licensed with good deeds in different domains, for example with green consumer behavior (Mazar and Zhong 2010). In this study participants had to fill a virtual shopping basket as they liked with items of up to \$25 in value. One group did this shopping in a conventional store, the other group in a store with green products. Afterwards participants performed a visual perception task where they had to watch a box with a diagonal line on the screen. For 1 second dots would show up that were scattered within this box and participants should indicate if more dots were on the left or on the right of the diagonal line. The dots were scattered in such a way that it was easy to identify the correct answer. The critical manipulation in this task was that participants were paid 0.5¢ if they indicated there were more dots on the left but 5¢ if they indicated there were more on the right. In 40% of the 90 trials more dots were indeed on the right side. While participants that shopped in the conventional store identified 42.5% of the trials having more dots on the right and did not significantly differ from the correct value, participants that shopped in the green store identified 51.4% of the trials as having more dots on the right side. These participants were intentionally lying to earn more money. In this

experiment lying was licensed by a good deed in a different domain (shopping in a green store). I try to explore if lying can be licensed by truth-telling and therefore within the same domain.

3. Experimental Design and Treatments

Parallel to similar experiments on licensing and cleansing I designed the experiment in two parts. Part 1 provides the mechanisms necessary to manipulate participants' moral balances, whereas in part 2 honesty as the dependent variable is established. In this experiment participants' moral balances are increased in *treatment truth* and decreased in *treatment lie*. The baseline treatment provides a neutral setup with no changes to moral balances against which the changes in treatment truth and lie are tested. In part 2 participants played a deception game (Gneezy 2005) in which they had to either tell the truth or tell a lie. This setup allows to assess how changes in moral balances affect honesty.

The central idea I employ to set up this experiment is to combine ideas from different directions of research. I bring together the experimental setup from Monin and Miller (2001) and Sachdeva et al. (2009) with Gneezy's (2005) deception game. The experimental design from the first two studies is borrowed in order to implement a mechanism that changes participants' moral balances. But I substitute the second part, which is usually a hypothetical scenario in social-psychological research, with a deception game (Gneezy 2005). This setup allows me to observe actual behavior rather than stated preferences for certain outcomes and to implement a licensing and cleansing mechanism into a behavioral experiment with controlled conditions in a laboratory environment.

Part 1 - Manipulation of moral credentials and moral credits

So far the literature on moral self-regulation provides no evidence if licensing or cleansing works via *moral credentials* or *moral credits* in the domain of honesty. Demonstrating honest behavior could provide the credentials for subsequent dishonesty. A lie may then be interpreted in a more favorable way if it was prefaced by honest behavior. Similar to the results on sexism and racism (Monin and Miller, 2001), participants may feel free to lie after having expressed their preference for honesty in general. On the other hand, licensing and cleansing could work through the balancing of moral credits. Honesty could build up moral credits that would later be spent on dishonest behavior and vice versa. Since no clear evidence exists which mechanism is at work, I decided to implement both mechanisms and manipulate participants' moral credentials as well as their credits. Much evidence exists on the fact that people are averse to lying and even abstain from lying if it results in a pareto-superior situation (Gneezy 2005; Lundquist et al. 2009; Erat and Gneezy 2012; Maggian and Villeval 2014). I hypothesize that in order to license lying a strong mechanism is required to generate a license. To create such a mechanism I developed one task that manipulated credentials and one task that manipulated credits. Participants in the experiment performed both tasks in sequence so that their credentials and their credits were manipulated at the same time. With this approach I want to identify if licensing (and possibly cleansing) exists at all rather than determining the individual mechanism that is at work. If this procedure does not produce a license for lying, this would provide strong evidence against same-domain licensing.

Manipulation of Moral Credentials

Following the manipulation of credentials that Monin and Miller (2001) implemented, all first-movers (the *senders* in the following deception game) were presented with five statements (see Appendix A1 for the exact wording of each statement) and asked to state their agreement to each statement. Monin and Miller (2001) constructed treatments by the use of different keywords ("some" vs. "most")⁶. Disagreement to sexist statements such as "Some/Most women are better off at home taking care of

⁶ This approach is based on Salancik and Conway (1975).

the children" (Monin and Miller 2001, p. 35) is easier to voice if statements are framed in a blatant sexist way using the word "most". By their disagreement participants acquired moral credentials of being not sexist. This in turn led them voice sexist behavior more easily afterwards as they were licensed to do so. In my experiment I used that general idea as well, but the wording was slightly different. The statements took the form of advice that one person gives to another. Five statements were constructed in such a way that for five different areas of life (family, politics, academia, journalism and business) a superior gives advice to a subordinate person suggesting to utilize dishonesty for personal gain. Furthermore, the statements in treatment truth were framed as strong as possible and therefore as easy to disagree to as possible by substituting the keyword "most" with "all"⁷. The critical manipulation in treatment lie was to insert the word "sometimes" in all sentences. This made the advice more ambiguous and is expected to lead to less disagreement as sometimes dishonesty might be a reasonable decision. This in turn means that less moral credentials are acquired. In the baseline treatment statements from the same five areas as above were employed but the content had no connection to honesty. Here, it was necessary to give participants a neutral task that would leave their moral credentials unaltered. At the same time this task should not be too easy so participants would not get bored and in turn experience negative emotions towards the experimental procedures which would bias the results. Furthermore, agreement and disagreement was required to balance out over all, so no bias was introduced by an overall stronger focus on positive or negative answers. According to these prerequisites I constructed five statements (see Appendix A1) covering five topics that were relevant in the news at the time the experiment took place. Participants were expected to have a personal opinion either in favor or against each of the given statements. In all treatments participants were asked to state their agreement on a 5-point scale ranging from "I completely agree" (1) to "I completely disagree" (5).⁸

Manipulation of Moral Credits

After providing participants with credentials for honesty, I went on to manipulate their moral credits in task 2. Sachdeva et al. (2009) require participants to write a self-relevant story with positive, negative or neutral trait words in order to change their moral self-worth. I build upon this idea and gave first-movers (senders) the following task: In treatment truth, participants were asked to write about a real or fictional situation, where it was difficult for them to tell the truth. In treatment lie, participants were asked to write about a situation in which they lied. In baseline participants had to describe a situation where they discussed any of the topics from task 1 (see Appendix A2 for the exact wording). Participants used pen and paper to work on this task in handwriting. I asked them to not write more than five sentences or bullet points and finish this task within five minutes.

A situation in which one behaved honestly, especially if it was hard to do so, would provide individuals with evidence of being an honest and upright person, thereby increasing their moral credits. By adding "although this was difficult for you" I forced participants to think about situations where honesty was not an easy option to choose. In doing so, I tried to rule out answers that did in fact reflect honest behavior but without any moral costs. Possible examples are all instances where someone tells honest facts, like the current time or daily routines. Reporting such facts of course qualifies as honest behavior, but it does not increase a moral balance because critically no moral credits are acquired by such behavior. I conjecture that in order to obtain a moral license it is not only important to tell the truth, but also that some difficulty was involved in doing this. Therefore, participants were asked to describe a situation that involved some obstacle.

⁷ The word "all" itself was dropped in the sentences for grammatical reasons.

⁸ The exact question for treatment truth and treatment lie was: "Please indicate on a scale from (1) to (5) how much you agree to the advice given on a moral level."

In treatment lie participants' moral credits are expected to decrease, as describing a situation where one told a lie would force participants to visualize and think about an immoral situation. Telling lies is generally considered immoral behavior and as such participants would be reminded of their negative moral balance. In baseline no change of the moral balance is expected, as participants described a neutral situation without any connection to moral topics. A discussion about family, politics or business should neither increase nor decrease moral credits. Using this approach, experimental procedures are identical over all treatments and make comparisons between treatments possible.

Scrambling Task for Second-Movers

Since I was interested in how changes in the moral self-worth changed first-movers' (senders') behavior, I tried to avoid any effects that would arise out of beliefs formed by first-movers on how second-movers (the *receiver* in the following deception game) might act. To avoid these effects (of first or higher order) every first-mover was informed that only he would deal with task 1 and 2. Furthermore, I explicitly informed first-movers about the fact that their matched partners would do an unrelated word-scrambling task in the meantime and that they were completely ignorant to either the statements (task 1) or the self-relevant story (task 2). Since first-movers needed some time to complete tasks 1 and 2 I devised a word-scrambling task for second-movers to keep them occupied for the same amount of time. For this scrambling task I used an idea from Khan and Dhar (2005, p. 11 and Appendix) and adapted the content to fit the experiment. All second-movers in all treatments used pen and paper to unscramble seven sentences in handwriting. The sentences used neutral words (see Appendix A3 for the exact wording) and had no connection to truth-telling or lying.

Part 2 - The Deception Game

The second part of the experiment contained the dependent variable of the experiment. Here, senders (first-movers) were to decide if they would tell the truth or lie. In order to measure the amount of truth-telling and lying, I closely follow Gneezy's (2005) setup. Participants played the same cheap talk sender-receiver game Gneezy (2005) used. In this game the sender alone holds the information about payoffs that result out of two possible options, while only the receiver (second-mover) decides what option gets implemented. To link information and decision, the sender communicates a true or a false message to the receiver. This message is cheap-talk because there is no way to verify it. This makes it possible to measure senders' honesty by analyzing the messages sent.

I chose to replicate relative payoffs from treatment 3 in Gneezy's (2005) game (see Table 1), since his results show a relative high number of liars (52%) for this treatment. This constitutes a nearly equal split of the population and provides a good starting point for my experiment, since deviations in both directions are possible. Table 1 shows the payoffs to sender and receiver that resulted out of the two possible options. Following the two tasks from part 1, all senders were presented with the following payoff matrix⁹:

	Option A	Option B
Sender	2€	6€
Receiver	6€	2€

Table 1: Payoffs for sender and receiver in all treatments

After having received this information senders were asked to send one of two possible messages to the receiver:

⁹ To avoid order effects that may bias the results, random selection was used during the experiment. Each sender was presented with one out of two payoff matrices with probability 0.5. For the second matrix, payoffs for options A and B were simply reversed.

Message 1: "Option A will earn you more money than Option B"

Message 2: "Option B will earn you more money than Option A"

Comparing the message sent to the payoff matrix makes it possible to identify if a sender was sending the truth or a lie.¹⁰ After receiving the message from the sender, the receiver had to decide which option he would like to implement. This option would determine the final payoffs for both players. The receiver was informed about the two possible messages but received no information about absolute or relative payoffs.

While the receiver considered his decision, the sender was asked to state his beliefs about this decision. I asked the simple question "What do you think: Which option will player 2 pick?", which was inspired by Sutter (2007). This lets me extract beliefs about the receiver's actions. After answering this question, the sender was told which option the receiver had actually decided to implement and what payoff would result out of this decision.

4. Experimental Hypotheses

I set up this experiment to test if moral licensing and moral cleansing influence behavior in the domain of honesty. More exactly, I ask if on the one hand lying can be licensed within the same domain (treatment truth) and on the other hand if participants feel the need to cleanse themselves of their dishonesty by telling the truth (treatment lie). To compare behavior, I constructed a baseline treatment with the exact same experimental procedures but without any connection to moral topics or specifically (dis)honesty. Participants' moral credentials as well as credits were manipulated which allows me to assess the fundamental question of the existence of same-domain moral self-regulation rather than to individually assess which mechanism is the reason for such behavior. I set up two hypotheses as follows:

H1 (moral licensing): In *treatment truth* significantly more lying is observed compared to the baseline treatment.

H2 (moral cleansing): In *treatment lie* significantly more truth-telling is observed compared to the baseline treatment.

According to past research on moral licensing and cleansing I expect to observe significantly more lying in treatment truth compared to the baseline and significantly more honesty in treatment lie compared to baseline. In treatment truth *licensed* individuals are expected to *lie more*, while individuals in treatment 2 would strive for *cleansing* and *lie less*.

292 students from the University of Passau participated in the experiment. All sessions were executed in the experimental laboratory at the University of Passau¹¹. All interactions took place anonymously via computer clients. Next to every computer a pen and a sheet of paper was placed and participants were told that they needed this material during the experiment to answer questions in writing. The experiment was programmed in z-Tree (Fischbacher 2007). The experiment took about 25 minutes to complete with an average payoff of 4 Euro. Data was gathered from 292 participants and resulted in 49 sender-receiver pairs in baseline, 48 pairs in treatment truth and 49 pairs in treatment lie. Overall, 70.3% of the participants were female and the mean age was 22.5 years. Participants were on average in their 4.9th semester.

¹⁰ In this case, message 1 would be the truth and message 2 would be a lie. If nature selected the other payoff matrix with $p=0.5$, message 1 would be a lie and message 2 would be the truth.

¹¹ For the recruitment of participants and providing laboratory resources I kindly thank PAULA – the Passau Experimental Laboratory.

5. Results

Task 1 - Manipulation of Credentials

The manipulation of the statements in task 1 worked as expected. In treatment truth and treatment lie the histograms in Figure 1 show a clear negative skew indicating strong disagreement overall. In the baseline treatment the distribution is far more symmetric without a tendency towards either agreement or disagreement. A two-sample Wilcoxon-Mann-Whitney-Test shows that disagreement in treatment truth is significantly stronger than in treatment lie ($p=0.039$). This indicates that the manipulation worked as intended. Most disagreement is observed in treatment truth ($M=4.6$, $SD=0.41$) where the statements suggested to use blatant dishonesty. In treatment lie less disagreement ($M=4.35$, $SD=0.63$) is observed as the statements were more ambiguous by the inclusion of the word "sometimes". In baseline ($M=2.95$, $SD=0.50$) arguments for both agreement and disagreement are equally valid and therefore no skewed distribution is observed. While in baseline participants' credentials remained unaltered, participants in treatment truth strongly disagreed to using dishonesty and consequently have acquired the credentials of honest individuals. In treatment lie, disagreement was weaker and participants therefore have acquired no credentials for honesty.

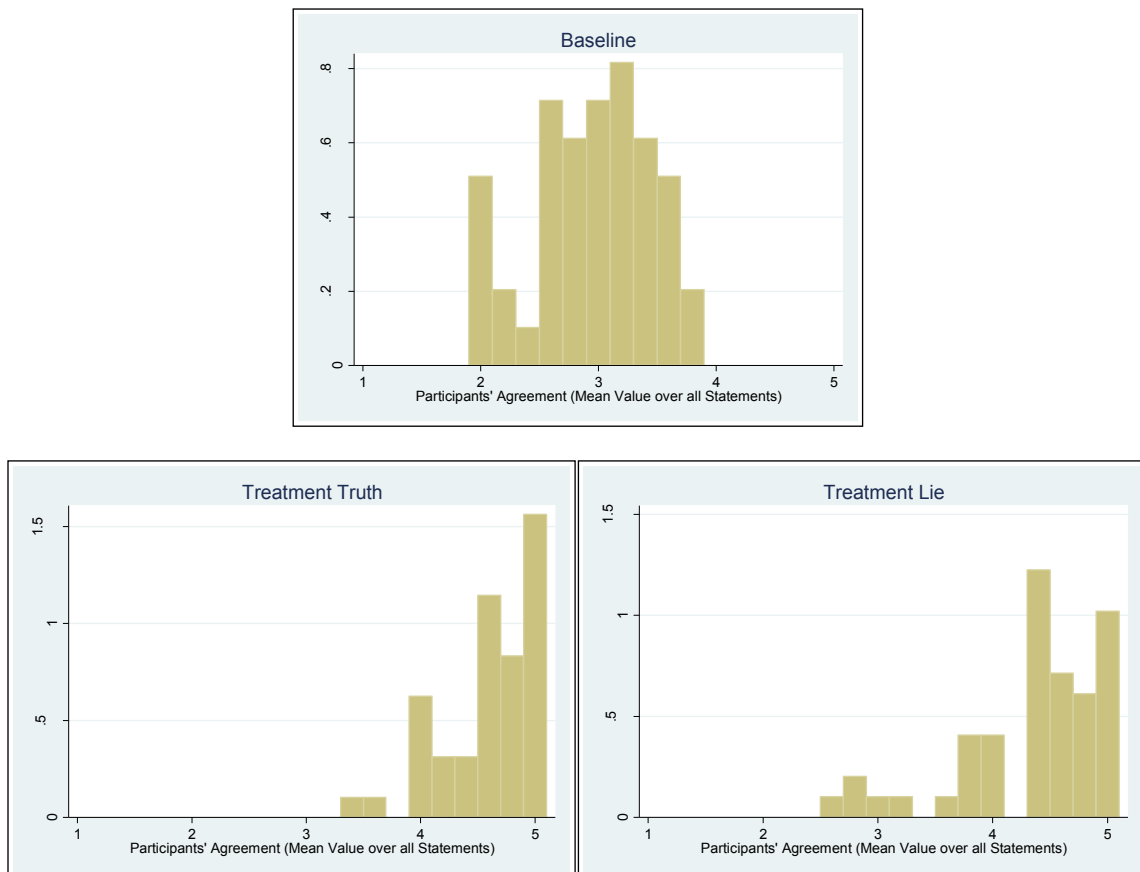


Figure 1: Density of means for task 1 (on the scale from (1): "I completely agree" to (5): "I completely disagree")

Task 2 - Manipulation of Credits

In task 2 participants were asked to describe a recent situation in which they either told the truth (treatment truth) or told a lie (treatment lie). In the baseline treatment participants were asked to describe a situation where they discussed any of the topics from task 1. All participants were able to complete this task within the given 5 minute timeframe. Nearly all participants indicated that they

described a real situation that happened to them recently¹². Reading the stories revealed that participants wrote about truth-telling and lying in a huge variety of situations. Stories included very different topics concerning university, family, dating, friends or leisure activities. Table 2 provides some examples of the handwritten stories for each treatment.

	<i>Examples</i>
<i>Baseline</i>	<p>Talked with friends about the abolishment of college tuition fees. We all agreed that this is a good thing.</p> <p>I do not agree that the Federal Government did a good job. I discuss this topic frequently with my roommate after the evening news. I remember that I tried to explain to her the pros and cons of collective labor agreements.</p> <p>As homework for an English language course I had to write an essay about my opinion on the salary of a top manager from the Royal Bank of Scotland. I pondered the pros and cons and in the end decided that his wage is too high. Therefore I like to opt for governmental regulation.</p>
<i>Treatment Truth</i>	<p>Borrowed something and partially broke it. When returning it, I told about it and offered compensation, even though I could have acted as if nothing had happened.</p> <p>Truthfully told my roommate in our shared flat that she had to change her behavior or else she would have to move out. Confronting her in that way was not easy.</p> <p>A friend of mine bought a new dress which she totally adored. I told her that it did not look nice on her and knew she would be sulky for the next days.</p>
<i>Treatment Lie</i>	<p>Told my friends I was too busy to go to a party with them. In reality I was just lazy.</p> <p>I was responsible for a scratch in my mother's car. She did not realize it at first and when she asked me about it two months later, I denied it.</p> <p>I told my girlfriend that I did attend a lecture even though I did not. I did not want her to nag at me.</p>

Table 2: Examples of participants' answers for task 2¹³

These examples demonstrate that participants provided lots of different answers in task 2. This was of course expected as the task was formulated on purpose in an open way so participants would not feel constrained and perhaps would not come up with a fitting answer in time. Therefore heterogeneity was expected¹⁴. But nevertheless all stories qualified as correct answers to task 2 and no participant was excluded from the analysis.

The results from both tasks indicate that the necessary manipulation of moral credentials as well as credits was achieved in the desired way. Next, I turn to the results from the follow-up deception game in order to assess how the manipulation of moral balances influenced honesty. Figure 2 shows the results for truth-telling and lying over all treatments. In baseline with neutral wording, I observe that 37% of participants tell a lie. In treatment truth the rate of lying increases to 40% and in treatment lie lying decreases to 35%. However, the differences between treatment truth and baseline ($p=0.386$) as

¹² In treatment truth, 1 participant stated to describe a fictional situation; in treatment lie, 4 stories were labeled fictional.

¹³ Examples are translated from German and are abridged to provide illustrative examples rather than exact typed transcripts.

¹⁴ For a more detailed analysis of participants' answers see the follow-up paper, where I take a look at this qualitative data in more detail (Nagel 2014, working paper).

well as between treatment lie and baseline ($p=0.583$) are not statistically significant. The difference between treatment truth and treatment lie ($p=0.309$) fails to meet significance as well¹⁵.

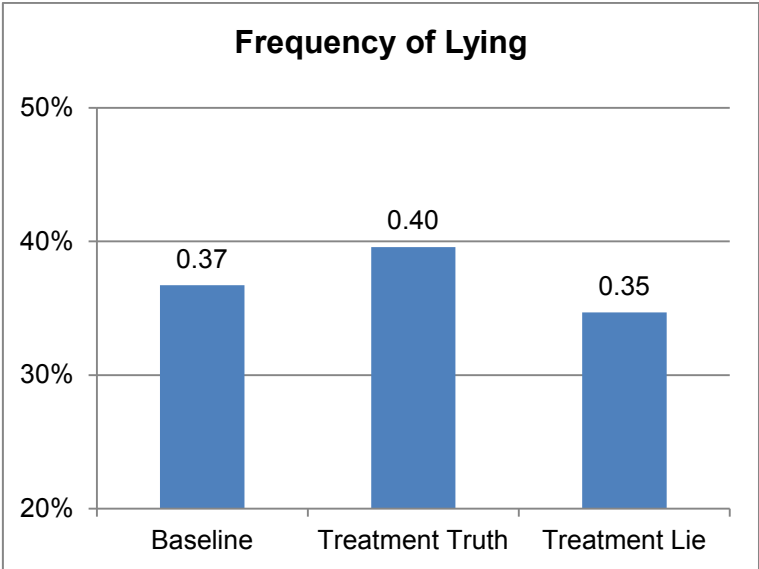


Figure 2: Frequency of lying (dishonest messages)

These results indicate that neither a licensing nor a cleansing effect exists for truth-telling and lying. If licensed behavior existed a significant increase in lying in treatment truth should be observed. Cleansing behavior on the other hand should lead to significantly more truth-telling in treatment lie. But compared to the baseline treatment I cannot identify any differences in the two treatments.

But Figure 2 looks only at actual messages sent and does not include players' expectations. The message sent and the expectation of what the receiver will do may not always align. Senders will not always expect receivers to trust them and implement their messages. Therefore Sutter (2007) proposed to extend the definition of deception beyond the one Gneezy (2005) used. He states that in order to measure the true extent of deception not only those players sending a lie need to be accounted for but also those sending the truth because they expect the other players to distrust them¹⁶. Such a sender would engage in one additional step of reasoning because he anticipates possible distrust by the receiver and accounts for that by sending the truth on purpose in order to get the lie implemented. By applying this broader definition of deception, I observe the following rates of lying as depicted in Figure 3.

¹⁵ p-values are calculated from one-tailed tests of the equality of proportions (z-test).

¹⁶ Sutter (2007, p. 9) uses the following definition: "Deception includes all cases where a sender sends either of the two messages, but expects [a lie] to be implemented by the receiver."

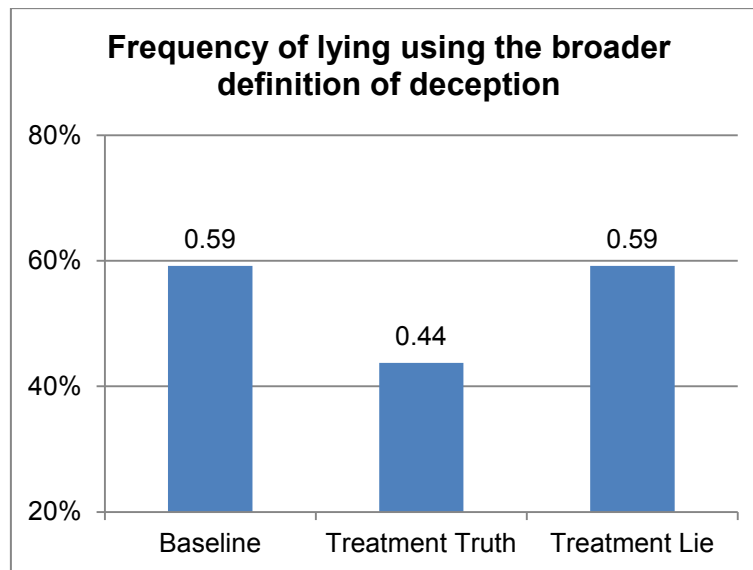


Figure 3: Frequency of lying (broader definition: sender tries to mislead receiver)

In Figure 3 the percentages of those senders are displayed who expected the receiver to implement the option that is favorable for the sender (the lie), regardless of the message they sent. A sender sending a lie and expecting this lie to be implemented falls into that category as well as a sender who sends the truth but expects the receiver to not believe his truthful message. Both types of senders try to mislead the receiver into picking the favorable option for the sender but differ in their expectations about the receiver and therefore about the message they need to send in order to achieve the higher payoff. In baseline as well as in treatment lie I observe 59% of senders who want to mislead the receiver, whereas in treatment truth only 44% try to do so (one-tailed z-test, $p=0.064$). Using this definition, lying decreases significantly in treatment truth. By applying this broader definition I observe that participants did not act according to what licensed behavior suggests but instead even increased honesty. This is the opposite of what is expected from licensed behavior and runs contrary to what hypothesis 1 suggested.

This observation provides additional support for the initial results. Applying the broader definition of deception fails to provide evidence for the existence of licensed or cleansing behavior as well. Neither if I look only at the message sent (Figure 2), nor if I account for senders' expectations (Figure 3) I find evidence in favor of moral self-regulation of truth-telling and lying. I therefore state that within my experimental setup neither same-domain licensing nor cleansing exists for truth-telling and lying. The two main results are:

Result 1 (licensing): Participants acquire no license for lying through honest behavior. Moral credentials as well as moral credits fail in providing a license for subsequent dishonesty.

Result 2 (cleansing): Participants do not increase truth-telling as a consequence of having no moral credentials for honesty or having lost moral credits through lying. No moral cleansing behavior that aims at balancing dishonesty with subsequent truth-telling is observed.

6. Additional Results

Based on these two main findings I want to address some additional results in the following section, since the experimental setup allows for more in-depth analysis of players' behavior. Especially, I want

to take another look at senders' expectations and assess how these differ compared to receivers' actual behavior. Furthermore I like to assess how expectations differ depending on message sent. Lastly, based on message sent and subsequent expectation I describe results on four different types of senders.

Expected Trust and Actual Trust

By comparing the message sent with the expectations about what the receiver will do it is possible to calculate how many senders expected the receiver to trust the message. Trusting a message means that the receiver implements the option which is suggested to be favorable for him. If a sender expects trust he believes that the receiver will implement the message by picking the option that is indicated to be more favorable for the receiver. This could either be the case if the truth is sent, resulting in the receiver getting the higher payoff, or if a lie was sent, resulting in the sender getting the higher payoff. The results show that 69% of senders expect the receiver to trust them in baseline, 58% expect their message to be trusted in treatment truth and 63% in treatment lie. Differences between treatments are not significant (one-tailed z-tests: $p=0.129$ and $p=0.261$ respectively).

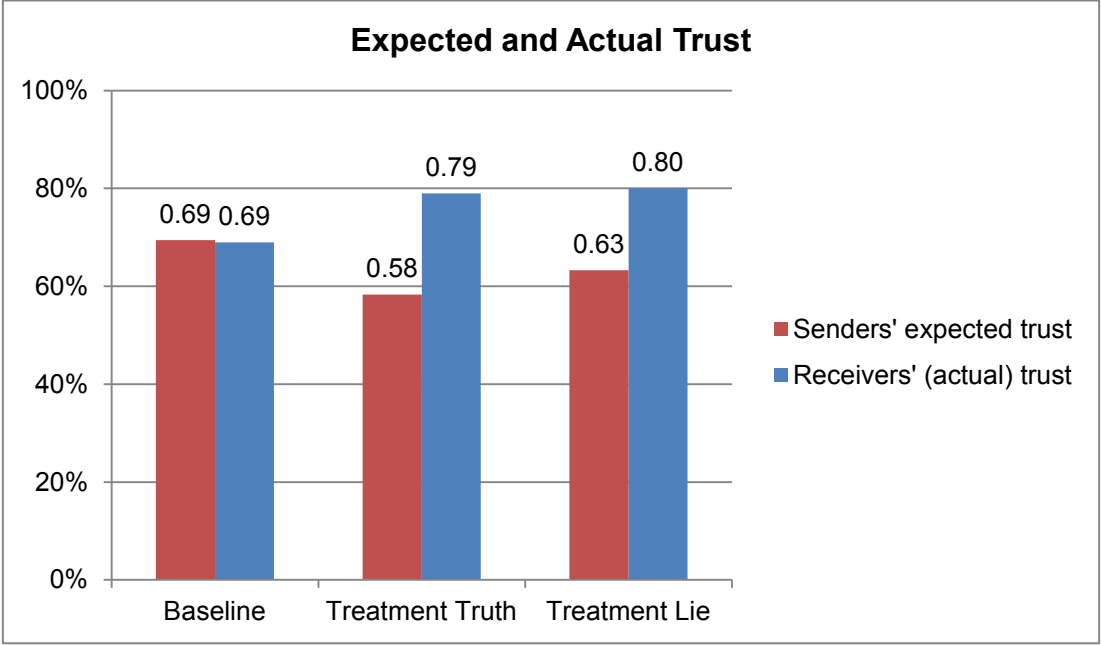


Figure 4: Senders' expected and receivers' actual trust

Receivers' actual trust is 69% in baseline, 79% in treatment truth and 80% in treatment lie. Again, treatments do not differ in the rates of actual trust. This comes as no surprise, as receivers conducted the exact same word scrambling task in all three treatments and were otherwise oblivious to any connection to honesty or morals or the fact that senders conducted different tasks. Therefore nearly identical rates of trust are observed in the three treatments.

It is worthy to note that the above presented rates of expected trust are different to other research. Sutter (2007, p. 6) reports "that senders' expectations match receivers' actions remarkably well in aggregate". This result does not hold true here, where expected trust is significantly smaller than actual trust in treatment truth ($p=0.013$) as well as in treatment lie ($p=0.037$). It seems that the two tasks, where participants' moral balances were manipulated with regard to truth-telling and lying, had the effect to make senders more doubtful about receivers' actions. Therefore, senders' expectations no longer match receivers' actions very well. Simply thinking about the own (dis)honesty influenced what behavior is expected from others.

Trust depending on message sent

Having established the general levels of trust between treatments, I further differentiate trust by the message sent. Figure 4 showed all receivers who expected trust from the receiver regardless of the message they sent. In Figure 5 I split these senders into two groups depending on message. In baseline I observe that more senders (89%) expect a dishonest message to be followed compared to a honest one (58%). The same pattern is visible in treatment lie where 82% expect a lie to be followed compared to only 53% who expect a truthful message to be followed. A different picture is visible in treatment truth, however. The amount of senders who expect trust after sending the truth is similar to the other treatments (62%) but trust after sending a lie drops considerably (53%). The difference to what is observed in baseline is significant ($p=0.008$). This means that after sending a lie significantly less senders expect this lie to be trusted and implemented by the receiver. Again, such behavior does not indicate licensed behavior. If participants had acquired a license for lying, it is to be assumed that they would not only behave more dishonestly (by sending lies) but furthermore expect to get away with their dishonesty. Here, I observe the exact opposite. Nearly half of all senders who sent a lie do not expect this lie to be implemented.

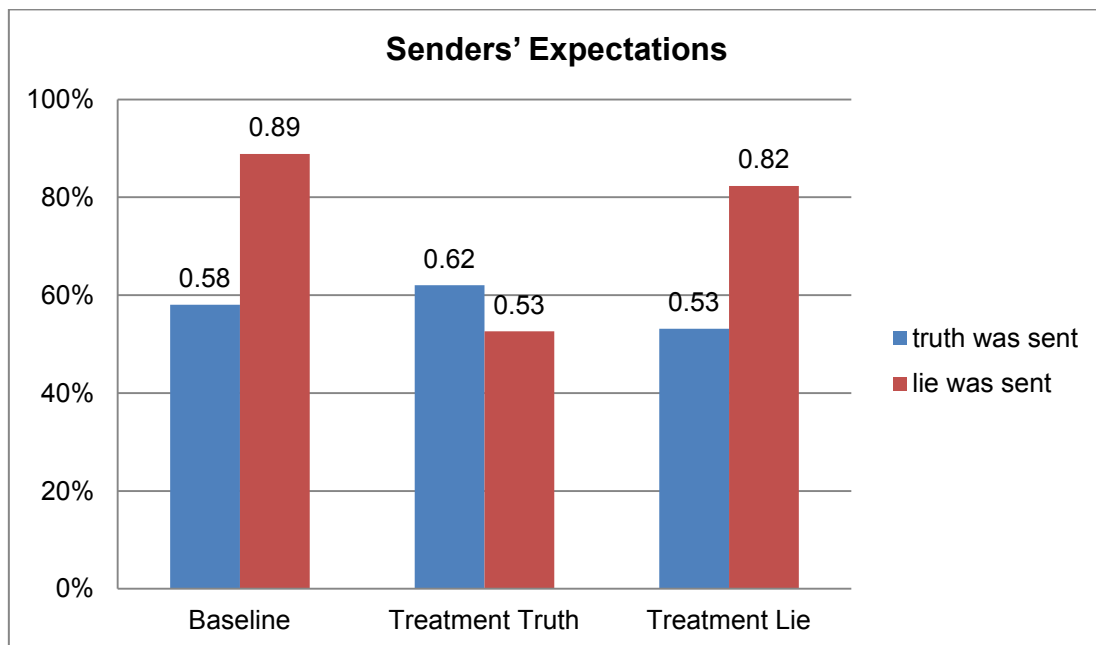


Figure 5: Senders' expectations of receivers trust depending on message sent

A possible explanation for this observation could be that in treatment truth I asked participants to take a firm position against lying in task 1. This might have made them alert to the dangers of lying and lead to their conjecture that receivers will weigh messages more carefully and actively expect deception. Although I explicitly informed senders that receivers would not know about the tasks related to honesty but instead perform an unrelated word scrambling task, senders might still (wrongly) hold such a conjecture. Yet, a similar effect is not observed in treatment lie. This suggests a difference in expected trust if one thinks about honesty on the one hand and dishonesty on the other hand. Furthermore, if participants thought about honesty in a situation where it was difficult to do so (task 2), they focused on the fact that honesty is often difficult to achieve. Since the task established the belief that honest behavior comes with costs and proves to be difficult in many situations, participants could assume that receivers think exactly this way as well. Being in such a setting, a message that cannot be verified to be honest would instantly evoke feelings of suspicion. Being more alert about dangers of lying could explain the higher rates of distrust after sending a lie. This could lead to the increased beliefs of distrust among senders in treatment truth. In treatment lie on the other

hand, it seems reasonable to assume that participants wrote about situations where they successfully got away with a lie. One tends to remember positive and successful events more often than negative or unsuccessful ones. Therefore, it could be possible that participants wrote stories mainly about successful lies. This in turn could have lead them to belief that as they got away with a lie in the past, this would work in this experiment as well. These results indicate that expectations about receivers' actions differ depending on treatment. This means that thinking about truth leads to more distrust, while thinking about lies did not change beliefs compared to baseline. These results point out that an increase in credentials and/or credits may lead to different beliefs about others.

Different types of senders

Combining the dimensions "message" (truth, lie) and "expectation" (trust, no trust), Sutter (2007) developed four categories of players. Someone who tells the truth can be either a *benevolent truth-teller* if he also expects the receiver to implement the suggested message. Or he can be a *sophisticated truth-teller*, because he sends the truth but expects the receiver to not follow his message and implement the opposite. For liars the same logic applies: a *liar* sends a lie and expects the receiver to follow this lie. A *benevolent liar* on the other hand sends a lie and believes in the receiver distrusting his message, ultimately resulting in the better outcome for the receiver.

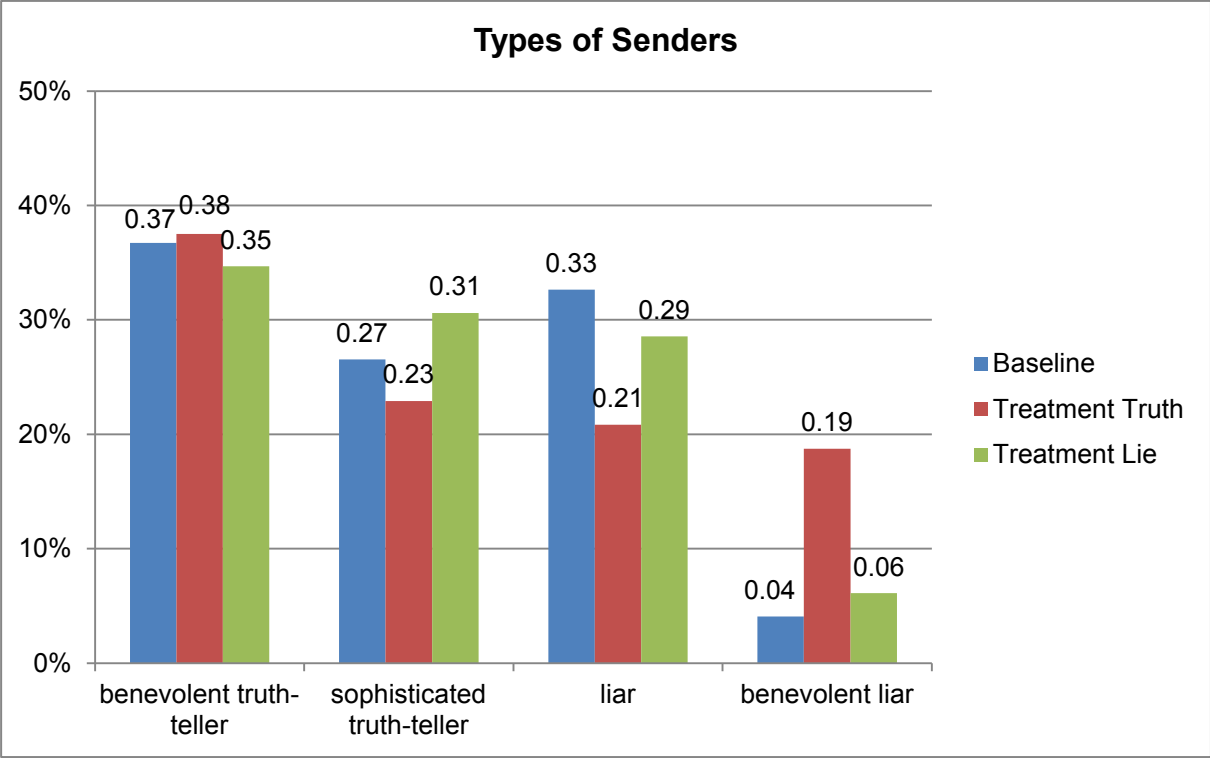


Figure 6: Different Types of Senders (based on message and expectation)

In Figure 6 the percentages of players in each category are compared over treatments. I observe similar amounts of benevolent truth-tellers, sophisticated truth-tellers and liars for the treatments truth and lie (all differences compared to baseline are not significant with $p > 0.1$). In treatment truth however, significantly more benevolent liars (19%) are observed compared to baseline (one-tailed z-test, $p = 0.011$)¹⁷. This means that considerably more senders chose to send a lie but distrusted the

¹⁷ I carried out different multinomial logistic regressions with type of sender as the dependent variable controlling for treatment as well as gender, age and study. The results are similar to what is reported here and provide no new insights.

receiver and expected him to implement the other option instead. This corresponds to the findings from Figure 5 where only 53% of senders expected the receiver to trust their lie.

7. Conclusion

Moral licensing and cleansing are well established mechanisms for explaining human behavior. While such behavior could be verified in many different domains, in this paper I find no support that licensing and cleansing can predict behavior within the domain of honesty. In this domain, I could not verify the existence of either licensed or cleansing behavior. I constructed an experiment to identify if same-domain licensing and cleansing exist for truth-telling and lying. Previous research has proven to successfully license transgressions in other domains. My experiment checks for dishonest behavior that is licensed by honesty and honest behavior that occurs out of the need for moral cleansing. I allow for a broad experimental approach by including the manipulation of moral credentials as well as credits through two tasks that have been proven to work in similar experiments. The results on the two tasks indicate that these tasks were indeed successful in the requirement to provide participants with moral credentials and increased (decreased) moral credits in treatment truth (lie). In the subsequent deception game however, I do not find that this has an influence on participants' dishonesty. I observe no different rates of lying than in a neutral baseline treatment. Since evidence for increased dishonesty due to licensed behavior or increased honesty due to moral cleansing is not found I argue that neither credentials nor credits are able to influence participants' honesty. This reveals that exceptions to licensing and cleansing exist. While in some areas licensing seems to work very well (e.g. political incorrectness) in others it does not (e.g. honesty). Trying to license lying proved to fail in my setup and instead of inconsistent behavior after a license I rather observed consistent behavior, especially in treatment truth where honesty increased according to the broader definition of deception. As additional results I observed that participants' expectations differed depending on treatments. I find that taking a firm position against lying (task 1) and writing a self-relevant story about honest behavior (task 2) leads participants to expect more distrust from the receivers of their message, especially after a dishonest message. To understand the forces that drive licensing and cleansing, future research could therefore explore if and how lying may be licensed by other good deeds that lie outside the realm of honesty and how people engage in moral cleansing for dishonesty. Furthermore an interesting question seems to be how moral credentials for honesty or an increase of moral credits through honest behavior influence subsequent trust.

Appendix

A1. Task 1 (manipulation of moral credentials)

Original statements were in German language.

Baseline

1. Instead of building daycare centers for children, families should receive direct monetary benefits.
2. It is a good thing that college tuition fees got abolished.
3. Overall the Federal Government did a good job in the legislative period from 2009 - 2013.
4. Public television and radio license fees should be abolished and the content of the media should be left to the discretion of private suppliers.
5. Excessive salaries for executives should be prohibited by law.

Treatment Truth

1. A father tells his son: "In order to have success in life, you need to lie and deceive."
2. A chemistry professor tells his laboratory assistant: "You need to rig your data to obtain a good publication."
3. A consultant tells a politician: "You need to lie to the voters. That will get you re-elected."
4. A chief editor tells a journalist: "You need to twist the facts for your article to sell well."
5. A manager tells his accountant: "You need to cook the books so we will get a better economic performance."

Treatment Lie

1. A father tells his son: "In order to have success in life, you sometimes need to lie and deceive."
2. A chemistry professor tells his laboratory assistant: "Sometimes you need to rig your data to obtain a good publication."
3. A consultant tells a politician: "Sometimes you need to lie to the voters. That will get you re-elected."
4. A chief editor tells a journalist: "Sometimes you need to twist the facts for your article to sell well."
5. A manager tells his accountant: "Sometimes you need to cook the books so we will get a better economic performance."

A2. Task 2 (manipulation of moral credits)

Original statements were in German language.

Baseline:

"Please describe a situation that happened lately, where you discussed one of the topics from task 1."

Treatment Truth

"Please describe a situation that happened lately, where you told the truth although this was difficult for you."

Treatment Lie

"Please describe a situation that happened lately, where you told a lie."

A3. Scrambling Task for Second-Movers (receivers) for all treatments

Original statements were in German language.

1. The University the river Inn is located at of Passau.
2. Weekly market to the Susi on goes Friday.
3. This is my not backpack new.
4. The cinema many movies good you can in see.
5. Dog brown has the hair.
6. Birthday party a Peter on is Sunday at.
7. Child mother the her buys ice cream an.

References

- Brañas-Garza, P., Bucheli, M., Espinosa, M. P., and García-Muñoz, T. (2013): *Moral Cleansing and Moral Licenses: Experimental Evidence*. Economics and Philosophy, Cambridge University Press 29(2), 199-212.
- Gneezy, U. (2005): *Deception: The Role of Consequences*. The American Economic Review 95(1), 384-394.
- Effron, D. A., Cameron, J. S., and Monin, B. (2009): *Endorsing Obama licenses favoring Whites*. Journal of Experimental Social Psychology 45, 590-593.
- Erat, S., and U. Gneezy (2012): *White Lies*. Management Science 58(4), 723-733.
- Fischbacher, U. (2007): *z-Tree: Zurich toolbox for ready-made economic experiments*. Experimental Economics 10(2), 171-178.
- Fishbach, A., and Dhar, R. (2005): *Goals as excuses or guides: The liberating effect of perceived goal progress on choice*. Journal of Consumer Research 32, 370–377.
- Hao, L., Houser, D. (2010): *Honest Lies*. GMU Working Paper in Economics No. 11-16. Available at SSRN: <http://ssrn.com/abstract=1801546> or <http://dx.doi.org/10.2139/ssrn.1801546>
- Jordan, J., Mullen, E., and Murnighan, J. K. (2011): *Striving for the Moral Self: The Effects of Recalling Past Moral Actions on Future Moral Behavior*. Personality and Social Psychology Bulletin 37(5), 701-713.
- Hollander, E. P. (1958): Conformity, status, and idiosyncrasy credit. Psychological Review 65, 117-127.
- Khan, U., and Dhar, R. (2006): *Licensing effect in consumer choice*. Journal of Marketing Research 43(2), 259–266.
- Lundquist, T., Ellingsen, T., Gribbe, E., and Johannesson, M. (2009): *The Aversion to Lying*. Journal of Economic Behavior & Organization 70(1-2): 81-92.
- Maggian, V., and Villeval, M. C. (2014): *Social Preferences and Lying Aversion in Children*. Available at SSRN: <http://ssrn.com/abstract=2368098> or <http://dx.doi.org/10.2139/ssrn.2368098>
- Mazar, N., and Zhong, C.-B. (2010): *Do Green Products Make Us Better People?* Psychological Science 21(4), 494-498.
- Merritt, A. C., Effron, D. A., and Monin, B. (2010): *Moral self-licensing: When being good frees us to be bad*. Social and Personality Psychology Compass 4, 344-357.
- Miller, D. T., and Effron, D. A. (2010): *Psychological license: When it is needed and how it functions*. In M. P. Zanna & J. M. Olson (Eds.), *Advances in experimental social psychology* 43, 117-158. San Diego, CA: Academic Press/Elsevier.
- Monin, B., and Miller, D. T. (2001): *Moral credentials and the expression of prejudice*. Journal of Personality and Social Psychology 81, 33-43.

Nisan, M. (1991): *The moral balance model: Theory and research extending our understanding of moral choice and deviation*. In W. M. Kurtines & J. L. Gewirtz (Eds.), *Handbook of moral behavior and development*, 213-249, Hillsdale, NJ: Erlbaum.

Sachdeva, S., Ilic, R., Medin, D. L. (2009): *Sinning saints and saintly sinners: The paradox of moral self-regulation*. *Psychological Science* 20, 523-528.

Salancik, G. R., and Conway, M. (1975): *Attitude inferences from salient and relevant cognitive content about behavior*. *Journal of Personality and Social Psychology* 32, 829-840.

Sutter, M. (2009): *Deception through telling the truth?! Experimental evidence from individuals and teams*. *Economic Journal* 119(534), 47-60.

Zhong, C.-B., Liljenquist, K. A., and Cain, D. M. (2009): *Moral Self-Regulation: Licensing & Compensation*. In D. De Cremer (Ed.), *Psychological Perspectives on Ethical Behavior and Decision Making*, 75-89, Charlotte, NC: Information Age Publishing.

Consistent Behavior in the Domain of Honesty: Why Moral Self-Regulation Fails

Volker Nagel[✉]

Abstract

In a previous experimental study (Nagel 2014) no evidence for moral self-regulation (moral licensing or moral cleansing) in the domain of honesty was found. In this paper, possible causes for this result are addressed and a coding procedure is employed to utilize data from participants' hand-written stories which resulted from the original experiment. Results reveal that participants acted consistent to what they indicated about themselves in their stories. Stories that suggested an increase in moral credits were followed by more honesty. Stories that suggested a decrease were followed by more dishonesty. This supports the initial finding that licensing and cleansing fail in the domain of honesty and sheds light on participants' behavior in more detail. The fear of appearing as a hypocrite as well as the desire to maintain an honest self-concept are discussed as possible reasons for such behavior.

Keywords: moral self-regulation, deception, honesty, lying, moral hypocrisy, self-concept

[✉] Volker Nagel is a doctoral scholar at the University of Passau, Germany. The author is grateful to the participants of the brownbag economics seminar at the University of Passau, Nov. 14, 2012, for helpful comments. Furthermore, the author thanks the student assistants who coded the experimental data.

1. Introduction

Moral self-regulation has been widely researched and two main mechanisms have been proven to influence behavior. Firstly, moral licensing occurs when individuals engage in morally questionable behavior, because their previous good deeds provide them with a license to do so. Secondly, individuals engage in actions that figuratively cleanse them of their previous transgressions. Licensing and cleansing explain why people work out hard in the gym to lose weight and afterwards - seemingly oblivious to their effort - snack chocolate bar or why the mere contemplation about personal transgressions in the past increases charitable giving. Licensing and cleansing both result in inconsistent behavior. Snacking a chocolate bar is inconsistent to the goal of losing weight. Giving to a charity is inconsistent if someone describes himself as an egoistic and selfish person. Research on moral licensing and cleansing emphasizes that such inconsistencies occur because any single behavior is not an isolated observation but is embedded in a series of previous and subsequent behavioral choices that are all interrelated. To explain how one action influences another within this framework, the moral credentials (e.g. Monin and Miller 2001, Effron et al. 2009) and credits model (e.g. Sachdeva et al. 2009, Brañas-Garza et al. 2013) have been established. While the moral credentials model explains how morally ambiguous behavior gets reinterpreted in the light of previously correct behavior, the moral credits model explains how good and bad actions are executed in such a way as to balance each other out.

In my previous study (Nagel 2014) I tried to identify if those models can explain self-regulatory behavior for truth-telling and lying as well. I set up an experiment to check if honesty leads to an increase in lying (licensing) and if dishonesty leads to an increase in truth-telling (cleansing). The results could not find support for either moral licensing or moral cleansing. While providing participants with (or without) credentials for honesty and increasing (or decreasing) their moral credits worked as expected, participants did not lie (or tell the truth) significantly more than compared to a baseline treatment. These results were robust when using the broader definition of deception, i.e. controlling for expectations (Sutter 2009), and revealed that honesty might even increase as a result of acquisition of credentials and credits (Nagel 2014, p. 12). This result runs contrary to what licensing predicts. These experimental findings suggest that moral self-regulation does not emerge for truth-telling and lying and that individuals rather stick to consistent behavior.

In this paper I want to analyze the determinants of the observed behavior in the initial experiment in more detail. By employing a coding procedure I am able to generate additional data from participants' hand-written stories and extend the original data set with a set of new variables from this coding procedure. These variables reveal new insights about participants' behavior and provide evidence why licensing and cleansing fails. This paper is structured as follows: in section 2 the relevant literature is discussed. In section 3 I shortly review the experimental design from Nagel (2014) and explain the coding process that is at the core of this study. Section 4 presents the results for inter-rater agreement and descriptive statistics from the coding procedure. Section 5 links these results to the data from the deception game and presents the results for consistent behavior. Section 6 discusses these results in light of the relevant literature and section 7 concludes.

2. Literature Review

Research on moral self-regulation provides ample evidence on when, why and how moral licensing and moral cleansing appear. Monin and Miller (2001) for example identified licensed behavior for sexism and racism. Participants freely voiced sexist or racist attitudes if prior to that they had been given the opportunity to state their general anti-sexist or anti-racist attitudes. Similarly, Sachdeva et al.

(2009) find evidence how charitable giving increases after participants write a self-relevant story about their negative traits (e.g. selfish, mean, greedy) as a way to engage in moral cleansing. Other studies replicated results on moral licensing and cleansing (Nisan 1991, Fishbach and Dhar 2005, Khan and Dhar 2006, Effron et al. 2009, Zhong et al. 2009, Merritt et al. 2010, Miller and Effron 2010, Mazar and Zhong 2010, Jordan et al. 2011, Monin and Merritt 2012, Brañas-Garza et al. 2013) and provided insights into other areas where behavior can be explained by the forces of moral self-regulation. But Effron and Monin (2010) point out that certain criteria have to be met for moral self-regulation to occur. In their study they focus on the question how and when observers license transgressions. Their results identify certain factors that suppress licensing. For example, a school principal's sexist behavior was strongly condemned when he had implemented a campaign against sexism at school before. If, on the other hand, that campaign was against drug-use, his behavior was excused more often. The transgression was identical in both studies, yet observers were willing to excuse it in one case but not in the other. Effron and Monin (2010) argue that hypocrisy is the reason for such behavior. In the first case the school principal appears as a hypocrite whereas in the second he does not. Observers were unwilling to excuse transgressions whenever the prior good deed made an individual look like a hypocrite.

Hypocrisy has been heavily researched by Batson et. al (1997, 1999, 2003, 2006) and is commonly defined as *saying one thing but doing another* (Barden et al. 2005) or not *practicing what you preach* (Stone and Fernandez 2008). Batson et al. (1997) showed that individuals behaved in a way as to appear moral yet avoid the costs of actually being moral. In his landmark experiment participants had to decide how to assign two tasks between themselves and another person. One task was pleasurable and fun (with the additional chance to win raffle tickets for a \$30 price), the other dull and boring. Participants could either assign the task directly or let a fair coin decide. But flipping the coin was not a binding decision. Participants would flip the coin in private and then only report the (non-verifiable) outcome. This introduced ambiguity into the decision process and allowed for outcomes to appear fair but still favor the participant himself. For example the consequences of the coin flip could be specified post hoc so the coin would assign the pleasurable task to the participant himself regardless of the outcome (e.g. "Heads, I win."; "Tails, you lose."). Or the participant could flip the coin multiple times if the first result was not the desired one (e.g. call a mulligan). In the experiment a significant amount of participants engaged in this kind of behavior: out of the 10 participants who flipped the coin, 9 assigned the fun task to themselves. This proportion differs significantly from the expected 50% proportion of a fair coin. Participants therefore engaged in moral hypocrisy. They stated to have flipped a fair coin (i.e. appear moral) but nevertheless assigned the fun task to themselves (act immoral). Another study (Batson 2006) finds similar results. Participants had to divide 12 raffle tickets between themselves and another person. 6 of these tickets were for a high-value raffle (a \$30 gift certificate), while 6 were for a low-value raffle (a \$5 gift certificate). Moral hypocrisy arose when only the sender was informed about the individual values of the tickets whereas the receiver was not. In this case, participants wanted to appear fair by giving a nearly equal split of tickets (4.75 on average) but most of these tickets were of low value (3.38). The amount of tickets sent was similar to a treatment where neither player was informed about values (4.12). This shows how participants tried to maintain an appearance of fairness but simultaneously acted selfish.

Similar results on hypocritical behavior are reported by Walkowitz et al. (2013), where participants acted fair as long as a scenario was hypothetical. But as soon as real money was at stake, participants again favored themselves. Participants liked to appear fair, yet actually acted unfair. The same is true for honesty, as Hao and Houser (2010) report. They find that participants cheated to the greatest possible degree after having established an appearance of honesty.

While experiments have proven numerous examples of hypocritical behavior, the literature points out factors that suppress moral hypocrisy and lead to consistent behavior instead. People usually refrain from behaving in ways that violate their moral standards or deeply held moral values (Bandura 1991, Aquino and Reed 2002, Effron and Monin 2010). A deviation from their moral identity creates self-condemnation, which is the greatest possible punishment according to Bandura (1991, p. 19). Similarly, Batson (2006) states that for hypocrisy to occur the desire for the preferred distribution has to outweigh the desire to uphold the relevant moral principle. Furthermore, he argues that ambiguity in the transgression encourages hypocrisy, as it grants wiggle room and the possibility to reinterpret behavior in the desired way.

Opposed to hypocritical behavior, Mazar et al. (2008) identified situations where participants did not say one thing and do the other, but rather stuck to their self-concept and acted consistent. They mention the attention to standards as an important factor that influences behavior. For example, their study shows that participants cheated less after trying to remember and write down the Ten Commandments or after being reminded of an honor-code. Such behavior relates to an individual's self-concept which states how that individual views and perceives himself. If someone has a strong belief in his own morality he wants to maintain this aspect of his self-concept. Individuals therefore try to comply with their internal standards even when doing so results in sacrificing financial gains. This research shows that when people are mindful of their own moral standards they tend to act according to these standards. Therefore, thinking about recent instances of honest behavior should make the self-concept with regard to honesty more salient and lead to behavior in line with this self-concept.

In my previous experiment, treatments differed by the manipulation of participants' moral credits. This manipulation was achieved by having participants write about recent situations where they either told the truth (treatment truth) or told a lie (treatment lie). Based on the design employed by Monin and Miller (2001) and Sachdeva et al. (2009) this manipulation assumed that moral credits increase in treatment truth and decrease in treatment lie. Yet, the research on dishonesty suggests that lying may not always reduce moral credits and that lying does not always constitute morally objectionable behavior. Mostly, the economic literature has looked at selfish black lies. These are lies that help the liar at the expense of the other side and it seems reasonable to assume that those lies decrease moral credits in the expected way as it is considered morally wrong to enrich oneself at the costs of others. But what about lies that are told to help other people? Often these so called white lies are told with good intentions and at no harm for the liar (Gneezy 2005). They represent only minor falsehoods told to protect others. Experimental evidence on white lies finds that people do indeed tell lies in order to increase another person's payoff, even if it decreases their own (Erat and Gneezy 2011). Fischbacher and Utikal (2013) even report that in rare instances lies are told which only harm the liar and do not even help another person.

DePaulo offers extensive resources on human lies (DePaulo and Bell 1996, DePaulo and Kashy 1996, DePaulo et al. 1996, DePaulo and Kashy 1998, DePaulo 2004, DePaulo et al. 2004). A similar distinction between white and black lies is provided there. DePaulo et al. (1996) performed a diary study where participants were instructed to write down every lie they told every day for the period of one week. Additionally they provided information about the how's and why's of each lie. This procedure revealed that about 45% of all lies were self-centered, i.e. for personal advantage, and 25% were other-oriented, i.e. told to protect another person. Furthermore, the research revealed that participants often felt little to no distress after telling a lie or indicated that both, they and the target, would have felt worse, had they told the truth instead. This suggests that a huge difference within lies exists and the general assumption that writing a story about an instance where a lie is told may not necessarily lead to a decrease of moral credits. Therefore a more detailed approach has to be taken by

looking at stories individually to determine what kind of lie (and similarly truth) was told and how this ultimately affected moral credits.

3. Experimental Design and Coding Procedure

In the previous paper (Nagel 2014) I followed the design from Monin and Miller (2001) for the manipulation of moral credentials (task 1) and Sachdeva et al. (2009) for the manipulation of moral credits (task 2). Afterwards participants engaged in a deception game (Gneezy 2005), where only a sender is informed about the payoffs resulting from two possible options. He sends one of two possible messages to the receiver (e.g. "You receive a higher payoff from option X (Y)"). The receiver then decides which option to implement. With this design it is possible to identify if the truth or a lie was sent. To assess how this decision was influenced by previous (dis)honesty, I manipulated participants' credentials and credits with regard to honesty. The two tasks asked participants to indicate their agreement to statements suggesting to utilize dishonesty (task 1) and furthermore to write a self-relevant story about themselves (task 2). The treatments differed by the framing of the statements and the content of the self-relevant story. In treatment truth, participants indicated (dis)agreement to statements that suggested using blatant dishonesty (e.g. you have to lie in order to be successful in life). This provided them with credentials of honest behavior, since strong disagreement to such statements is expected. Afterwards they wrote a story about a situation where they recently told the truth even so this was difficult. Remembering a situation of honest behavior was expected to increase moral credits. In treatment lie, participants indicated their agreement to statements that were more ambiguous (e.g. you have to lie sometimes in order to be successful in life). The ambiguity introduced by the word "sometimes" made strict honesty harder to justify and resulted in more agreement. Consequently, no credentials of honesty were acquired. Afterwards participants wrote about a situation where they recently told a lie. This was expected to decrease their moral credits as they thought about a situation where they personally behaved immoral as they told a lie. A third treatment had neutrally framed tasks that required participants to indicate agreement and write a story about topics that were relevant in the news during the time of the experiment. This provided a baseline treatment against which the two manipulation treatments could be tested. This setup resulted in three treatments with (1) no change of credentials and credits (baseline), (2) the provision of credentials for honesty and an increase of moral credits (treatment truth) and (3) no provision of credentials and a decrease of moral credits (treatment lie). The results show that the manipulation of credentials and credits worked as expected, but still I was unable to identify evidence in favor of licensing or cleansing. Rates of truth-telling and lying remained stable over all three treatments and did not differ significantly. Yet, as the literature on lying suggests, not every lie may necessarily decrease moral credits. As participants were free to write about any lie that came to their mind, it is possible that some participants wrote about lies which were not morally condemnable as they helped or protected others. Therefore, I employ a coding procedure to identify such differences in lies (and possible truths as well) and to obtain a better understanding about participants' behavior.

For this coding process an exact typed transcript of each hand-written story was produced. All stories were then transferred to an online platform¹⁸ to allow for fast and easy coding. The coders were five student assistants who were paid for the coding procedure. They were blind to the hypotheses of this study. The coders had to access the webpage where they received some general information and instructions about how the coding process would work. Each story was presented on an individual page followed by eight questions. These questions were developed to quantify the stories with regard to certain dimensions. As the only additional information, participants' gender was stated at the top of

¹⁸ www.socisurvey.de

each page since in many cases this reduced ambiguity about who the participant was talking about (e.g. "my friend" could either indicate a good friend or indicate a relationship). The stories and questions were presented to each of the five coders individually and coding could be paused and resumed whenever necessary.

To quantify the stories, relevant dimensions were identified that needed to be assessed by the coders. Based on the literature on lying (DePaulo and Bell 1996, DePaulo and Kashy 1996, DePaulo et al. 1996, DePaulo and Kashy 1998, DePaulo 2004, DePaulo et al. 2004, Erat and Gneezy 2011) and the taxonomies presented there, I developed a set of eight questions which would help in obtaining additional information about the intent of the truth-teller/liar and possibly identify how moral credits had changed in unexpected ways. Each individual question assessed one of the following categories: recipient of truth/lie (Q1), subject of truth/lie (Q2), consequences for participant himself (Q3), consequences for another person (Q4), hypothetical consequences of alternative behavior (Q5), level of morality (Q6), feelings of having done something good (Q7) and feelings of pride/shame (Q8)¹⁹. Coders assessed the questions on a 7-point scale in case of question 1 and on 5-point scales for questions 2-8. Table 1 summarizes the questions and answers used for the coding procedure.

Question	Treatment Truth	Treatment Lie	Answers
Q1	The truth was told to...	The lie was told to...	Partner in life; best friend; parents; other members of family; friends; other known person; unknown person
Q2	The subject told the truth about...	The subject lied about...	him/herself; another person (5 point scale)
Q3	For the subject himself the consequences of telling the truth were...	For the subject himself the consequences of the lie were...	very bad; very good (5 point scale)
Q4	For another person the consequences of telling the truth were...	For another person the consequences of the lie were...	very bad; very good (5 point scale)
Q5	If the subject had told a lie instead of telling the truth he would have...	If the subject had told the truth instead of telling a lie he would have...	felt worse; felt better (5 point scale)
Q6	Would it have been better to lie from a moral perspective?	Would it have been better to tell the truth from a moral perspective?	no, not better at all; yes, much better (5 point scale)
Q7	Does the subject perceive telling the truth to have done something good?	Does the subject perceive the lie to have done something good?	no; yes (5 point scale)
Q8	In the described situation the subject all in all is...	In the described situation the subject all in all is...	ashamed; proud (5 point scale)

Table 1: Questions and answers used for the coding procedure (original questions were in German)

¹⁹ Q1, Q2 and Q5 are inspired by DePaulo et al. (1996). Q3 and Q4 are based on Erat and Gneezy (2011).

Overall, 292 students from the University of Passau participated in the original experiment²⁰. For the qualitative analysis in this study only the data from the senders in treatment truth and treatment lie can be used, reducing the sample to a total of 97 observations (48 senders in treatment truth and 49 senders in treatment lie). Of these, 70.1% were female and the mean age was 22.6 years.

4. Inter-Rater Agreement and Descriptive Statistics

As a measure for inter-rater agreement the intra-class correlation coefficient $ICC(3,k)$ ²¹ was calculated and to obtain one single value for each question, individual ratings from the five raters were averaged. This resulted in eight unique values to quantify each story on the given dimensions. The values for inter-rater agreement as well as the descriptive statistics of each question are reported in Table 2.

	Treatment Truth		Treatment Lie	
	<i>ICC(3,k)</i>	<i>Mean Values</i>	<i>ICC(3,k)</i>	<i>Mean Values</i>
Q1	0.98	M=4.67 SD=1.55	0.99	M=4.92 SD=1.57
Q2	0.94	M=2.45 SD=1.57	0.98	M=1.40 SD=1.03
Q3	0.83	M=2.5 SD=0.72	0.80	M=4.21 SD=0.54
Q4	0.87	M=2.89 SD=0.90	0.86	M=2.55 SD= 0.77
Q5	0.69	M=1.85 SD=0.62	0.61	M=2.34 SD=0.56
Q6	0.30	M=1.43 SD=0.36	0.67	M=3.49 SD=0.75
Q7	0.73	M=3.78 SD=0.77	0.77	M=3.03 SD=0.97
Q8	0.74	M=3.5 SD=0.64	0.71	M=2.9 SD=0.64

Table 2: Inter-rater-agreement and mean values in both treatments

Highest agreement among raters is observed for Q1 and Q2 ($ICC \geq 0.94$). This is not surprising as the receiver of the truth/lie (Q1) as well as the subject (Q2) is in most cases unambiguous and therefore easy to assess. The lowest agreement is observed for Q6 ($ICC=0.30$ in treatment truth). Here it was asked if it would have been better to lie (tell the truth) instead of telling the truth (lying) from a moral perspective. The low agreement reveals that this question was difficult to assess and resulted in heterogeneity in raters' answers. This could be the case because of the hypothetical nature of the question which involved raters to guess about participants' feelings if they had acted differently.

²⁰ For the recruitment of participants and providing laboratory resources I kindly thank PAULA – the Passau Experimental Laboratory.

²¹ Ultimately, my goal was to obtain one single value for each question and participant calculated as the average over all raters. Furthermore, raters are fixed and not randomly selected from a larger group of raters. In this case the $ICC(3,k)$ applies (Shrout and Fleiss 1979).

Overall, low agreement is observed for this question only. Raters' agreement for all other questions is sufficiently high. This shows that for the most part raters agreed in their assessments of the individual stories. Values are therefore reliable for the categorization of stories.

Comparing the mean values for treatment truth and treatment lie confirms that the manipulation of treatments worked as expected. The differences between treatments reflect the general assumption that honest behavior is morally desirable whereas lying is morally objectionable. Except for Q1, where no difference between treatments is to be expected, all other values differ significantly between treatment truth and treatment lie (t-tests, $p < 0.05$). The truth is equally told about oneself as about another person (Q2, $M = 2.45$), whereas a lie is told mostly about oneself ($M = 1.40$). Lying results in favorable consequences for the liar (Q3, $M = 4.21$) and neither very good nor very bad consequences for others (Q4, $M = 2.55$), whereas consequences of truth-telling are mixed for the participant ($M = 2.5$) and for others ($M = 2.89$). When telling the truth, participants would not have felt better had they lied instead (Q5, $M = 1.85$), compared to liars, who would have felt a bit better when telling the truth instead ($M = 2.34$). Truth-tellers perceived themselves to do more good (Q7, $M = 3.78$) than liars ($M = 3.03$) and lastly, telling the truth resulted in more feelings of pride (Q8, $M = 3.5$) compared to telling a lie ($M = 2.9$).

5. Results for Consistent Behavior

Having obtained mean values for each participant's story, these can be matched with the data from the deception game. This lets me identify how individual differences in participants' stories within a treatment result in differences of subsequent truth-telling behavior. I ran individual logit regressions with truth as the dependent variable and the mean value of each individual question as the independent variable (see Appendix A for the full regression results).

In treatment truth no significant coefficients ($p < 0.1$) are observed. This indicates that truth-telling in the deception game cannot be explained by differences between participants' stories. The only coefficient that comes close to significant levels is observed for Q1 ($\beta = -0.28$, $p = 0.184$), where raters indicated whom the truth was told to. An additional one-sided t-test ($p = 0.09$) reveals that a participant who told a lie in the deception game described to tell the truth to a person not as closely related to him ($M = 5.04$) compared to participants sending a truthful message ($M = 4.43$). This result indicates consistent behavior if a truth told to a closely related person is assumed to increase moral credits more than a truth told to a person only distantly related. DePaulo and Kashy (1998) found that participants lied less to closely related persons and felt more uncomfortable when doing so. These results give reason to assume that lies told to closely related persons reduce moral credits relatively more and that likewise a truth told to a closely related person increases moral credits relatively more. Therefore, relationship closeness may act as a proxy for moral credits. The closer the relationship, the more moral credits are gained by truth-telling or lost by lying. A truth told to a close family member is expected to increase moral credits to a higher degree than a truth told to a stranger. The results therefore show that participants did not engage in moral licensing. Rather, those participants who had acquired the most moral credits, engaged in truth-telling most often.

In treatment lie, the only coefficient significantly different from zero is observed for Q6 ($\beta = -1.06$, $p = 0.035$). This question asked whether it would have been better to tell the truth instead of lying from a moral perspective. The negative coefficient shows that if raters indicated that it would have been much better to tell the truth from a moral perspective, participants lied more often in the subsequent deception game. Thus, if participants wrote about situations that were considered highly immoral by the raters (indicated by the fact that telling the truth would have been much better), they sent a truthful message less often. Or, turning this the other way around, participants who told a lie which was rated

to be on the same level of morality as telling the truth, subsequently told the truth more often. I argue that this provides evidence for consistent behavior rather than for moral cleansing, since highly immoral behavior is followed by a lie in the deception game, whereas behavior that is regarded as morally acceptable (even if it resulted from telling a lie) is followed by the truth. Looking at marginal effects shows that a story where raters indicated that telling the truth would not have been better from a moral perspective (indicated by M=1 for Q6) resulted in 98.8% of truth-telling in the deception game. A highly immoral story on the other hand, where telling the truth was rated to be much better instead of lying (M=5 for Q6), resulted in only 29.4% of truth-telling²².

Coefficients for the other questions in treatment lie fail to meet standard levels of significance ($p < 0.1$), nonetheless looking at the signs provides no reason to bring the previous results to question. Q7 asked whether the participant perceived his behavior as having done something good. The positive coefficient ($\beta = 0.31$, $p = 0.34$) indicates that the more a participant perceived his lie as having done something good, the more truth is told in the deception game. A similar observation is obtained for Q8. An increase in pride for the lie told leads to an increase in truth-telling ($\beta = 0.54$, $p = 0.273$). Q5 asked how the participant would have felt had he told the truth instead of a lie. Higher values of Q5 indicate that the participant would have felt better with alternative behavior, therefore feeling bad with his current behavior. This indicates that in line with the result from Q6, participants who told a lie that did something good or had good feelings or even feelings of pride for telling this lie, all told the truth more often in the deception game.

Q3 and Q4 provide additional evidence of participants' consistent behavior. Erat and Gneezy's (2011) taxonomy can be used to categorize stories according to the consequences for the participant and for another person²³. With this taxonomy, four distinct categories of truth-tellers and liars can be distinguished²⁴. I define the category *Pareto* as resulting in positive consequences for the participant as well as for another person. *Altruistic* includes observations where consequences for another person were positive, but consequences for the participant were negative. *Selfish* includes stories where consequences are positive for the participant but negative for another person. Lastly, when consequences are negative for the participant as well as for another person stories are categorized as *Spiteful*. All four categories can result either after telling the truth (treatment truth) or after telling a lie (treatment lie). The scatter plot in Figure 1 shows the distribution of stories within these four categories.

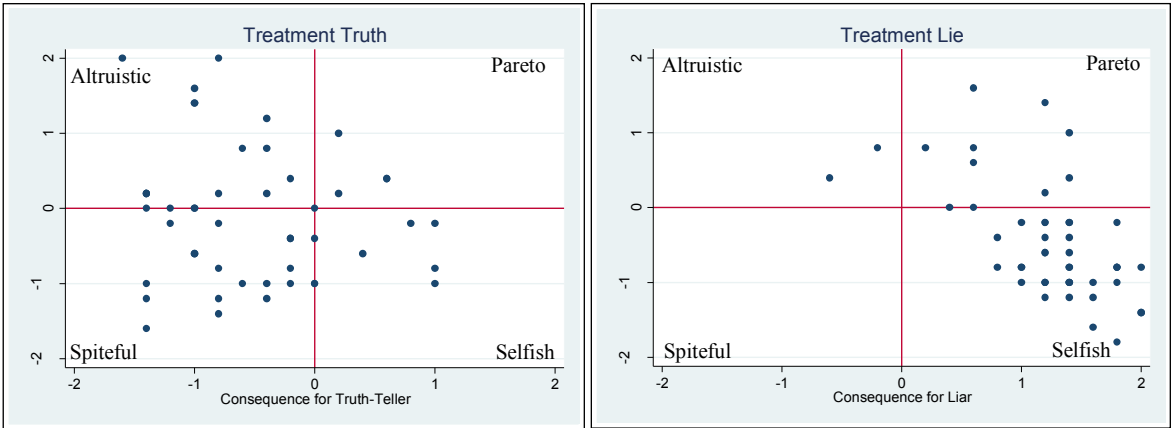


Figure 1: Categories of truths/lies based on Erat and Gneezy (2011)

²² This result is robust to additionally controlling for receiver (Q1) and subject (Q2) of the lie.
²³ As in my experiment participants' did not describe any monetary payoffs, I substituted "payoff" with the more general term "consequence" for the coding procedure.
²⁴ Originally, this taxonomy was developed for lies, but I use it to categorize truth-telling likewise.

As is expected, the patterns differ distinctively between treatments. In treatment truth, most stories are coded as spiteful (35%), followed by altruistic (29%), selfish (10%) and lastly pareto (8%). In treatment lie, most lies are rated as selfish (76%), followed by pareto (16%) and altruistic (4%). Spiteful lies are not observed²⁵. Matching these categories with behavior in the subsequent deception game reveals that truth-telling differs distinctly between categories. Figure 2 shows how often an honest message was sent in the deception game depending on the category of participants' stories.

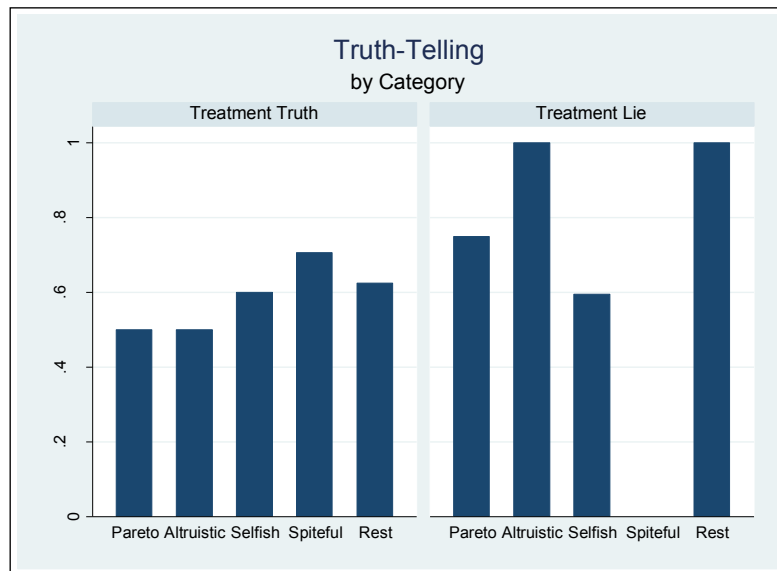


Figure 2: Amount of truth- by category of truth/lie in the hand-written story

In treatment truth, truth-telling amounts to 50% if a story was rated as either pareto or altruistic. Truth-telling increases to 60% if the story described a selfish truth and to 71% if it described a spiteful one. In treatment lie on the other hand, truth-telling amounts to 75% after a pareto lie, 100% after an altruistic lie²⁶ and 59% after a selfish lie. This shows that in treatment lie the lowest amount of truth-telling is observed after participants described a selfish lie. The proportion of truth-telling in this category is significantly smaller compared to all other stories (one-tailed z-test, $p=0.066$). Thus, truth-telling decreases after participants described a selfish lie. These are participants who reported a situation where they told a lie that had positive consequences for themselves but negative consequences for another person, i.e. told a lie for personal gain. Consistent with the selfish lie they described, these participants had the highest rate of telling a lie in the deception game as well. Instead of engaging in moral cleansing, these participants acted consistent to what they described in their stories and continued to lie.

In treatment truth, the proportion of truth-telling after a spiteful truth is not significantly larger compared to the other stories, but standard levels of significance are missed only slightly (one-tailed z-test, $p=0.143$). This may be interpreted as additional (weak) evidence for consistent behavior. Telling a spiteful truth obviously required a lot of courage. In order to encourage participants to think about situations where the truth was not the easiest path to take, the task explicitly stated to describe a situation where telling the truth was difficult. This difficulty is reflected in the negative consequences for the truth-teller as indicated by Q3. If furthermore consequences for another person are negative, it was probably very difficult to tell this truth in the given circumstances. Choosing this option therefore should result in a large increase of moral credits, since the morally right path was taken at high costs.

²⁵ Percentages do not add to 100% because some stories (18% in treatment truth and 4% in treatment lie) could not be uniquely attributed to a single category. These are the points lying directly on the y- or x-axis in Figure 1.

²⁶ This result has to be interpreted carefully, as the number of observations for altruistic lies is very small ($n=2$)

Participants describing such situations are those with the highest rates of truth-telling in the deception game. Even after having described the story and being aware of their large amount of moral credits, participants did not engage in licensed behavior and felt free to lie, but instead continued to tell the truth.

In summary, these results provide no support for licensed or cleansing behavior but rather suggest that participants acted consistent to what they described in their stories. Of course, participants did not literally act consistent. First, they described a lie, afterwards they told the truth. But this simple black and white view of truth-telling and lying falls short of identifying differences between different kinds of truths and lies and how these impact the moral balance. By performing a qualitative analysis, it was possible to identify these differences. A considerable amount of participants (20%) had no selfish motives in mind when telling a lie, but instead told lies that had positive consequences for other persons. Telling such a lie questions the initial assumption that dishonesty decreases moral credits. Obviously those participants told a lie, but their ultimate goal was to help or protect others. It seems questionable to argue that such behavior leads to a decrease of moral credits. Differentiating truths and lies more precisely showed a better picture of participants' behavior and suggests that instead of moral self-regulation consistent behavior emerges.

6. Discussion

Possible reasons for the suppression of moral self-regulation can be attributed to the fear of appearing hypocritical. As Effron and Monin (2010) have shown, observers refused to license hypocritical behavior. Similarly, participants themselves might want to behave in a way to avoid the uncomfortable feelings of hypocrisy. While a license frees a person to act inconsistently, the forces of hypocrisy constrain one to consistent behavior. A licensing effect of a good deed can be suppressed if feelings of hypocrisy arise. Miller and Effron (2010) name three conditions that make hypocrisy especially likely. Firstly, hypocrisy arises if participants' good behavior, that is supposed to create the license, reflects deeply held moral values. If this is the case, a deviation to opposite behavior is less likely even if a license to do so may be present. Honesty has to be considered as such a basic moral value. Participants therefore felt reluctant to lie because with writing a story about truth-telling they were reminded about the moral importance of honesty. Secondly, hypocrisy arises if the good and bad deed are within the same domain. If a person states to be honest, sincere, and upright and later makes a sexist remark that may be considered dismissive and inappropriate but not hypocritical. If on the other hand the same person tells a lie he will be considered a hypocrite. In my experiment the good deed, as well as the transgression, were in the domain of honesty, therefore making hypocrisy especially likely. Thirdly, the nature of the transgression plays an important role. In Monin and Miller's (2001) experiment the transgression was to voice a sexist or racist opinion by favoring a male or white person in hypothetical scenarios. Sachdeva et al. (2009) let participants chose to give money to a charity or abstain from it. But charitable giving as well as politically incorrect statements are acts that need to be seen in context and are often hard to endorse or denounce on their own. They are ambiguous acts and are rated depending on context. The transgression in my experiment was sending a lie in the deception game. Consequences of this behavior are rather straightforward and not open to much interpretation. Sending a lie aims at enriching oneself at the cost of another person, since the gain of the sender is the loss of the receiver. In this situation hypocrisy arises more easily. Since all three conditions apply to my experiment it seems reasonable to assume that participants had feelings of hypocrisy to some extent. Even so they were anonymous during the experiment and no one could observe their actions, these feelings might have led them to consistent behavior as they did not want to appear hypocritical, not even to themselves.

Another motivation to behave consistent may be the desire to maintain the positive self-concept as an honest individual. According to the theory of self-concept maintenance (Mazar et al. 2008) people are dishonest, as long as they can uphold their honest self-concept without being forced to update it. Categorization and attention to standards are identified to be the two mechanisms influencing the process of updating the self-concept. A dishonest action that can be categorized in more compatible terms (e.g. "I did not steal the pen, I merely borrowed it.") does not force an update of the self-concept and lets people still view themselves as honest individuals. Attention to the own moral standards on the other hand enforces a faster update of the self-concept. When participants are mindful of moral standards there is less room to elude these standards. Thus, if made aware of standards, people are more likely to stick to these standards as well. The possibility to categorize dishonesty in my experiment was low. There is little to no possibility to reinterpret a lie sent to the receiver in the deception game. Therefore, sending a lie forces participants to update their self-concept with this act of dishonesty. Additionally, especially in treatment truth, standards of honesty are made salient for participants. In task 1, participants were asked to agree to statements about honesty, where one person gives advice to another person. In this situation they revealed their general attitude towards honesty and voiced their opinion about how others should behave. This procedure made standards of honesty more salient. A high salience of standards and the missing possibility to categorize dishonesty made participants especially aware of possible dishonesty in the deception game. If they wanted to send a lie they would have been forced to update this self-concept and openly acknowledge their dishonest behavior. As the costs of updating may have been higher than the gains from lying, participants decided to act honest instead.

7. Conclusion

In this study, the results from Nagel (2014) are discussed and the analysis is enriched by additional data from participants' hand written stories. Instead of only differentiating by treatments and processing all stories about truth-telling and lying in an identical fashion, the approach presented in this study looks at individual differences within treatments. The experimental data suggested that truth-telling can lead to different amounts of moral credits and even lying may lead to an increase in moral credits in some circumstances. Based on the relevant literature I developed and employed a coding procedure which made it possible to categorize stories on multiple dimensions. Results from this coding procedure confirm the evidence found in the previous study. Looking at the regression results does not support the hypothesis that participants engaged in moral self-regulation and balanced truth-telling with lying or the other way around. Results rather suggest consistent behavior, if truth-telling and lying are analyzed more carefully and not only at face value.

In treatment truth, looking at the person the truth was told to supports the hypothesis of consistent behavior. Participants telling the truth to a closely related person and in consequence having increased their moral balance relatively more choose a truthful message more often. Instead of counter-balancing their good behavior with lying, those participants engaged in honest behavior even though moral credits for possible dishonesty are present. Consistent behavior is also observed in treatment lie. Telling a lie is immoral in most cases, but some lies are morally permissible or even considered necessary from a moral perspective. If participants wrote about lies that were rated to be morally acceptable, lies that were considered to do something good, or lies that result in feelings of pride, truth-telling increased in the deception game. All these instances relate to lies that are not expected to have a negative impact on the moral balance but rather are told to help or protect others, possibly at costs for the liar himself. Furthermore, the highest rate of lying was observed after participants described a selfish lie. A selfish lie resulted in a positive consequence for the liar and a negative consequence for another person. Consistent to describing such a lie, participants engaged in a selfish

lie in the deception game. There, telling a lie is expected to increase the sender's payoff at the cost of the receiver's payoff. As possible reasons for the failure of moral self-regulation the fear of appearing hypocritical as well as the desire to maintain an honest self-concept could be identified. Describing honest behavior, but afterwards engaging in dishonesty, may have induced feelings of hypocrisy and in consequence self-contempt. These feelings possibly outbalanced the moral license acquired from honest behavior. Furthermore, it can be argued that the experimental tasks made participants aware of the moral standard of honest behavior. Subsequently engaging in dishonesty would have contradicted these standards and threatened the self-concept. Instead of taking this burden, participants complied to standards and acted (consistently) honest.

To gain more insights into moral self-regulation in the domain of honesty, future research could take a more careful look at how truth-telling and lying affects the moral balance. Motivation for dishonesty can have multiple reasons, the obvious one being selfishness. But people tell lies out of altruistic motives as well. Looking at the underlying motivation of dishonesty and how this motivation changes moral credits could offer further insights for moral self-regulation of truth-telling and lying.

Appendix A

Regression Results for Treatment Truth

truth	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Q1	-0.28 (0.21)							
Q2		0.16 (0.20)						
Q3			-0.12 (0.41)					
Q4				-0.12 (0.33)				
Q5					-0.33 (0.48)			
Q6						0.34 (0.85)		
Q7							-0.12 (0.39)	
Q8								-0.09 (0.47)
Constant	1.77 (1.08)	0.04 (0.55)	0.72 (1.08)	0.76 (1.01)	1.03 (0.94)	-0.06 (1.25)	0.88 (1.52)	0.74 (1.66)
Observations	48	48	48	48	48	48	48	48
Pseudo R-squared	0.030	0.010	0.001	0.002	0.007	0.002	0.002	0.000

Standard errors in parentheses
 *** p<0.01, ** p<0.05, * p<0.1

Regression Results for Treatment Lie

truth	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Q1	-0.01 (0.19)							
Q2		0.05 (0.30)						
Q3			-1.07 (0.70)					
Q4				0.30 (0.41)				
Q5					-0.26 (0.54)			
Q6						-1.06** (0.50)		
Q7							0.31 (0.32)	
Q8								0.54 (0.50)
Constant	0.69 (1.00)	0.57 (0.51)	5.18* (3.03)	-0.13 (1.07)	1.25 (1.32)	4.43** (1.86)	-0.28 (0.99)	-0.93 (1.44)
Observations	49	49	49	49	49	49	49	49
Pseudo R-squared	0.000	0.000	0.044	0.009	0.004	0.085	0.015	0.020

Standard errors in parentheses
 *** p<0.01, ** p<0.05, * p<0.1

References

- Aquino, K., and Reed II, A. (2002): *The self-importance of moral identity*, Journal of Personality and Social Psychology, 83(6), 1423-1440.
- Bandura, A. (1991): *Social cognitive theory of moral thought and action*. In W. M. Kurtines and J. L. Gewirtz (Eds.), Handbook of moral behavior and development 1, 45-103, Hillsdale, NJ: Erlbaum.
- Barden, J., Rucker, D. D., and Petty, R. E. (2005): “*Saying one thing and doing another*”: Examining the impact of event order on hypocrisy judgments of others. Personality and Social Psychology Bulletin 31, 1463-1474.
- Batson, C. D., Kobrynowicz, D., Dinnerstein, J. L., Kampf, H. C., and Wilson, D. W. (1997): *In a very different voice: Unmasking moral hypocrisy*. Journal of Personality and Social Psychology 72, 1335–1348.
- Batson, C. D., Thompson, E. R., Seufferling, G., Whitney, H., and Strongman, J. (1999): *Moral hypocrisy: Appearing moral to oneself without being so*. Journal of Personality and Social Psychology 77, 525–537.
- Batson, C. D., Thompson, E. R., and Chen, H. (2002): *Moral hypocrisy: Addressing some alternatives*. Journal of Personality and Social Psychology 83, 330–339.
- Batson, C. D., Lishner, D. A., Carpenter, A., Dulin, L., Harjusola-Webb, S., and Stocks, E. L. (2003): “. . . As you would have them do unto you”: Does imagining yourself in the other’s place stimulate moral action?. Personality and Social Psychology Bulletin 29, 1190-1201.
- Batson, C. D., Collins, E. C., and Powell, A. A. (2006): *Doing business after the fall: The virtue of moral hypocrisy*. Journal of Business Ethics 66, 321-335.
- Brañas-Garza, P., Bucheli, M., Espinosa, M. P., and García-Muñoz, T. (2013): *Moral Cleansing and Moral Licenses: Experimental Evidence*. Economics and Philosophy, Cambridge University Press 29(2), 199-212.
- DePaulo, B. M., and Bell K. L. (1996): *Truth and investment: lies are told to those who care*. Journal of Personality and Social Psychology 71(4), 703-716.
- DePaulo, B. M. and Kashy, D. A. (1996): Who lies?, Journal of Personality and Social Psychology 70(5), 1037-1051.
- DePaulo, B. M., Kashy, D. A., Kirkendol, S. E., Wyer, M. M., and Epstein, J. A. (1996): *Lying in everyday life*. Journal of Personality and Social Psychology 70(5), 979-995.
- DePaulo, B. M., and Kashy, D. A. (1998): *Everyday lies in close and casual relationships*. Journal of Personality and Social Psychology 74, 63-79.
- DePaulo, B. M. (2004): *The Many Faces of Lies*. In A. G. Miller (Ed.), The Social Psychology of Good and Evil, New York: Guilford Press. Chapter 12, 303-326.
- DePaulo, B. M., Ansfield, M. E., Kirkendol, S. E., and Boden, J. M. (2004): *Serious lies*. Basic and Applied Social Psychology 26, 147-167.

- Effron, D. A., Cameron, J. S., and Monin, B. (2009): *Endorsing Obama licenses favoring Whites*. Journal of Experimental Social Psychology 45, 590-593.
- Effron, D. A. and Monin, B. (2010): *Letting People Off The Hook: When Do Good Deeds Excuse Transgressions?*. Personality and Social Psychology Bulletin 36(12), 1618-1634.
- Erat, S. and Gneezy, U. (2011): *White Lies*. Management Science 58(4), 723-733, available at <http://dx.doi.org/10.1287/mnsc.1110.1449>
- Fishbach, A., and Dhar, R. (2005): *Goals as excuses or guides: The liberating effect of perceived goal progress on choice*. Journal of Consumer Research 32, 370–377.
- Fischbacher, U., and Utikal, V. (2013): *Disadvantageous lies in individual decisions*. Journal of Economic Behavior & Organization 85, 108–111
- Gneezy U (2005): *Deception: The role of consequences*. The American Economic Review 95(1), 384-394.
- Hao, L. and Houser, D. (2010): *Honest Lies*. GMU Working Paper in Economics No. 11-16, available at: <http://ssrn.com/abstract=1801546> or <http://dx.doi.org/10.2139/ssrn.1801546>
- Jordan, J., Mullen, E., and Murnighan, J. K. (2011): *Striving for the Moral Self: The Effects of Recalling Past Moral Actions on Future Moral Behavior*. Personality and Social Psychology Bulletin 37(5), 701-713.
- Khan, U., and Dhar, R. (2006): *Licensing effect in consumer choice*. Journal of Marketing Research 43(2), 259-266.
- Mazar, N., Amir, O. and Ariely, D. (2008): *The Dishonesty of Honest People: A Theory of Self-Concept Maintenance*. Journal of Marketing Research 45(6), 633-644, available at: <http://ssrn.com/abstract=979648>
- Mazar, N., and Zhong, C.-B. (2010): *Do Green Products Make Us Better People?* Psychological Science 21(4), 494-498.
- Merritt, A. C., Effron, D. A., and Monin, B. (2010): *Moral Self-licensing: When being good frees us to be bad*. Social and Personality Psychology Compass 4, 344-357.
- Miller, D. T., and Effron, D. A. (2010): *Psychological license: When it is needed and how it functions*. In M. P. Zanna & J. M. Olson (Eds.), *Advances in experimental social psychology* 43, 117-158, San Diego, CA: Academic Press/Elsevier.
- Monin, B., and Miller, D. T. (2001): *Moral credentials and the expression of prejudice*. Journal of Personality and Social Psychology 81, 33-43.
- Monin, B., and Merritt, A. (2012): *Moral hypocrisy, moral inconsistency, and the struggle for moral integrity*. In Mikulincer, M. (Ed); Shaver, P. R. (Ed). *The social psychology of morality: Exploring the causes of good and evil*. Herzliya series on personality and social psychology, 167-184.
- Nagel, V. (2014): *Licensed to Lie? Moral Self-Regulation of Truth-Telling and Lying*. Working Paper, University of Passau

Nisan, M. (1991): *The moral balance model: Theory and research extending our understanding of moral choice and deviation*. In W. M. Kurtines & J. L. Gewirtz (Eds.), *Handbook of moral behavior and development*, 213-249, Hillsdale, NJ: Erlbaum.

Sachdeva, S., Iliev, R., and Medin, D. L. (2009): *Sinning saints and saintly sinners: The paradox of moral self-regulation*. *Psychological Science* 20, 523-528.

Shrout, P. E. and Fleiss, J. L. (1979): *Intraclass correlations: uses in assessing rater reliability*. *Psychological Bulletin* 86, 420-3428.

Stone, J. and Fernandez, N. C. (2008): *To Practice What We Preach: The Use of Hypocrisy and Cognitive Dissonance to Motivate Behavior Change*. *Social and Personality Psychology Compass* 2, 1024-1051.

Sutter, M. (2009): *Deception through telling the truth?! Experimental evidence from individuals and teams*. *Economic Journal* 119(534), 47-60.

Walkowitz, G., Lönnqvist, J.-E., and Irlenbusch, B. (2013): *Moral hypocrisy: Self-deception or impression management?*, Beiträge zur Jahrestagung des Vereins für Socialpolitik 2013: Wettbewerbspolitik und Regulierung in einer globalen Wirtschaftsordnung - Session: Social Values and Political Attitudes, No. A03-V2

Zhong, C.-B., Liljenquist, K. A., and Cain, D. M. (2009): *Moral Self-Regulation: Licensing & Compensation*. In D. De Cremer (Ed.), *Psychological Perspectives on Ethical Behavior and Decision Making*, 75-89, Charlotte, NC: Information Age Publishing.

What Motivates Impression Management? Between Good Business and Moral Cleansing

Johann Graf Lambsdorff


Volker Nagel 

Abstract

Is there a financial return to a favorable impression? We run an experiment to investigate whether looters can reduce punishment by manipulating their record. They decide between a high and a low probability of looting money from a collection box designated to a charity. An observer decides on the level of altruistic punishment, noticing the looters past record and whether looting took place. We find that looters are willing to pay a fee for dressing up their record. But observers do not reduce punishment in response and looters do not even expect them to do so. Impression management is thus motivated intrinsically and not financially. From a purely monetary perspective investments on impression management are illspent.

JEL Classification: K42, D03, C91

Keywords: Self-Image, Corporate Social Responsibility, Virtue Ethics, Altruistic Punishment.

 Johann Graf Lambsdorff is full professor in economic theory at the University of Passau, Germany. Contact address: Innstrasse 27, D-94032 Passau, jlambsd@uni-passau.de. Volker Nagel is doctoral scholar at the University of Passau, Germany. The authors are grateful to Manuel Schubert and Marcus Giamattei and to participants of the brownbag economics seminar at the University of Passau, April 16, 2014, for helpful comments.

1. Introduction

Beyond pure profit maximization companies engage, to varying degrees, in activities that are not *prima facie* related to their advantage. They show interest in corporate social responsibility (CSR), cooperate with others to advance public goods, preserve the commons, or abstain from paying bribes. Such engagement is often actively marketed, sometimes even exaggerated or falsely claimed. Portraying a related corporate commitment has the potential of advancing also profits. But do the benefits of such an impression management outbalance the costs?

These questions are empirically difficult to assess. Evidence is sometimes reported of a positive correlation between public measures of CSR and financial performance. But empirically the dividing line between CSR and impression management is difficult to draw. Is self-reported CSR truthful or part of impression management? Positive correlations between impressions gathered on CSR and financial performance might be driven by reverse causality: Good financial results provide more room for a company to increase CSR activities. At the same time we are not aware of robust evidence from the field. Requests for experiments with increased control of variables and causation have thus been raised (Schmitz and Schrader 2013). Given this lack of evidence we design a laboratory experiment. Participants can engage in CSR (by avoiding looting), can engage in impression management (dress up their past performance) and are confronted with counterparts who can engage in altruistic punishment. This allows us to cover new ground in determining the financial rewards to impression management and address a variety of questions: Who engages in impression management, the ones who engage in CSR or rather the self-seeking firms? How would companies react to public punishment that may follow a scandal? Does impression management increase revenues and profits?

There exists a widespread, but not uncontested, viewpoint that corporate social responsibility is only justified if it increases profits. It should align with shareholders' interests and provide strategic advantages vis-à-vis consumers, employees, investors and regulators (Friedman 1970; Gallagher 2005; Amalric and Hauser 2005; Hooghiemstra 2000). Such advantages are diverse. CSR might signal high profits, superior quality, or favorable working environments. It might dampen public protests targeted at environmental misconduct, ease governmental regulation, or improve access to capital (Schmitz and Schrader 2013). But CSR may also go along with an agency conflict. Managers may over-invest in CSR while seeking private non-monetary benefits (Barnea and Rubin 2006, Cheng et al. 2014).

Debate arose on whether companies should market their good deeds and impress others. Marketing would help reap the strategic advantages, turning CSR into good business (Lantos 2001). But critics have warned of "managerial capture". Impression management, the desire to be seen favorably, entails that engagement in CSR is exploited for strategic purposes, companies collect and disseminate only favorable information, thus painting a biased image of their engagement. This bias might be anticipated by target groups who adjust their notion of CSR accordingly. A favorable impression may no longer be seen as a truthful commitment to act ethically (Valor 2005; Owen et al. 2000). Whether to engage in CSR and whether to engage in impression management can thus be seen as two distinct decisions. To illustrate these considerations, take the three following examples.

Foxconn is a supplier for Apple that assembles the iPhone and the iPad and was reported to force workers to work more than 60 hours a week. Workplace safety and living conditions at factory-owned dorms were criticized as being inhuman (Forbes: Apple's Supplier Labor Practices In China Scrutinized After Foxconn, Pegatron Reviews, 12 December 2013; New York Times: In China, Human Costs Are Built Into an iPad, 25 January 2012). This is supposed to have caused the suicide of more than a dozen workers between 2010 and 2012. In 2012 a petition was signed by hundreds of

thousands of consumers, requesting Apple to improve working conditions in China. Apple's approach since 2007 has been to request suppliers to abandon such practices. This went along with requirements to sign up to codes of conduct, carry out repeated audits and publish responsibility reports. But given the persistent violations of human rights, critics have argued that Apple lacks the will to pay the price for rigorously improved working conditions. The core interest of Apple, it is argued, is to maintain the status quo while avoiding the associated embarrassment. In our terms, Apple is criticized for a narrow focus on impression management. But how would consumers react if no such impression management would take place?

Mining is a business that involves close ties to local politicians and substantial risks of environmental damage. Dilemmas between profit and corporate social responsibility are thus standard to this sector. In West Papua, Indonesia, a joint-venture by Freeport and Rio Tinto mines the world's third-largest copper deposit and the world's biggest gold mine (New York Times: *Below a Mountain of Wealth, a River of Waste*, 27 December 2005) faced such a dilemma. Since 1997, the Indonesian Environment Ministry repeatedly warned the company that Freeport was breaching environmental laws and that its waste killed all life in the rivers. Other allegations link human rights abuses and the killing of an estimated 160 people by the Indonesian military to the mining interests. But contrary to the case of Apple there appears to be little effort to paint more favorable impressions. No corporate excuses can be spotted, hardly any improvements are announced and documentation on audits remain scarce or inaccessible. Instead, the company supported the local Indonesian military with multimillion dollars in order to corroborate control and suppression. But is it a wise decision to devote so few resources to the company's impression? The answer will largely depend on whether companies with an unfavorable impression are punished by stakeholders.

Such a link between a company's appearance and punishment can be observed for the case of bribery. Companies run into conflicts of interest when profitable contracts can only be achieved by help of bribes to public officials and politicians. The last two decades have seen many companies take a strict approach towards abstaining from the payment of bribes, which parallels the increased risks, fines and penalties imposed by the SEC in the USA (Shearman and Sterling 2013; Lambsdorff 2013). Recent legislation discussed how companies should be penalized for wrongdoing. In the UK legislation holds a company C guilty of an offence if an associated person bribes another person unless the company can "prove that C had in place adequate procedures designed to prevent persons associated with C from undertaking such conduct" (UK Bribery Act 2010). But criticism has been raised that proved procedures are a type of impression management and that they may be ineffective. Laufer (2006: 99-129) warns against the type of leniency implemented by the UK Bribery Act. This type of leniency reduces the repressive pressure of the legal code and induces firms to invest in potentially useless compliance systems rather than in eliminating actual misconduct. Such compliance systems represent methods of impression management rather than organizational methods that reduce actual bribery. Is it thus a wise decision to grant leniency to companies that impress judges with their procedures?

This study assesses the causes and returns to impression management. For this purpose we designed an experiment where observers (just like customers) were willing to engage in altruistic punishment and made the level of punishment dependent on observed infractions of a potential perpetrator (a company) but also the intentions (and virtuousness) they assigned to the perpetrator. This resembles the behavior of customers who might be willing to boycott a company that lacks CSR. A perpetrator decides between a high and a low probability of looting money from a collection box designated to a charity. Observers are endowed with resources, observe the looters' past record and whether looting occurred in the current round. They can allocate some of their resources to punishment. Depending on the treatment of the experiment looters can manipulate their past record for a fee.

We observe, first, that punishment motivates perpetrators to reduce looting unless they can manipulate their record. In the latter case, perpetrators continue with their preferred level of looting and spend more for impression management. Second, the well performing companies engage less in impression management, those who choose the low probability of looting. Perpetrators who choose the high probability of looting devote more resources to impression management. Third, in line with this finding, observers were able to correctly infer the chosen probability (in a sense, the underlying intention). Impression management is thus not capable of completely diffusing the informativeness of a perpetrators past record. Fourth, observers substantially punish looting when observing a history that cannot be manipulated. But they care little for a history that is up to manipulation. This brings about our fifth finding: money spent for impression management does not have a financial return. Perpetrators do not even expect lower punishment after manipulating their history. This suggests that the advantage from impression management is largely intrinsic. An improved self-image is cultivated for its own sake.

Section 2 reviews the relevant experimental literature and explains why a favorable self-image may have an intrinsic value. Section 3 explains our experimental design. Section 4 derives our hypotheses. We report the experimental proceedings and our findings in section 5 and section 6. We discuss why our conclusions might be relevant for corporate policy in the concluding section 7.

2. Literature Review

Recent experimental studies suggest that impression managements yields little return. Observers tend to give little credit to a perpetrator's history. One such piece of evidence relates to apologies, which can be seen as a type of impression management. Ho (2012) investigates the role of apologies as signals for improving one's trustworthiness. He employs a trust game where trustors can transfer up to 10 tokens to a trustee who receives triple this amount and can reciprocate by purchasing lottery tickets, each increasing the trustor's chance to win 20 tokens by 5%. Trustors are informed only about the winning of the lottery but not about how many tickets the trustee purchased. In case of not winning the lottery, the trustee can send a costly signal "I am sorry". These signals improve future transfers to the trustee. But for 1 token spent for an apology future transfers increase only by 0.7.²⁷ Expenses for improved impressions thus fall short of their return.

Tiedens (2001) investigates the returns to apologies in politics. Former U.S. president Bill Clinton was seen to have lied with respect to his relationship with Monica Lewinsky. In an experiment, videos where shown to participants on Clinton's grand jury testimony, one where Clinton appears to apologize regarding the Lewinsky affair, and one where he is angry about the cause of the investigation. Participants were then asked whether Clinton should maintain his status and power. This was approved to significantly higher degrees by those who saw the angry Clinton. An apology thus comes with a cost of losing status, while having a negative return to a politician's future career.

²⁷ Ho (2012: 153) writes with respect to agents (trustees): "Regressing the continuation value on the number of tokens an agent spends on an apology (including period fixed effects) shows that each token spent on an apology yields the agent 1.37 tokens (standard error = 0.35) in future profit; this result is robust to adding controls for prior beliefs and agent fixed effects." This coefficient of 1.37, however, is biased upward because trustees that return many tokens have a higher tendency to apologize. The actual yield from a token spent for an apology is obtained when controlling for agents (trustees) fixed effects. In personal correspondence Ben Ho reported the outcome from such a controlled regression to be 0.692 (standard error = 0.33). Returns to investments into apologies are thus below 1.

Psychological research on impression management has focused on human (rather than corporate) behavior. Humans have a tendency to portray themselves in a favorable light and devote resources in order to convince others of their job-related capacities, their willingness to devote effort, the goodness of their intentions, or the virtuousness of their character and the flawlessness of their lifestyle. Impression management is defined as the situational manipulation of the image portrayed towards others with the goal of getting approval (Snyder 1977; Schlenker 1980; Bromley 1993). Impression management has been investigated lately in relation to job interviews, where applicants involve tactics to distort their attributes while interviewers seek to anticipate these tactics in order to maintain a good person-job fit (Ingold et al. 2014; Kristof-Brown et al. 2002). Another line of research has set out to determine the cognitive costs. Vohs et al. (2005) find that impression management can deplete cognitive capacities by requiring subjects to override their dispositional tendencies.

This early research on impression management implied that there should be a return to such an investment, the approval obtained by others. But it was widely recognized that there can be an intrinsic value to a favorable self-image. In particular, more recent research recognized that humans cultivate a stable perception of themselves, which assures not only others but even themselves of the goodness of their deeds and virtues (Swann and Bosson 2010). This type of assurance has its own value. From this perspective a favorable impression might be sought even if it is not expected to bring about a monetary return.

Yamagishi et al. (2009) present evidence that individuals are willing to forgo financial rewards when these are considered to be unfair. They run an altered version of the ultimatum game where a proposer offers a split of money which can be accepted or rejected by the responder. Contrary to a standard ultimatum game however, rejection only reduces the responder's payoff to zero leaving the proposer with the money allocated to him. Furthermore, the rejection of an offer is not communicated to the proposer, eliminating the possibility to express anger towards the proposer and at least symbolically punish him. Even without such symbolic punishment, a substantial rejection rate is observed. This behavior by responders might be explained by their desire to maintain their self-image.

Substantial experimental evidence has been gathered on the conflict between selfish interests and self-image. Batson et al. (1997) let subjects assign a pleasant and a dull task between themselves and another player. Subjects were given the option to delegate the decision to a coin-flip. 10 subjects decided in favor of the coin but of these only 1 reported that the coin assigned the dull task to themselves. The authors conclude (p. 1342): "Apparently, some of those flipping the coin took advantage of this ambiguity to hide self-interest in the guise of morality." In a subsequent publication (Batson et al. 1999) participants played the same game but saw themselves in a mirror as a method for increasing self-awareness. In this case, 5 out of 10 participants who flipped the coin assigned the dull task to themselves. This is suggestive of the idea that participants deceived themselves when tossing the coin. They believe to act morally but can no longer do so when a mirror increases their self-awareness.

One method for aligning self-interest with a favorable self-image is to engage in self-deception. Subjects have been found to carry out immoral acts while failing to update their self-image (Chugh et al. 2005; Benabou and Tirole 2006). They uphold their positive self-image – towards others and towards themselves – by denying responsibility for unfair or immoral outcomes. This denial of responsibility can arise if the link between a decision and the immoral or unfair outcome is less straightforward. Subjects can fabricate excuses and bias downward their own responsibility by shifting it to random procedures such as the tossing of a coin (Batson et al. 1997, Shalvi et al. 2011, Lönnquist et al. 2013), to the ambiguity of the environment (Murnighan et al. 2001; Mazar et al. 2008, Dana et

al. 2006 and 2007, Haisley and Weber 2010) or to other subjects (Gino et al. 2009, Mazar et al. 2008, Paharia et al. 2009, Mazar and Aggarwal 2011, Conrads et al. 2013, Bartling and Fischbacher 2012, Hamman et al. 2010, Coffmann 2011).

It will be difficult for humans to fabricate excuses if acts of self-seeking are blunt and straightforward. But they may retain their desire to wash away their sins by manipulating their history. Experimental evidence that would be supportive of such a motivation has been reported by Zhong and Liljenquist (2006) who observe that subjects that were confronted with a morally objectionable memory were more likely to wash their hands afterwards. Subjects were asked to recall an unethical deed or copy a story of an unethical act. This might impact their desire for moral purity and trigger a corresponding act of cleansing, which was significantly more pronounced as compared to a control group who recalled an ethical deed or copied a story of a selfless act. This type of moral cleansing has been linked to charitable engagement by Sachdeva et al. (2009). They equally prime subjects by recalling either a selfless or an unethical act, the latter group of subjects donating significantly more. Impression management may thus be an act of moral cleansing, restoring the sense of purity vis-à-vis others or oneself.

Humans would thus be willing to spend money for an improved self-image. They may not care that such money does not bring about financial advantages. A favorable self-image has its own intrinsic value and investments would be carried out even if outsiders provide no monetary return. The costs and benefits of impression management are thus indicative of the underlying motivation. If returns are high, investments might be driven by the desire to reap financial benefits. If expected returns are poor, humans (and companies) care for their self-image as an intrinsic value or form biased beliefs about their capacity to influence their impression.

3. Experimental Design and Treatments

Subjects are willing to exercise altruistic punishment. Substantial experimental evidence has been gathered, revealing that humans devote resources to punish the misdeeds of perpetrators even if their personal interests are largely unaffected by a transgression (Fehr and Gächter 2002; Egas and Riedl 2008; Turillo et al. 2002; Kahneman et al. 1986; Eckel and Grossman 1996). This behavior is in particular related to the willingness to enforce a social norm, for example related to fairness or the protection of public goods. We thus employ these methods to detect how perpetrators interact with observers who might engage in altruistic punishment. This parallels a company that might increase profits by violating human rights and consumers that might boycott the company's products.

A collection box contains 1000 Taler per round, designated to the charity Doctors without Borders. Across 40 rounds a perpetrator P who is initially endowed with 100 Taler chooses between urn A where the probability of looting is 80% (4 red balls who imply looting and 1 white ball) and urn B with the probability amounting to 20% (1 red ball and 4 white balls). In case of looting 100 Taler are transferred from a collection box, increasing P's payoffs to 200. A laboratory session consists of 10 players in the role of P, such that potentially all money in the collection box is looted. P is randomly matched to one of 10 observers O who can devote resources to punishing P. O is initially endowed with 150 Taler and can spend up to 20 Taler for punishment. Each Taler spent for punishment reduces P's payoff by 5 Taler. The reduction generates a loss to both players and the money is not transferred back to the collection box. O observes whether in a current round P looted money, but not the choice of the urn. O may want to punish P and let the size of punishment depend on conjectures about P's intentions or virtues. For this purpose O observes the signals about P's history of looting from the last five periods. Each round ends with P deciding whether to manipulate the last signal, potentially

turning a red ball into a white ball. In case of manipulation only the signals of past behavior are modified, but the incidence of looting is not made undone. We run three different treatments and differentiate by costs for changing the signal. In the baseline treatment a change of signal is not possible, reflecting prohibitively high costs. P has no chance to engage in impression management because each ball drawn enters his history as it is. In the high cost treatment changing a signal costs 20 Taler, in the low cost treatment costs are 10 Taler. Both players are informed about these costs.

One reason why the costs for manipulating ones record exceed the benefits would arise if perpetrators hold biased beliefs on how their past record may impress observers. We run our experiment for 40 rounds so subjects have sufficient time to learn about actual levels of punishment. Beliefs about the returns to a clean record should thus converge towards realized levels. To determine how beliefs about punishment are formed, we asked perpetrators to state their expected punishment every period based on the draw of the ball and their current history. Perpetrators were asked to answer the question for expected punishment while observers decided for actual punishment levels. This makes it easy to compare beliefs and realized values. To incentivize answers, we rewarded perpetrators with a bonus payment of 10 Taler if their expectations matched the realized punishment (+/- 1 Taler). This allows us to detect whether belief formation is indeed rational or whether expected returns to a manipulated record are biased upward.

Likewise, we asked observers to state their expectations about the perpetrator's choice of urn in every period. They were asked to state their beliefs about the perpetrator's pick of urn before they had decided for their punishment and before a new period began. We were especially interested in how observers rate a perpetrator's history and infer the choice of urn from the information a history provides. When remunerating observers for correct expectations we needed to take into account that the draw of the ball indicates the chosen urn with 80% probability correctly. We constructed incentives in such a way, that the expected bonus was the same for both urns regardless of the draw of the ball. After a red ball was drawn, observers got 3 Taler if they expected this draw to result from urn A and 12 Taler if it resulted from urn B. The expected bonus from both options is therefore identical ($0.8 \cdot 3 = 0.2 \cdot 12$). After a white ball was drawn, payoffs were reversed. Thus, we constrained observers to additionally evaluate a perpetrators history in order to form beliefs about his choice of urn.

The experiment was programmed in z-Tree (Fischbacher, 2007) and subjects were recruited via ORSEE (Greiner, 2004). The experiment took place in the classEx laboratory at the University of Passau in January and March 2014. We ran 6 sessions with 20 participants each. This resulted in a total of 120 participants, 20 perpetrators and 20 observers for each of the three treatments. Across all treatments, participants were on average 23 years old and 61% of all participants were female. The average length of a session was 83 minutes with the fastest session being completed after 71 minutes and the longest session taking 100 minutes. To determine participants' payoffs, one period was randomly selected in every session. Participants earned 13.20€ on average (6.50€ min, 17.80€ max). On top of their individual payoffs, participants donated 168€ to Doctors without Borders, because in total 24 white balls were drawn in the payoff relevant periods and each white ball resulted in a donation of 7€.

4. Hypotheses

Our experiment lets us evaluate causes and returns of impression management and provides insights into the rationale of participants. Based on the related literature we derive a total of seven hypotheses that predict participants' behavior. In the course of one experimental period, perpetrators first decide which urn to pick and second if they want to manipulate a red ball if one was drawn. Observers on the

other hand decide on the magnitude of punishment they would like to impose on a perpetrator. But punishment comes with no monetary reward for observers. Instead, observers could be motivated by a desire to express their negative emotions towards looting of the charity box or might try to enforce the social norm that one does not grab charity money to enrich oneself. The purpose of observers' punishment could be to deter perpetrators from picking urn A in future periods and instead encourage them to opt for the lower probability of looting by picking urn B. Observers might potentially obtain the role of a moral authority which is in charge of preventing perpetrators from looting the charity box. We hypothesize that lower returns resulting from observers' punishment will discourage perpetrators from picking urn A. This means that we expect punishment to successfully deter perpetrators from looting with high probability (H1).

Hypothesis 1 (P behavior): Perpetrators choose urn A less often after experiencing punishment.

When observers engage in altruistic punishment, their punishment decision can be based on two observable variables: the draw of the ball and the history. When analyzing observers' punishment we can therefore differentiate between two types of punishment, which we label *outcome-based* and *virtue-based*. Outcome-based punishment takes place when observers base their punishment on the draw of the ball. In this case punishment will be higher if a red ball is drawn. Additional to punishing just the draw of the current ball, observers may look at the history in order to assess a perpetrator's intentions. Encountering an unfavorable history with a high amount of red balls signals repeated misconduct in the past and hints at the intention to loot the charity box for personal gain. If punishment takes a perpetrator's history into account we call this kind of punishment virtue-based. We expect observers to not only punish the draw of a red ball (outcome) but as well punish the number of red balls in the history (virtue). We thus hypothesize that we will observe a positive correlation between the number of red balls in a perpetrator's history and punishment (H2).

Hypothesis 2 (O behavior): Punishment increases with the number of red balls in the observed history.

When observers are presented the two observable variables, we expect them to form rational expectations about a perpetrator's choice of urn and therefore his intentions. From the current draw of the ball and the history observers will try to infer what urn a perpetrator has picked in the current period. Beliefs about the choice of the urn will then determine the punishment. The belief that urn A was picked goes along with the belief about this perpetrator's intentions to loot the charity box. We expect observers to punish such an intention more severely compared to the belief that urn B was picked and the intention to not loot the box. Therefore, if observers believe that urn A was picked a higher punishment is to be expected compared to the belief that urn B was picked (H3).

Hypothesis 3 (O behavior): Punishment increases with the belief that urn A was picked.

For perpetrators, manipulation makes it possible to decrease the number of red balls that is shown in their history. From (H2) it follows that a lower number of red balls will result in less punishment. Less punishment in turn increases a perpetrator's payoff. Therefore, by manipulating his history a perpetrator is able to generate a positive return. But as manipulation is costly to the perpetrator, the question arises if the returns generated by manipulation outbalance the required costs to change a signal. We hypothesize that perpetrators engage in impression management because by doing so they are able to make a profit (H4).

Hypothesis 4 (Impression Management): Manipulation increases perpetrators' profits.

Furthermore, we expect (H4) to be reflected in perpetrators' beliefs as well. When engaging in manipulation, perpetrators do so because they expect a signal change to reduce the punishment they receive. This reduction of punishment occurs because a more favorable history is punished less by observers as stated in (H2). We expect perpetrators to engage in manipulation because they believe this will lead to lower punishment (H5).

Hypothesis 5 (P behavior): Expected punishment decreases in response to manipulated signals.

Additionally, we conjecture that when manipulation is possible in the high cost and low cost treatment, signal change offers a substitute to switching from the bad urn to the good one. Instead of switching urns, perpetrators are now offered the possibility to maintain their initial selection of the urn and simultaneously build a more favorable history. Manipulation makes it possible to modify the history without changing the choice of urn. We assume that after experiencing punishment perpetrators react to the possibility of manipulation by substituting a switch of urns (as H1 stated) with a signal change (H6)

Hypothesis 6 (P behavior): Perpetrators increase manipulation after experiencing punishment, if manipulation is possible.

Our last hypothesis looks at the causes for impression management. Perpetrators will manipulate their histories to paint a more favorable picture of their past behavior. Similar to Zhong and Liljenquist (2006), where moral purity was achieved by washing one's hands, we argue that perpetrators are driven by a similar desire in our experiment. Perpetrators who engage in manipulation try to establish a clean image of themselves. In doing so, the manipulation of a red ball offers decreasing benefits depending on the number of red balls already present in the history. If a history shows 5 white balls, the manipulation of the first red ball offers the highest benefit, as this manipulation makes the difference between a complete clean history and a history with a red ball. These marginal benefits of manipulation decrease, as the number of red balls increases. A perpetrator whose history already shows five red balls, gains relatively little by manipulating the fifth red ball and changing his history into showing 4 red balls and 1 white ball instead. This logic implies that manipulating a signal is more beneficial if the actual history is closer to a clean image (H7). This idea is also described by Wilhelm Busch's wise maxim "*Once your worldly reputation is in tatters, the opinion of others hardly matters*".

Hypothesis 7 (P behavior): Manipulation decreases as the number of red balls in the history increases.

5. Results on the Low Returns to Impression Management

Table 1 shows descriptive results for the three treatments over 35 periods and all perpetrators. Periods 1-5 are excluded, since these periods are required to build up the initial history first. In baseline, urn A (where the probability of looting was 80%) was chosen in 72% of cases. The frequency of red balls amounted to 0.63 and the average history displayed 3.13 red balls. We evaluate the five period history by adding up all red balls in an individual history. We assume that the relative position of a red ball within the history makes no difference and therefore process data on the total number of red balls only. A history can have values ranging from 0 (five white balls) to 5 (five red balls). When we introduce manipulation at high costs, urn A was picked in 66% of cases which lead to a frequency of red balls of

0.60. In the low cost treatment the choice of urn A increased to 79% and with it the frequency of red balls to 0.66. Perpetrators manipulated a considerable amount of signals. Out of all red balls drawn, nearly every 5th red ball (0.18) was manipulated to show a white ball in high cost. In low cost, manipulation increased to 25% which means that every 4th red ball was manipulated on average. The frequencies of red balls drawn as well as the manipulation of balls resulted in an average history of 2.37 red balls in high cost and 2.46 in low cost. For our analysis we restrict manipulation of signals to those cases where red balls were changed into white balls, reflecting behavior targeted at improving a perpetrator's history²⁸. On average, observers spent 4.00 Taler on punishment in treatment baseline. These expenditures decreased to 3.94 Taler in treatment high cost and to 3.53 Taler in treatment low cost.

	<i>Baseline</i>	<i>High Cost</i>	<i>Low Cost</i>
Urn A	0.72	0.66	0.79
Frequency of red balls	0.63	0.60	0.66
Change Signal (over red balls)		0.18	0.25
5 Signal History	3.13	2.37	2.46
Observers' Punishment	4.00	3.94	3.53

Table 1: Descriptive results for periods 6-40

To assess how perpetrators react to punishment and adjust their choice of urn (H1) we ran a regression with choice of urn as the dependent and the aggregate punishment that perpetrators received over the last five periods as an independent variable. The aggregate punishment sums up the amount of Taler that was spent on punishment during the last five periods for each individual perpetrator. The variable ranges from 0 Taler to 100 Taler. Since history has a five period timeframe by design, we decided to use that same interval for punishment received as well²⁹. Furthermore we add the ball and the history from the previous period as controls in the regression since both variables influence the choice of urn in the actual period. The dummy variable *red ball in previous period* indicates if a red ball was drawn in the previous period. *History* denotes the number of red balls that are present in a perpetrator's history before he makes his choice of urn in the current period. Perpetrators might follow simple heuristics regarding the number of red balls in their history and react to deviations from that heuristic by adjusting their choice of urn accordingly. By controlling for ball and history we account for perpetrators' different strategies when picking an urn³⁰. Furthermore, the choice between urn A and urn B could be either driven by a perpetrator's type, e.g. someone who favors looting versus someone who favors altruism, or by the reaction to experimental incentives, e.g. the punishment received, the desire to build up a certain history or the success of looting in past periods. To separate these two influences, we take a perpetrator's choice of urn in the very first period and include it as a dummy variable (*urn A in period 1*). This variable serves as a proxy for his type. In the regression models (2), (4) and (6) we carry out additional robustness checks by including the variables *female* and *period*.

²⁸ Opposite behavior existed, i.e. changing a white ball into a red one, but those cases were rare and did not offer enough data for analysis. In total, we find 9 such observations for periods 6-40 (3 in treatment high cost and 6 in treatment low cost). Quite possibly, these changes could be due to players who wanted to just check out what happens after such a change or players simply making mistakes. When analyzing signal changes we exclude all these observations and focus on manipulation from red balls to white balls only.

²⁹ It may be possible that perpetrators are more short-sighted and react only to the last punishment received. We checked for that possibility by repeating the regression with punishment from the previous period instead of the five-period aggregate. The results remain robust for that change.

³⁰ Including 1-period lagged variables as controls limits observations for choice of urn to the periods 7-40. This results in 680 observations.

	(1)	(2)	(3)	(4)	(5)	(6)
Urn A	<i>Baseline</i>	<i>Baseline</i>	<i>High Cost</i>	<i>High Cost</i>	<i>Low Cost</i>	<i>Low Cost</i>
Sum of Punishment/100	-2.01** (0.89)	-1.97** (0.94)	0.89 (1.31)	0.83 (1.23)	-0.02 (1.22)	-0.20 (1.12)
Red ball in previous period	0.47 (0.30)	0.41 (0.29)	0.50 (0.36)	0.44 (0.35)	0.56* (0.33)	0.55 (0.34)
History	0.84*** (0.19)	0.78*** (0.17)	0.39* (0.22)	0.32 (0.21)	-0.16 (0.18)	-0.14 (0.18)
Urn A in period 1	0.97 (0.66)	1.30* (0.70)	2.55*** (0.49)	2.67*** (0.60)	4.30*** (0.57)	4.49*** (0.55)
Female		0.59 (0.58)		-0.70 (0.63)		-0.47 (0.60)
Period		-0.00 (0.01)		0.02 (0.01)		0.01 (0.02)
Constant	-2.18*** (0.82)	-2.39** (0.99)	-1.31* (0.76)	-1.06 (0.99)	-1.43*** (0.51)	-1.56** (0.66)
Observations	680	680	680	680	680	680
Pseudo R-squared	0.17	0.17	0.31	0.32	0.47	0.47

Logit regression with robust standard errors in parentheses (clustered by individual)

*** p<0.01, ** p<0.05, * p<0.1

Table 2: Perpetrators' choice of urn

Table 2 shows the regression results. Since we find that the coefficients of *sum of punishment* are quite small, we calibrated this variable relative to the maximum value of 100. This means that the maximum punishment of 100 Taler over five periods obtains the value 1. Consequently, coefficients are now 100-times as big as the original coefficients for the variable *sum of punishment*.

We see that in baseline the sum of punishments received over the last five periods has a significant negative effect on the choice of urn (model (1)). But this effect disappears when manipulation is possible (model (3)-(6)). This indicates that altruistic punishment is indeed successful in reducing looting in treatment baseline. When evaluating marginal effects of the regression we see that the probability of choosing urn A without any prior punishment (when sum of punishment is equal to 0) is 82%³¹. Every additional 10 Taler of punishment during the last five periods decrease that probability by about 4.4%. The highest possible punishment of 100 Taler over five periods results in a probability for picking urn A of only 38%. While increased punishment reduces the probability of choosing urn A considerably in treatment baseline, this is not observed in the manipulation treatments. In treatment high cost, the probability of picking urn A even increases from 74% without any punishment received to 87% with the highest possible punishment. In treatment low cost the probability of picking urn A does not change with punishment (88%). Therefore we state as the first result:

Result 1 (P behavior): In baseline, punishment deters perpetrators from picking urn A. In both manipulation treatments, punishment fails to influence the choice of urn.

To answer hypothesis 2, we look at observers' punishment behavior. To assess how observers engage in altruistic punishment, we run a regression with realized punishment as the dependent variable

³¹ Marginal effects are evaluated at the mean of every other control variable.

controlling for the ball drawn and the history, because observers base their punishment decision on these two variables.

	(1)	(2)	(3)	(4)	(5)	(6)
Punishment	<i>Baseline</i>	<i>Baseline</i>	<i>High Cost</i>	<i>High Cost</i>	<i>Low Cost</i>	<i>Low Cost</i>
Red ball	4.48*** (1.28)	4.46*** (1.29)	3.65*** (1.01)	3.69*** (1.00)	4.26*** (1.43)	4.31*** (1.44)
History	0.60* (0.30)	0.56* (0.31)	0.09 (0.29)	0.12 (0.29)	0.27 (0.26)	0.28 (0.27)
Period		0.00 (0.02)		-0.02 (0.02)		-0.00 (0.01)
Female		-1.15 (1.90)		-0.72 (1.56)		-1.55 (2.96)
Constant	-0.73 (0.92)	-0.13 (1.53)	1.54** (0.60)	2.53** (1.16)	0.04 (0.72)	1.47 (2.55)
Observations	700	700	700	700	700	700
R-squared	0.13	0.14	0.09	0.09	0.11	0.12

Linear regression with robust standard errors in parentheses (clustered by individual)

*** p<0.01, ** p<0.05, * p<0.1

Table 3: Observers' Punishment

As no surprise, observers' punishment increases significantly if a red ball was drawn. In baseline this markup amounts to 4.5 Taler, reducing a perpetrator's payoff by 22.5 Taler, since 1 Taler spent by observers resulted in a payoff reduction of 5 Taler. While *red ball* is highly significant across all treatments, *history* is significant in baseline only (p=0.06). This indicates observers' willingness to take the signal history into account but only if they know the history to be free from manipulation. As soon as manipulation is possible, observers recognize the possibly flawed type of information the history provides and condition their punishment on the draw of the ball alone.

To corroborate this result we looked at observers at an individual level to learn what strategies they chose in order to decide for their level of punishment. While some observers may have a more narrow focus on the ball drawn, others may prioritize the history. In order to assess how observers differ in their punishment strategies, we ran individual regressions on the player level, checking if ball or history or both have a significant influence on punishment. For every observer we regress punishment as the dependent variable on ball and history as independent variables. We allocate observers to the two groups *outcome-based* and *virtue-based* punishment, which were described previously in chapter 4. A perpetrator is allocated to *outcome-based* punishment if only ball has a significant (p<0.1) and positive coefficient in the regression, but history has not. This group includes all observers who put their focus on punishing the outcome, i.e. if looting took place or not, without having the history influencing their decision. *Virtue-based* punishment on the other hand reacts to history. Observers in this group are responsive to history by increasing their punishment the more red balls the history presents. When we observe a positive and significant (p<0.1) coefficient for history, while the coefficient for ball may or may not be significant at the same time, we allocate an individual to this group. All individuals who do not match one of these two groups are either disengaged, meaning they do not punish at all, or cannot be matched to one of the two above groups because coefficients fail to meet the required criteria. We allocate those individuals to the group *others*. Figure 1 shows percentages of these groups over our three treatments.

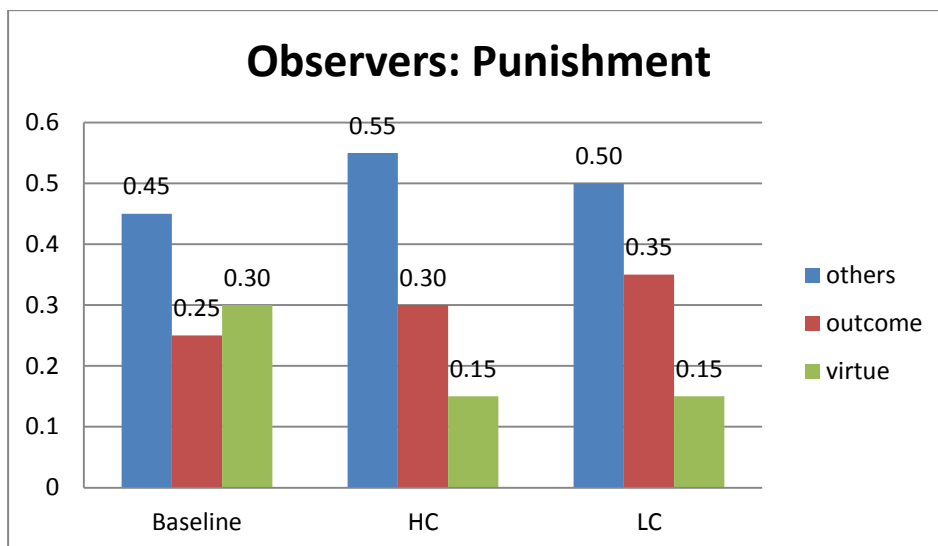


Figure 1: Outcome- and virtue-based punishment

In each treatment, we can assign roughly 50% of observers to either the outcome-based or virtue-based punishment group. Between these two groups, observers follow a virtue-based punishment approach more often in baseline, compared to high cost and low cost. In the manipulation treatments, outcome-based punishment is more frequently executed. This change of distribution between treatments supplies additional evidence on hypothesis 2: in baseline more observers take the history into account compared to the manipulation treatments. As costs for manipulation decrease and the history becomes more noisy, observers progress towards outcome-based punishment as a favorable strategy.

Result 2 (O behavior): In baseline, punishment increases by 0.6 Taler for every additional red ball in the history. In the manipulation treatments, additional red balls in the history do not increase punishment.

As result 2 shows, in the manipulation treatments observers do not react to additional red balls in the history by increasing punishment. This indicates that observers recognize the possibility of manipulation and react to that by not looking at the manipulated history when engaging in punishment. Yet, this does not necessarily mean that observers do not want to punish perpetrators for their intentions to loot the charity box. In (H3) we hypothesized that punishment increases with the belief that urn A was picked. While in baseline the history offered an unadulterated picture of past draws and therefore was a valid proxy for actual intentions, with manipulation observers need to rely more on their formed beliefs about a perpetrator's intentions. For hypothesis 3, we look at observers' beliefs about the perpetrator's choice of urn. After having implemented the punishment decision, observers were asked to state their beliefs about the perpetrator's choice of urn. We constructed this question in such a way that the current draw of the ball did not influence observers' beliefs, but only the observed history (for details see section *Experimental Design*). Therefore, *history* significantly influences *expected urn A* in all treatments³². Because of this we substitute *history* with *expected urn A* in the regression from Table 3 to take a look at how beliefs influence the punishment decision.

³² Coefficients for a logit regression are 0.61 ($p < 0.01$) in baseline, 0.34 ($p < 0.01$) in high cost and 0.56 ($p < 0.01$) in low cost

Punishment	(1) <i>Baseline</i>	(2) <i>Baseline</i>	(3) <i>High Cost</i>	(4) <i>High Cost</i>	(5) <i>Low Cost</i>	(6) <i>Low Cost</i>
Red ball	4.68*** (1.24)	4.64*** (1.25)	3.54*** (1.02)	3.60*** (1.00)	4.13*** (1.38)	4.17*** (1.39)
Expected urn A	1.87 (1.18)	1.87 (1.18)	1.12 (0.82)	1.16 (0.83)	2.72** (1.16)	2.81** (1.16)
Period		0.00 (0.02)		-0.03 (0.03)		-0.00 (0.01)
Female		-1.27 (1.92)		-0.75 (1.57)		-1.79 (2.90)
Constant	-0.19 (0.91)	0.34 (1.43)	1.18** (0.52)	2.23* (1.09)	-0.75 (0.59)	0.75 (2.52)
Observations	700	700	700	700	700	700
R-squared	0.14	0.15	0.09	0.10	0.15	0.16

Linear regression, with robust standard errors in parentheses (clustered by individual)

*** p<0.01, ** p<0.05, * p<0.1

Table 4: Observers' Punishment controlling for expectations about chosen urn

We observe that *expected urn A* obtains positive coefficients for all treatments, and even a significantly positive one in treatment low cost. This indicates that when manipulation is cheap, observers do no longer rely on the history in order to deduce a perpetrator's intentions but instead rely on their beliefs about the choice of urn. In treatment baseline the un-manipulated history offers observers a direct way to punish perpetrators intentions (result 2). But with manipulation observers rely less on this observable history, as they acknowledge the possibly flawed nature of this history. Instead they form beliefs about a perpetrator's intentions and make their punishment dependent on these beliefs.

Additionally, we ran regressions (not reported here) on whether expected urns were significantly affected by the chosen urns, which cannot be observed by the observer. Our results are significant for the treatments high cost and low cost³³. This reveals that perpetrators left traces of their intentions in their history, which were correctly interpreted by observers. In spite of the possibility of manipulation, observers thus formed rational expectations. The unwillingness to make the level of punishment dependent on the history is thus not related to a failure to correctly derive intentions from the history. Instead, observers may be uncertain about the quality of a manipulated signal and prefer to base their decision on less uncertain criteria.

Result 3 (O behavior): In treatment low cost, punishment increases significantly with the belief that urn A was picked. With manipulation, instead of punishing based on history, observers punish based on their beliefs about perpetrator's choice of urn.

As we have shown, observers' punishment increases with the history in the baseline treatment (Result 2). Based on this observation it seems possible that by manipulating a history into showing less red balls, perpetrators are able to generate positive returns in the form of reduced punishments in the manipulation treatments. Next, we want to address the question if these returns outbalance the required costs for the signal change. By answering this question we are able to determine if impression

³³ Coefficients for a logit regression are 0.41 (p>0.1) in baseline, 0.76 (p<0.05) in high cost and 0.91 (p<0.05) in low cost

management is a monetary profitable activity. To obtain the gains from manipulation we start by calculating the revenue a single signal change generates. After obtaining the revenue from a signal change we can compare that value to the costs required and calculate if the resulting change in payoff is positive or negative. Our dependent variable for the regression in Table 5 is the revenue obtained in one period. Different to the total payoff a perpetrator receives at the end of a period, we generate the variable *revenue* by excluding costs of a signal change. Without these costs we acquire a perpetrator's revenue in each period which includes initial endowment, transfers from the charity box and punishment. Revenue ranges from 0 (a white ball was drawn and the perpetrator was punished with 20 Taler, resulting in a 100 Taler reduction) to 200 (a red ball was drawn and the perpetrator did not receive any punishment). Having obtained this value, we can determine what impact a signal change has. We need to consider that any changed signal remains in the history for 5 periods. Afterwards it drops out, as the history includes the last five signals only. Therefore a signal change in period p_t affects all revenues from p_{t+1} to p_{t+5} . Vice versa, the revenue in period p_t is influenced by all previous signal changes in p_{t-5} to p_{t-1} . Thus, if we want to analyze how the revenue of any given period is influenced by signal changes, we have to ask how many of the last five signals were changed and how many were not. To obtain this value, we first calculate a perpetrator's un-manipulated history (variable *history w/o change*). This value depicts the number of red balls that were actually drawn during the last five periods. Secondly, we calculate the number of signal changes that took place during those five periods (variable *sum of changes*). These two variables provide us with the information how a history would have looked without any manipulation and how many of the red balls are actually manipulated. Lastly, it is important to note that we exclude periods 1-5 and 36-40 from the regression to avoid distortions. In periods 1-5 we are unable to calculate history values for all previous 5 periods as these are not fully present yet. In the last 5 periods a signal change affects less than 5 future periods, as the experiment ends after period 40. Table 5 presents the results including the two above described variables as controls. Additionally we include the dummy variable *urn A* in the regression to control for a perpetrator's choice of urn in each period³⁴.

Revenue	(1) <i>High Cost</i>	(2) <i>High Cost</i>	(3) <i>Low Cost</i>	(4) <i>Low Cost</i>
Urn A	54.29*** (3.65)	54.39*** (3.64)	56.00*** (5.40)	56.01*** (5.44)
History w/o change	-0.13 (1.60)	-0.53 (1.53)	-0.60 (1.42)	-0.61 (1.44)
Sum of changes	0.45 (1.63)	0.80 (1.44)	0.50 (1.52)	0.50 (1.52)
Period		0.31 (0.19)		-0.01 (0.18)
Constant	104.2*** (3.43)	98.82*** (5.00)	104.5*** (4.63)	104.7*** (6.22)
Observations	600	600	600	600
R-squared	0.27	0.27	0.22	0.22

Linear regression with robust standard errors in parentheses (clustered by individual)

*** p<0.01, ** p<0.05, * p<0.1

Table 5: Revenue from impression management

³⁴ Additionally controlling for *gender* would only make sense if observers could react to it. Since gender is not revealed in the experiment and it seems implausible that observers implicitly gain clues about perpetrators' gender and react to that, we dropped gender from the regression.

In order to make a profit by manipulating a signal, the revenue from that manipulation would have to outbalance the costs. In treatment high (low) cost the costs for changing one signal were 20 (10) Taler and an increase in revenue would have to compensate this value for manipulation to be profitable. We have to remember that a manipulated signal remains in the history for five periods and influences revenue in each period. Changing a signal therefore does not need to compensate for 20 (10) Taler in an individual period. In order for manipulation to be profitable it is sufficient for a signal change to increase revenue by a fifth of the total costs in any single period. This means that a signal change would need to increase revenue in a single period by 4 Taler in treatment high cost and 2 Taler in treatment low cost. We conduct a Wald-Test to identify if the coefficients of *sum of changes* in Table 5 are significantly different from these values. In treatment high cost 0.45 is significantly smaller than 4 ($p=0.04$), in treatment low cost we cannot identify a significant difference between the necessary revenue of 2 and the coefficient of 0.50 ($p=0.34$). Based on these results we can state that impression management offers only a low profitability in monetary terms. In treatment low costs, evidence for profits is weak, whereas in treatment high cost we even find evidence that impression management results in losses. This is summarized in result 4:

Result 4 (Impression Management): Impression management is not profitable. In treatment high cost perpetrators even incur losses by engaging in the manipulation of signals.

6. Results for Moral Cleansing

Result 4 illustrates the low profitability of impression management. There are two reasons that could explain this finding. First, perpetrators could be subject to erroneous beliefs about the profits from impression management. If this would be the case perpetrators would engage in impression management because they (wrongly) expected a high profitability. A second possibility is that perpetrators form correct expectations about the low profits from impression management but still engage in manipulation. In this case the motivation for impression management is not monetary profit, but other, intrinsic motives. To identify which of these two possibilities holds true, we ran a regression with perpetrators' expected punishment (Table 6). We include the ball and the history in the regression as controls, as these are the two variables determining punishment and subsequently expectations about punishment.

	(1)	(2)	(3)	(4)	(5)	(6)
Expected punishment	<i>Baseline</i>	<i>Baseline</i>	<i>High Cost</i>	<i>High Cost</i>	<i>Low Cost</i>	<i>Low Cost</i>
Urn A	0.33 (0.99)	0.28 (0.98)	-1.34* (0.78)	-1.44** (0.65)	-1.53 (0.95)	-1.85** (0.80)
Red ball	4.65*** (1.06)	4.84*** (1.11)	5.98*** (1.15)	6.02*** (1.27)	4.49*** (0.87)	4.50*** (0.86)
History w/o change	0.13 (0.49)	0.33 (0.52)	-0.15 (0.30)	-0.15 (0.29)	0.05 (0.41)	-0.02 (0.39)
Sum of changes			0.97 (0.57)	0.98* (0.54)	-0.28 (0.46)	-0.07 (0.47)
Female		-2.44 (2.25)		-0.39 (1.94)		1.54 (1.21)
Period		-0.08** (0.03)		0.01 (0.03)		0.06* (0.03)
Constant	1.80	3.93***	0.95	1.15	1.67	-0.22

	(1.19)	(1.00)	(0.69)	(1.96)	(1.09)	(1.37)
Observations	700	700	700	700	700	700
R-squared	0.10	0.13	0.23	0.23	0.09	0.12

Linear regression with robust standard errors in parentheses (clustered by individual)

*** p<0.01, ** p<0.05, * p<0.1

Table 6: Perpetrators' expected punishment

Table 6 shows that the coefficient of *red ball* is highly significant in all treatments. Comparing the coefficients here with those for actual punishment in Table 3, we observe that perpetrators' expectations match actual punishment rather well. In baseline as well as low cost, the draw of a red ball is expected to increase punishment by 4.65 Taler and 4.49 Taler respectively (models (1) and (5)). Observers actually increase punishment by 4.48 in treatment baseline and 4.26 Taler in treatment low cost (see Table 3). In treatment high cost perpetrators expect to be punished more harshly for a red ball (5.98 Taler, model (3)), whereas observers' punishment is only 3.65 Taler (Table 3, model (3)). While expectations about the punishment of a red ball are quite close to the actual values, expectations for history are not. All coefficients for *history w/o change* fail to be significantly different from 0. In baseline (models (1) and (2)), the coefficients of 0.13 and 0.33 at least display the correct positive impact of actual punishment on signal change because actual punishment increased by 0.60 Taler for each additional red ball in history (Table 3). In the manipulation treatments however, coefficients turn negative (models (3), (4) and (6)). This indicates that as a history gets more and more unfavorable by additional red balls, perpetrators expect even less punishment.

Furthermore, for the variable *sum of changes* we would expect negative coefficients that indicate a decrease of expected punishment after changing signals. This follows from (H2), were we stated that a lower number of red balls in the history leads to less punishment. Thus, when changing a red ball into a white ball and thereby reducing the number of red balls in the history we would expect that perpetrators subsequently expect lower punishment. Yet, results reveal that this is not the case. In high cost, we observe a significant positive coefficient (0.98) in model (4). This means, that perpetrators expect their punishment to increase by nearly 1 Taler for every signal they manipulated in the last five periods. This observation runs contrary to the assumption that signals are manipulated for monetary reasons, as assumed in (H4). The regression presented here (Table 6) provides no evidence that perpetrators expect their punishment to decrease as a result from changing signals.

Result 5 (P behavior): Perpetrators do not expect a lower punishment after changing signals.

If we combine the results from Table 5 and Table 6 we see that perpetrators do not profit from impression management in monetary terms and do not even expect to receive less punishment after manipulating signals. Therefore, we conclude that perpetrators who manipulate are driven by an intrinsic motivation. Manipulation is not implemented to influence observers into punishing less but rather because manipulation offers a way to paint a favorable picture of oneself. This behavior can be considered as a manifestation of moral cleansing (Zhong and Liljenquist 2006, Sachdeva et al. 2009, Brañas-Garza et. al 2011). Looting a charity box imposes a moral burden on perpetrators because this is considered as a transgression. Changing signals and building a positive history seems to provide perpetrators with a mechanism to reduce this moral burden.

Having identified the reasons for impression management, we want to look into when and why perpetrators change their signals in more detail. Table 7 shows the results. Here, we ask how perpetrators' probability of changing signals is influenced by punishment. Result 1 has shown that in

the baseline treatment perpetrators react to punishment by looting less often (i.e. picking urn B more often). In the manipulation treatments this effect was not visible. Since manipulation is possible there, it may offer perpetrators an alternative to switching urns. Instead of changing their choice of urn, perpetrators can now change the ball through manipulation. Thus, manipulation could be a substitute to choice of urn.

Signal change	(1) <i>High Cost</i>	(2) <i>High Cost</i>	(3) <i>Low Cost</i>	(4) <i>Low Cost</i>
Sum of punishment/100	2.46** (1.03)	2.51** (1.02)	1.39 (1.25)	1.00 (1.31)
Urn A	0.85 (0.82)	1.02 (0.81)	0.19 (0.75)	0.18 (0.72)
history	-0.57*** (0.22)	-0.55** (0.24)	-0.86*** (0.20)	-0.79*** (0.22)
urn A in period 1	-0.86 (0.56)	-0.99** (0.48)	0.66 (0.58)	0.69 (0.69)
female		0.59 (0.69)		-0.53 (0.40)
period		-0.01 (0.01)		-0.01 (0.01)
Constant	-1.13 (0.85)	-1.47 (0.95)	-0.08 (0.79)	0.30 (0.85)
Observations ³⁵	418	418	463	463
Pseudo R-squared	0.15	0.16	0.15	0.16

Logit regression with robust standard errors in parentheses (clustered by individual)

*** p<0.01, ** p<0.05, * p<0.1

Table 7: Perpetrators' Change of Signal

Coefficients for *sum of punishment* are positive in both treatments, yet only significant in treatment high cost (p=0.02 in model (1), p=0.01 in model (2))³⁶. These positive coefficients indicate that for an increase in punishment an increase in the probability of a signal change is observed as well. Analyzing marginal effects of the logit regression reveals that additional 10 Taler of punishment increase the probability of a signal change by roughly 4% in treatment high cost and 3% in treatment low cost. For example, in treatment high cost, the probability to change a signal is 8% without any prior punishment but increases to about 52% with the maximum punishment of 100 Taler. In treatment low cost the probabilities are 16% without punishment and 44% with maximum punishment.

As a result, we can state that the reaction to punishment exhibits different characteristics. Misconduct (picking urn A) can be reduced through punishment in the baseline treatment, where no manipulation is possible (result 1). When manipulation is possible on the other hand, punishment does not move perpetrators towards better conduct. Perpetrators rather increase manipulation instead (H6). This shows that good conduct (picking urn B) and changing signals are indeed substitutes. Yet, this result has to be interpreted cautiously, as it is not robust for both treatments.

³⁵ Observations include red balls only.

³⁶ Again, we calibrate *sum of punishment* relative to the maximum value of 100, as coefficients are otherwise very small. The variable used in the regression obtains values from 0 to 1.

Result 6 (P behavior): The probability of changing a signal increases with punishment. In treatment high (low) cost every 10 Taler of punishment increase the probability of changing a signal by about 4% (3%).

Additionally to perpetrators' signal change as a reaction to punishment, Table 7 identifies how perpetrators' change their signals dependent on their current history. The negative coefficients for history (-0.57 in model (1), -0.86 in model (2)) indicate that perpetrators change less signals as a history includes more red balls. This effect is strongly significant over both treatments. This means that as a history accumulates more red balls, perpetrators are less likely to change their signals. Evaluating marginal effects for *history* at the means of all other control variables shows that when the history shows white balls only, the probability of changing a signal is 42.15% in treatment high cost. With five red balls on the other hand, the probability is only 0.40%. In treatment low costs, the probability drops from 72.73% with five white balls to 0.35% with 5 red balls. This shows that perpetrators tend to manipulate signals in situations where a favorable history is threatened by an additional red ball. As a history contains more and more red balls, the propensity to change one additional red ball on top of all the others decreases. As (H7) stated, this reflects behavior of feeling free to loot, once additional evidence of looting does no longer deteriorate the history.

Result 7 (P behavior): Manipulation is highest for the first red ball that would enter the history and decreases with any additional red ball.

In summary, our results show how and why perpetrators engage in impression management and how observers react to the possibility of manipulation. Without the possibility of manipulation our results indicate that punishment reduces looting (result 1) and that an unfavorable history gets punished more severely (result 2). With manipulation on the other hand, we see observers no longer reacting to perpetrators' histories in their punishment decision. Observers seem to fully understand the consequence of manipulated signals and in turn no longer react to the history itself. Rather, observers form beliefs about perpetrators' intentions and base their punishment decision on these beliefs (result 3). This behavior is rational as the information provided by the history loses its value, the cheaper manipulation is. Interestingly, we find that perpetrators gain relatively little by manipulating signals in monetary terms (result 4). In treatment high cost, impression management even results in monetary losses. Still, perpetrators continue to engage in this activity. We argue that moral cleansing is the reason for this behavior, as perpetrators correctly anticipate to not receive a more lenient punishment after manipulation (result 5). The reason for impression management seems to be intrinsically motivated by the desire to simply have a good-looking history. Perpetrators thus continue to behave badly but substitute their change of behavior with engagement in impression management (result 6). Furthermore, those whose history shows no red balls are the ones who engage in impression management the most. For perpetrators with very unfavorable histories, little impression management is observed (result 7).

7. Conclusions

Our experiment showed how experimental subjects looted a charity box and how observers were willing to altruistically punish such behavior. We observed that looters were willing to pay for a clean history but did not expect a subsequent reduction in punishment. This expectation was in line with realized punishments, which also were not reduced for looters with favorable signals. A contrasting finding was obtained in a baseline treatment. Punishment was responsive to history if signals could not be manipulated. Observers were thus seen to care about the intentions and virtue of looters. Our findings reveal that a history that can be costly manipulated fails to be taken seriously by observers.

Still, looters continued to allocate money for manipulating signals. We thus conclude that a clean record is cultivated for its own sake.

These results provide a cautious warning for activities that relate to impression management. While it might seem profitable to take up an opportunity to manipulate signals and generate a more favorable image of oneself, the advantages of impression management are observed to be minor in our experiment. If the record of past behavior can be manipulated, our results suggest that observers are not fooled by such manipulation. Observers correctly identified the possibilities of impression management and in turn disregarded information that was easily and cheaply manipulated. And even without such information, observers discovered perpetrators' true intentions remarkably well. Efforts to disguise the underlying intentions through impression management failed in our experiment. This implies that expenditures on impression management are probably ill spent.

Our findings are largely restricted to experimental subjects and conclusions with respect to real world impression management are notoriously difficult. External validity is thus problematic. However, the speed with which experimental subjects discounted signals provides a warning towards arguments on the profitability of impression management. Given that a clean record was not even expected to reduce punishment, our study hints at an alternative motivation to the favorable impression. Companies, just as our experimental subjects, might equally expect little financial return to their social and environmental engagement. They might be motivated intrinsically instead, seeking to portray themselves in a more favorable light and be seen as such by others.

One potential conclusion from our findings would be that resources spent for impression management violate shareholders' interests. Widespread claims that improved impressions are profitable in the long run do not find support in our experiment. The intrinsic desire to dress up one's self-image is the actual driving force behind impression management. The costs of impression management might also be seen as a non-monetary benefit. Just as some jobs besmirch workers physically, others do so mentally. Shareholders may allow managers and corporate staff to dedicate resources to impression management not because it increases profits but because it serves the desire of moral cleansing.

Appendix

General Instructions (both players)

Welcome to our experiment and thank you for participating. This experiment consists of 40 periods, which are identical. You will be randomly assigned either the role of "Player 1" or the role of "Player 2".

The role that was assigned to you is: Player 1 (Player 2).

In each period you will interact with another Player 2 (Player 1). You will be matched randomly by the computer at the start of every period.

There are 1000 Taler in a charity box at the start of every period. This amount may be reduced depending on the players' behavior. The amount that remains in the charity box at the end of a period will be converted into Euros and donated fully to *Doctors without Borders* if this period gets randomly selected at the end of the experiment. Possible donations are between 0 and 1000 Taler. Negative amounts are not possible. The charity box is located outside of this laboratory. After you have received your payoff, please feel free to take a look at the charity box and the additional information brochures about *Doctors without Borders*.

At the end of the experiment, out of all 40 periods one period will be randomly selected. This period will determine the payoffs for all participants and the amount of money that remains in the charity box. All Taler that you earned in the selected period will be converted into Euros (100 Taler = 7€) and paid to you outside of the laboratory. Periods that are not selected do not influence your payoff.

Instructions for Player 1 (Perpetrator)

At the start of each period you receive 100 Taler. Player 2 receives 150 Taler.

You decide between urn A and urn B. There are 4 red balls and 1 white ball in urn A and 4 white balls and 1 red ball in urn B. After having selected an urn, one ball will be drawn at random. If a red ball was drawn, 100 Taler will be transferred from the charity box to you. If a white ball was drawn, no Taler will be transferred.

You and player 2 will observe the ball that was drawn and if 100 Taler were transferred to you or not. But player 2 will not be told which urn you picked. If a red ball was drawn, this could therefore either result from you picking urn A or urn B. The same is true if a white ball was drawn.



After a ball was drawn, player 2 can reduce your payoff. Every Taler player 2 spends reduces your payoff by 5 Taler. Player 2 can spend up to 20 Taler and reduce your payoff by 100 Taler at a maximum.

Additional to the ball drawn, player 2 will receive information about your last 5 signals. The last 5 balls drawn will be stored and displayed in a table as shown below. These are your signals. There will be either a red or a white ball in each cell.



You will observe by how many Taler your payoff got reduced. Afterwards a new period begins and you will be randomly matched with another player 2.

Additional Instructions for treatment High Cost (Low Cost)

After observing the reduction of your payoff you can change your signal (red ball or white ball) from the current period. If a red ball was drawn and you change this signal, instead of this red ball a white ball will be stored in the table for the next five periods. If a white ball was drawn and you change this signal, a red ball will be stored instead. Changing your signal does not change the amount of money transferred from the charity box to you, instead only a different signal will be stored in the table.

Changing a signal costs 20 (10) Taler.

After having changed your signal, this new table will be displayed to the randomly selected player 2 in the next period. All players know that changing a signal is possible and that it costs 20 (10) Taler.

Instructions for Player 2 (Observer)

At the start of each period you receive 150 Taler. Player 1, who is randomly selected, receives 100 Taler.

Player 1 decides between urn A and urn B. There are 4 red balls and 1 white ball in Urn A and 4 white balls and 1 red ball in urn B. After having selected an urn, one ball will be drawn at random. If a red ball was drawn, 100 Taler will be transferred from the charity box to player 1. If a white ball was drawn, no Taler will be transferred to player 1.

You will observe what ball was drawn and if money was transferred to player 1 or not. But you will not be told what urn player 1 picked. If a red ball was drawn this could either result from player 1 picking urn A or urn B. The same is true for a white ball.



After a ball was drawn and you observe if money was transferred or not, you can reduce player 1's payoff. Every Taler you spend reduces the payoff of player 1 by 5 Taler. You can spend up to 20 Taler and reduce the payoff by 100 Taler at a maximum.

Additional to the ball drawn, you will receive information about player 1's last 5 signals. The last 5 balls drawn will be stored and displayed in a table as shown below. These are the signals from player 1. There will be either a red or a white ball in each cell.



After you have decided for the reduction of player 1's payoff a new period starts and you will be randomly matched with another player 1.

Additional Instructions for treatment High Cost (Low Cost)

After you reduced player 1's payoff he can change his signal (red ball or white ball) from the current period. If a red ball was drawn and he changes this signal, instead of this red ball a white ball will be stored in the table for the next five periods. If a white ball was drawn and he changes this signal, a red ball will be stored instead. Changing a signal does not change the amount of money transferred from the charity box to player 1, instead only a different signal will be stored in the table.

Changing a signal costs player 1 20 (10) Taler.

After having changed the signal, the new table will be displayed to the randomly selected player 2 in the next period. All players know that changing a signal is possible and that it costs 20 (10) Taler.

References

- Amalric, F. and Hauser, J. (2005): *Economic Drivers of Corporate Responsibility Activities*, Journal of Corporate Citizenship 20, 27-38.
- Barnea, A. and Rubin, A. (2006): *Corporate social responsibility as a conflict between shareholders*, Working Paper Series, EFA 2006 Zurich Meetings.
- Bartling, B. and Fischbacher, U. (2012): *Shifting the Blame: On Delegation and Responsibility*, Review of Economic Studies 79(1), 67-87.
- Batson, C., D. Kobrynowicz, J. Dinnerstein, H. Kampf and Wilson, A. (1997): *In a Very Different Voice: Unmasking Moral Hypocrisy*, Journal of Personality and Social Psychology 72, 1335–1348.
- Batson, C., E. Thompson, G. Seufferling, H. Whitney and Strongman J. (1999): *Moral Hypocrisy: Appearing Moral to Oneself Without being so*, Journal of Personality and Social Psychology 77, 525-537.
- Benabou, R. and Tirole, J. (2006): *Incentives and Prosocial Behavior*, American Economic Review 96(5), 1652-1678.
- Bromley D. B. (1993): *Reputation, Image and Impression Management*, John Wiley & Sons Ltd.
- Cheng, I.-H, Hong, J.G. and Shue, K. (2014): *Do managers do good with other peoples' money?*, Chicago Booth Research Paper.
- Chugh, D., M. R. Banaji and Bazerman, M. H. (2005): *Bounded Ethicality as a Psychological Barrier to Recognizing Conflicts of Interest*, in D. A. Moore, M. Cain, G. Loewenstein and M. Bazerman (Eds.), *Conflicts of Interest: Problems and Solutions from Law, Medicine, and Organizational Settings*, Cambridge University Press, London.
- Coffman, L. (2011): *Intermediation Reduces Punishment (and Reward)*, American Economic Journal, Microeconomics 3, 77-106.
- Conrads, J., Irlenbusch, B., Rilke, R. M., and Walkowitz, G., (2013): *Lying and team incentives*, Journal of Economic Psychology 34, 1-7.
- Dana, J., Cain, D. M. and Dawes, R. M. (2006): *What you don't know won't hurt me: Costly (but quiet) exit in dictator games*, Organizational Behavior and Human Decision Processes 100, 193–201.
- Dana, J., Weber, R. A. and Kuang, J. X. (2007): *Exploiting moral wiggle room: experiments demonstrating an illusory preference for fairness*, Economic Theory 33, 67–80.
- Egas, M. and Riedl, A. (2008): *The economics of altruistic punishment and the maintenance of cooperation*, Proceedings of the Royal Society B 275, 871–878.
- Eckel, C. C., and Grossman, P. J. (1996): *The relative price of fairness: Gender differences in a punishment game*, Journal of Economic Behavior and Organization 30, 143–158.
- Fehr, E. and Gächter, S. (2002): *Altruistic punishment in humans*, Nature 415, 137-140.
- Fischbacher, U. (2007): *z -Tree: Zurich toolbox for ready-made economic experiments*, Experimental Economics 10(2), 171-178.

- Friedman, M. (1970): *The social responsibility of business is to increase its profits*, The New York Times Magazine, September 13.
- Gallagher, S. (2005): *A strategic response to Friedman's critique of Business Ethics*, Journal of Business Strategy 26(6), 55-60.
- Gino, F. A., Ayal, S. and Ariely, D. (2009): *Contagion and Differentiation in Unethical Behavior: The Effect of One Bad Apple on the Barrel*, Psychological Science 20(3), 393-309.
- Greiner, B. (2004): *The Online Recruitment System ORSEE 2.0 - A Guide for the Organization of Experiments in Economics*, Working Paper Series in Economics 10, University of Cologne, Department of Economics.
- Haisley, E. und Weber, R. A. (2010): *Self-Serving Interpretations of Ambiguity in Other-Regarding Behavior*, Games and Economic Behavior 68(2), 634-645.
- Hamman, J.R., Loewenstein, G. and Weber, R. (2010): *Self-Interest through Delegation: An Additional Rationale for the Principal-Agent Relationship*, The American Economic Review 100(4), 1826–1846.
- Ho, B. (2012): *Apologies as Signals: With Evidence from a Trust Game*, Management Science 58(1), 141-158.
- Hooghiemstra, R. (2000): *Corporate Communication and Impression Management - New Perspectives Why Companies Engage in Corporate Social Reporting*, Journal of Business Ethics 27(1/2), 55-68.
- Ingold, P.V., Kleinmann, M., König, C.J. and Melchers, K.G. (2014): *Shall we Continue or Stop Disapproving of Self-Presentation? Evidence on Impression Management and Faking in a Selection Context and their Relation to Job Performance*, European Journal of Work and Organizational Psychology, forthcoming.
- Kahneman, D., Knetsch, J. L. and Thaler, R. H. (1986): *Fairness and the assumptions of economics*, Journal of Business 59, 285-300.
- Kristof-Brown, A., Barrick, M. R. and Franke, M. (2002): *Applicant Impression Management: Dispositional Influences and Consequences for Recruiter Perceptions of Fit and Similarity*, Journal of Management 28(1), 27-46.
- Lambsdorff, J. Graf (2013): *Corrupt Intermediaries in International Business Transactions: Between Make, Buy and Reform*, European Journal of Law and Economics 35, 349-366.
- Lantos, G.P. (2001): *The boundaries of strategic corporate social responsibility*, Journal of Consumer Marketing 18(7), 595-630.
- Laufer, W. S. (2006): *Corporate Bodies and Guilty Minds: The Failure of Corporate Criminal Liability*, Chicago.
- Lönnqvist, J.-E., Irlenbusch, B. and Walkowitz, G. (2013): *Moral Hypocrisy and the Use of (Un)Fair Decision Procedures*, Unpublished Working Paper, University of Cologne, Germany.
- Mazar, N. and Aggarwal, P. (2011): *Greasing the Palm: Can Collectivism Promote Bribery?* Psychological Science 22(7), 843-848.

- Mazar, N., Amir, O. and Ariely, D. (2008): *The dishonesty of honest people: A theory of self-concept maintenance*, Journal of Marketing Research 45, 633–644.
- Murnighan, J.K., Oesch, J.M. and Pilutlla, M. (2001): *Player Types and Self-Impression Management in Dictatorship Games: Two Experiments*, Games and Economic Behavior 37, 388–414.
- Owen, D.L., Swift, T., Humphrey, C. and Bowerman, M. (2000): *The New Social Audits: Accountability, Managerial Capture or The Agenda of Social Champions?*, European Accounting Review 9(1), 81-98.
- Paharia, N., Kassam, K., Greene, J. and Bazerman, M. (2009): *Dirty Work, Clean Hands: The Moral Psychology of Indirect Agency*, Organizational Behavior and Human Decision Processes 109, 134-141.
- Sachdeva, S., Iliev, R. and Medin, D.L. (2009): *Sinning saints and saintly sinners: The paradox of moral self-regulation*, Psychological Science 20, 523-528.
- Schmitz, J. and Schrader, J. (2013): *Corporate Social Responsibility: A Microeconomic Review of the Literature*, Journal of Economic Surveys, available at <http://dx.doi.org/10.1111/joes.12043>.
- Schlenker, B.R. (1980): *Impression Management: the self-Concept, Social Identity, and interpersonal relations*, Belmont, Calif: Crooks and Cole.
- Shalvi, S., Dana, J., Handgraaf, M., and De Dreu, C. (2011): *Justified ethicality: Observing desired counterfactuals modifies ethical perceptions and behavior*, Organizational Behavior and Human Decision Processes 115, 181-190.
- Shearman and Sterling (2014): *FCPA Digest. Cases and Review Releases Relating to Bribes to Foreign Officials under the Foreign Corrupt Practices Act of 1977*, available at <http://www.shearman.com/en/sitesearch?q=FCPA%20digest> (accessed April 17, 2014).
- Snyder, N. (1977): *Impression Management*, in Social Psychology, ed. by L.S. Wrightsman, Monterey, Calif. Brooks and Cole, 115-145.
- Swann, W. and Bosson, J. (2010): *Self and Identity*, in S. Fiske, D. Gilbert and G. Lindzey (eds.), Handbook of Social Psychology, 5th ed. New York: McGraw-Hill, 589-628.
- Tiedens, L. Z. (2001): *Anger and Advancement Versus Sadness and Subjugation: The Effect of Negative Emotion Expressions on Social Status Conferral*, Journal of Personality and Social Psychology 80(1), 86-94.
- Turillo, C.J., Folger, R., Lavelle, J.J., Umphress, E.E. and Gee, J.O. (2002): *Is virtue its own reward? Self-sacrificial decisions for the sake of fairness*, Organizational Behavior and Human Decision Processes 89, 839–865.
- UK Bribery Act (2010), available at http://www.legislation.gov.uk/ukpga/2010/23/pdfs/ukpga_20100023_en.pdf (accessed April 17, 2014).
- Valor, C. (2005): *Corporate Social Responsibility and Corporate Citizenship: Towards Corporate Accountability*, Business and Society Review 110(2), 191-212.

Vohs, K. D., Baumeister, R. F. and Ciarocco, N. J. (2005): *Self-Regulation and Self-Presentation: Regulatory Resource Depletion Impairs Impression Management and Effortful Self-Presentation Depletes Regulatory Resources*, Journal of Personality and Social Psychology 88(4), 632–657.

Yamagishi T., Horita, Y., Takagishi, H., Shinada, M., Tanida, S. and Cook, K.S. (2009): *The private rejection of unfair offers and emotional commitment*, Proceedings of the National Academy of Sciences of the United States of America 106(28), 11520-11523, PMID: 19564602.

Zhong, C.-B. and Liljenquist, K. (2006): *Washing Away Your Sins: Threatened Morality and Physical Cleansing*, Science 313(5792), 1451-1452.

Acknowledgements

I am grateful for the time as a research assistant and wish to express my thanks to my supervisor, Prof. Dr. Johann Graf Lambsdorff, for his sustained assistance and the many discussions that followed the initial draft of my research. Furthermore, I received constant support and helpful advice from my research team while I worked at this thesis. Especially, I am indebted to Dr. Manuel Schubert, who always had time to discuss new ideas, the experimental setup or the results. He also provided invaluable help with the intricacies of z-Tree. Another big thanks goes towards the teaching assistants who coded my data and without whom study 2 would not have been possible. Lastly, I wish to thank my parents and my girlfriend for their never ending encouragement and support.