Self-image and strategic ignorance in moral dilemmas

Grossman, Z.; van der Weele, J.

Self-Image and Strategic Ignorance in Moral Dilemmas*

Zachary Grossman† Joël van der Weele‡

July 11, 2013

Abstract

Avoiding information about adverse welfare consequences of self-interested decisions, or *strategic ignorance*, is an important source of corruption, anti-social behavior and even atrocities. We model an agent who cares about self-image and has the opportunity to learn the social benefits of a personally costly action. The trade-off between self-image concerns and material payoffs can lead the agent to use ignorance as an excuse, even if it is deliberately chosen. Two experiments, modeled after Dana, Weber, and Kuang (2007), show that a) many people will reveal relevant information about others' payoffs after making an ethical decision, but not before, and b) some people are willing to pay for ignorance. These results corroborate the idea that Bayesian self-signaling drives people to avoid inconvenient facts in moral decisions.

JEL-codes: D83, C72, C91.
Keywords: prosocial behavior, dictator games, strategic ignorance, self-signaling.

*We are grateful to Gary Charness, Aldo Rustichini, Jeroen van de Ven, Roland Bénabou, Mark Le Quement, Joel Sobel, Ferdinand von Siemens, Heiner Schumacher, Elisabeth Schulte, Karine Nyborg, Tobias Broer, Leonie Gerhards, Tobias Brünner, Lydia Geijtenbeek and numerous seminar participants for useful comments, and Sebastian Schäfer, Matthias Heinz, Karin Hettwer and Julija Kulisa for helping out with practical matters. Joël van der Weele is indebted to the Vereinigung der Freunde und Förderer der Goethe-Universität for financial support.

†University of California, Santa Barbara. Email: grossman@econ.ucsb.edu.
‡Goethe University Frankfurt. Email: vdweele@econ.uni-frankfurt.de.

"Living is easy with eyes closed."

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

1 Introduction

Willful avoidance of evidence about the negative social impact of one's own decisions, or 'strategic ignorance', plays an important role in political and corporate corruption, the perpetuation of conflicts and even genocide. Participants in the Watergate scandal are said to have shown "intense faith in the immunizing power of deliberate ignorance" (Simon, 2005, p.5). Top Enron executives argued at trial that their ignorance of any fraud should exonerate them, despite the fact that they had explicitly instructed Enron's lawyers to abstain from inquiries into Enron's accounting practices (Simon, 2005). Nowhere does one find a starker example of how strategic ignorance can lead to moral failure in all levels of society than in the Holocaust (Cohen, 2001). Many bystanders sought to remain ignorant of the atrocities committed in their communities (Bankier, 1996; Horwitz, 1991) and at the top of the hierarchy, Nazi government minister Albert Speer claimed at the Nuremberg trials that he did not know of the mass killings, although he admitted that he should and could have.[1]

Complementing the field evidence on the importance and prevalence of strategic ignorance is recent evidence from the laboratory. In experimental allocation problems where participants can choose whether or not to reveal costlessly how their choice will affect others, many choose not to know (Dana, Weber, and Kuang, 2007; Larson and Capra, 2009; Feiler, 2007; Matthey and Regner, 2011; Ehrich and Irwin, 2005). Paradoxically, these experiments also show that most people choose a fair outcome if they know that a selfish choice hurts the other participants. These findings can not be explained by standard models of distributional preferences, which predict that those who sacrifice to avoid the adverse consequences of their actions should also acquire costless information about these consequences. Nor can they be explained by social-image or reputational concerns, since subjects were anonymous and no participant observed whether or not the decision maker actually chose to be ignorant.

Instead, several authors suggest that ignorance serves the purpose of protecting the decision-maker's self-image (e.g., Dana, Weber, and Kuang, 2007; Bénabou and Tirole, 2006). Indeed, the notion of ignorance as a tool for maintaining happiness with one's own situation or behavior has long been acknowledged in cultural works. Essayist Michel de Montaigne

[1] In his autobiography (Speer, 1970), he writes about the extermination camps

> "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."

(1533 – 1592) wrote, "ignorance is the softest pillow on which a man can rest his head," and poet Thomas Gray (1716 – 1771) gave us the saying, "where ignorance is bliss, 'tis folly to be wise." In the context of the Holocaust, Bankier (1996) and Horwitz (1991) forcefully argue that bystanders' desire to remain ignorant of Nazi atrocities was aimed at least in part at avoiding accountability to their own conscience.[2] Self-image has long been accepted by psychologists as an important source of motivation (e.g., Bem, 1972; Baumeister, 1998; Bersoff, 2002) and was first incorporated into a dual-self signaling model by Bodner and Prelec (2003). Such models assume that a person cannot perfectly introspect or recall the motivation underlying her own behavior. An observer-self learns the individual's moral preferences from her actions, so the decision-maker-self strategically distorts her choices so as to manage her self-image.

Although plausible, a self-image explanation for the laboratory evidence raises the question how ignorance can succeed as an exonerating strategy if it is clear that the actor *chose* to be ignorant. Should not willfully chosen ignorance undermine its own strategic value? Indeed, the judge in the Enron case instructed jurors to be skeptical of ignorance-based excuses: "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..." Without understanding the value of deliberately chosen ignorance in signaling equilibrium, it is difficult to see how a self-image model can explain strategic ignorance.

In this paper, we make three contributions to this debate. The first is to show that a standard Bayesian preference-signaling model can explain self-image in moral dilemmas. The model, presented in Section 2, features a decision-maker who cares about her own material well-being and—to some degree—that of others, as well as her self-image as an altruistic person. The decision-maker first chooses whether to inform herself about the social benefits of an action that is personally costly and then chooses whether to engage in that action. We prove the existence of an equilibrium in which ignorance is strictly preferred by agents who care about their self-image, but are not very altruistic. Although consciously avoiding information leads to a diminished self-image, the stigma attached to knowingly engaging in harmful actions is worse, since this involves pooling only with completely selfish types. Thus, remaining ignorant serves to avoid a choice between two evils: to take a personally costly action, or to be revealed as immoral.

Our second contribution is to show that the ignorance equilibrium can explain several

[2]Bankier (1996) writes that "many deliberately sought refuge from the consciousness of genocide and tried to remain as ignorant as possible: because it salved their conscience. Knowledge generated guilt since it entailed responsibility, and many believed that they could preserve their dignity by avoiding the horrible truth". Horwitz (1991) interviewed residents of the Austrian village of Mauthausen who lived next to a cluster of concentration camps. Although presented with strong cues about the killings going on there, they made no effort to learn what was going on in the camp. 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".

puzzling behavioral patterns that have been observed in previous experiments. Most prominently, we show how the morally ambiguous option of remaining ignorant dilutes self-signaling incentives, which can explain the puzzling finding that some subjects who act prosocially under full information about the consequences of their actions will nevertheless not acquire such information voluntary.

Finally, we test new predictions following from our model that are not compatible with models of outcome-based preferences or social image concern. To do so, we provide two new experiments, reported in Section 3, which are modeled after Dana, Weber, and Kuang (2007, henceforth, DWK). Participants play a binary dictator game, in which the dictator is initially uninformed whether choosing the option with the higher personal payoff hurts or helps the recipient. As in previous experiments, the informational environment rules out social signaling.

In Experiment 1 we vary whether the dictator can obtain the information about the recipient's payoffs *before* she makes the allocation choice or only *after* she has already made a contingent allocation choice. Models of outcome-based motivations predict a lower ignorance rate in the first case, since the information can be used to make a more informed choice. In contrast, we observe a lower ignorance rate in the second case. In other words, many subjects like to know the payoffs of the other player, but only *after* they made their allocation choice. The self-signaling model explains this result, since it predicts that ignorance loses its strategic value to protect self-image after a moral choice is made.

In Experiment 2 we vary the costs of obtaining information. We show that people choose ignorance more frequently as the cost of information climbs, and that some people are willing to *pay* to remain ignorant.[3] The latter finding is inconsistent with models of outcome-based preferences, but not with the self-signaling model that predicts that in the ignorance equilibrium some types are strictly better off without information.

Taken together, we believe our theoretical and experimental results provide clear evidence that Bayesian self-signaling is an important driver of behavior in moral dilemmas. Although there have been many previous indications of the importance of self-image (Murnigham, Oesch, and Pillutla, 2001; Mazar, Amir, and Ariely, 2009; Fischbacher and Heusi, 2008; Gneezy, Gneezy, Riener, and Nelson, 2012), more direct tests have not found evidence in support of self-signaling models of giving behavior (Grossman, 2009).

Our paper relates to a growing theoretical and empirical literature on self-signaling. The model in Section 2 is adapted from that of Bénabou and Tirole (2006) and closely related to that of Grossman (2009). Despite the emphasis on *self*-image concerns as a driver of prosocial behavior, it is technically similar to the social-signaling models of Ellingsen and Johannesson (2008), Andreoni and Bernheim (2009), and Tadelis (2011). Bénabou and Tirole (2011) use

[3]Independent recent work by Cain and Dana (2012) finds similar results.

a similar signaling model to analyze strategic ignorance in the context of taboos, but do not explicitly model both the decision to remain ignorant and the decision to take an ethical action, and cannot compare the social image of behaving badly unknowingly with that of knowingly behaving badly, which is the focus of this paper. The reasoning used by Andreoni and Bernheim (2009, online appendix) to explain why some people are willing to pay to avoid a dictator game is similar to the reasoning we invoke: opting out helps to avoid the low image resulting from the decision not to share. However, they model the image related to the 'outside option' as exogenous, whereas the central theoretical exercise in this paper is to endogenously derive the image associated with ignorance.

Our analysis of strategic ignorance in moral dilemmas contributes to a broader literature of non-signaling models examining strategic ignorance in various contexts. Ignorance can be rational in the presence of anticipatory utility (Caplin and Leahy, 2001) and time-inconsistency (Carillo and Mariotti, 2000; Bénabou and Tirole, 2011), or among duty-oriented consumers contributing to a public good (Nyborg, 2011). It may also be employed by managers to better provide incentives, either by maintaining subordinates' de facto authority and thus, their incentive to gather information about project quality (Aghion and Tirole, 1997) or by reducing moral hazard (Crémer, 1995).[4] The common thread among these models is the analysis of ignorance as a way to avoid to making 'wrong' decisions in the future. Our model model shares this feature, in the sense that an agent's ignorance about the consequences of her actions serves to avoid large signaling investments in prosocial behavior that would be necessary to maintain a good self-image under full information.

Finally, our research also relates to a large literature on 'motivated cognition' in social psychology (e.g. Kunda, 1990) and a growing theoretical and empirical 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 (Köszegi, 1996; Möbius, Niederle, Niehaus, and Rosenblat, 2011; Eil and Rao, 2011). Although that literature has some intuitive parallels with this study, it does not address the moral trade-off that is at the heart of our analysis.

2 Signaling Equilibrium in a Model of Image Concerns

Why do people choose not to know the consequences of their own actions? To answer this question, we apply a model which combines preferences over material payoffs with a) an intrinsic concern for social welfare and b) a preference for a (self) image as a prosocial actor.

[4] Domingues-Martinez, Sloof, and von Siemens (2010) provide experimental evidence in support of Aghion and Tirole (1997).

2.1 The model

An agent chooses whether to take a prosocial action ($a = 1$) or not ($a = 0$). Taking the prosocial action incurs a material cost c and the agent is uncertain of the action's impact on social welfare, W. She knows that $W = w$, where $w > c$, with prior probability p, and that $W = 0$ with complementary probability.

Before the agent decides, she has the opportunity to inform herself ($I = 1$) about the true welfare impact at a cost k, or to remain uninformed ($I = 0$). We 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 a 'high signal' ($W = w$), σ_0 denotes a 'low signal' ($W = 0$), and with some abuse of notation $\emptyset$ denotes the case in which no information is acquired.

Timing. Thus, 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 about the level of W.
3. The agent chooses whether to take the prosocial action ($a = 1$) or not ($a = 0$).
4. The agent's actions a and the signal content σ are perceived by an observer and payoffs are realized.

Preferences. The agent has preferences that can be represented by the following utility function

$$u(\theta, a, I, \sigma) = a\,(\theta W - c) - kI + \mu E[\theta \mid \sigma, a; s]. \tag{1}$$

The first term denotes the material payoff from her outcome choice, which, if she takes the action, consists of the welfare benefit multiplied by a parameter θ, the 'type' of the agent, minus the costs, and otherwise is zero. The type θ is private information of the agent, and can be interpreted as the degree of altruism or prosocial motivation of the agent. The fact that low types care less about welfare ensures the single crossing-property that underlies the separating equilibrium in the next section. The second term of the utility function is the cost of information k. This cost could be negative, when information is presented in a way that makes it hard to avoid.

The last term denotes the payoffs from image concerns, where $E[\theta \mid \sigma, a; s]$ is the inference made by the observer about the type of the agent. This inference can be conditioned on the observed choice a, the content of the signal σ (and therefore the information acquisition decision), and the (equilibrium) strategy s of the agent. For simplicity, we 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 μ captures the importance of image concerns in the utility function of the agent.

We assume two categories of agents:

1. With probability ε the agent is a *homo economicus*, who only cares about her own material payoff, i.e. $\theta = \mu = 0$. The existence of such an agent is suggested by studies on prosocial behavior which typically find one-fifth to one-third of the population to be "selfish types" (e.g. Fischbacher, Gächter, and Fehr (2001), Kurzban and Houser (2005), and Burlando and Guala (2004)).

2. With probability $1 - \varepsilon$ she is a *social agent*, who cares about her image and about social welfare. Her type θ is distributed according to $F(\theta)$ with full support on $[0, 1]$. We assume that image concerns are small relative to material concerns ($0 < \mu < c$), which rules out the agent taking the prosocial action purely for image reasons.

This distribution is consistent with evidence from dictator games, where we typically see a large spike of selfish choices as well as a dispersed distribution of more generous behavior (see Engel, 2011, for an overview). Our main result, establishing the existence of an equilibrium with strategic ignorance, can be obtained in the absence of the *homo economicus*, i.e. $\varepsilon = 0$. However, $\varepsilon > 0$ is necessary for Proposition 2, which establishes that allowing decision-makers to remain ignorant of the consequences of their choices can actually lead to more selfish behavior overall

Self-image While preference-signaling models admit two distinct interpretations—as social-signaling model with an external observer or as a self-signaling model with an internal observer—our focus is squarely on the latter. Social-signaling is likely to play a role in decisions that can be expected to come under legal and therefore public scrutiny. In the Nuremberg, Enron, and Watergate trials, a central issue to the prosecution was who knew what when. In such cases, the assumption that the observer sees the information obtained by the agent would be quite strong. However, when the agent's self-image is at play, the importance of image concerns does not rely on observability of actions by others and it is natural to assume that the observer knows the content of the signal.

Bodner and Prelec (2003) introduced the dual-self signaling-model approach adopted by others such as Bénabou and Tirole (2006); Grossman (2009); Bénabou and Tirole (2011) and used herein. The decision-maker and observer roles are viewed as two aspects of a divided self, rather than separate people. The informed, decision-making self can access and act upon her preferences. Her aim is to impress an uninformed self who lacks such introspective

insight, and can be interpreted as a Smithian "imagined spectator" or "man in the breast", or a Freudian "super-ego". One view of self-signaling is as an attempt to influence the beliefs of a future self who, in retrospect, cannot recall the original motivation for the behavior.

2.2 Strategic ignorance in equilibrium

We focus on a perfect Bayesian equilibrium in which 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. We assume the tie-breaking rule that a fraction $0 < \alpha < 1$ of agents who are indifferent between acquiring information or not will acquire information. The proofs for the results in this section can be found in Appendix A. The main theoretical result of the paper is the following.

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

a) the homo economicus chooses $a = 0$ and acquires information if and only if $k \leq 0$,

b) all social types $\theta < \theta^*$ remain ignorant and choose $a = 0$, while all social types $\theta \geq \theta^*$ acquire information and choose $a = 1$ if and only if the signal is high.

To understand the equilibrium, consider the trade-off for a social agent of a relatively low type. 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 stemming from social preferences. On the upside, she avoids paying the cost of prosocial behavior c.

If she acquires information, she faces a lottery. When the signal is low, she can pool with the high types without any material sacrifice. When the signal is high, she faces a choice between two evils. She can take the prosocial action and obtain a high image at a price of c. Or she can be selfish and end up with the lowest possible image. The latter result obtains because either (if $k \leq 0$) she pools with the *homo economicus* or (if $k > 0$) beliefs for this off-equilibrium action are assumed to be 0.[5] When the probability of a high signal is sufficiently large, the low type strictly prefers to remain ignorant.

It is worth noting that the strategy of the marginal types just below θ^* specifies that these types behave pro-socially in the subgame where they receive a high signal. Thus, ignorance protects (self) image, because an ignorant person can credibly make the counterfactual statement that "if I had found out, I would have behaved prosocially." The reason she does

[5] With respect to refinements, because payoffs depend directly on beliefs, the standard refinements do not apply. However, in the appendix we show that these beliefs satisfy a refinement akin to the intuitive criterion. These negative beliefs could also be justified for $k > 0$ if the *homo economicus* would be willing to pay for information out of curiosity or if spiteful types ($\theta < 0$) existed who are keen on not being prosocial.

not find out is the anticipation that following a high signal, the strong image incentives to be prosocial would cause her to 'overinvest' (from an ex-ante perspective) in prosocial behavior.

In general, θ^* need not be unique, but a sufficient condition for uniqueness and stability of equilibrium is that

$$\frac{d\delta(\theta^*)}{d\theta^*} > -\frac{pw}{\mu}, \tag{2}$$

where $\delta(\theta^*)$ is the increase in (expected) image when choosing information rather than remaining ignorant.[6] Although our analysis concentrates on a 'strategic ignorance equilibrium', the model admits another pure-strategy 'no-ignorance equilibrium' with very low (off-equilibrium) image for ignorant agents. In this equilibrium the image associated with acting selfishly under ignorance is lower than doing so under full information. Therefore ignorance is never chosen by the social agents if k is small. Which, if any, of these equilibria will be played is an empirical question. However, the existence of the equilibrium described in Proposition 1 provides a rational explanation for the findings of previous experiments, as well as the experiments reported in this paper.

2.3 Relation to Previous Experimental Results

The equilibrium established in Proposition 1 can explain three behavioral patterns found in previous experiments.

Sorting. DWK finds higher rates of giving among dictators who reveal how their choice affects the recipient, than among the full sample of participants. Furthermore, Fong and Oberholzer-Gee (2011) demonstrate that subjects in a dictator game who purchase information about the 'worthiness' of a recipient (poor and disabled versus a drug addict) are on average more generous towards a worthy recipient than subjects who are confronted with the information exogenously. Conversely, those who choose not to acquire information are less generous than those who do not have the option to obtain it. This behavior is consistent with ignorance equilibrium. Since the homo economicus will remain ignorant if $k > 0$, positive information costs result in sorting: those who remain ignorant will be less generous types than those who inform themselves.

Exculpation. Krupka and Weber (2008) show that experimental subjects assign a higher 'social appropriateness' to a selfish action if the decision maker was (intentionally) unaware of the consequences for others. Similarly, Conrads and Irlenbusch (2011) find that unequal

[6] More formally, $\delta \equiv p\phi_w^1(\theta^*) + (1-p)\phi_0^0(\theta^*) - \phi_\emptyset^0(\theta^*)$, where ϕ is defined in Appendix A. This condition implies that the density $f(\theta)$ should not increase too steeply anywhere on its domain (see also Bénabou and Tirole, 2006, p. 1668).

proposals in an ultimatum game are rejected less often if the proposer chose to be ignorant of the payoffs for the responder. Furthermore, our own experiment, reported below, finds that people evaluate decision-makers who choose self-interestedly more positively if they did so under self-imposed ignorance.

This is consistent with the equilibrium described in Proposition 1, where the image value of taking a selfish action with full knowledge of the adverse consequences is lower than the value of doing so while ignorant, even if it is known that the decision maker chose not to know. The reason is that an agent who remains ignorant pools with at least some of the social types $[0, \theta^*)$, whereas agent who takes a selfish action after knowing that $W = w$ pools only with the *homo economicus*. Thus, self-imposed ignorance induces a partial exculpation from egoistic motives in signaling equilibrium.

Evasion. Ehrich and Irwin (2005) find that people are reluctant to ask for ethical attributes of consumer products, but will use the information if it is made available exogenously. Furthermore, the most puzzling result of the DWK experiment is that while only 26% choose selfishly in a binary dictator game with full information, 44% choose to remain ignorant of the potentially adverse consequences to the recipient of a self-interested choice.[7] Including the informed participants who fail to sacrifice to help the recipient, a total of 53% of subjects act in a way inconsistent with a preference for the fair outcome in the ignorance game.

These results are inconsistent with theories of outcome-based preferences. Given the results of the full-information baseline game, these theories predict that at most 26% of the subjects would choose ignorance. Social signaling does not account for these results either, since the experiment was anonymous and the recipients did not learn whether the dictator remained ignorant.

To analyze this situation with the self-signaling model, denote by Γ the 'ignorance game' explained above and denote by $\hat{\Gamma}$ the simpler game of DWK's *Baseline* treatment, in which it is common knowledge that $W = w$, and the only choice is whether or not to take the prosocial action. Using the signaling model, we compare the share of people who choose selfishly in $\hat{\Gamma}$ with the share that remains ignorant in game Γ when $k = 0$ (as is the case in the DWK experiment).

Proposition 2 *In game $\hat{\Gamma}$, there exists an equilibrium with a threshold type $\hat{\theta}$, such that all types $\theta < \hat{\theta}$ and the homo economicus choose $a = 0$, and all types $\theta \geq \hat{\theta}$ choose $a = 1$.*

Moreover, there exist $\bar{\alpha} < 1$, $\bar{\mu}$ and $0 < \underline{\varepsilon} < \bar{\varepsilon} < 1$ such that if $\mu > \bar{\mu}$ and $\alpha > \bar{\alpha}$ and $\varepsilon \in [\underline{\varepsilon}, \bar{\varepsilon}]$, then the share of people who choose ignorance in Γ is higher than the share of people who act selfishly in $\hat{\Gamma}$.

[7] We replicate this finding in Experiment 1, reported below. The numbers in the replication of Larson and Capra (2009) are 22% and 53% respectively.

Proposition 2 shows that the signaling model can explain evasion.[8] The intuition for this result is that image concerns drive social types to avoid pooling with the *homo economicus*. In game $\hat{\Gamma}$, the *homo economicus* chooses $a = 0$. This *increases* the signaling value of a pro-social action and induces some marginal social types to behave prosocially. By contrast, in game Γ, a fraction α of the indifferent *homo economicus* will choose to inform themselves. This *decreases* the signaling value of acquiring information. As a consequence, some marginal social types in game Γ switch to ignorance, thus increasing the total amount of selfish choices.

The conditions on the parameters are intuitive in this context. In order to deter the social types from choosing information in game Γ we need a sufficiently large fraction of *homo economicus* to do so, which explains why ε and α (the fraction of indifferent agents that reveals) need to be large enough. Note that a high α is consistent with the idea that people have preferences for information, evidence for which is presented in Domingues-Martinez, Sloof, and von Siemens (2010) and Loewenstein, Moore, and Weber (2006). Moreover, the importance of image concerns, as measured by μ, needs to be high enough so that the shift of the social types is large enough to outweigh the effect of the *homo economicus* acquiring information.

Summarizing, choices under full information produce a clear signal of who the selfish types are. This is not necessarily the case for the more ambiguous information acquisition decision, where the *homo economicus* may pool with those who inform themselves. The dilution of signaling incentives in the information acquisition decision explains why selfish behavior in the ignorance game can exceed such behavior in the setting without uncertainty.

3 Experimental Tests of the Self-Signaling Model

The theoretical results establish that self-signaling can explain the puzzling experimental results on strategic ignorance that are inconsistent with other popular models. In this section we report two experiments designed to falsify the self-signaling model, testing additional hypotheses that are consistent with neither outcome-based preferences nor social-image concern. Both experiments are modeled closely after the "hidden information" treatment of DWK.[9]

Subjects were instructed that they would be playing a simple game with one other person with whom they had been randomly and anonymously matched. One of the players was assigned the role of a dictator whose choices determined the payoffs of both players. In the experiment, the dictator was referred to as 'Player X' and the recipient as 'Player Y'. The dictator had to choose between an action A and an action B, which in Experiment 1 yielded

[8] In the proof of Proposition 2 we assumed that (2) holds, so that θ^* is stable and unique, and assumed a similar condition to guarantee the uniqueness of equilibrium threshold $\hat{\theta}$. Although these conditions are not necessary to derive the qualitative result, they greatly simplify the proof.

[9] The experiments were originally described in two separate working papers by the two individual authors, Experiment 1 in Grossman (2010) and Experiment 2 in Van der Weele (2012).

the dictator $6 and $5, respectively. The recipient's payoff varied between two different payoff states. In the 'conflicting interests game' (CIG) the recipient's payoffs from A and B were $1 and $5, respectively, while in the 'aligned interests game' (AIG) version the recipient's payoffs were flipped and the recipient obtained $5 and $1, respectively. The dictator was told that each of these two games had been randomly selected with equal probability at the start of the experiment. Before the dictator chose A or B she could choose to find out which game was being played (i.e. the recipient's payoffs from each action) by clicking a button labeled 'reveal game'.

Each subject participated in only one treatment and was not aware of the other treatments. After participants read instructions describing a generic payoff table, they completed a short quiz to ensure that they understood the task. Next they were shown the actual payoffs for the experiment and any other information relevant to their particular experimental condition, before taking another short quiz. The sessions lasted approximately 30 minutes in each experiment. Upon completion of the experiment, participants were paid privately in cash as they exited the room. The interface for both experiments was programmed using the Z-Tree software package (Fischbacher, 2007) and subjects were recruited using the ORSEE system (Greiner, 2003).

In both experiments, the dictator was anonymous. Moreover, both roles were informed that the dictator's decision of whether to reveal would be kept private. The dictator could remain ignorant of the payoffs, and the recipient would not know her information state. Thus, while the model has a dual interpretation in terms of either social-image or self-image, the experiments are most plausibly interpreted as a test of the the self-signaling interpretation. The descriptive statistics, and the number of participants for both experiments are provided in Appendix B and instructions are provided in Appendix D.

3.1 Experiment 1: Timing of the Revelation Decision

Ignorance protects self-image because it allows one to avoid the tradeoff between taking a costly prosocial action or being revealed as a selfish individual. If the allocation choice is elicited as a contingent strategy that specifies the chosen behavior in each state, the dictator is forced to reveal her type, and ignorance of the true state no longer protects the decision maker's image. In the first experiment, we examine whether ignorance is less attractive after having made such a contingent choice. Specifically, we vary whether the decision maker decides to reveal the information before or after the choice to be prosocial. So as to provide a complete replication of DWK's "hidden information" experiment, we also reproduce the baseline dictator game with full information of DWK. The experiment was carried out at the Experimental and Behavioral Economics Laboratory (EBEL) at the University of California, Santa Barbara.

Treatments. The three treatments of Experiment 1 are described below. The *CIG Only* and *Reveal Before* exactly replicate the DWK experiment while the *Reveal After* condition highlights the dictator's information choice when she cannot avoid revealing her preferences in the CIG.

1. *CIG Only*: This exactly replicated the DWK baseline treatment. Dictators played the CIG game with certainty, so the link between actions and outcomes was transparent.

2. *Reveal Before*: This exactly replicated the "hidden information" treatment of DWK. The participants were presented with the two versions of the game and told that the true payoffs were equally likely and would never be revealed publicly, but that the dictator could reveal them by clicking a button on the same screen labeled "Reveal Game".

3. *Reveal After*: This condition differed from the *Reveal Before* condition only in that the dictator entered her outcome choice for each of the two payoff schemes, with the outcome determined by her choice in the game version actually being played. As in the *Reveal Before* condition, the dictator could reveal the payoffs by clicking a button on the same screen.

Before participants were told to which role they had been assigned and were allowed to make a choice, they were given sixty seconds –during which the payoff matrix or matrices were displayed on the screen– to consider their choice. In general, the screen progression and layout reproduced the DWK interface as faithfully as possible. The text of the general instructions were reproduced almost verbatim, as were the condition-specific instructions in the replication conditions.[10]

Hypotheses. In the model, we assumed that a fraction α of the indifferent subjects acquires information. In the *Reveal Before* treatment where only the *homo economicus* is indifferent, the signaling model therefore predicts the ignorance rate to be $(1-\varepsilon)F(\theta^*)+\varepsilon(1-\alpha)$. In the *Reveal After* treatment, where there is no signaling value from being ignorant, the model predicts an ignorance rate of $1-\alpha$. Thus, ignorance is higher in *Reveal After* if $F(\theta^*) > 1-\alpha$, i.e., when most indifferent people reveal.

The prediction of the model thus depends on α, and may be considered somewhat ambiguous. However, as we argued in the discussion of Proposition 2, there is evidence from multiple studies that people are curious and have a preferences for information. In very different setup from ours, Loewenstein, Moore, and Weber (2006) and Domingues-Martinez, Sloof, and von Siemens (2010) find that many people invest in information acquisition even

[10]We are grateful to Jason Dana for sharing the software used in DWK. Minor differences in layout arose because the DWK experiment was programmed using a different software package.

if this is not in their own interest. Thus, we expect α to be high, which yields the following hypothesis.

Hypothesis 1 *Ignorance will be lower in the* Reveal After *than in the* Reveal Before *treatment.*

By contrast, consider the predictions of any model of outcome-based preferences. Suppose the distribution of preferences is such that there is a fraction $\beta \in (0, 1)$ of 'fair' types who prefer the fair outcome in the CIG. All other types prefer same action in both states and will always be indifferent about acquiring information. Such a model predicts that the ignorance rate will be $(1-\beta)(1-\alpha)$ in the *Reveal Before* treatment, and $1-\alpha$ in *Reveal After* treatment. Thus, there is an unambiguous prediction that ignorance should be lower in the *Reveal Before* condition, simply because the fair types value the information to make a more informed decision.

Results. On average participants earned \$9.72, including a \$5 show-up fee, with dictators earning slightly more (\$10.6) than recipients (\$8.84). The results are shown in the left panel of Figure 1, descriptive statistics can be found in Appendix B. First, while only nine out of 26 (35%) dictators in the *CIG Only* condition chose A, in the *Reveal Before* condition 23 of 39 (59%) chose in a manner inconsistent with a preference for the fair outcome, i.e., either choosing to remain ignorant of the recipient's payoffs or, conditional on revealing and being in the *CIG* game, choosing A. This 24 percentage point difference is significant at the 5% level ($p = 0.047$).[11] Thus, the main result of DWK is replicated.[12]
Comparing our main treatments, the overall ignorance rate in the *Reveal After* treatment was 0.26. This is significantly lower at the 5% level ($p = 0.013$) than the 0.54 rate in the *Reveal Before* condition, thus confirming Hypothesis 1. Note that in the *Reveal After* treatment, information acquisition does not vary much with the conditional allocation choices. Among the 17 dictators who chose A in both versions of the game, 29% choose ignorance. Among the 15 who chose B only in the *CIG* game, this rate was 27%. Note that this last group would learn the payoff state simply by observing their own payoff at the end of the session, so the high reveal rate supports the idea that subjects were generally curious to learn the outcome of the game in absence of strategic considerations.

[11] Unless otherwise indicated, all results reported hold for a one-sided Fisher's exact test, and the somewhat more powerful a one-sided exact z-test for equal proportions.

[12] Interestingly, there is a difference between the *CIG only* treatment where 35% choose selfishly, and the conditional choices in the CIG game in the *Reveal After* treatment, in which 54% choose selfishly. This may indicate that the conditional choice is perceived somewhat differently in the context of the ignorance game than it is in the simpler CIG only game. However, a two-sided test (which is appropriate since there is no directional hypothesis) does not find this difference to be significant ($p = 0.194$).

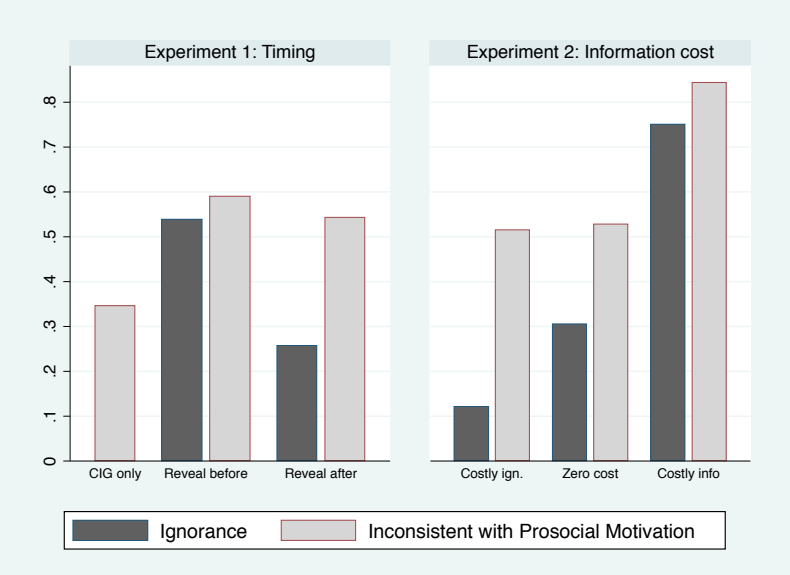

Figure 1: Ignorance and selfish behavior by treatment. "Ignorance" is the fraction of subjects choosing ignorance. "Inconsistent with Prosocial Motivation" is the fraction of ignorant agents plus the fraction who knew they played the CIG and chose A.

3.2 Experiment 2: Varying Information Costs

The self-signaling model generates some specific predictions about how people react to the cost of information. Specifically, it predicts that some of the social types are strictly better off when they remain ignorant, and would therefore be willing to pay for ignorance. In this experiment we vary the cost of obtaining information in the setting where subjects can find out the payoffs of the recipient before they make the allocation decision. The experiment was carried out at the Frankfurt Laboratory for EXperimental economics (FLEX) at the Goethe University Frankfurt.

Treatments. This experiment differed from Experiment 1 in some details. First, ignorance had to be chosen actively. Whereas in Experiment 1 a Player X who wanted to remain ignorant could simply abstain from clicking the "Reveal Game" button, in this experiment she actively had to click a "Not Reveal" button before being able to make an allocation choice. A second difference is that the highest payoff for the dictator was €10 (instead of \$6) and the lowest was €6 (instead of \$5). The highest payoff for the dictator was €6 (instead of \$5) and the lowest was €1 (instead of \$1). Finally, payoffs were expressed in experimental currency, where 10 EC is €1. Although none of these differences alters the structure of the game, they disallow direct comparisons between treatments across the two experiments.

The treatments in this experiment are straightforward manipulations of the game de-

scribed above. In the *Zero Cost* treatment, information was free, as it was DWK. The *Costly Information* treatment was equivalent to the *Zero Cost* treatment in all aspects, except that the dictator had to pay €0.50 to obtain information. By contrast, in the *Costly Ignorance treatment (Costly Ign.)* the dictator had to pay €0.50 to remain ignorant.

Hypotheses. The predictions of models of outcome-based preferences and the comparative statics derived from the signaling model for the cost of information (both for $k < 0$ and for $k > 0$) all suggest that the ignorance rate in the *Zero Cost* treatment will be higher than in the *Costly Ignorance* treatment and lower than in the *Costly Information* treatment.[13] However, when it comes to the *level* of ignorance in the *Costly Ignorance* treatment, the predictions of the signaling model and outcome-based models diverge. While outcome-based models predict that no one would pay to know less, the signaling model predicts that all social types with $\theta < \theta^*$ will strictly prefer ignorance over the trade-off between material costs and image that may result from acquiring information. This implies that some agents are willing to pay for ignorance.

Hypothesis 2 *In the* Costly Ignorance *treatment there will be a positive fraction of people who will pay to remain ignorant.*

Finally, a crucial prediction of the model is that people who choose A in the CIG under full knowledge have a worse image than those who choose ignorance and then choose A. It is this feature of the equilibrium that drives the social types $\theta < \theta^*$ to strictly prefer ignorance. In one of the treatments, we asked recipients to answer the question "How social [German: "sozial"] do you rate Player X, based on each of the following actions ...", where actions included the joint decision to become informed or not and to choose A or B. Answers were given on a 5-point scale from "very anti-social" (1) to "very social" (5).[14]

Hypothesis 3 *Subjects will judge a dictator who acts self-interestedly as less 'social' when she does so under ignorance than when she does so in the CIG.*

Results. On average participants earned €10.6, including a €4 show-up fee, with dictators earning €13.40 and recipients €7.82. The right panel of Figure 1 shows the results, descriptive

[13] As a qualification, we note that for some parameter values the signaling model allows for the possibility that switching from zero information cost to a (very small) positive information cost leads to a *decrease* in the predicted ignorance rate. This would require the influx of *homo economicus* into the ignorance pool as a result of the positive cost to decrease its image value enough to generate a compensating exodus of social types away from ignorance. Similarly going from zero cost to a very small cost of ignorance could theoretically lead to more ignorance. These effects are still consistent with the comparative statics derived from either (A.8) or (A.16) as they derive from the discontinuity from switching from the $k \leq 0$ case to the $k > 0$ case.

[14] Because the experiment was anonymous, the beliefs of the recipient need not matter directly to the dictator. However, we elicited the recipient's beliefs because in this way the elicitation cannot interfere with decision making, and beliefs are less likely to be biased by self-serving motives than those of the dictators.

statistics can be found in Appendix B. Clearly, the changes in the ignorance rate are in line with the idea that ignorance increases with the relative cost of information. Introducing a cost for information more than doubles the ignorance rate, while a cost for ignorance more than halves it. The hypothesis of equal proportions can be rejected at 1% ($p = 0.000$) for the former case, and at 10% ($p = 0.058$) for the latter case.

Moving to Hypothesis 2, 12% of the subjects are willing to pay to remain ignorant. Unfortunately, it is not straightforward to test statistically whether this fraction is different from 0, since any test with the null-hypothesis that the true probability is 0 will give a significant result. A two-sided binomial test finds that the interval for which we cannot reject at the 5% level that the true probability is p is $[0.04, 0.28]$. When we compare the result with a fictional sample featuring zero successes out of 33 trials, we find that the difference is significant at 10% (p=0.058) for a Fisher exact test and at 5% ($p = 0.0195$) for a z-test of equal proportions.[15]

In addition, we observe that the proportion of selfish choices increases as relative information costs go up. The difference between the *Zero cost* and *Costly info* treatments is significant at the 1% ($p = 0.005$) level. The small drop in selfishness in the *Costly Ignorance* treatment is not significant.

Finally, Figure 2 in Appendix C shows the elicited normative evaluations of the dictator (conditional on her strategy). In line with Hypothesis 3, recipients on average judge Player X to be more social when she chooses A under self-imposed ignorance, than when she does so knowingly in the CIG. Conversely, Player X is judged to be more social when she chooses B in the CIG, rather than under ignorance. A Mann-Whitney test shows that the distributions of responses in both cases differ significantly at the 1% level.

4 Discussion and Conclusion

The results described in the previous section are consistent with the self-signaling theory outlined in Section 2. In Experiment 1 we found that more people reveal after a decision has been made, even though the information is no longer useful to inform the decision. The signaling model predicts that agents avoid information exactly *because* it may influence the decision, since an increase in transparency generates strong image incentives that 'force' the agent to be prosocial. Thus, the self-signaling model can explain not only strategic ignorance, but also why ignorance is not strategic when it does not shield one from confronting a moral tradeoff. In Experiment 2 we found evidence that higher costs of information decrease prosocial behavior and that some people are willing to pay for ignorance. The signaling

[15]Note that Cain and Dana (2012) independently find a similar result that a minority is willing to pay to remain ignorant.

model explains this because low social types are strictly better-off avoiding the strong trade-off between material and image concerns that may result from acquiring information.

Are there are other theories that could also explain these results? There is evidence that outcome-based 'social' preferences matter in our experiment. In Experiment 2, participants were asked if they chose to reveal the information, and why (not). About 31% participants who reveal information express concerns for fairness in the questionnaire.[16] However, as we have emphasized, models of outcome based preferences cannot explain the results of our two experiments, as well as the earlier results in DWK.

Second, even though the experiments were conducted anonymously and the recipient did not observe the information-acquisition decision, one may wonder whether social-image concerns could play a role. Although we have stressed the importance of self-image, the model is entirely consistent with an interpretation in terms of social image. The questionnaire responses provide indications that social-image concerns towards the experimenter play some role. Three informed subjects chose the selfish action in the CIG and report (falsely) that they had been ignorant. However, the questionnaire responses also provide evidence of self-signaling. Of the eleven subjects who chose to remain ignorant in the *Zero Cost* treatment, three answered that they did so not to have a "bad conscience", and two subjects chose ignorance to avoid "having to be nice". We are therefore confident that the results of our experiments are at least partly due to self-image concerns.

Finally, one can speculate about presence of broader psychological motives relating to *cognitive dissonance*, the psychological cost that arises from the knowledge of having acted contrary to what is morally right (see Matthey and Regner, 2011). Self-image concerns can be one component of such dissonance, but emotional responses such as guilt or pity may play a role too. If the utility loss from such emotions is convex in the (perceived) probability of harming the other person, such a model may generate information aversion. An advantage of the Bayesian image-based theory presented here is that it provides such convexity without further assumptions on the relation between information and the strength of the emotion.

On the basis of these considerations, we believe that Bayesian self-signaling is an important factor in moral decision making. We do not, however, dismiss models of fairness or social-signaling concerns, the importance of which has been established by decades of research. Rather, we argue that these models are incomplete.

Our findings have consequences for policy makers who wish to promote prosocial behavior. The bad news is that while we demonstrate the importance of self-image in giving, our study also highlights how frail this motivation really is. Trivial excuses like self-imposed ignorance can neutralize the demands of our moral conscience and provide us with the excuses we need to behave selfishly. On a more positive note, both the model and the experiment show that

[16] This percentage was taken over the 90% of answers that actually provided a discernable reason.

much can be achieved by decreasing information costs, for example, through information campaigns. Making information harder to avoid increases self-signaling incentives and is an effective way to encourage prosocial behavior. The self-image model also provides an argument to increase the salience or weight of self-image concerns through the use of moral appeals, codes of conduct, self-evaluations, etc.

In an ideal world, people inform themselves as well as possible. In reality, people are aware that beliefs are reasons for action, and rationally and willfully manipulate their belief systems to support the behavioral patterns from which they benefit. How such signaling behavior is impacted by behavioral biases, by characteristics of the choice environment, or is manipulated by authorities and other social actors is an important topic for future research.

References

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

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

Bankier, D. (1996): *The Germans and the Final Solution: Public Opinion Under Nazism*. Wiley-Blackwell.

Baumeister, R. (1998): "The Self," in *The Handbook of Social Psychology*, ed. by D. Gilbert, S. Fiske, and G. Lindzsey. McGraw-Hill.

Bem, D. J. (1972): "Self-Perception Theory," in *Advances in Experimental Social Psychology, Vol 6*, ed. by L. Berkowitz, pp. 1–62. McGraw-Hill.

Bénabou, R., and J. Tirole (2006): "Incentives and Prosocial Behavior," *American Economic Review*, 96(5), 1652–1678.

——— (2011): "Identity, Morals and Taboos: Beliefs as Assets," *The Quarterly Journal of Economics*, 126, 805–855.

Bersoff, D. M. (2002): "Explaining Unethical Behavior Among People Motivated to Act Prosocially," *Journal of Moral Education*, 28(4), 413–428.

Bodner, R., and D. Prelec (2003): "Self-signaling and Diagnostic Utility in Everyday Decision Making," in *The Psychology of Economic Decisions. Vol. 1. Rationality and Well-being*, ed. by I. Brocas, and J. Carillo, pp. 105–26. Oxford University Press.

Burlando, R. M., and F. Guala (2004): "Heterogeneous Agents in Public Goods Experiments," *Experimental Economics*, 8, 35–54.

Cain, D., and J. Dana (2012): "Paying People to Look at the Consequences of Their Actions," Working paper.

Caplin, A., and J. Leahy (2001): "Psychological Expected Utility Theory and Anticipatory Feelings," *The Quarterly Journal of Economics*, pp. 55–79.

Carillo, J. D., and T. Mariotti (2000): "Strategic Ignorance as a Self-Disciplining Device," *Review of Economic Studies*, 67(3), 529–544.

Cohen, S. (2001): *States of Denial*. Cambridge: Blackwell.

Conrads, J., and B. Irlenbusch (2011): "Strategic Ignorance in Bargaining," Discussion paper 6087, IZA.

Crémer, J. (1995): "Arm's Length Relationships," *The Quarterly Journal of Economics*, 110(2), 275–295.

Dana, J., R. Weber, and J. X. Kuang (2007): "Exploiting Moral Wiggle Room: Experiments Demonstrating an Illusory Preference for Fairness," *Economic Theory*, 33(1), 67–80.

Domingues-Martinez, S., R. Sloof, and F. von Siemens (2010): "Monitoring your Friends, Not your Foes: Strategic Ignorance and the Delegation of Real Authority," Discussion Paper 2010-101/1, Tinbergen Institute.

Ehrich, K. R., and J. R. Irwin (2005): "Willful Ignorance in the Request for Product Attribute Information," *Journal of Marketing Research*, XLII, 266–277.

Eil, D., and J. M. Rao (2011): "The Good News-Bad News Effect: Asymmetric Processing of Objective Information about Yourself," *American Economic Journal: Microeconomics*, 3(May), 114–138.

Ellingsen, T., and M. Johannesson (2008): "Pride and Prejudice: The Human Side of Incentive Theory," *American Economic Review*, 98, 990–1008.

Engel, C. (2011): "Dictator Games: A Meta Study," *Experimental Economics*, 14, 583–610.

Feiler, L. (2007): "Behavioral Biases in Information Acquisition," Ph.D. thesis, Caltech.

Fischbacher, U. (2007): "z-Tree: Zurich Toolbox for Ready-made Economic Experiments," *Experimental Economics*, 10(2), 171–178.

Fischbacher, U., S. Gächter, and E. Fehr (2001): "Are People Conditionally Cooperative," *Economics Letters*, 71(3), 397–404.

Fischbacher, U., and F. Heusi (2008): "Lies in Disguise. An experimental study on cheating," Discussion paper.

Fong, C., and F. Oberholzer-Gee (2011): "Truth in Giving: Experimental Evidence on the Welfare Effects of Informed Giving to the Poor," *Journal of Public Economics*, 95(5-6), 436–444.

Gneezy, A., U. Gneezy, G. Riener, and L. D. Nelson (2012): "Pay-what-you-want, identity, and self-signaling in markets," *Proceedings of the National Academy of Sciences*.

Greiner, B. (2003): "An Online Recruitment System for Economic Experiments," *Forschung und wissenschaftliches Rechnen*, 63, 79–93.

Grossman, Z. (2009): "Self-signaling versus Social Signaling in Giving," Working paper, UC Santa Barbara.

——— (2010): "Strategic Ignorance and the Robustness of Social Preferences," Working paper, UC Santa Barbara.

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

Kőszegi, B. (1996): "Ego Utility, Overconfidence, and Task Choice," *Journal of the European Economic Association*, 4(4), 673–707.

Krupka, E. L., and R. Weber (2008): "Identifying Social Norms Using Coordination Games" Why Does Dictator Game Sharing Vary?," Discussion Paper 3860, IZA.

Kunda, Z. (1990): "The Case for Motivated Reasoning," *Psychological Bulletin*, 108(3), 480–498.

Kurzban, R., and D. Houser (2005): "Experiments investigating cooperative types in humans: A complement to evolutionary theory and simulations," *Proceedings of the National Academy of Sciences*, 102(5), 1803–1807.

Larson, T., and M. C. Capra (2009): "Exploiting moral wiggle room: Illusory preference for fairness? A comment," *Judgement and Decision Making*, 4(6), 467–474.

Loewenstein, G., D. Moore, and R. Weber (2006): "Misperceiving the value of information in predicting the performance of others," *Experimental Economics*, pp. 281–295.

Matthey, A., and T. Regner (2011): "Do I really want to konw? A cognitive dissonance-based explanation of other-regarding behavior," *Games*, 2, 114–135.

Mazar, N., O. Amir, and D. Ariely (2009): "The Dishonesty of Honest People: A Theory of Self-Concept Maintenance," *Journal of Marketing Research*, XLV, 633–644.

Möbius, M., M. Niederle, P. Niehaus, and T. Rosenblat (2011): "Managing Self-Confidence: Theory and Experimental Evidence," Mimeo, Stanford.

Murnigham, K., J. M. Oesch, and M. Pillutla (2001): "Player Types and Self-Impression Management in Dictatorship Games: Two Experiments," *Games and Economic Behavior*, 37, 388–414.

Nyborg, K. (2011): "I Dont Want to Hear About it: Rational Ignorance among Duty-Oriented Consumers," *Journal of Economic Behavior and Organization*, 79, 263–274.

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

Speer, A. (1970): *Inside the Third Reich*. New York: MacMillan.

Tadelis, S. (2011): "The power of shame and the rationality of trust," Mimeo, Berkeley, Haas School of Business.

Van der Weele, J. J. (2012): "When Ignorance is Innocence: On Information Avoidance in Moral Dilemmas.," *SSRN working paper, available at http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1844702*.

Appendix A: Proofs

Proof of Proposition 1. We start with the case where $k \leq 0$, and the homo economicus chooses information acquisition. Subsequently, we study the case where $k > 0$. In each case, the proof proceeds in two steps. First, we confirm that the decisions to (not) take the prosocial action are indeed optimal, given proposed off-equilibrium beliefs. Second, given these decisions, we establish which types will acquire information. Finally, we discuss whether the proposed off-equilibrium beliefs are reasonable.

The case where $k \leq 0$. We start with some notation. For the social agents, let $\theta^* \in (0,1)$ be the threshold type who is indifferent between acquiring information and not. To ease notation, let $\phi_\sigma^a = E[\theta \mid a, \sigma; s]$ denote the equilibrium expectation conditional on the equilibrium strategy profile,

the chosen action a and information σ:

$$\phi^0_\emptyset = \phi^0_\emptyset(\theta^*) \equiv E[\theta \mid 0, \emptyset; \theta^*] = \int_0^{\theta^*} \frac{(1-\varepsilon)\theta dF(\theta)}{(1-\varepsilon)F(\theta^*) + (1-\alpha)\varepsilon} \tag{A.1}$$

$$\phi^0_0 = \phi^0_0(\theta^*) \equiv E[\theta \mid 0, \sigma_0; \theta^*] = \int_{\theta^*}^1 \frac{(1-\varepsilon)\theta}{(1-\varepsilon)\left(1-F(\theta^*)\right) + \alpha\varepsilon} dF(\theta) \tag{A.2}$$

$$\phi^1_w = \phi^1_w(\theta^*) \equiv E[\theta \mid 1, \sigma_w; \theta^*] = \int_{\theta^*}^1 \frac{\theta dF(\theta)}{1-F(\theta^*)} \tag{A.3}$$

$$\phi^0_w \equiv E[\theta \mid 0, \sigma_w] = 0. \tag{A.4}$$

Note that the tie-breaking rule that a fraction α of the homo-economicus chooses information only applies when $k = 0$. When $k < 0$, then all homo-economicus strictly prefer to be informed and α has no impact on their behavior. With some abuse of notation, we treat α as being equal to 1 in that case.

Step 1. We now verify whether the proposed decisions to be prosocial or not 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. The *homo economicus* always chooses $a = 0$, so we concentrate on the social agents.

Step 1a) If $\sigma = \sigma_w$, an agent of type θ will take the prosocial action iff

$$\begin{aligned} u(a = 1 \mid \sigma = \sigma_w; \theta^*) &\geq u(a = 0 \mid \sigma = \sigma_w; \theta^*) \\ \theta w - c + \mu\phi^1_w &\geq \mu\phi^0_w \\ \theta &\geq \frac{c - \mu\phi^1_w}{w} \equiv \bar{\theta}. \end{aligned} \tag{A.5}$$

It is immediate that in equilibrium all types $\theta \geq \bar{\theta}$ who observed $\sigma = \sigma_w$ take the prosocial action, and all $\theta < \bar{\theta}$ do not. Note that $\bar{\theta} > 0$, since we assumed that $c > \mu$.

Step 1b) Next, consider the case in which $\sigma = \sigma_0$. It is optimal for the agent not to take the prosocial action iff

$$\begin{aligned} u(a = 0 \mid \sigma = \sigma_0; \theta^*) &> u(a = 1 \mid \sigma = \sigma_0; \theta^*) \\ \mu\phi^0_0 &> -c + \mu\phi^1_0 \\ c &> \mu(\phi^1_0 - \phi^0_0). \end{aligned} \tag{A.6}$$

which is satisfied for any belief since we assumed that $c > \mu$.

Step 1c) Consider now the uninformed agent, $\sigma = \emptyset$. She will take the self-interested action iff

$$\begin{aligned} u(a = 0 \mid \sigma = \emptyset; \theta^*) &> u(a = 1 \mid \sigma = \emptyset; \theta^*) \\ \mu\phi_\emptyset^0 - p\theta w &> -c + \mu\phi_\emptyset^1 \\ \theta &< \frac{c - \mu(\phi_\emptyset^1 - \phi_\emptyset^0)}{pw}. \equiv \tilde{\theta} \end{aligned} \tag{A.7}$$

Step 2. We now check which type will acquire information. Since the *homo economicus* cares only about his own material payoffs, it is obvious that she will acquire information as long as $k < 0$ (where the case of $k = 0$ is covered by our tie-break rule). We know the equilibrium action of the social agents upon (not) acquiring information. Keeping in mind that the equilibrium beliefs depend on θ^*, we can derive that θ^* is given implicitly by the fixed point equation

$$\begin{aligned} Eu(\text{acquire info}) &= Eu(\text{not acquire info}) \\ (1-p)\mu\phi_0^0 + p\left(\theta^* w - c + \mu\phi_w^1\right) - k &= \mu\phi_\emptyset^0 \\ \theta^* &= \frac{pc + k - \mu\left(p\phi_w^1 + (1-p)\phi_0^0 - \phi_\emptyset^0\right)}{pw}. \end{aligned} \tag{A.8}$$

It is straightforward that all types $\theta < \theta^*$ remain ignorant and all types $\theta \geq \theta^*$ acquire information.

Only if $\theta^* < \tilde{\theta}$ do all ignorant types take the self-interested action and only if $\bar{\theta} < \theta^*$ will all types who observe $\sigma = \sigma_w$ indeed take the prosocial action. Next, we will establish sufficient conditions for the existence of a $\theta^* \in (0, 1)$ such that $\bar{\theta} < \theta^* < \tilde{\theta}$.

Some algebra shows that $\theta^* < \tilde{\theta}$ iff

$$k < (1-p)c - \mu(p\phi_w^1 + (1-p)\phi_0^0 - \phi_\emptyset^1). \tag{A.9}$$

Since $k \leq 0$, it is sufficient that the RHS of this inequality is positive. It is easy to show that this is the case if off-equilibrium beliefs satisfy $\phi_\emptyset^1 \leq \phi_w^1$.

Similar algebra comparing (A.5) and (A.8) yields that $\bar{\theta} < \theta^*$ if and only if

$$k > \mu\left((1-p)\phi_0^0 - \phi_\emptyset^0\right) \equiv \underline{k}. \tag{A.10}$$

A necessary and sufficient condition for $\underline{k} < 0$ is that $p > 1 - \frac{\phi_\emptyset^0}{\phi_0^0}$.

It remains to check that θ^* exists and is in the interior. We have already verified that $\theta^* > 0$, since $\theta^* > \bar{\theta} > 0$. To check that $\theta^* < 1$, note that if the threshold type is $\theta^* = 1$, we have $\phi_w^1 = 1$, $\phi_0^0 = 0$, and $\phi_\emptyset^0 = E_F\theta$. Plugging this into (A.8), it is straightforward to show that $p > \frac{\mu E_F\theta}{w - c + \mu} \Rightarrow \theta^* < 1$. Existence follows from the continuity of both sides of (A.8).

Combining arguments, an interior equilibrium exists if $\underline{k} < k \leq 0$ and

$$p > \max\left\{1 - \frac{\phi_\emptyset^0}{\phi_0^0}, \frac{\mu E_F\theta}{w - c + \mu}\right\}. \tag{A.11}$$

Since $\frac{\mu E_F\theta}{w-c+\mu} < 1$ and $\frac{\phi_\emptyset^0}{\phi_0^0}$ is positive for any interior value of θ^* (i.e. whatever the value of p), there exists a $\tilde{p} < 1$ such that this condition will be satisfied for some $p > \tilde{p}$.

Comparative statics with respect to ignorance levels can be obtained by implicit differentiation of (A.8) with respect to θ^* and c, w or k respectively (and using (2)).

The case where $k > 0$. In this case, the homo-economicus will not acquire information. The equilibrium beliefs become:

$$\phi^0_\emptyset = \phi^0_\emptyset(\theta^*) \equiv E[\theta \mid 0, \emptyset; \theta^*] = \int_0^{\theta^*} \frac{(1-\varepsilon)\theta}{(1-\varepsilon)F(\theta^*) + \varepsilon} dF(\theta) \tag{A.12}$$

$$\phi^1_w = \phi^0_0 = \phi^0_0(\theta^*) \equiv E[\theta \mid 0, \sigma_0; \theta^*] = \int_{\theta^*}^1 \frac{\theta dF(\theta)}{1 - F(\theta^*)}. \tag{A.13}$$

Moreover, we will assume that the (now off-equilibrium) belief $\phi^0_w = 0$.

The analysis proceeds like before, and is not reconstructed here in detail for reasons of space. We obtain

$$\bar{\theta} = \frac{c - \mu\phi^1_w}{w}, \tag{A.14}$$

$$\tilde{\theta} = \frac{c - \mu(\phi^1_\emptyset - \phi^0_\emptyset)}{pw} \tag{A.15}$$

$$\theta^* = \frac{pc + k - \mu\left(\phi^1_w - \phi^0_\emptyset\right)}{pw}. \tag{A.16}$$

Some algebra yields that $\bar{\theta} \leq \theta^*$ iff $k \geq \mu(\phi^1_w(1-p) - \phi^0_\emptyset)$, which is satisfied $p \geq 1 - \frac{\phi^0_\emptyset}{\phi^1_w}$. Furthermore, $\theta^* < \tilde{\theta}$ iff

$$k < c(1-p) + \mu(\phi^1_w - \phi^1_\emptyset) \equiv \bar{k}. \tag{A.17}$$

A sufficient condition for $\bar{k} > 0$ is that off-equilibrium beliefs satisfy $\phi^1_w \geq \phi^1_\emptyset$.

It remains to check that $\theta^* < 1$. By substituting $\theta^* = 1$ into (A.16), we can find the sufficient condition $p > \frac{k - \mu(1 - E_F\theta)}{w - c}$. Using (A.17) we obtain $p > \frac{c - \mu(1 - \phi^1_w + \phi^0_\emptyset - E_F\theta)}{w}$.

Collecting arguments, we have shown that an interior equilibrium exists if $0 < k < \bar{k}$ and

$$p > \max\left\{1 - \frac{\phi^0_\emptyset}{\phi^1_w}, \frac{c - \mu(1 - \phi^1_w + \phi^0_\emptyset - E_F\theta)}{w}\right\}. \tag{A.18}$$

Since $c < w$ and $\phi^0_\emptyset < \phi^1_w$ for any value of θ^*, there exists a $\hat{p} < 1$ such that this condition will be satisfied for some $p > \hat{p}$.

Reasonableness of off-equilibrium beliefs. We need to check that the assumptions on off-equilibrium beliefs 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 (IC'): we require that off-equilibrium beliefs upon observing the deviation (σ', a') place zero weight on type θ', if equilibrium payoffs of θ' dominate the deviation payoffs when observer beliefs equal $E[\theta|\sigma', a'] = 1$ (i.e. are maximally optimistic about the sender's type). Thus, off-equilibrium beliefs do not place

weight on types that would never deviate, even if this would give them the best possible image.[17]

In the case where $k > 0$, we assumed that $\phi_w^0 = 0$. Type $\theta = 0$ would be willing to deviate to acquiring information if $\mu\phi_\emptyset^0 < \mu(p + (1-p)\phi_w^1) - k$. Using (A.18), we obtain the sufficient condition $k < \mu\left(1 - \frac{\phi_\emptyset^0}{\phi_w^0}\right)$. Thus, there exists a $k > 0$ such that the off-equilibrium beliefs do not violate the IC'.

We also assumed that $\phi_\emptyset^1 \leq \phi_w^1$. A sufficient condition is that type θ^* is willing to deviate to the prosocial action under ignorance, since this justifies the assumption that $\phi_\emptyset^1 = \theta^* < \phi_w^1$. This is the case when $\mu - c \geq \mu\theta^* - pc - k \Leftrightarrow k > c(1-p) - \mu(1-\theta^*)$. For this to be satisfied for a $k \leq 0$, a necessary condition is $p \geq 1 - \frac{\mu(1-\theta^*)}{c}$. Note that the empirical evidence presented in the right panel of Figure 2 supports the assumption that $\phi_\emptyset^1 \leq \phi_w^1$.

■

Proof of Proposition 2. First, consider the equilibrium in game $\hat{\Gamma}$. Define

$$\hat{\phi}_w^0 = \hat{\phi}_w^0(\hat{\theta}) \equiv E[\theta \mid 0; \hat{\theta}] = \int_0^{\hat{\theta}} \frac{(1-\varepsilon)\theta}{(1-\varepsilon)F(\hat{\theta}) + \varepsilon} dF(\theta). \tag{A.19}$$

$$\hat{\phi}_w^1 = \hat{\phi}_w^1(\hat{\theta}) \equiv E[\theta \mid 0; \hat{\theta}] = \int_{\hat{\theta}}^1 \frac{\theta dF(\theta)}{1 - F(\hat{\theta})}. \tag{A.20}$$

The equilibrium threshold $\hat{\theta}$ is given implicitly by the fixed point equation

$$\begin{aligned} Eu(a=1) &= Eu(a=0) \\ \hat{\theta}w - c + \mu\hat{\phi}_w^1 &= \mu\hat{\phi}_w^0 \\ \hat{\theta} &= \frac{c - \mu\left(\hat{\phi}_w^1 - \hat{\phi}_w^0\right)}{w}. \end{aligned} \tag{A.21}$$

Note that the assumptions $c < w$ and $\mu < c$ guarantee that $\hat{\theta}$ is always in the interior, and existence follows from the continuity of both sides of (A.21).

We now turn to the second part of Proposition 2. We can denote the fraction of people who choose prosocially $\hat{\Gamma}$ by $(1-\varepsilon)(1-F(\hat{\theta}(\varepsilon)))$. The fraction of people who choose to reveal information in game Γ is $\alpha\varepsilon + (1-\varepsilon)(1 - F(\theta^*(\varepsilon)))$. Thus, we want to prove that

$$\begin{aligned} &(1-\varepsilon)(1-F(\hat{\theta}(\varepsilon))) > \alpha\varepsilon + (1-\varepsilon)(1-F(\theta^*(\varepsilon))) \\ &F(\theta^*(\varepsilon)) - F\left(\hat{\theta}(\varepsilon)\right) > \frac{\alpha\varepsilon}{1-\varepsilon}. \end{aligned} \tag{A.22}$$

Let us first consider the LHS of (A.22). To ease notation, define $\Delta(\varepsilon) \equiv \theta^*(\varepsilon) - \hat{\theta}(\varepsilon)$. Substituting in the expressions for θ^* from (A.8) and $\hat{\theta}$ from (A.21) and setting $k = 0$ we find

[17] Other authors have applied standard refinements to games where people care about other's beliefs. Andreoni and Bernheim (2009) and Ellingsen and Johannesson (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. We thank Martin Dufwenberg for pointing this out.

$$\Delta(\varepsilon) = \frac{pc - \mu\left(p\phi_w^1 + (1-p)\phi_0^0(\varepsilon) - \phi_\emptyset^0(\varepsilon)\right)}{pw} - \frac{c - \mu\left(\hat{\phi}_w^1 - \hat{\phi}_w^0(\varepsilon)\right)}{w}$$
$$= \frac{\mu\left(p(\hat{\phi}_w^1 - \hat{\phi}_w^0(\varepsilon)) - (p\phi_w^1 + (1-p)\phi_0^0(\varepsilon) - \phi_\emptyset^0(\varepsilon))\right)}{pw}, \tag{A.23}$$

which shows that $\Delta(\varepsilon)$ depends on the relative strength of image concerns in the two games.

We now show that $\Delta(\varepsilon) > 0$ when ε and α are large enough. First note that $\Delta(0) < 0$, because $\phi_0^0(0) = \phi_w^1$, and so

$$\Delta(0) = \frac{\mu\left(p(\hat{\phi}_w^1 - \hat{\phi}_w^0(0)) - (\phi_w^1 - \phi_\emptyset^0(0))\right)}{pw}. \tag{A.24}$$

Consider now the case where ε is close to 1. Suppose that $\Delta(\varepsilon) \leq 0$, i.e. $\hat{\theta} \geq \theta^*$. This implies that $\hat{\phi}_w^1 \geq \phi_w^1$, so it must be that

$$p\hat{\phi}_w^0 + (1-p)\phi_0^0(\varepsilon) \geq \phi_\emptyset^0(\varepsilon)$$
$$p\int_0^{\hat{\theta}} \frac{(1-\varepsilon)\theta dF(\theta)}{(1-\varepsilon)F(\hat{\theta}) + \varepsilon} + (1-p)\int_{\theta^*}^1 \frac{(1-\varepsilon)\theta dF(\theta)}{(1-\varepsilon)(1-F(\theta^*)) + \alpha\varepsilon} \geq \int_0^{\theta^*} \frac{(1-\varepsilon)\theta dF(\theta)}{(1-\varepsilon)F(\theta^*) + (1-\alpha)\varepsilon}. \tag{A.25}$$

If $\alpha = 0$, we have that $\phi_0^0(\varepsilon)$ and $\hat{\phi}_w^1$ are both larger than $\phi_\emptyset^0$, so (A.25) is always satisfied. The LHS of (A.25) is decreasing in α, while the RHS is increasing. For $\alpha = 1$, the LHS (A.25) approaches 0 when ε gets large, while the RHS is positive. Thus, there exists some $\bar{\alpha}$, such that if $\alpha > \bar{\alpha}$, (A.25) is violated when ε is high. This means that $\hat{\theta} \geq \theta^*$ leads to a contradiction, and we must have $\hat{\theta} < \theta^*$, i.e. $\Delta(\varepsilon) > 0$.

By implicit differentiation of (A.8) and (A.21), we can derive

$$\frac{d\Delta(\varepsilon)}{d\varepsilon} = \frac{d\theta^*}{d\varepsilon} - \frac{d\hat{\theta}}{d\varepsilon}$$
$$= \left(\frac{\mu\left(\frac{d\phi_\emptyset^0}{d\varepsilon} - (1-p)\frac{d\phi_0^0}{d\varepsilon}\right)}{pw + \mu\left(p\frac{d\phi_w^1}{d\theta^*} + (1-p)\frac{d\phi_0^0}{d\theta^*} - \frac{d\phi_\emptyset^0}{d\theta^*}\right)}\right) - \left(\frac{\mu\frac{d\hat{\phi}_w^0}{d\varepsilon}}{w + \mu\left(\frac{d\hat{\phi}_w^1}{d\theta^*} - \frac{d\hat{\phi}_w^0}{d\theta^*}\right)}\right). \tag{A.26}$$

It is easy to show that $\frac{d\phi_\emptyset^0}{d\varepsilon}$, $\frac{d\phi_0^0}{d\varepsilon}$, $\frac{d\hat{\phi}_w^0}{d\varepsilon} < 0$. We can guarantee that the denominators of both terms on the RHS are positive by invoking the uniqueness condition (2) for θ^* as well as an equivalent condition for $\hat{\theta}$

$$\frac{d\left(\hat{\phi}_w^1(\hat{\theta}) - \hat{\phi}_w^0(\hat{\theta})\right)}{d\hat{\theta}} > -\frac{w}{\mu}, \tag{A.27}$$

Therefore, this expression is positive if $(1-p)\frac{d\phi_0^0}{d\varepsilon} < \frac{d\phi_\emptyset^0}{d\varepsilon}$. This is satisfied when α is large, since in that case $\frac{d\phi_\emptyset^0}{d\varepsilon}$ approaches 0, while $\frac{d\phi_0^0}{d\varepsilon}$ is negative.

Summarizing, we have established that $\exists\, \hat{\varepsilon} < 1$ and $\bar{\alpha} < 1$, such that if $\alpha > \bar{\alpha}$ and $\varepsilon > \hat{\varepsilon}$, then the

LHS of (A.22) is positive and strictly increasing on $[0, 1)$.

We now show that $\Delta > 0$ implies that $\frac{d\Delta}{d\mu} > 0$. By implicit differentiation of (A.8) and (A.21), we find that

$$\begin{aligned}\frac{d\Delta(\varepsilon)}{d\mu} &= \frac{d\theta^*}{d\mu} - \frac{d\hat{\theta}}{d\mu} \\ &= \left(\frac{-(p\phi_w^1 + (1-p)\phi_0^0(\varepsilon) - \phi_\emptyset^0(\varepsilon))}{pw + \mu\left(p\frac{d\phi_w^1}{d\theta^*} + (1-p)\frac{d\phi_0^0}{d\theta^*} - \frac{d\phi_\emptyset^0}{d\theta^*}\right)} \right) - \left(\frac{-(\hat{\phi}_w^1 - \hat{\phi}_w^0(\varepsilon))}{w + \mu\left(\frac{d\hat{\phi}_w^1}{d\hat{\theta}} - \frac{d\hat{\phi}_w^0}{d\hat{\theta}}\right)} \right) \end{aligned} \tag{A.28}$$

Suppose that ε is large, so that $\frac{d\phi_\emptyset^0}{d\theta^*}$, $\frac{d\phi_0^0}{d\theta^*}$ and $\frac{d\hat{\phi}_w^0}{d\hat{\theta}}$ are small and can be ignored. Stability and uniqueness require that $f(\cdot)$ is relatively flat, so that $\frac{d\hat{\phi}_w^1}{d\hat{\theta}} \approx \frac{d\phi_w^1}{d\theta^*}$. Using these findings, we obtain that (A.28) is positive when

$$p(\hat{\phi}_w^1 - \hat{\phi}_w^0(\varepsilon)) - (p\phi_w^1 + (1-p)\phi_0^0(\varepsilon) - \phi_\emptyset^0(\varepsilon)) > 0, \tag{A.29}$$

i.e. if and only if $\Delta(\varepsilon) > 0$. Above, we have shown that this is the case for α and ε sufficiently large.

Let us now consider the RHS of (A.22). Since $\lim_{\varepsilon \uparrow 1} \frac{\alpha\varepsilon}{1-\varepsilon} = \infty$, $\exists\, \bar{\varepsilon}$ such that if $\varepsilon > \bar{\varepsilon}$, then (A.22) does not hold. Since we have shown that the LHS of (A.22) is positive for α and ε sufficiently large, and increasing in μ, we obtain the statement in Proposition 2. ■

Appendix B: Summary Statistics

Treatment	N	Chose ignorance	Chose A		
			Ignorant	AIG	CIG
Reveal first	39	54% (21/39)	76% (16/21)	100% (7/7)	18% (2/11)
Reveal last	35	26% (9/35)	-	91% (32/35)	54% (19/35)
CIG only	26	-	-	-	35% (9/26)
Costly Ign.	33	12% (4/33)	75% (3/4)	100% (10/10)	68% (13/19)
Zero cost	36	31% (11/36)	100% (11/11)	100% (12/12)	62% (8/13)
Costly Inf.	32	75% (24/32)	100% (24/24)	100% (2/2)	50% (3/6)

Table 1: Dictators' decisions.

Appendix C: Normative evaluations of actions

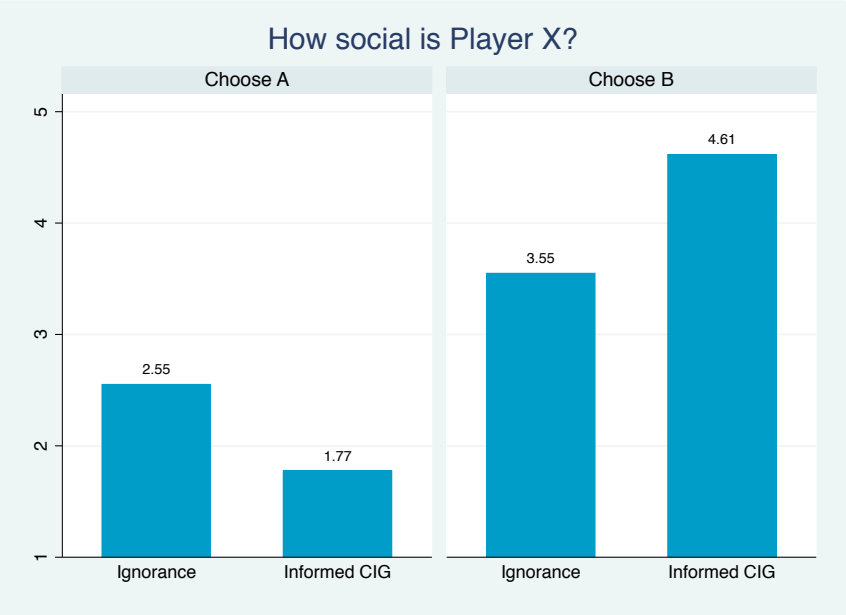

Figure 2: Answers to the question "How social ("sozial") do you think Player X is, based on each of the following actions ...", where actions included the joint decision to be informed or not and to choose €10 or €6. Answers were given on a 5-point scale from "very anti-social" (1) to "very social" (5). Average ratings are based on 31 participants, and are displayed on top of the bars.
N.B. The answers were elicited in the *Costly Information* treatment, but subjects were explicitly asked to abstract from the cost of information in their evaluation of Player X.

Appendix D: Instructions and screenshots [NOT FOR PUBLICATION]

Below we reproduce the instructions for Experiment 1. The instructions for Experiment 2 differ only in small and predictable details, and screenshots can be found in Van der Weele (2012).

Experiment

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

This is an experiment in the economics of decision-making. Several research institutions have provided funds for this research. You will be paid for your participation in the experiment. The exact amount you will be paid will depend on your and/or others' decisions. Your payment will consist of the amount you accumulate plus a $5 participation bonus. You will be paid privately in cash at the conclusion of the experiment.

If you have a question during the experiment, raise your hand and an experimenter will assist you. Please do not talk, exclaim, or try to communicate with other participants during the experiment. Please put away all outside materials (such as book bags, notebooks, cellphones) before starting the experiment. Participants violating the rules will be asked to leave the experiment and will not be paid.

OK

Experiment

Description of the Game

In this experiment, each of you will play a game with one other person in the room. Before playing, we will randomly match people into pairs. The grouping will be anonymous, meaning that no one will ever know which person in the room they played with. Each of you will be randomly assigned a role in this game. Your role will be player X or player Y. This role will also be kept anonymous. The difference between these roles will be described below. Thus, exactly one half of you will be a Player X and one half a Player Y. Also, each of you will be in a pair that includes exactly one of each of these types.

The game your pair will play will be like the one pictured below. Player X will privately choose one of two options: "A" or "B". Player Y will not make any choice. Both players will receive payments based 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 dollars and Player Y receives 4 dollars.

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

OK

Experiment

Check your understanding

At this point, to make sure that everyone understands the game, please answer the two questions below.

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

CIG Only

Experiment

You are Player X. To make your choice, please select one of the options below, then click OK to confirm your choice.

I choose ○ A
○ B

OK

Player X chooses	Player X receives	Player Y receives
A	6	1
B	5	5

Experiment

Two possible versions of the actual game

The actual game you will play will be one of the two pictured below. Notice that both games are the same except that two of Player Y's payments have been switched. Note that in both games, Player X gets his or her highest payment of $6 by choosing A. In the game on the left, this gives Player Y his or her lowest payment of $1. In the game on the right this gives Player Y his or her highest payment of $5. In both games, if Player X chooses B, he or she gets a lower payment of $5. In the game on the left, this gives Player Y the highest payment of $5. In the game on the right, this gives Player Y the lowest payment of $1.

Note that you will not know which of the games that you are playing, but for Player X, the payments will be identical. The only thing that differs are the payments for Player Y.

Which of these games will you actually play? That was determined randomly by the computer at the beginning of the experiment, with each game being equally likely. However, we will not reveal publicly which game you are actually playing. Player X can choose to find out which game is being played if he or she wants to do so by clicking a button. This choice will be anonymous, thus Player Y will not know if X knows which game is being played. However, Player X is not required to find out and may choose not to. Player Y will not have this option. When the game ends, we will pay each player privately.

Player X chooses	Player X receives	Player Y receives
A	6	1
B	5	5

Player X chooses	Player X receives	Player Y receives
A	6	5
B	5	1

OK

Experiment

Check your understanding

To make sure that everyone understands the game, please answer the two questions below. Remember that each of the two possible versions of the game shown below are equally likely.

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

If Player X chooses B, then Player Y receives
- $5
- $1
- either $5 or $1

OK

Player X chooses	Player X receives	Player Y receives
A	6	1
B	5	5

Player X chooses	Player X receives	Player Y receives
A	6	5
B	5	1

Thinking delay

Time remaining 56

Thinking time

Please take a minute to think about your decision. After one minute you will be told whether you are Player X or Player Y. If you are Player X, you will then make your decision.

Player X chooses	Player X receives	Player Y receives
A	6	1
B	5	5

Player X chooses	Player X receives	Player Y receives
A	6	5
B	5	1

Reveal Before

Experiment

If there are no further questions, we will begin the game.

You are Player X. To make your choice, click the corresponding button below. If you wish to reveal which of the two games is actually being played, click "Reveal Game".

Player Y will not find out whether or not you learned the actual version of the game.

I choose:
- A
- B
- Reveal Game

OK

Player X chooses	Player X receives	Player Y receives
A	6	?
B	5	?

Reveal After

Experiment

If there are no further questions, we will begin the game.

You are Player X. Though you have not been told the actual version of the game that is being played, please maker your choice for each version by clicking the corresponding button below. If the actual game is the one on the left, your payment and the payment of Player Y will be determined by the choice you indicated for the left game. If the actual game is the one on the right, the payments will be determined by the choice you indicated for the game on the right.

For the game on the left, I choose ○ A ○ B

For the game on the right, I choose ○ A ○ B

To confirm your choices, please click OK when you are done. If you would to see which of the two games you actually played, click the "Reveal Game" button before clicking OK.

○ Reveal Game

OK

Player X chooses	Player X receives	Player Y receives
A	6	?
B	5	?