When ignorance is innocence: on information avoidance in moral dilemmas.

Joël van der Weele*

February 20, 2012

Abstract

Deliberately avoiding information about potentially harmful consequences of self-interested decisions, or 'strategic ignorance', is an important source of corruption, anti-social behavior and even atrocities. This paper uses the signaling model by Bénabou and Tirole (2006) to explain strategic ignorance as the outcome of a rational trade-off between image concerns and material desires. The model accommodates the findings of several existing experiments and generates new predictions about the determinants of strategic ignorance. An experimental test of these predictions shows that strategic ignorance depends on the payoff of the subsequent moral dilemma. Specifically, ignorance increases when helping others comes at a higher personal cost, decreases when potential harm to others is higher, and a minority is willing to pay to remain ignorant. The results provide clear evidence that people are more likely to learn 'convenient' facts.

JEL-codes: D83, C72, C91.

Keywords: strategic ignorance, pro-social behavior, image concerns, experimental economics.

*I would like to thank Mark Le Quement, Joel Sobel, Ferdinand von Siemens, Michael Kosfeld, Karine Nyborg, Tobias Broer, Zachary Grossman, Elisabeth Schulte, Heiner Schumacher, Gary Charness, Tore Ellingsen, Martin Dufwenberg, Matthias Blonski, Avichai Snir, Karl Schlag, Matthias Heinz, Leonie Gerhards, and seminar participants in Frankfurt, Amsterdam and Tilburg for useful comments, Zachary Grossman again for generously sharing his programs with me, and Julija Kulisa for helping me out with technical matters. I am indebted to the Vereinigung der Freunde und Förderer der Goethe-Universität for financial support. All errors are mine.
Email: vdweele@econ.uni-frankfurt.de

"Living is easy with eyes closed."

The Beatles, "Strawberry Fields Forever" (1967).

1 Introduction

People often avoid or ignore evidence about the negative social impact of decisions for which they are responsible. Such 'strategic ignorance' plays an important role in political and corporate corruption, the perpetuation of conflicts and even genocide. Strategic ignorance is a means to reduce liability and/or the demands of consciousness by avoiding information about suspected wrongdoings.

As examples, consider two of the largest corruption cases in U.S. history: Watergate and Enron. In an analysis of the Watergate scandal, Simon (2005:4) writes "The most salient theme in the unsavory moral world of the Watergate participants is not amorality or ruthlessness, but rather aversion to accountability. The participants showed intense faith in the immunizing power of deliberate ignorance and calculated ambiguity." This came to the fore during the trial, when the treasurer of Nixon's re-election committee testified that he queried campaign finance chairman Maurice Stans on why (Watergate burglar) Gordon Liddy had been given large sums of money. Stans replied "I do not want to know, and you do not want to know".

During the Enron trial, top executives Lay and Skilling argued that they were ignorant of any fraud. However, Lay had explicitly instructed the committee of Enron's lawyers looking into internal allegation of fraud, to abstain from inquiries into the accountants' practices or any lower level employees identified as witnesses to the fraud (Simon 2005). This prompted the judge in the Enron case to explain the concept of 'willful blindness' to the jury members: "You may find that a defendant had knowledge of a fact if you find that the defendant deliberately closed his eyes to what would otherwise have been obvious to him..."

Cohen (2001) shows how strategies of information avoidance and denial also play a crucial role in the perpetuation of wars and even genocide. Much research has centered on the atrocities of the Holocaust, which involved large-scale strategic ignorance from both bystanders and perpetrators. Horwitz (1991) interviewed residents of the Austrian village of Mauthausen who lived next to a cluster of labor/extermination facilities. Although presented with strong cues, they made no effort to learn what was going on in the camps. Horwitz (p. 175) writes: "Blindness was willed [...] By remaining ignorant the residents would be spared the agony of worrying what was happening inside the camp constituted a violation of humane behavior against which their conscience might demand they object".

At the top of the hierarchy, the case of Nazi government minister Albert Speer is a much cited instance of strategic ignorance. At the Nuremberg trials Speer claimed that he did not know of

the mass killings, although he admitted that he should and could have. In his autobiography, Speer relates how one of his colleagues advised him never to inspect the Auschwitz concentration camp, because of the indescribable things going on there. Speer writes[1]

> "I did not query him. I did not query Himmler, I did not query Hitler, I did not speak with personal friends. I did not investigate - for I did not want to know what was happening there. [...] [F]rom fear of discovering something which might have made me turn from my course, I had closed my eyes." A. Speer (1970: 376).

These examples demonstrate that strategic ignorance is an important problem in the real world. Less disturbing but more systematic evidence comes from economic laboratory experiments. Dana, Kuang and Weber (2007, henceforth DKW) matches subjects into pairs, consisting of a 'dictator' and a 'receiver'. The dictator chooses between two options A and B. Option A yields a slightly higher payoff for the dictator, but she knows that it leads with some probability to a relatively large monetary loss for the receiver. Before the dictator makes her choice, she can find out whether the loss to the receiver will actually occur by clicking a button on the screen. Even though it is virtually costless to get information, 44% of the people choose not to do so and choose action A.[2] The authors argue that people use ignorance of the consequences for player B as an 'excuse' to behave selfishly.

On the surface, the reason for strategic ignorance in the above examples seems straightforward. The excuse of not-knowing allows the pursuit of self-interest at a reduced guilt and/or legal repercussions. On second thought however, it is puzzling how the excuse of not-knowing can be valuable given that it is obvious that the subject *chose* to be ignorant. Under these circumstances, how can the excuse of ignorance have any exonerating effect? And what determines whether people choose to avoid information in the first place?

In this paper, I address these questions with the help of a theoretical model and a laboratory experiment. In the theoretical part in Section 3, I formally analyze a situation in which people first choose whether to inform themselves about the social costs of an action that benefits them personally, and then choose whether to engage in that action. The analysis is based on the model by Bénabou and Tirole (2006), and departs from the assumption that people care about their own material payoffs, the payoffs of others (to a heterogeneous degree), and their image as a prosocial actor. Such image concerns can reflect either the value of having a reputation as a moral person amongst outside observers, or the internalized demands of one's own moral conscience or 'man-in-the-breast'.

[1] What Speer really did know about the holocaust has been the subject of great controversy, described at length in Sereny (1996).

[2] These results has been replicated by Larson and Capra (2009) and Grossman (2010).

I show that there exists a pure strategy equilibrium featuring strategic ignorance. In this equilibrium, people with weaker social preferences avoid information about the consequences of their actions. This implies that there is a social stigma attached to avoiding information. However, the stigma attached to knowingly engaging in harmful actions is even worse. Thus, low types prefer to remain ignorant in order to avoid an explicit moral decision in which they may have to choose between two evils: to abstain from an action with high personal payoffs, or to reveal themselves as immoral individuals. In other words, people anticipate that if they inform themselves, they will feel forced to engage in high signaling investments which they prefer to avoid.

In Section 4, I report the results of a laboratory experiment designed to test the comparative statics of the model. In the experiment, subject have to decide between two actions, one of which brings them higher payoffs but involves potential losses for another subject. Before doing so, they have to choose whether to find out about those losses or not. In line with the theory, I find that actively chosen ignorance about such losses rises with the material sacrifice associated with avoiding losses to others. Furthermore, ignorance decreases if the potential loss inflicted on others increases but stays constant if the *probability* of imposing losses on others goes up. Finally, I find that a small but significant minority is willing to pay to remain ignorant. Thus, the experiment provides unambiguous evidence that people are biased to the acquisition of 'convenient' information.

In Section 5, I show that the model can accommodate previous experimental results on strategic ignorance found by DKW, Krupka and Weber (2008), Fong and Oberholzer-Gee (2011) and Conrads and Irlenbusch (2011). In the conclusion, I discuss implications for the existence of self-serving biases and some consequences for economic behavior.

2 Literature

In terms of modeling, the paper applies the model by Bénabou and Tirole (2006) that stresses (self-) image concerns as a main driver of pro-social behavior. Related models are presented by Ellingsen and Johannesson (2008), Andreoni and Bernheim (2009), and Tadelis (2011).

Bénabou and Tirole (2011) use a signaling model to analyze strategic ignorance in the context of taboos. The authors argue that refusing to think about or inform oneself of the payoffs of deviant actions can be a signal of virtuous character. They do not not explicitly model both the decision to remain ignorant and the decision to take an ethical action, and therefore cannot compare the social image of behaving badly unknowingly with behaving badly knowingly, which is the focus of this paper.

Andreoni and Bernheim (2009, online appendix) use signaling to explain the finding that

some people are willing to pay not to play a dictator game. The logic is similar to the one in this paper: opting out helps to avoid a low social image resulting from the decision not to share. In the model, the image related to the 'outside option' (opting out of the dictator game) is given exogenously, whereas the central theoretical exercise in this paper is to derive the image associated with the outside option (remaining ignorant) endogenously.

There are several non-signaling models that generate strategic ignorance. Nyborg (2011) assumes that people are 'duty-oriented', meaning that they suffer disutility (e.g. a loss of self-image) if their contributions in a public good game are far away from some 'ideal' contribution level. The model implies that a higher ideal contribution (weakly) raises the contribution of a duty oriented agent, but lowers her overall utility. Therefore, the agent is willing to pay to avoid information that may raise the ideal contribution, for example information that contributions have a high social value. The model in this paper differs by explicitly modeling the image formation process, which leads to additional comparative statics.

Several papers have modeled strategic ignorance as a commitment device. Carillo and Mariotti (2000) show that ignorance about the health risks of certain activities like smoking, can be rational for a time-inconsistent agent. Maintaining a high (although not necessarily accurate) perception of health risks will help the agent to reduce overconsumption. Similarly, Bénabou and Tirole (2002) argue that an agent may prefer not get feedback on her abilities, because being overconfident about those abilities reduces under-investments resulting from present-bias. Aghion and Tirole (1997) show that managers may want to stay ignorant of the payoffs of different projects in the firm. Their ignorance increases the de facto authority of subordinates, and thereby raises their incentives to gather information. Dominguez-Martinez *et al.* (2010) provide experimental evidence for such strategic ignorance. Crémer (1995) argues that strategic ignorance of the personal circumstances of the agent can help a principal to credibly implement stronger incentives for the agent, because he reduces the moral hazard associated with conditional incentives. Although the model in the current paper has a very different structure, it also allows the interpretation that ignorance works like a commitment to avoid to making 'wrong' decisions in the future. The commitment is useful because if it is known that the agent knows that her actions have adverse consequences for social welfare, social pressures lead her to over-invest in 'nice' actions.

The model also relates to a large literature in psychology and a growing literature in economics showing that people downplay negative feedback about their competence and are sometimes willing to pay not to receive any feedback at all (Möbius *et al.* 2011). Explanations of this phenomenon assume that people either derive direct (anticipatory) utility from a positive self-image (Köszegi 2006, Möbius *et al.* 2011) or instrumental benefits from combatting time inconsistency (Bénabou and Tirole 2002). Although this literature has some intuitive parallels

with this study, it does not address the moral trade-off that is at the heart of the current paper.[3]

On the experimental side, the paper builds on and complements existing studies on self-image and self-serving biases (Bersoff 1999, Mazar *et al.* 2009), and more specifically about ignorance in moral dilemmas (Dana *et al.* 2007, Krupka and Weber 2008, Larsson and Capra 2009, Grossman 2010, Fong and Oberholzer-Gee 2011, Matthey and Regner 2011, Conrads and Irlenbusch 2011). The relation to this latter literature is explained in more detail in Sections 4 and 5. The results of the experiment go beyond this literature by providing very precise evidence on the determinants of information avoidance and the value of moral excuses, and doing so in the context of a theoretical model.

3 Theory

Why do people choose not to know the consequences of their own actions? The answer that I develop in this section applies the model by Bénabou and Tirole (2006). This model augments standard preferences for material payoffs with a) an intrinsic preference for social welfare and b) a preference for an image as a pro-social actor.[4] There are two reasons why a model of image concerns is appropriate. First, it is intuitively plausible that the excuse "I did not know" is an attempt to uphold an image as a moral person. Second, as I explain in more detail in Sections 5, some results from the existing experiments (as well as the current one) are difficult to explain by simpler models.

3.1 The model

Consider an agent who chooses whether to engage ($a = 1$) in an activity or not ($a = 0$). Engaging in the activity yields a payoff of v to the agent, but causes potentially negative external (welfare) effects, denoted by W. The exact value of W is initially unknown to the agent. As examples of such activities one can think about the consumption of cheap products made by children or slave labor in developing countries, using cosmetics tested on animals, but also engaging in dubious accounting practices, or the provision of political support to dubious regimes. All these activities have in common that they may have negative effects for other individuals or the environment, but the exact nature of these effects is typically not immediately transparent.

[3]Other non-signaling models include Caplin and Leahy (2001), who show that anticipatory utility can lead to strategic ignorance, and Fershtman *et al.* (2011) who provide a model of taboos.

[4]The model by Ellingsen and Johannesson (2008) includes Bënabou and Tirole (2006), and therefore the model in this paper, as special case.

More specifically, before the agent takes any decisions, she expects that

$$W = \begin{cases} w & \text{with probability } p, \\ 0 & \text{with probability } 1-p. \end{cases} \tag{1}$$

To make things interesting from a welfare perspective, I assume that $w > v$.

Before the agent decides to engage in the activity (or not), she has the opportunity to inform herself about the true welfare cost ($I = 1$) or to remain uninformed ($I = 0$) at a cost c. I call the latter decision or state 'strategic ignorance'. For simplicity, information takes the form of a perfectly informative signal $\sigma \in \{\sigma_w, \sigma_0, \emptyset\}$, where σ_w denotes 'bad news' ($W = w$), σ_0 denotes 'good news' ($W = 0$), and with some abuse of notation $\emptyset$ denotes the case in which no information is acquired. Examples of such information may be direct observation, newspaper articles, flyers, tv documentaries or conversations with others.

The timing of the game is as follows:

1. Nature selects the level of $W \in \{0, w\}$ associated with activity a.
2. The agent chooses whether to receive a signal ($I = 1$) or not ($I = 0$) about the level of W.
3. The agent chooses whether to engage ($a = 1$) or not ($a = 0$) and payoffs from engaging are realized.
4. The agent's actions a and I and the signal content σ are observed, and image payoffs are realized.

Let s denote a (pure) strategy for the agent, which is a function from the agent's type into her action space. The agent has preferences that can be represented by the following von-Neumann-Morgenstern utility function

$$u(\theta, a, I, \sigma) = a\,(v - \theta W) - cI + \mu E[\theta \mid I, \sigma, a; s]. \tag{2}$$

Here, the first term denotes the payoff of the activity, which consists of v, minus the welfare cost of the activity multiplied by a parameter θ, the 'type' of the agent. The higher θ, the more altruistic or pro-social the agent. I assume that the type of the agent is known only to her, and the observer has a prior distribution $F(\theta)$ with full support on $[0, 1]$.

The second term of the utility function is the cost of information c. This cost may be positive, but also negative when information is presented in a way that makes it hard to avoid.

The last term denotes the payoffs from image concerns, where $E[\theta \mid I, \sigma, a; s]$ denotes the inference made by an observer about the type of the agent. In equilibrium, this inference is formed by the application of Bayes' rule to the observed choices I and a, the content of the

signal σ, and the equilibrium strategy s^* of the agent. For simplicity, I assume that the agent cares directly about her image, but one could view this as a reduced form representation of a model where the agent derives material benefits from a positive inference by the observer, e.g. by engaging in surplus-generating future interactions. The parameter $\mu > 0$ captures the degree to which being viewed as someone who cares about others takes a central place in the agents self-worth or identity. In Section 4.4 I discuss briefly the possibility of heterogeneity in μ.

With respect to the relative importance of the different motives of the agent, I will assume that $v > \mu$, i.e. image concerns are small relative to material concerns. This rules out that agents abstain purely for image reasons, and is sufficient (although not necessary) to guarantee that in equilibrium there will always be some types who engage in action a.

Two points deserve elaboration. First, the reader may be surprised that the agent's preferences depends both on her type and the expectation of her type. Second, it may strike the reader as unrealistic that both the choice to acquire information and its content are observable.

I address both concerns by offering two distinct interpretations of who exactly the observer is in this game. First, there is a standard social signaling interpretation in which the observer may be one or more other people. In this case, the assumption that the observer sees the information obtained by the agent is strong. Nevertheless, it will play a role in decisions that may come under legal and therefore public scrutiny: in the Nuremberg, Enron and Watergate trials referred to in the introduction, a central issue to the prosecution was 'who knew what when'. Note that although the importance of social signaling concerns may be somewhat limited in anonymous laboratory environments, there is convincing evidence that even in such situations subjects care about the opinion of others (Dana *et al.* 2006) as well as that of the experimenter (Hoffman *et al.* 1996).

Second, image concerns also have a dual-self interpretation in which the agent tries to impress a Smithian "imagined spectator" or "man in the breast", or a Freudian "super-ego" or future selves.[5] These entities pass moral judgements very much like an external observer. In this interpretation the importance of image concerns does not rely on observability of actions by others and the assumption that observer knows the information acquisition decision and the content of the signal is natural.[6]

[5]Signaling to future selves, based on the work of Bem (1972), rests on the assumption that people do not always have introspective access to their own deep values. Instead, they sometimes use their past actions to form an opinion of their own character or 'identity', much like an outsider would do. The implicit assumption is that the agent maximizes the self-image of future selves, and μ may be interpreted as a discount rate. This idea has been used by Bénabou and Tirole (2004, 2006, 2010) in several applications.

[6]In psychology, there is a large literature on self-image enhancement in the context of moral decisions (e.g. Baumeister 1998, Bersoff (1999), Mazar *et al.* 2009). In economics, Bodner and Prelec (2003) give an analytic exposition of self-signaling. Murnighan *et al.* (2001) and DKW (the plausible deniability treatment) find evidence for self-image concerns in economic games, Grossman (2009) does not. Bénabou and Tirole (2004, 2006, 2011)

3.2 Strategic ignorance in equilibrium

In this section, I will focus on the existence of a particular perfect Bayesian equilibrium outlined in the proposition below. In this equilibrium, all types play a strategy s^* that maximizes their utility given the behavior of the other types, and beliefs are formed by the application of Bayes' rule wherever possible. Although there are other pure-strategy equilibria in the model, ignorance is not part of the equilibrium strategy profile. Since these equilibria cannot explain the results from this or earlier experiments (see Section 5), I will not discuss them any further.

The main theoretical result of the paper is the following.

Proposition 1 *There exist a $\bar{p}$ and $\underline{c} < 0 < \bar{c}$ such that if $p > \bar{p}$ and $c \in [\underline{c}, \bar{c}]$ there exists a semi-separating equilibrium characterized by $\theta^* \in (0,1)$, in which*

a) all types $\theta < \theta^*$ choose not to inform themselves and subsequently choose $a = 1$,

b) all types $\theta \geq \theta^*$ choose to inform themselves and choose $a = 1$ if and only if the news is good,

c) θ^* is increasing in v, c and p, and decreasing in w.

The proof is in Appendix A. To understand these results, consider the trade-off for the agent. On the one hand, if she remains ignorant, she pools with the lower types $\theta < \theta^*$ and reduces her image relative to the prior expectation. She also suffers a utility cost of $p\theta w$, which can be interpreted as disutility from 'guilt'. On the upside, she is sure to get the engagement payoffs v. If, on the other hand, she acquires information, she faces a lottery. If the news is good, she pools with the high types and reaps the benefits v of engaging, the best possible outcome. If the news is bad, she faces a choice between two evils. She can pool with the high types at the cost of v. If she engages instead, she suffers guilt of θw, and ends up with the lowest possible image, because beliefs for this off-equilibrium action are 0.[7] When the probability of bad news p is relatively high, the lottery associated with information acquisition is unattractive to a low type, who chooses to remain ignorant. The single-crossing condition is satisfied because low types suffer less from guilt when they remain ignorant.

Part c) of the Proposition supplies comparative statics on the threshold type θ^*, that I will test in the second part of this paper. An increase in v makes engaging in the action more attractive. Since engaging is guaranteed in equilibrium only through strategic ignorance, θ^* increases. An increase in w has the opposite effect: expected higher social costs raise the 'guilt' $p\theta w$ of the agent when she chooses strategic ignorance. She can avoid these costs by choosing to

cite additional evidence for self-image concerns.

[7] Because payoffs depend directly on beliefs, the standard refinements do not apply. However, in the appendix I show that the equilibrium satisfies a refinement akin to the intuitive criterion.

inform herself, and therefore θ^* decreases in w. Obviously, an increase in the cost of information c raises the attractiveness of ignorance. Note that strategic ignorance may exist even if there is a cost of ignorance (i.e. $c < 0$).

An increase in p has contradictory effects. On the one hand, it worsens the 'guilt' of an agent and makes ignorance less attractive. On the other hand, it raises the probability of bad news, which decreases the expected utility of informing oneself. Thus, an increase in p unambiguously lowers the utility of every type $\theta > 0$. Moreover, I can show that the second effect dominates the first, and a higher p results in a higher θ^* and more strategic ignorance. Note that while a higher potential welfare cost w leads to *less* strategic ignorance, a higher probability of that cost leads to *more* of it.

Proposition 1 explains why ignorance is an 'excuse' for anti-social behavior, because the image associated with acting nasty unknowingly is higher than that of acting nasty knowingly. Moreover, an agent who remains ignorant can credibly make the counterfactual statement that "if I had known, I would have behaved nice." To understand this, note that the strategy of some of the agents who remain ignorant specifies that they would abstain if they had chosen information and were confronted with bad news. Or, to say this is a different way: suppose that an agent who chose to be ignorant would be informed 'by surprise' that the news was bad. Technically, one can model this as a move of nature that occurs with probability 0 in equilibrium (and therefore leaves the equilibrium beliefs unaltered). The following result obtains.

Remark 1 *In the semi-separating equilibrium of Proposition 1, there exists a type $\bar{\theta} < \theta^*$, such that all types in $\theta \geq \bar{\theta}$ would abstain if they were exogenously informed that $W = w$.*

The reason is that the image associated with engaging knowingly is lower than that of engaging unknowingly, so social pressures to abstain are stronger once bad news has arrived. These social pressures cause some types to abstain, even if from the ex-ante viewpoint of the agent this represents an 'over-investment' in signaling. One can thus interpret strategic ignorance as a commitment device to avoid a status competition for the moral high ground.

It is instructive to illustrate this result graphically. Figure 1 shows the value of the signal for each type, which is computed by comparing the expected utility from acquiring information with that of remaining ignorant, assuming each type subsequently makes optimal decisions given the equilibrium beliefs. Figure 1 shows that there are three kinds of decision makers. Borrowing the terminology from Lazear *et al.* (2010), the types $\theta \in \left[0, \bar{\theta}\right)$ can be called "non-sharers", who will engage regardless of the signal's content, and therefore information is irrelevant for their decision. However, because there is an expected image loss from acquiring information (assuming p is high enough), the value of the signal is negative. The types $\left[\bar{\theta}, \theta^*\right)$ are "reluctant sharers", who would abstain if they received a signal σ_w. Therefore, acquiring information will raise their image. However, these benefits are outweighed by the expected material loss pv.

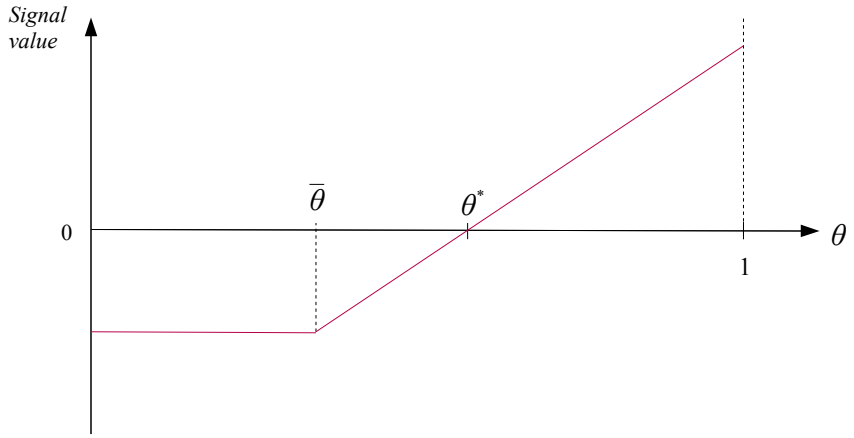

Figure 1: The 'value' of the signal (the ex-ante expected utility of informing oneself minus the utility of remaining ignorant, given equilibrium beliefs).

Thus, both non-sharers and reluctant sharers are willing to pay to remain ignorant. Finally, for the "willing sharers" $\theta \in [\theta^*, 1]$, the value of the signal is positive, because they internalize the expected costs from engaging unknowingly.

4 An experimental test

Proposition 1c) offers several comparative statics with respect to the occurrence of strategic ignorance. In this section I describe the results of an experiment designed to test these predictions.

4.1 Experimental design

The design is a generalization of the "hidden information treatment" in DKW.[8] In all treatments subjects are randomly paired in groups of two, consisting of a 'dictator' (Player X) and a receiver (Player Y). Both players are paid according to the decision of the dictator. The receiver is passive and does not make any decision. The players receive a show-up fee of 4 euros, and all amounts are framed in terms of experimental currency (EC), where 10 EC = 1 euro.

The dictator is facing the following situation. She can choose between two actions A and B, resulting in a payoff (net of information costs) for the dictator of y_A or y_B respectively, where

[8] That treatment is a special case of the design below with $y_A = \$6$, $y_B = y_H = \$5$, $y_L = \$1$ and $p = \frac{1}{2}$. Direct comparison with the current experiment is difficult however, since payoffs are different and 'ignorance' has to be chosen actively, rather than being the default option as in DKW.

y_A or y_B are known quantities and $y_A > y_B$. The dictator is told that the potential payoffs of the observer are y_H and y_L, with $y_H > y_L$, but that a computer randomly decided before the start of the experiment which payoff is associated with which action. More precisely, let (y^X, y^Y) denote the payoffs of the two players (net of a cost of information). With probability p, the computer decided action A results in payoffs (y_A, y_L) and action B yields (y_B, y_H). Since action A yields the highest payoff for Player X, while action B yields the highest payoff for player Y, I will refer to this situation in the remainder as the Conflicting Interest Game (CIG). With probability $(1-p)$, the computer decided action A yields payoffs (y_A, y_H) and action B yields (y_B, y_L). Since action A now yields the highest payoff for both players, I call this the Aligned Interest Game (AIG). I implement $y_B = y_H$ to rule out jealousy by the dictator.

The dictator makes two choices. First, she has to decide whether to acquire information. She faces a screen with a payoff matrix where her own payoffs y_A and y_B are revealed, but y_H and y_L are replaced by a question mark. The screen features two buttons saying "Reveal game" or "Don't reveal". If the dictator decides to reveal the game she moves to the next screen where the full matrix is shown , as well as two buttons for choosing A and B. If she decides to not reveal, the question marks remain. The subject then proceeds to choose A or B. The decisions are made in an anonymous fashion. After each treatment, dictators were asked in a questionnaire: "Did you choose to reveal the payoffs of the observer? Why (not)?"

Screen shots of the instructions and the experiment can be found in Appendix D. Programming was done[9] in z-Tree (Fischbacher 2007) and carried out at the Frankfurt Laboratory for Experimental economics (FLEX).

4.2 Treatments and hypotheses

Denote by m the fraction of subjects that choose to remain ignorant. The objective of the experiment is to compare m across the different treatments outlined below. The hypotheses are derived from Proposition 1. We can relate the parameters in this experiment to the model above as follows: $v = y_A - y_B$ is the gain of the dictator from choosing action A. Similarly $w = y_H - y_L$ is the loss to the observer if the dictator chooses the action associated with y_L rather than with y_H. The probability p translates directly from the model to the experimental setting.[10]

To mimic most real-world information decisions, and in keeping with the previous experiments on this topic, the decisions in the experiment are made anonymously. Most importantly, the observer does not learn the dictator's decision to reveal, and the dictator knows this. This

[9] I am indebted to Zachary Grossman for generously sharing his code with me.

[10] The setting is not exactly the same as in the model, where abstaining can never involve a loss to the observer. However, in practice this difference does not matter, since only two subjects abstained without knowing that abstaining was indeed the action that benefitted the observer. It would be trivial to reformulate the model to exactly fit the experiment.

means that image concerns towards the self and the experimenter are more relevant than towards the receiver.

In the baseline treatment, $p = \frac{1}{2}$ and payoffs (in EC) are $y_A = 100$, $y_B = y_H = 60$ and $y_L = 10$. Thus, in this payoff structure, $v = 40$ and $w = 50$. Figure 1 shows the graphical presentation as given in the instructions of the CIG ("Game 1") and AIG ("Game 2"), see also the screenshots in Appendix D.

Player X chooses	Player X receives	Player Y receives
A	100	10
B	60	60

(a) Game 1

Player X chooses	Player X receives	Player Y receives
A	100	60
B	60	10

(b) Game 2

Table 1: The experimental games. Each game has been chosen with 50% probability.

The v-treatment is equivalent to the baseline treatment, except that $y_A = 70$, so that choosing a 'fair' action in the CIG becomes less costly. In this payoff structure, $v = 10$ and $w = 50$. Since v has decreased relative to the baseline treatment, the model predicts that

Hypothesis 1 *In the v-Treatment, m will fall relative to the baseline treatment.*

The w-treatment is equivalent to the baseline treatment, except that $y_L = -20$, i.e. the receiver may lose part of her show-up fee. In this payoff structure, $v = 40$ and $w = 80$. Since w has increased relative to the baseline treatment, the model predicts that m goes down.

Hypothesis 2 *In the w-Treatment, m will fall relative to the baseline treatment.*

The p-treatment is equivalent to the baseline treatment, except that $p = 0.8$, so that the CIG is more likely. Because p has increased relative to the baseline treatment, choosing information becomes more 'risky' and the model predicts that

Hypothesis 3 *In the p-Treatment, m will rise relative to the baseline treatment.*

The c-treatment is equivalent to the baseline treatment, except that the dictator has to pay 5 EC (€.50) to remain ignorant. The theory predicts that the m will go down and also, as we learn from Figure 1, that there will be some people who are willing to pay to remain ignorant.

Hypothesis 4 *In the c-Treatment, m will fall relative to the baseline treatment, but there will be a positive fraction of people who will pay to remain ignorant.*

In addition to these hypotheses, the model also predicts whether subjects will choose A or B. These predictions are the same for all treatments:

Hypothesis 5

a) Subjects who remained ignorant will choose A.

b) Informed subjects who face the aligned interest game (AIG) will choose A.

c) Informed subjects who face the conflicting interest game (CIG) will choose B.

Finally, note that the theory is not strong enough to predict the level of ignorance or the relative size of the comparative static effects, which depends on the (unobserved) value of μ, and the shape of the distribution $F(\theta)$.

4.3 Experimental results

330 Subjects participated in the main experimental treatments[11], which lasted approximately 35 minutes, earning on average 10,28 euros (dictators 12,66 euros, receivers 7,89 euros). The data of the four experimental treatments are summarized in Table 2 in Appendix C, and presented graphically below.

Figure 2 shows the shares of subjects who choose to remain ignorant in the different treatments. The first observation of interest is that a substantial fraction of the subjects, 31% , choose to remain uninformed in the baseline treatment. This result is remarkable, because unlike the studies of DKW and Larson and Capra (2009), the present study did not have ignorance as a default option but required it to be actively chosen. Grossman (2010) varies default settings, and finds that requiring ignorance to be an active choice reduces ignorance levels by almost 50%.

In line with Hypothesis 1, we see a drop in ignorance in the v-treatment, which is significant at the 1% level.[12] The theory predicts that ignorance should also drop in the w-treatment, where the observer suffers a greater loss, and choosing action A under ignorance thus makes the dictator feel more guilty. The evidence here is weaker: the drop in ignorance relative to the baseline is significant only at the 10% level.

By contrast, we cannot confirm Hypothesis 3 that in the p-treatment where the CIG is more likely to be played, the level of ignorance should go up. We do not observe any shift relative to the baseline treatment and consequently cannot reject the hypothesis that the change in probability has no effect on ignorance levels.

[11] I carried out an additional treatment on image concerns reported in the discussion, with an additional 70 subjects.

[12] All results reported hold for a one-sided Fisher's exact test (with mid-P correction), and a one-sided exact z-test for equal proportions (Suissa and Shuster 1985).

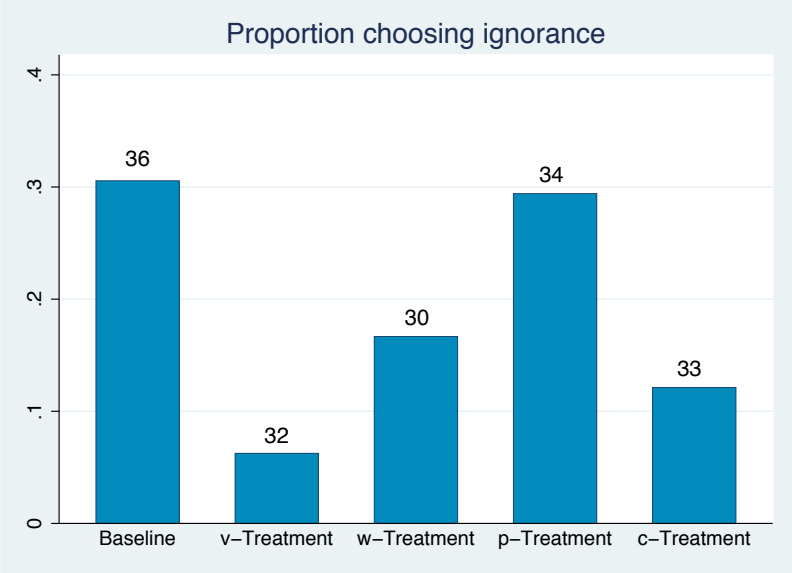

Figure 2: Proportion choosing ignorance by treatment. Number of observations in each treatment at the top of the bar.

The cost treatment shows a significant (at 5%) drop in the ignorance level. Perhaps more interestingly, there is a small but significant[13] minority who are willing to pay for ignorance. Both facts are consistent with Hypothesis 4.

Summary 1 *The data support Hypothesis 1,2, and 4: ignorance increases if it is more expensive to be fair, and decreases if the potential losses to other parties are bigger or if it is more expensive to be ignorant. Moreover, a small but significant minority is willing to pay to remain ignorant. The data do not confirm Hypothesis 3: we do not find evidence that the probability of inducing losses on others affects ignorance.*

Moving to testing Hypothesis 5, Figure 3 shows the amount of people choosing A in each treatment (the last column of Table 2). It is clear from the figure that almost all subject who are uninformed or face AIG choose action A, as predicted by the theory. It is equally clear that contrary to the theory, not all subjects in the CIG choose action B. Note that treatment fluctuations of behavior in the CIG are not statistically significant at the 10% level (2 sided z-test). Altogether, we can summarize as follows:

Summary 2 *The data corroborate Hypothesis 5a) and 5b), but falsify hypothesis 5c), because a substantial fraction of the subjects choose the selfish option in the CIG.*

[13] A z-test for the proportion being different from 0 yields significance at 5%.

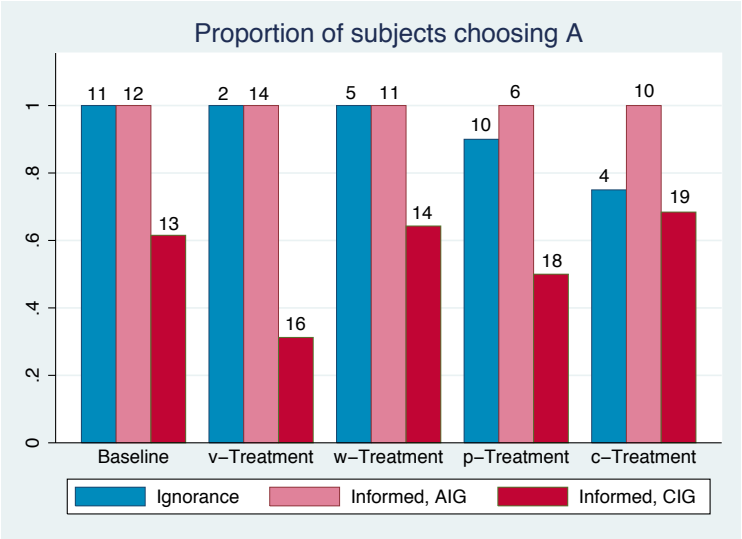

Figure 3: Proportion choosing A by treatment. Number of observations at the top of the bar.

Motives revealed in questionnaire responses

After the experiment, the subjects filled out a short questionnaire, in which dictators were asked: "Did you reveal the payoffs of Player Y? Why (not)?". Although these answers are not incentivized, the questionnaire nevertheless provides support for the importance of the motives put forward by the theory.[14]

First of all, people seem to differ in their fairness concerns, along the lines of theory. Of the 32 people who chose to remain ignorant in the experiment, the majority (55%) says that they do not care much about the payoffs of Player Y and/or that more information would not have mattered for their decision. On the other hand, many (34%) of those who reveal explicitly say they did so out of fairness considerations, and their choices are consistent with this. Second, conscience or self-image seem to play a role: 5 of those who chose to remain ignorant answered that they did so not to have a bad conscience ("schlechtes Gewissen"), and one subject chose ignorance to avoid 'having to be nice'. Moreover, there are 7 informed subjects who chose the selfish action in the CIG and report (falsely) that they had been ignorant. My interpretation is that these subjects claim ignorance to excuse their selfish actions towards the experimenter, which supports the idea that ignorance is seen to be image-enhancing.

Finally, many participants who informed themselves (32%) cite curiosity as a reason. A common response is "I intended to play A in any case, but I revealed because I was interested in what Player Y would earn." This suggests that people may have something resembling a

[14] About 11% of the subjects only replied yes/no without explanation. Percentages in this section are taken over those who provided some kind of explanation.

preference for information.[15] Note however that the percentage citing curiosity is higher in the CIG (33%) than in the AIG (25%), which may indicate that curiosity is used as an ex-post rationalization.

4.4 Discussion

With two exceptions that I discuss below, the experimental data confirm the hypotheses. Most importantly, the decision to remain ignorant is dependent on subsequent payoffs in ways that are predicted by the model.

What can we say about the motives underlying the observed behavior? Fairness considerations have a prominent presence in the questionnaire responses, and their existence is in line with a large experimental literature. However, social preferences alone (i.e. if $\mu = 0$, or other models based on payoffs distributions only) cannot explain the fact that some people are willing to pay to remain ignorant. Such models would predict that low types, who always play A, would be indifferent about being ignorant or not, and would therefore never be willing to pay for ignorance. In addition, fairness concerns cannot explain the experimental results in some of the previous literature (see Section 5).

With respect to image concerns, the fact that the experiment is anonymous but not double blind means that both self-image and image towards the experimenter may be relevant. The questionnaire responses provides evidence that both play a role. To see if social image concerns towards the receiver (Player Y) may have an additional effect and to check the robustness of the results, I conducted three sessions without anonymity. After completion of the instructions, but before the start of the actual experiment, the experimenter asked the subjects in each pair to stand up and briefly look at each other. It was thus common knowledge to both the members of the pair who they were playing with. In this "face-to-face" treatment, levels of ignorance were slightly higher than in the baseline, but none of the behavioral differences were significant.[16] Social image concerns towards the receiver do not have a large additional effect in this experiment.

Two results are contrary to the predictions of the model. First, varying the probability of

[15] This is consistent with earlier evidence. In one treatment of GM's version of the DKW experiment, choices for both the CIG and AIG are elicited using the strategy method. Afterwards, people choose whether they wish to learn which game was actually played. In this case, information only serves to reveal ex-post payoffs and not as an excuse, since actual choices have been revealed. GM finds that ignorance drops considerably relative to the baseline, consistent with "a general desire to resolve uncertainty or with a desire to learn the outcome of ones choice" (Grossman 2010:10).

[16] In the "face-to-face" treatment, 12 out of 35 (34%) subjects remained ignorant, 12 out of 12 (100%) uninformed subjects chose A, 11 out of 11 (100%) subjects in the AIG game chose A, and 6 out of 12 (50%) subjects chose A in the CIG.

harming others does not have any noticeable effect on behavior. This may point to a defect of the model, or it may be that the probabilities were not salient enough in the minds of the subjects. Since the model prediction is the result of two opposing forces, any effect was likely to be small.

Second, the theory predicts that people are deterred from choosing A in the CIG by low image payoffs, which is clearly rejected by the data. This is somewhat surprising, since there is strong evidence in the existing literature that the social appropriateness of this action is rated very low, in line with the predictions of the model (see Section 5). So why do people nevertheless engage in it?

One explanation is that the model is incomplete, and omits relevant motives of the subjects. Rather than speculating about additional motives, it is also possible to explain the data by introducing more heterogeneity. Specifically, the model is consistent with the data if we assume that there is a fraction subjects do not care about fairness or image payoffs (i.e. the infamous *homo economicus* with $\theta = \mu = 0$).

To see this, note that a homo economicus will always play A, and is therefore indifferent between acquiring information or not. We may assume that a number of them choose information and subsequently play A. This means that in contrast to the equilibrium in Proposition 1, choosing A in the CIG is now no longer an off-equilibrium action. Note that this does not change the equilibrium strategy of the other agents. Because only a homo economicus with $\theta = 0$ would choose to be knowingly nasty, image payoffs associated with choosing A in the CIG have not changed. To explain the data quantitatively, we would need homo economicus to have a share of at least 44% in the population.[17] This is only a little higher than the 36% found by Engel (2011) in a meta-study of the dictator game.

A second explanation may be that not all subjects coordinate on playing the same (strategic ignorance) equilibrium. Note that there exists also an equilibrium where subjects pool on information acquisition, supported by low off-equilibrium beliefs for ignorance. In this equilibrium, low types choose selfishly when the news is bad. Although evidence from existing experiments does not support the beliefs associated with this equilibrium (see below), it may be that some subjects play in accordance with this equilibrium strategy profile.

[17] Of subjects who play the CIG knowingly, 44 out of 80 (55%) play A. Assuming that the homo economicus has an equal presence in the CIG and AIG (since the game was chosen randomly), one can estimate that 29 out of 53 subjects in the AIG condition should be a homo economicus. Consistent with questionnaire answers, we can further assume that curiosity functions as a tie-breaking rule that drives the indifferent homo economicus to acquire information. Thus, we arrive at a fraction $\frac{44+29}{165} \approx .44$.

5 Relation to existing experiments

The theory presented in Section 3 fits the results of existing experiments on strategic ignorance. First, there is strong evidence supporting the ranking of (off-) equilibrium beliefs implied by Proposition 1. In Krupka and Weber (2008, Figure 4, reproduced in Appendix C), subjects provide normative evaluations of the different actions in DKW and in this study. Acquiring information and subsequently choosing the unfair action A is rated very socially inappropriate. Remaining uninformed and choosing A is considered only moderately inappropriate.

Further evidence for the equilibrium beliefs comes from a bargaining experiment by Conrads and Irlenbusch (2011). A proposer proposes one of two payoff distributions, A or B. A responder can accept the proposal or punish the proposer by rejecting it, in which case both players get nothing. The proposer knows her own payoffs from A and B, but treatments differ with respect to the information she has or can obtain about the responder's payoffs. The results show that if the proposer chooses to be informed and makes an unfavorable proposal, it is accepted 40% of the time. Subject who choose to remain ignorant and make unfavorable offers are accepted 58% of the time, indicating that strategic ignorance reduces 'punishment' by the responder. In addition, the experimenters also asked non-revealers what proposal they would have made under full information. Their stated choices are substantially more selfish than those of the revealers, indicating that fair-minded subjects sort into revelation.

Second, Proposition 1 accounts for the finding of DKW that 44% of the subjects chose to remain ignorant. DKW also finds that if there is no option to remain ignorant and subjects are asked to make a choice in the conflicting interest game directly, only 26% of subjects choose the selfish option.[18] Although the robustness of this result is debated[19], it suggests that some of the people who act fairly under complete information choose ignorance ignorance if given the chance. This is hard to explain with standard 'social preference' models, which predict that those who demonstrate a preference for fairness will also reveal. Remark 1 shows that by including image concerns, the model can replicate the observation by DKW.[20]

Fong and Oberholzer-Gee (2011) studies a $10 dictator game where the dictator knows that

[18] In the replication studies, the corresponding percentages of ignorance and selfishness in the full-information treatment were 43% and 22% in Larson and Capra (2009), and 46% and 35% in Grossman (2010) respectively.

[19] (Grossman 2010) finds that when ignorance is no longer a default setting but has to be chosen actively, the percentage of non-revealers is similar to those who choose selfishly in the full-information setting. However, the study also concludes that strategic ignorance is not just a default effect, since many people choose to reveal (and overcome the default) when they are able to do so *after* having chosen action A or B.

[20] Note that Remark 1 holds equilibrium beliefs constant when subjects become (exogenously) informed. This may be theoretically questionable when predicting cross-treatment differences, but is again in line with the results of Krupka and Weber (2008), who show that selfish behavior is valued equally negatively, no matter if the information was acquired endogenously or provided exogenously by the experimenter.

the real world recipient is equally likely to be deserving (a poor, disabled person), or not (a drug addict). In the 'choice' treatment, subjects can purchase knowledge about the recipient's type at a cost of \$1. This is contrasted with an 'exogenous no-info' treatment, where dictators do not have the option to purchase information, and an 'exogenous info' treatment in which dictators are informed exogenously. The results show that ignorant dictators are less generous if their ignorance was chosen than if it was imposed exogenously. In addition, dictators who know they are paired with a disabled person are more generous if they obtained the information by choice. The model in this paper explains both findings as it predicts that the decision to buy information in the choice treatment leads to sorting of generous and non-generous types.

Finally, the study has some overlap with 'exit'-experiments (Dana *et al.* (2006), Lazear *et al.* (2010), Broberg *et al.* (2007), Jacobsen *et al.* 2011). These experiments use a standard dictator game as a control treatment. In a manipulation, after the subjects have specified their choice in the dictator game, they have the option to exit from the game. If they do so, the dictator choice is not implemented and subjects receive the dictator endowment minus a small fee. Moreover, the observer will never learn that a dictator game was to be played. In this game, Lazear *et al.* (2010) find the same taxonomy of types as in the discussion of Figure 1. Specifically, the exit-studies find that roughly one third of the people choose to pay the fee and exit the dictator game if they have the possibility to do so (the "reluctant sharers").

6 Conclusion

To my knowledge this study presents the first formal theory of strategic ignorance in the context of pro-social behavior, a pervasive phenomenon that lies at the root of collective and individual moral failures. The model shows that strategic ignorance is a way to have your cake and eat it too. The strategically ignorant benefit materially from their behavior, but do not bear the full image cost, since they can credibly claim that if they had known that there was a trade-off, they would have done the right thing.

The experiment shows that the amount of strategic ignorance depends on the payoffs of the subsequent interaction. People are most likely to look the other way when material gains from self-interested actions are high. When potential costs for other parties rise, there is some evidence that people are more inquisitive into the consequences of their actions for others. However, when the probability of harming others through self-interested decisions goes up, the level of ignorance remains stable. A minority of people is willing to pay to remain ignorant. Note that these results are obtained in an abstract environment, and the choice for ignorance is presented in a very explicit way. By contrast, real world choices are embedded in a social and cultural context that may facilitate self-serving rationalizations (Norgaard 2006). Therefore, I

believe the experimental results are more likely to under- than to overstate the occurrence of strategic ignorance.

Assuming some external validity, the experiment supports the constructivist conclusion that people are more likely to know 'convenient' facts. This has some important implications. First, it may be difficult to make people understand the social benefits of sustainable or welfare enhancing practices if these violate their self-interest. An example are beliefs about climate change. Norgaard (2006) conducts participatory research, media analysis and in-depth interviews to study attitudes to climate change in Norway, a country that derives much of its wealth from oil revenues. She finds that non-responsiveness to climate change stems in large part from self-interested denial.

> "Given that Norwegian economic prosperity and way of life are intimately tied to the production of oil, denial of the issue of climate change serves to maintain Norwegian global economic interests. [...] Within this context to "not know" too much about climate change maintains the sense that if one did know one would have acted more responsibly."
>
> Norgaard (2006: 366).

More generally, people will be reluctant to acquire or process information about potential externalities associated with the consumption of cheap products. The higher the price difference between those products and more 'ethical' (i.e. fair-trade or eco-labeled) goods, the more pronounced this tendency will be. On the upside, the results of the cost treatment suggest that the authorities can increase awareness by making information about welfare costs harder avoid, e.g. by public awareness campaigns.

A second implication is that authorities with the power to change payoffs or relative prices will affect not only the behavior of individuals, but also their beliefs. For example, authoritarian dictators can raise the price of opposition by threatening punishment. This may induce people to look away from the regime's wrongdoings, in order to avoid feeling obliged to engage in costly acts of protest. Cohen (2001) provides specific examples of this phenomenon from the Argentinean junta and the National Socialist era in Germany.

In an ideal world, people inform themselves as well as possible in order to take socially optimal actions. In reality, people are aware that beliefs are reasons for action, and rationally and willfully bias their belief systems to support the behavioral patterns from which they benefit. How such biases are manipulated by authorities and other social actors is a fascinating topic for future research.

7 References

Aghion, Philip and Jean Tirole. 1997. "Formal and Real Authority in Organizations", *Journal of Political Economy*, 105:1, 1-29.

Andreoni, James and Douglas B. Bernheim. 2009. "Social Image and the 50-50 Norm: a Theoretical and Experimental Analysis of Audience Effects", *Econometrica*, 77:5, 1607-1636.

Baumeister, R. 1998. The Self", in *The Handbook of Social Psychology*, D. Gilbert, S. Fiske, and G. Lindzey, eds. Boston, MA: McGraw-Hill.

Bem, D. J. 1972. "Self-Perception Theory", in L. Berkowitz, ed., *Advances in Experimental Social Psychology, Vol. 6*, 1-62. New York: Academic Press.

Bénabou, Roland and Jean Tirole. 2002. "Self-Confidence and Personal Motivation", *Quarterly Journal of Economics*, 117:3, 871-915.

Bénabou, Roland and Jean Tirole. 2004. "Willpower and Personal Rules," *Journal of Political Economy*, 112:4, 848-87.

Bénabou, Roland and Jean Tirole. 2006. "Incentives and Pro-social Behavior", *American Economic Review*, 96:5, 1652-78.

Bénabou, Roland and Jean Tirole. 2011. "Identity, Morals and Taboos: Beliefs as Assets", *Quarterly Journal of Economics*, forthcoming.

Bersoff, David M. 1999. "Explaining Unethical Behavior Among People Motivated to Act Prosocially", *Journal of Moral Education*, 28:4, 413-28.

Bodner, R. and D. Prelec. 2003. "Self-signaling and Diagnostic Utility in Everyday Decision Making", in I. Brocas and J. Carrillo eds. *The Psychology of Economic Decisions. Vol. 1: Rationality and Well-being*, Oxford University Press, 105-126.

Broberg, Tomas, Tore Ellingsen, and Magnus Johannesson. 2007. "Is generosity involuntary? *Economic Letters*, 94:1, 32-37.

Caplin, Andrew and John Leahy. 2001. "Psychological expected utility theory and anticipatory feelings", *Quarterly Journal of Economics*, 55-79.

Carillo, Juan D. and Thomas Mariotti. 2000. "Strategic Ignorance as a Self-Disciplining Device", *Review of Economic Studies*, 67:3, 529-44.

Cohen, Stanley. 2001. *States of denial.* Cambridge: Blackwell.

Conrads, Julian and Bernd Irlenbusch. 2011. "Strategic Ignorance in Bargaining", *IZA discussion paper* 6087.

Crémer, Jacques. 1995. "Arms Length Relationships", *Quarterly Journal of Economics*, 110:2, 275-95.

Dana, Jason D., Daylian M. Cain, and Robyn M. Dawes. 2006. "What you dont know wont hurt me: Costly (but quiet) exit in dictator games", *Organizational Behavior and Human Decision Processes*, 100:2, 193 - 201.

Dana, Jason, Kuang, Jason, and Roberto Weber. 2007. "Exploiting Moral Wriggle Room: Experiments Demonstrating an Illusory Preference for Fairness," *Economic Theory*, 33:1, 67-80.

Dominguez-Martinez, Silvia, Sloof, Randolph and Ferdinand von Siemens. 2010. "Monitoring your Friends, Not your Foes: Strategic Ignorance and the Delegation of Real Authority", Tinbergen Institute Discussion Paper TI 2010-101/1.

Ellingsen, Tore, and Magnus Johannesson. 2008. "Pride and Prejudice: The Human Side of Incentive Theory", *American Economic Review*, 98, 9901008.

Engel, Christoph. 2011. "Dictator Games: A Meta Study", *Experimental Economics*, forthcoming.

Fershtman, Chaim, Uri Gneezy and Moshe Hoffman. 2011. "Taboos and Identity. Considering the Unthinkable", *American Economic Journal: Micro-economics*, 3, 139164.

Fischbacher, Urs. 2007. "z-Tree: Zurich Toolbox for Ready-made Economic Experiments", *Experimental Economics*, 10:2, 171 - 78.

Fong, Christina, and Felix Oberholzer-Gee. 2011. "Truth in Giving: Experimental Evidence on the Welfare Effects of Informed Giving to the Poor", *Journal of Public Economics.* Forthcoming.

Grossman, Zachary. 2009. "Self-signaling versus Social Signaling in Giving", *UC Santa Barbara working paper.*

Grossman, Zachary. 2010. "Strategic Ignorance and the Robustness of Social Preferences", *UC Santa Barbara working paper.*

Hoffman, Elizabeth, Kevin A. McCabe, and Vernon L. Smith. 1996. "Social distance and other-regarding behavior in dictator games", *American Economic Review*, 86:3, 653-60.

Horwitz, Gordon J. 1991. *In the Shadow of Death: Living Outside the Gates of Mauthausen*, London: I.B. Taurus.

Jacobsen, Karin J., Kari H. Eika, Leif Helland, Jo Thori Lind and Karine Nyborg. 2011. "Are nurses more altruistic than real estate brokers?", *Journal of Economic Psychology*, 32:5, 818-31.

Jewitt, Ian. 2004. "Notes on the Shape of Distributions", Mimeo, Oxford University.

Köszegi, Botond. 2006. "Ego Utility, Overconfidence, and Task Choice", *Journal of the European Economic Association*, 4:4, 673 - 707.

Krupka, Erin L., and Roberto A. Weber. 2008."Identifying Social Norms Using Coordination Games: Why Does Dictator Game Sharing Vary?" IZA Discussion Paper 3860.

Larson, Tara and Monica C. Capra. 2009. "Exploiting moral wiggle room: Illusory preference for fairness? A comment", *Judgment and Decision Making*, 4: 6, 467 - 74.

Lazear, Edward P., Ulrike Malmendier, and Roberto A. Weber. 2010. "Sorting, Prices, and Social Preferences", manuscript.

Matthey, Astrid and Tobias Regner. 2011. "Do I really want to know? A cognitive dissonance-based explanation of other-regarding behavior", *Games*, 2, 114-135.

Mazar, Nina, On Amir, and Dan Ariely. 2009. "The Dishonesty of Honest People: A Theory of Self-Concept Maintenance", *Journal of Marketing Research*, XLV, 633 - 44.

Möbius, Markus M., Muriel Niederle, Paul Niehaus and Tanya Rosenblat. 2011. "Managing Self-Confidence: Theory and Experimental Evidence", Mimeo, Stanford university.

Murnigham, Keith, John M. Oesch and Madan Pillutla. "Player Types and Self-Impression Management in Dictatorship Games: Two Experiments", *Games and Economic Behavior*, 37, 388-414.

Norgaard, Kari Marie. 2006. ""We Don't Really Want to Know". Environmental Justice and Socially Organized Denial of Global Warming in Norway", *Organization and Environment*, 19:3, 347-70.

Nyborg, Karine. 2011. "I Dont Want to Hear About it: Rational Ignorance among Duty-Oriented Consumers", *Journal of Economic Behavior and Organization*, 79, 263-74.

Sereny, Gitta. 1996. *Albert Speer: his Battle with Truth.* London: Picador.

Simon, William H. 2005. "Wrongs of Ignorance and Ambiguity: Lawyer Responsibility for Collective Misconduct", *Yale Journal of Regulation*, 22:1, 1-35.

Speer, Albert. 1970. *Inside the Third Reich.* New York and Toronto: Macmillan.

Suissa, Samy and Jonathan J. Shuster. 1985. "Exact Unconditional Sample Sizes for the 2x2 Binomial Trial", *Journal of the Royal Statistical Society A*, 148, 317-27.

Tadelis. Steve. 2011. "The power of shame and the rationality of trust", Berkeley, Haas school of Business, manuscript.

Appendix A: Proofs

Proof of Proposition 1. This proof proceeds in three steps. First, given the proposed equilibrium engagement decisions conditional on information acquisition and content, I establish which types will acquire information. Second, I confirm that the subsequent engagement decisions are indeed optimal, given proposed off-equilibrium beliefs. Third, I discuss whether the proposed off-equilibrium beliefs are reasonable. I employ the tie-break condition that all indifferent types acquire information/abstain.

Step 1. Let $\theta^* \in (0,1)$ be the threshold type who is indifferent between acquiring information and not. To ease notation, I define some ex-post equilibrium beliefs $\phi(I,\sigma,a;s) = E[\theta \mid I,\sigma,a;s]$ as well as a useful function

$$\phi_- = \phi_-(\theta^*) \equiv E[\theta \mid \theta < \theta^*] = \int_0^{\theta^*} \frac{\theta dF(\theta)}{F(\theta^*)} \tag{A.1}$$

$$\phi_+ = \phi_+(\theta^*) \equiv E[\theta \mid \theta \geq \theta^*] = \int_{\theta^*}^1 \frac{\theta dF(\theta)}{1-F(\theta^*)} \tag{A.2}$$

$$\Delta(\theta^*) \equiv \phi_+(\theta^*) - \phi_-(\theta^*) \tag{A.3}$$

Some of its properties of $\Delta(\theta^*)$ have been derived by Jewitt (2004) and Bénabou and Tirole (2006). We know what the agent will do in equilibrium upon (not) acquiring information, so we can derive that θ^* is given implicitly by the fixed point equation

$$\begin{aligned} Eu(\text{inform}) &= Eu(\text{not inform}) \\ (1-p)(v+\mu\phi_+) + p\mu\phi_+ - c &= v + \mu\phi_- - pw\theta^* \\ \theta^* &= \frac{pv + c - \mu\Delta(\theta^*)}{pw} \end{aligned} \tag{A.4}$$

It is easy to check that all types $\theta < \theta^*$ acquire information and all types $\theta \geq \theta^*$ abstain. It remains to check that θ^* is in the interior. Since we assumed $v < w$ it follows that $\theta^* < 1$ if $c \leq \mu\phi_+ \equiv \bar{c}$. Below I verify that $\theta^* > 0$.

Existence follows from the continuity of both sides of (A.4). Moreover, θ^* is unique if the left hand side of (A.4) increases faster than the right hand side. Taking the total derivative wrt θ^* yields that θ^* is unique if for all $\theta^* \in [0,1]$

$$\frac{d\Delta(\theta^*)}{d\theta^*} > -\frac{pw}{\mu}. \tag{A.5}$$

Intuitively, (A.5) implies that the density $f(\theta)$ should not increase too steeply anywhere on its domain, which we assume to be the case. With the stability condition (A.5) in hand, the comparative statics in Proposition 1c) are obtained by implicit differentiation of (A.4).

Step 2. We now verify whether the proposed engagement decisions are optimal, in case a) an informed agent observes $\sigma = \sigma_w$, b) an informed agent observes $\sigma = \sigma_0$, and c) an agent is uninformed.

Step 2a) If $\sigma = \sigma_w$, an agent of type θ will abstain iff

$$\begin{aligned} u(\text{abstain} \mid I=1, \sigma=\sigma_w; \theta^*) &\geq u(\text{engage} \mid I=1, \sigma=\sigma_w; \theta^*) \\ \mu\phi_+(\theta^*) &\geq v + \mu\phi(1,\sigma_w,1) - \theta w \\ \theta &\geq \frac{v - \mu(\phi_+ - \phi(1,\sigma_w,1))}{w} \\ \theta &\geq \frac{v - \mu\phi_+}{w} \equiv \bar{\theta} \end{aligned} \tag{A.6}$$

where in the last step I assumed that off-equilibrium beliefs satisfy $\phi(1,\sigma_w,1) = 0$. Thus, it is clear that in equilibrium all $\theta < \bar{\theta}$ engage and all $\theta \geq \bar{\theta}$ abstain. A condition similar to (A.5) guarantees uniqueness of $\bar{\theta}$. Note that $\bar{\theta} > 0$ if the lowest type $\theta = 0$ finds it optimal to engage, which is the case iff

$$v > \mu\phi_+. \tag{A.7}$$

This is satisfied, since I assumed that $v > \mu$.

We need to verify that $\bar{\theta} \leq \theta^*$, because only then will all types who observe $\sigma = \sigma_w$ indeed abstain. Comparing (A.6) and (A.4) and doing some algebra yields that $\bar{\theta} < \theta^*$ if and only if

$$p > \frac{\mu(\phi_+ - \phi_-) - c}{\mu\phi_+} \equiv \bar{p} \tag{A.8}$$

Since $\phi_+ > \phi_-$, it is clear that $\bar{p} < 1$ as long as $c > -\mu\phi_- \equiv \underline{c}$. Furthermore, $v > \mu$ implies that $\bar{\theta} > 0$, which together with equation (A.8) implies that $\theta^* > 0$.

Step 2b) Next, consider the case in which $\sigma = \sigma_0$. It is optimal for the agent to engage iff

$$\begin{aligned} u(\text{engage} \mid I = 1, \sigma = \sigma_0; \theta^*) &> u(\text{abstain} \mid I = 1, \sigma = \sigma_0; \theta^*) \\ v + \mu\phi_+ &> \mu\phi(1, \sigma_0, 0) \\ v &> \mu(\phi(1, \sigma_0, 0) - \phi_+) \end{aligned} \tag{A.9}$$

which is always satisfied since $v > \mu$.

Step 2c) Consider now the non-informed agent. She will engage if

$$\begin{aligned} Eu(\text{abstain} \mid I = 0; \theta^*) &< Eu(\text{engage} \mid I = 0; \theta^*) \\ \mu\phi(0, \emptyset, 0) &< v + \mu\phi_- - p\theta w \\ \theta &< \frac{v - \mu(\phi(0, \emptyset, 0) - \phi_-)}{pw} \equiv \tilde{\theta} \end{aligned} \tag{A.10}$$

We need to check if all types who are non-informed do indeed satisfy this condition, which is the case if $\theta^* < \tilde{\theta}$. Some algebra shows that this is satisfied iff $\mu\phi(0, \emptyset, 0) < v(1-p) + \mu\phi_+ - c + \mu\Delta(\theta^*)$, which I assume to be the case.

Step 3. We need to check that the assumptions on off-equilibrium beliefs that I made in step 2a) and 2c) above, are not unreasonable. The standard refinement for such games, the intuitive criterion (Cho and Kreps 1987) does not technically apply to this game, because payoffs depend directly on off-equilibrium beliefs. However, we can use logic akin to the intuitive criterion: I require that off-equilibrium beliefs upon observing the deviation (I', σ', a') place zero weight on type θ, if equilibrium payoffs dominate the deviation payoffs when observer beliefs equal $E[\theta|I', \sigma', a'] = 1$ (i.e. are maximally optimistic about the sender's type). This guarantees that off-equilibrium beliefs do not place weight on types that would never deviate, even if this would give them the best possible image.[21]

Consider $\phi(1, \sigma_w, 1) = 0$ from step 2a). Type $\theta = 0$ would be willing to deviate if $\phi(1, \sigma_w, 1) \geq \phi_-$, because this would bring the expected image payoff from acquiring information above the equilibrium value ϕ_-, while the expected material payoffs would be v in both cases. Since $\phi_- < 1$, setting $\phi(1, \sigma_w, 1) = 0$ does not violate the requirement.

Finally, considering $\phi(0, \emptyset, 0)$, the lowest type willing to deviate when $\phi(0, \emptyset, 0) = 1$ is given by $\theta = \frac{v - \mu(1-\phi_-)}{pw}$. It is easy to show that the proposed off-equilibrium beliefs satisfy the requirement if v and w are large and c is relatively small. (Note that one can show that if c and p are low, the deviation is dominated by the expected payoffs of getting information for all types, so no type should be expected to follow this deviation.) ■

Proof of Remark 1. Consider now a type who becomes exogenously informed that the news was bad, keeping constant the equilibrium beliefs. The proof follows from the above, where we have shown that $\bar{\theta} < \theta^*$ if $p > \bar{p}$. ■

[21] Note that other authors have similarly applied standard refinements to games where people care about other's beliefs. Andreoni and Bernheim (2009) and Ellingsen and Johannesen (2008) both apply the D1 criterion, and evaluate sets of off-equilibrium beliefs by the observer for which the sender would be willing to deviate. I thank Martin Dufwenberg for pointing this out to me.

Appendix B: Summary Statistics

Treatment	N	Chose ignorance	Chose A		
			Uninformed	AIG	CIG
Baseline	36	31% (11/36)	100% (11/11)	100% (12/12)	62% (8/13)
v-treatment	32	6% (2/32)	100% (2/2)	100% (14/14)	31% (5/16)
w-treatment	30	17% (5/30)	100% (5/5)	100% (11/11)	64% (9/14)
p-treatment	34	29% (10/34)	90% (9/10)	100% (6/6)	50% (9/18)
c-treatment	33	12% (4/33)	75% (3/4)	100% (10/10)	68% (13/19)
Total	165	19% (32/165)	94% (30/32)	100% (53/53)	55% (44/80)

Table 2: Dictators' decisions

Appendix C: Normative evaluations of actions in the DKW study

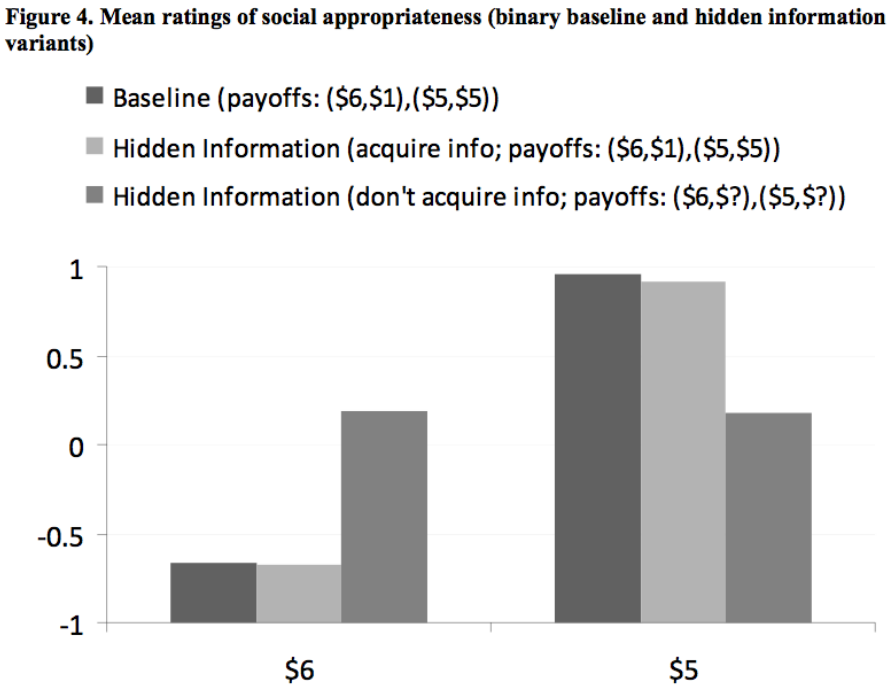

Figure 4: Figure 4 from Krupka and Weber (2008: 40). Measuring normative evaluations of actions in the DKW study. Evaluations are measured on a 4 point scale "very socially inappropriate", "somewhat socially inappropriate", "somewhat socially appropriate", and "very socially appropriate", which is then converted to a numerical scale.

Appendix D: Instructions and screenshots

This is a translation of the instructions for the baseline treatment (the original instructions are in German). The instructions for the other treatments differ only in small and predictable details, and are omitted here.

Introduction

INSTRUCTIONS: PLEASE READ VERY CAREFULLY. IF YOU HAVE A QUESTION, PLEASE RAISE YOUR HAND AND WAIT FOR ASSISTANCE.

Welcome! You will participate in an economic experiment, funded by several research institutitons. In what follows we will use male pronouns to shorten the text. We ask for your understanding.

Please read the instructions carefully. They contain everything you need to know for the experiment. If you have questions, please raise your hand and we will come to your seat to answer your question. During the experiment it is forbidden to talk to the other participants.

Every participant receives a show-up fee of 4 euros, which will be paid at the end of the experiment. During the experiment you can earn additional money. During the experiment we do not speak of money, but of points. Your payoffs will be represented in points, where

1 Point= 0,10 Euro

At then end of the experiment you will receive the money that you made during the experiment plus the 4 euro show-up fee in cash. The payment will be made privately, and no other participant will see how much you are paid.

OK

Instructions

Description of the game

During the experiment you are matched with another participant in a group of two. The other participant is assigned randomly by the computer. Neither before nor after the experiment do you learn the identity of the other participant. Nor does the other participant learn your identity. The choices in this experiment are made anonymously.

Each of you is assigned a role in the experiment. You are either player X or player Y. In each pair there is one player X and one player Y. The difference between these roles will be described below.

The game that you play in pairs, looks as depicted below. Player X will choose one of two options "A" or "B". Player Y will not make any choice. The payments both players receive depend only on the choice of player X.

The numbers in the table are the payments players receive. The payments in this table were chosen only to demonstrate how the game works. In the actual game, the payments will be different. For example, if player X chooses "B", then we should look in the bottom square for the earnings. Here, Player X receives 3 points and Player Y receives 4 points.

Player X chooses	Player X receives	Player Y receives
A	1	2
B	3	4

OK

Control questions

Please answer the following control questions. Your answers only serve to establish if you understood the game and do not influence the payments you will receive.

Player X chooses	Player X receives	Player Y receives
A	1	2
B	3	4

In this example, if Player X chooses "B" then:

Player X receives

Player Y receives

In this example, if Player X chooses "A" then:

Player X receives

Player Y receives

OK

Instructions

In the actual experiment you will play one of the two games pictured below. The payment of player X is the same in both games. The only thing that distinguishes both games is the income of Player Y, which has been swapped between the games. If Player X chooses "A" he receives the highest payment of 100 points in both games. In the first game, Player Y will then receive a payment of 10 points. In the second game Player Y will receive a payment of 60 points. If Player X chooses "B" he receives the lowest payment of 60 points in both games. In the first game, Player Y will then receive a payment of 60 points. In the second game Player Y will receive a payment of 10 points.

Which game you will actually be playing will not be revealed publicly. The payoffs of Player Y will initially be represented by a questionmark. The actual game is determined randomly by the computer before the start of the experiment. Both games have an equal probability (50%) of being played.

Before Player X chooses "A" or "B", he can find out which game is actually being played, by clicking a button "REVEAL GAME". Player X can also click a button "DON'T REVEAL", in which case he does not learn anything. This choice will be made anonymously, Player Y does not know whether Player X knows which game is being played. Player Y will not have a possibility to find out which game is being played.

While Player X makes his choice, Player Y will be asked to answer a few questions.

Game 1

Player X chooses	Player X receives	Player Y receives
A	100	10
B	60	60

Game 2

Player X chooses	Player X receives	Player Y receives
A	100	60
B	60	10

OK

Control questions

Please answer the following control questions. Keep in mind that both games are equally likely to be played.

Which option gives Player X his or her highest payment in both games?
- A
- B

If Player X chooses B, then Player Y receives
- 60 points
- 10 points
- either 60 or 10 points

OK

Player X chooses	Player X receives	Player Y receives
A	100	10
B	60	60

Player X chooses	Player X receives	Player Y receives
A	100	60
B	60	10

Instructions

Time remaining 114

Summary:

If you are player X, you will decide in one of the two games below.

The actual game was randomly decided before the experiment by the computer. Each game has a 50% probability of being played.

You choose whether to find out which of the games is actually being played.

Subsequently you will choose "A" or "B".

At the end of the experiment we will pay both players privately.

50% Probability

Player X chooses	Player X receives	Player Y receives
A	100	10
B	60	60

50% Probability

Player X chooses	Player X receives	Player Y receives
A	100	60
B	60	10

OK

Experiment

DON'T REVEAL

Please choose if you wish to reveal which of the two games is actually being played. Player Y will not observe this decision.

REVEAL

Player X chooses	Player X receives	Player Y receives
A	100	?
B	60	?

Experiment

A

Please make your choice.

B

Player X chooses	Player X receives	Player Y receives
A	100	10
B	60	60

Experiment

You will not make a choice. However, we ask that you pick which option, A or B, you would choose if you were player X in each of the two versions of the game. For each version, please select one of the two options. Then click OK to confirm your choices. The choices you make will not affect the payment you receive.

In the game on the left, I would choose ○ A ○ B

In the game on the right, I would choose ○ A ○ B

OK

Player X chooses	Player X receives	Player Y receives
A	100	10
B	60	60

Player X chooses	Player X receives	Player Y receives
A	100	60
B	60	10