Self-Image and Willful Ignorance in Social Decisions

Zachary Grossman\textsuperscript{a}, Joël J. van der Weele\textsuperscript{b}

\textsuperscript{a}Department of Economics, 2127 North Hall, University of California, Santa Barbara, CA 93106, Tel: 805-893-7309, Fax: 805-893-8830

\textsuperscript{b}CREED, Department of Economics, University of Amsterdam, Roeterstraat 11, 1018WB Amsterdam, the Netherlands. Tel. +31 (0)20 5254213.

Abstract

Avoiding information about adverse welfare consequences of self-interested decisions, or \textit{willful ignorance}, is an important source of socially harmful behavior. We analyze a Bayesian signaling model of an agent who cares about self-image and has the opportunity to learn the social benefits of a personally costly action. We show that willful ignorance can serve as an excuse for selfish behavior by obfuscating the signal about the decision maker’s preferences, and serves to maintain the idea that the agent would have acted virtuously under full information. We derive several behavioral predictions that are inconsistent with either outcome-based preferences or social-image concern and conduct experiments to test them. Our findings, as well as a number of previous experimental results, offer support for these predictions and thus, the broader theory of self-signaling.

\textit{Keywords:} prosocial behavior, dictator games, willful ignorance, self-signaling

\textit{JEL:} D83, C72, C91

\footnote{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.}

Email addresses: grossman@econ.ucsb.edu (Zachary Grossman), vdweele@uva.nl (Joël J. van der Weele)
“Living is easy with eyes closed.”


1. Introduction

Deliberate avoidance of evidence about the social impact of one’s decisions, or “willful ignorance”, plays an important role in prosocial behavior.\(^1\) For example, Niehaus (2014) argues that donors to charitable causes remain willfully ignorant of the effectiveness of the charity in order to maintain a feeling of warm glow about their contributions. Another example relates to the phenomenon of climate change. Despite the longstanding claim by the Intergovernmental Panel on Climate Change (IPCC) that the scientific evidence for anthropogenic climate change is “unequivocal” (IPCC, 2007), substantial minorities in many countries do not believe that climate change is man-made, or even real (Hobson and Niemeyer, 2013). Sociological studies based on in-depth interviews and focus-group interactions show that many people are reluctant to learn about climate change or to engage with their existing knowledge (Stoll-Kleemann \textit{et al.}, 2001; Norgaard, 2006a).\(^2\)

More structured evidence for willful ignorance in a social context comes from the economic laboratory. Experimental participants are reluctant to learn how their choices affect others, even if that information can be obtained without cost. At the same time, when provided with the knowledge that a self-interested choice hurts others, most people are willing to sacrifice personal payoffs to implement a fairer outcome, as documented in experiments by Dana \textit{et al.} (2007), Ehrich and Irwin (2005) and several others. These paradoxical results indicate that people cultivate uncertainty about the outcomes of their actions for others in order to justify self-interested decisions.

These findings pose a puzzle for standard economic models of social behavior, which rely on preferences over distributions of outcomes. These models predict that if people reveal a willingness to sacrifice for a fair distribution of outcomes, they should also acquire costless information about the consequences of their actions. Concerns for reputation or social image cannot explain the participants’ decision to ignore information either, as decisions were one-shot, anonymous and no participant observed whether or not the decision maker actually chose to be ignorant.

Instead, several authors suggest that ignorance can serve the purpose of protecting the decision-maker’s self-image (e.g., Dana \textit{et al.}, 2007; Bénabou and Tirole, 2006), a source of motivation that

---

1 Several popular overviews of instances in which willful ignorance occurs have been published in the last 15 years, including “States of Denial” (Cohen, 2001), “The elephant in the room” (Zerubavel, 2007), and “Willful blindness” (Heffernan, 2012).

2 In the U.S., such strategies have occasionally made their way into public policy. Lawmakers in North-Carolina passed a bill restricting local planning agencies’ ability to use climate change science to predict sea level rise in several coastal counties (Schwartz, 2012). In Florida, governor Rick Scott forbade environmental officials to use the terms “climate change” or “global warming” in official communications or reports, see http://www.miamiherald.com/news/state/florida/article12983720.html#storylink-cpy.
has long been recognized by psychologists (e.g., Bem, 1972; Baumeister, 1998; Fiske, 2013). This motive is echoed in the literature on climate change perceptions. On the basis of her interviews with Norwegian subjects, Norgaard (2006a) cites the “fear of being a bad person” as a reason to avoid knowing about difficult tradeoffs associated with climate change.

Although plausible, a self-image explanation for willful ignorance 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 image value? To address this question, we systematically study willful ignorance in social decisions, using both theoretical analysis of a Bayesian preference-signaling model and a set of experiments designed to examine its implications. Our theory builds on Bodner and Prelec (2003) and Bénabou and Tirole (2006), who incorporated self-image concerns in economics in the context of a dual-self signaling model. The key idea is that a decision-maker self uses her choices to optimally manage her image vis-a-vis an observer self, who cannot directly introspect or recall the motivations underlying the behavior. We assume that the decision-maker self cares about her own material well-being and—to some degree—that of others, as well as her image as an altruistic person in the eyes of the observer self. The decision-maker self faces two decisions: whether to inform herself about the social benefits of an action that is personally costly, and, subsequently, whether to engage in that action. This setting captures the essential features of willful ignorance in social decisions and as such generalizes the structure of experiments on willful ignorance (specifically Dana et al., 2007, henceforth, DWK).

Our first contribution is to prove the existence of an “ignorance equilibrium” in which ignorance is strictly preferred by agents who are not very altruistic but care about their self-image. The value of ignorance in this equilibrium arises from a novel mechanism, namely that people prefer to avoid environments with high incentives for costly signaling. Ignorance serves to obfuscate the choice the agent would have made in such an environment, and limits the inferences that can be made about the motivations of the decision-maker. Although the choice to avoid information itself reduces self-image compared to a common prior, the agent can plausibly maintain that she would have acted virtuously under full information. According to Norgaard (2006b), this psychological mechanism is indeed active in willful ignorance about climate change, as she concludes on the basis of her interviews that “[T]o “not know” too much about climate change maintains the sense that if one did know one would have acted more responsibly” (p. 365).

Our second contribution is to derive a number of testable implications from this self-image theory. In Section 3, we show that the ignorance equilibrium implies five distinct behavioral patterns. First, signaling incentives to behave pro-socially are weakened when there is initial uncertainty about the outcomes for others, even if there is an option to lift the uncertainty. This can lead to Information Avoidance as observed in the experiments of DWK, implying that the same people who act unselfishly under full information choose not to obtain relevant information even if it is free. Information avoidance may also be one of the reasons for the slow diffusion of the scientific
consensus on climate change in some communities.

Second, information is mostly acquired by relatively altruistic types who have a lower cost of prosociality, a phenomenon we call Sorting. Sorting implies that people who actively chose to obtain information will on average behave more prosocially than those who obtained the same information passively. Campbell and Kay (2014) show that climate change perceptions demonstrate such sorting. In their experiments, people who perceive higher economic or ideological costs of environmental regulations are more likely to be skeptical of scientific statements about climate change.

Third, willful ignorance provides exculpation, as people who act selfishly are judged more harshly when they acted with full information of the adverse social consequences than when they were ignorant, even if the ignorance was willfully chosen. Fourth, if the agent has already revealed her type in some way or another, ignorance can no longer serve to obfuscate the signals about the decision maker’s character. Such a situation occurs when the information acquisition choice follows the action choice instead of preceding it. One would expect the resulting non-willful ignorance to be less prevalent than willful ignorance. Finally, infra-marginal agents choosing ignorance in equilibrium are strictly better off without information, so paying for ignorance is consistent with the self-image model.

Our third contribution is to provide evidence for each of these five implications of the ignorance-equilibrium, derived from a series of new experiments, presented in Section 4. These experiments, built around the moral wiggle-room game of DWK, support all five predicted behavioral patterns. While previous studies provided evidence for the implied behavioral patterns in scattered environments, our experiments are designed to examine the implications directly in a unified framework. In Section 5, we use our theoretical results to organize the existing evidence and show systematically that other explanations cannot adequately account for the observed findings.

Taken together, our theoretical and experimental results provide strong evidence that self-signaling is an important driver of behavior in social situations, as well as support for the broader theory of self-signaling. There have been many previous indications of the importance of self-image (Murnighan et al., 2001; Mazar et al., 2009; Fischbacher and Föllmi-Heusi, 2013; Gneezy et al., 2012). These papers typically invoke self-signaling after more mainstream explanations fail, rather than experimentally manipulating a self-signaling model. In contrast, more direct tests of predictions derived from signaling models have not found evidence in support of self-signaling in giving behavior (Grossman, 2015; Van der Weele and Von Siemens, 2014). Through the focus on the information acquisition choice, we are able to provide a compelling and theoretically informed explanation of the role of Bayesian self-signaling in social decisions.

This paper thus also contributes to a growing theoretical literature on self-signaling. The model is adapted from that of Bénabou and Tirole (2006) and closely related to that of Grossman (2015). Despite the emphasis on self-image concerns as a driver of prosocial behavior, it is technically sim-
ilar to the social-signaling models of Ellingsen and Johannesson (2008), Andreoni and Bernheim (2009), and Tadelis (2011). Bénabou and Tirole (2011) use a similar signaling model to analyze willful 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. Bénabou (2013) combines a self-signaling model with anticipatory utility to explain collective delusions in groups, whereas we focus on individual information avoidance in the absence of anticipatory utility. 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 ours: 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 willful ignorance in social decisions also contributes to a broader literature of non-signaling models examining willful ignorance in various contexts. Nyborg (2011) provides a theoretical investigation of public-good contributions where duty-oriented agents experience disutility if their contribution falls short of an exogenously defined ‘ideal’ contribution. She shows that this can lead people to ignore information about the true value of a public good. While our approach has some similarities, our use of a canonical self-signaling model endogenizes the self-image value of behavior and obviates the need for additional assumptions. Furthermore, our model organizes a broad constellation of behavior and generates new, testable hypotheses for which we provide experimental evidence.

We argue that ignorance serves to avoid large signaling investments in prosocial behavior that would be necessary to maintain a good self-image under full information. Other papers have explored ignorance as a commitment device in different contexts, for example in the presence of time-inconsistency (Carillo and Mariotti, 2000; Bénabou and Tirole, 2011). Ignorance 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; Domingues-Martinez et al., 2014) or by reducing moral hazard (Crémer, 1995). In addition, a growing theoretical and empirical literature in economics shows that people downplay negative feedback and are sometimes willing to pay not to receive any feedback at all (Köszegi, 1996; Möbius et al., 2011; Eil and Rao, 2011; Grossman and Owens, 2012). Although these papers have some intuitive parallels with this study, they do not address the moral trade-off that is at the heart of our analysis.

Finally, our research relates to the large literature in social psychology on ‘moral disengagement’ (Bandura et al., 1996) and ‘motivated cognition’ (e.g. Kunda, 1990), and more recently, work on information avoidance in individual decisions (Sweeny et al., 2010). In sociology, there is also a nascent and growing literature on ignorance resulting from varying motivations (e.g. McGoey, 2012).
Our paper relates to these literatures by highlighting self-image as a source of willful ignorance, and formalizing the information acquisition decision in a model of Bayesian rationality.

2. Equilibrium ignorance in a model of self-image concerns

Why do people choose not to know the consequences of their own actions? To answer this question, we apply a signaling 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. Preference-signaling models admit two distinct interpretations that are technically equivalent — a social-signaling interpretation with an external observer, and a self-signaling interpretation with an internal observer. In light of the experimental evidence that we will discuss in more detail below, our focus is on the latter.

In the self-signaling interpretation, the decision-maker and observer roles are viewed as two aspects of a divided self, rather than separate people. The informed, decision-making self knows her preferences and acts upon them. Her aim is to impress an uninformed observer self who lacks introspective knowledge of her preferences, 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. Bodner and Prelec (2003) introduced the dual-self signaling-model approach and it was subsequently adopted by others such as Bénabou and Tirole (2006, 2011) and Grossman (2015).

2.1. The model

The outline of the model is as follows. An agent chooses whether to take a prosocial action \( (a = 1) \) or not \( (a = 0) \). When taking the prosocial action she incurs a material cost \( c \), and the agent is uncertain of the action’s impact on social welfare, \( W \). She knows that \( W = w \) with prior probability \( p \), where \( w > c \), 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 ‘willful ignorance’. For simplicity, information takes the form of a perfectly informative signal \( \sigma \in \{\sigma_w, \sigma_0, \emptyset\} \), where \( \sigma_w \) denotes a ‘high signal’ \( (W = w) \), \( \sigma_0 \) denotes a ‘low signal’ \( (W = 0) \), and with some abuse of notation \( \emptyset \) denotes the case in which no information is acquired.

Preferences. Each agent is characterized by a preference type \( (\theta, \mu) \), which is drawn by nature at the start of the game from a commonly known distribution described below. The type is known to the decision-maker self, but not to the observer self. Preferences are represented by the following
utility function

\[ u(\theta, a, I, \sigma) = a (\theta E[W|\sigma] - c) - kI + \mu E[\theta | \sigma, a; s]. \]  

The first term represents the material payoff from her outcome choice and the last term represents utility from image concerns. If the agent takes the action, her material payoff consists of the expected welfare benefit \( E[W|\sigma] \) weighted by her social-preference parameter \( \theta \), minus the cost of the action, and it is zero if she does not take the action. Thus, \( \theta \) can be interpreted as the degree of altruism or prosocial motivation of the agent. The fact that types with low values of \( \theta \) (low types) care less about welfare implies a single crossing-property that underlies the semi-separating equilibrium in the next section. Note that if the agent acquires information, she will learn the actual welfare benefit. Depending on the signal’s content, the expectation collapses to either \( E[W|\sigma_w] = w \) or \( E[W|\sigma_0] = 0 \). If she remains ignorant, her payoff is the expectation \( E[W|\emptyset] = pw \). 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 final term \( E[\theta | \sigma, a; s] \) reflects the utility from (self-)image, given by the expectation of the agent’s social-preference parameter, \( \theta \), taken over the observer’s posterior beliefs. Bodner and Prelec (2003) call this the “diagnostic” part of the utility function, as it reflects an assessment of the character of the agent. Following Bodner and Prelec (2003), we assume that the decision-making self knows and acts upon her true preferences \( \theta \) and \( \mu \). This knowledge may be based on experiences during the act of the decision, for example the agent’s emotional reaction to appeals for charity or images of human suffering. While these emotions are informative about the agents type, they are non-verifiable, and hard to imagine for anyone not involved in the actual decision making situation. Since the observing or future self lacks this knowledge of the agent’s type, the best she can do is to update beliefs about \( \theta \) and \( \mu \) based upon the available information, namely the agent’s choices and the signal \( \sigma \). Note that although both \( \theta \) and \( \mu \) are subject to the observer’s inference, we assume that image utility depends only on beliefs about the agent’s altruism \( \theta \). In doing so, we follow a large literature on the costly signaling of altruistic preferences. In the analysis, we will not explicitly discuss the simultaneous inferences the observer makes about \( \mu \).3

We assume the observer can condition her posterior beliefs on both the action taken and the signal obtained by the agent, but not on the counterfactual actions that the agent would have taken if she had observed a different signal. The assumption that the observer knows the information obtained by the agent would be quite strong in a social-signaling model, but is natural for an

---

3Other conceptions of self-image concern are certainly possible, but our objective is to examine whether existing, previously-proposed models can explain behavior. By hewing closely to standard assumptions for self-signaling models, we retain emphasis on what this category of model can explain, rather than the novelty of our particular model.
internal observer.\textsuperscript{4} Similarly, since the dual-self model is premised upon the inability to perfectly introspect, it is natural that the observer sees only those actions that actually occurred, even though the strategy for the decision-making self also specifies behavior for counterfactual realizations of the signal.\textsuperscript{5}

Our interpretation of the weighting parameter $\mu$ is that it captures directly the importance of image concerns to the agent. Indeed, psychologists emphasize the purely psychological rewards of positive self-esteem (Fiske, 2013). Alternatively, 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. For example, a good self-image can generate future payoffs as it may lead to increased motivation or social benefits that are internalized by the decision maker self.

\textit{Heterogeneity.} With respect to the heterogeneity of types, we assume a common prior of the following distribution:

1. With probability $1 - \varepsilon$ the agent is a social agent, who cares about her image and about social welfare. For social agents, $\theta$ is distributed according to $F(\theta)$ with full support on $[0, 1]$. Following the practice of a large line of preference-signaling models, we take $\mu$ to be homogeneous for social agents.\textsuperscript{6} We assume that image concerns are small relative to material concerns ($0 < \mu < c$), which rules out the agent choosing the prosocial action purely for image reasons and situations in which any action could be sustained as a pooling equilibrium.

2. With probability $\varepsilon$ the agent is non-social agent, or homo economicus, who only cares about her own material payoff, i.e. $\theta = \mu = 0$. The existence of homo economicus is suggested by studies on prosocial behavior in for example the public goods game, where a minority of subjects act like free riders (e.g. Fischbacher \textit{et al.} (2001), Kurzban and Houser (2005), and Burlando and Guala (2004)).

The resulting distribution of types is consistent with evidence from dictator games, which typically features a large spike of selfish choices as well as a dispersed distribution of more generous behavior (see Engel, 2011, for an overview).\textsuperscript{7}

\textsuperscript{4}This assumption may be satisfied in a social-image model in situations where behavior comes under legal scrutiny. Examples are the Nuremberg, Enron, and Watergate trials, where a central issue to the prosecution was who knew what when. If this assumption is not satisfied and beliefs cannot be conditioned on the signal, ignorance cannot serve as an excuse.

\textsuperscript{5}In this sense, the model can be said to involve ‘self-deception’. However, as in standard signaling models, the choices of the both the decision-maker self and the inferences of the observer self are completely rational.

\textsuperscript{6}While modeling heterogenous degrees of image-concern may yield interesting general results for preference-signaling models, homogenous $\mu$ better serves our aim to explain willful ignorance in social decisions with a tractable and well-precedented model of self-image concern.

\textsuperscript{7}Our main result, establishing the existence of an equilibrium with willful 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.
**Timing.** Summarizing, the timing of the game is as follows:

1. Nature selects the level of social benefit, \( W \in \{0, w\} \), associated with activity \( a \), as well as the agent’s type, \( (\theta, \mu) \).
2. The agent chooses whether to receive a signal informing her about the level of \( W \) or an empty signal and observes that signal.
3. The agent chooses whether to take the prosocial action \( (a = 1) \) or not \( (a = 0) \).
4. The agent’s action \( a \) and the signal content \( \sigma \) are perceived by the observer self, who updates beliefs about the agent’s type, and payoffs are realized.

Note that there are two kinds of uncertainty at play here. First, the observer self is uncertain of the agent’s type. Second, both selves are initially uncertain of the state \( W \) and whether or not that uncertainty is resolved is determined endogenously (for both selves) by the decision-maker’s choice.

### 2.2. Solution concept

We make predictions using perfect Bayesian equilibrium, which requires that all types play a strategy \( s^* \) that maximizes their utility given the behavior of the other types, and that beliefs are formed by the application of Bayes’ rule wherever possible. We assume the tie-breaking rule that all agents who are indifferent between acquiring information or not will acquire information. Given that we want to explain willful ignorance, this represents the most conservative assumption and ensures our results are not driven by indifferent agents. While this tie-breaking assumption is not necessary for our main propositions, it greatly simplifies the exposition.\(^8\)

Though we consider potential pooling equilibria in Section 2.4, our main focus is on semi-separating equilibria, which are equilibria with a cut-off value \( \theta^* \in [0, 1] \), such that all social types \( \theta \geq \theta^* \) acquire information, and social types \( \theta < \theta^* \) do not. The possibility of multiple equilibrium thresholds and unstable equilibria arises here, because the threshold \( \theta^* \) depends on the beliefs of the observer, which are themselves a function of the equilibrium threshold (and the behavior of the homo economicus, which depends on \( k \)).

Under what conditions can we rule out such multiplicity and instability? Consider the beliefs of the observer about \( \theta \). Upon seeing information acquisition and ignorance, these beliefs are \( E[\theta \mid \theta \geq \theta^*, k] \) and \( E[\theta \mid \theta < \theta^*, k] \), respectively, and we can define the image reward of information acquisition as \( \delta(\theta^*, k) := E[\theta \mid \theta \geq \theta^*, k] - E[\theta \mid \theta > \theta^*, k] \). In any ignorance equilibrium, the cut-off type must be indifferent between the expected cost of information \( k - p(\theta^* w - c) \) and the

\(^8\)A previous version of this paper provided proofs of more general versions of the propositions featuring weaker assumptions about the tie-breaking rule. Notably, Propositions 1 and 3 hold for any assumption about the probability that an indifferent agent will choose information and the main claim of Proposition 2 holds as long as the fraction of indifferent agents choosing information is sufficiently high.
image reward $\mu \delta(\theta^*, k)$, so the net benefit is $k - p(\theta^* w - c) - \mu \delta(\theta^*, k) = 0$. A sufficient condition for stability of the threshold is that the total derivative of this expression with respect to $\theta^*$ is positive.\(^9\) Moreover, a sufficient (but not necessary) condition for uniqueness of the threshold is that the total derivative is positive for all $\theta$. This implies that the expected cost of information is always decreasing in more steeply than the image reward, which implies that there is at most one $\theta$ for which they are equal.

For the remainder of this paper, we will assume that this is the case, i.e.

$$\frac{d\delta(\theta, k)}{d\theta} > -\frac{pw}{\mu}, \quad \forall \theta. \quad (2)$$

Here, the left-hand side depends on the density $f(\theta)$, and will be small if $f(\theta)$ does not increase too steeply anywhere on its domain. In this case, and if $\mu$ is not too large, (2) is likely to be satisfied (see Bénabou and Tirole, 2006, p. 1667 - 68 for a more extensive discussion). In Section 5.3 we show that the threshold is stable and unique in an example featuring a uniform distribution of social types.

With respect to beliefs about out-of-equilibrium actions, we require that these beliefs satisfy a refinement we call “pD1”, defined as follows. Denote the observer’s beliefs by $\phi(a, \sigma; \theta) \in [0, 1]$, and the equilibrium payoffs of the decision maker with altruism level $\theta$ by $u^*(\theta)$. We define $D(\theta) := \{\phi : u(\theta, a, I; \phi) > u^*(\theta)\}$, i.e. the set of observer beliefs for which a type with altruism $\theta$ would like to deviate from the equilibrium. pD1 requires that $\phi(a, \sigma; \theta')$ is positive only if $D(\theta')$ is maximal, i.e. not a proper subset of any $D(\theta)$. pD1 straightforwardly adapts the D1 criterion to psychological games by requiring that the receiver’s beliefs put weight only on types that deviate for the “largest” set of off-equilibrium beliefs.\(^{10}\)

The proofs for all theoretical results are found in Appendix A. In Section 5 we provide a broader discussion of the plausibility of the assumptions and conditions upon which our results depend.

2.3. Existence of a willful ignorance equilibrium

The main theoretical result of the paper is the existence of an equilibrium in which some social types prefer to remain ignorant.

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

\(^9\)Stability here means that if the threshold type switched behavior, it is optimal for other types to stick to their strategies.

\(^{10}\)The traditional D1 refinement defines the set $D(\theta)$ over the set of best responses of the observer (Banks and Sobel, 1987). In psychological games like ours, the sender’s payoffs depend on the receiver’s beliefs. By defining $D(\theta)$ over the receiver’s beliefs we follow Ellingsen and Johannesson (2008) and Andreoni and Bernheim (2009). Kolpin (1992) discusses the interpretation and theoretical underpinning of this idea.
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 why social types would like to remain ignorant even if this is costly, consider the trade-offs in the information-acquisition decision. In equilibrium, the material payoffs for the decision maker and the image that accompanies ignorance are certain. Willful ignorance is always followed by not taking the action, so the agent avoids the cost of prosocial behavior. However, she suffers a utility cost of \( p\theta w \), the expected disutility stemming from social preferences. Because she pools with the lower types, ignorance also lowers her image relative to the prior expectation.

On the other hand, the material payoffs and image that result from acquiring information are uncertain. When the realized signal shows no welfare benefit of the action \( a \), the agent can pool with the highly altruistic types without any material sacrifice. However, when the signal shows that there is a welfare benefit, she faces a choice between two evils. She can pay the price \( c \) to take the action and obtain a high image or she can decline to take the costly action and end up with the lowest possible image. This image is low either because only the homo economicus chooses this action (if \( k \leq 0 \)), or because equilibrium beliefs for this out-of-equilibrium action are specified to be 0 (if \( k > 0 \)).\(^{11}\) Between these options, social agents who care about image but only little about welfare (\( \theta < \theta^* \)) strictly prefer to remain ignorant when the probability of a high signal is sufficiently large.

Thus, in this equilibrium, ignorance serves to avoid an environment with strong signaling incentives to behave prosocially and to obfuscate the choices that the agent would make in such an environment. To see this, note that the strategy of the types just below \( \theta^* \) specifies that they behave pro-socially after observing a high signal. Willful ignorance thus protects self-image because the observer self cannot refute the claim that if the agent had chosen to inform herself, she would have been sufficiently concerned with social welfare to behave pro-socially.

The equilibrium poses several restrictions: if \( p \) is too low, then all social types will want to acquire information, because it is not very likely that they will face pressure to be prosocial. The resulting equilibrium is similar to the “information seeking” equilibrium described in Section 2.4. Similarly, if \( k \) is too high or too low, no-one or everyone will acquire information. Note also that our earlier restriction that \( \mu < c \) rules out the possibility that the image reward of information outweighs the material cost for even the least socially-minded people and all types will pool on information seeking. Conversely, in the absence of image concerns (i.e. if \( \mu = 0 \)), social types would follow their instrumental interests. If \( k < 0 \), all type would pool on information and sufficiently large \( k \) or \( c \)

\(^{11}\)These beliefs satisfy the pDI refinement. Negative beliefs could also be justified on the equilibrium path 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.
would lead to pooling on ignorance.

2.4. Multiplicity and information-seeking equilibrium

Like many signaling games, this game can feature multiple equilibria. Specifically, the model admits another pure-strategy equilibrium where all social types acquire information. In this “information-seeking” equilibrium, beliefs associated with ignorance are very low, as either the homo economicus is the only one to choose ignorance (if $k > 0$) or off-equilibrium beliefs specify a low image for this action (if $k \leq 0$). Since the image associated with acting selfishly under ignorance is lower than doing so under full information, ignorance is never chosen by the social agents. The information-seeking equilibrium exists when the cost of information $k$ is not too high, and satisfies the same refinement as the ignorance equilibrium.

The coexistence of an ignorance equilibrium and an information-seeking equilibrium suggests the possibility of different ‘social norms’. One norm, associated with the ignorance equilibrium, punishes those who refuse to choose the prosocial action after having become fully cognizant of its social benefits. Another norm, associated with the information-seeking equilibrium, punishes ‘hypocritical’ agents who would rather stick their heads in the sand than face up to the consequences of their decisions.

Both these equilibria seem theoretically plausible, and which of them, if any, will be played is an empirical question. This question can only be settled by deriving behavioral predictions that distinguish between the two equilibria. As we will show below in detail, the ignorance equilibrium described in Proposition 1 can explain the otherwise puzzling findings of previous experiments, as well as the experiments reported in this paper. By contrast, the information-seeking equilibrium cannot do so, as we will argue in more detail in Section 5.2.

3. Empirical Implications

In this section, we derive testable implications from the ignorance equilibrium discussed in Proposition 1. In the next section, we present the results of experiments designed to test these implications. In Section 5 we discuss competing predictions derived from the information-seeking equilibrium identified above, and from models of social or distributional preferences.

3.1. Information avoidance

One of the most striking results in the literature is the finding by DWK that while 74% of their subjects sacrifice payoffs to obtain a fair outcome in the binary dictator game with full information, only 56% choose to acquire free information about the potentially adverse consequences to the recipient of a self-interested choice. 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. Furthermore, Feiler (2014) shows that when subjects are
confronted with both a binary dictator game with full information and with several variations of the information choice game described in DWK, 30% choose fairness under full information but remain ignorant at least in some of information choice games.

Thus, a first test of the self-signaling model is whether it can explain these results, which are inconsistent with theories of outcome-based preferences and social signaling as we will show in Section 5.2. To investigate this, we denote by $\Gamma_C$ the information “choice game” explained in Section 2.1 (where subscript $C$ stands for choice) and denote by $\Gamma_I$ the simpler game in which it is common knowledge that $W = w$, and the only choice is whether or not to take the prosocial action (where subscript $I$ stands for informed). This latter game represents DWK’s baseline treatment. To compare the fraction of those who behave selfishly in $\Gamma_I$ with the share that remains ignorant in game $\Gamma_C$, we first derive the equilibrium in game $\Gamma_I$. As we did in game $\Gamma_C$, we focus on a semi-separating equilibrium characterized by a unique cut-off type.

Following the same reasoning that led to condition (2), we assume that a stability condition holds for this game to ensure uniqueness of the threshold type. Let $\zeta$ be the image reward of being prosocial as a function of the threshold type $\hat{\theta}$ and the fraction of homo economicus, i.e. $\zeta(\hat{\theta},\varepsilon) := E[\theta|\theta > \hat{\theta},\varepsilon] - E[\theta|\theta > \hat{\theta},\varepsilon]$. The threshold type is such that the net cost of the prosocial action equals the image benefits, i.e. $\hat{\theta}w - c + \mu\zeta(\hat{\theta},s) = 0$. A sufficient condition for uniqueness and stability is that the total derivative of this expression is positive, i.e. $\frac{d\zeta(\theta,s)}{d\theta} > -\frac{w}{\mu}$ for all $\theta$.

Lemma 1. In game $\Gamma_I$, 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 $a = 1$.

We compare this equilibrium with the ignorance equilibrium in the choice game $\Gamma_C$ where, consistent with the experimental literature we assume $k = 0$, and obtain the following result.

Proposition 2 (Information avoidance). There exist $\bar{\mu}$ and $0 < \xi < \bar{\varepsilon} < 1$ such that if $\mu > \bar{\mu}$ and $\varepsilon \in [\xi,\bar{\varepsilon}]$, then the fraction of all agents—including both social agents and homo economicus—who choose ignorance in the ignorance equilibrium of $\Gamma_C$ is higher than the total fraction of all agents who act selfishly in $\Gamma_I$.

Proposition 2 shows that the signaling model can explain information avoidance, given two conditions on preferences. The intuition is that social types want to avoid pooling with the homo economicus, which has different effects on their behavior in the two games. In game $\Gamma_I$, the homo economicus chooses $a = 0$. This increases the signaling value of a prosocial action and induces some marginal social types to behave prosocially. By contrast, in game $\Gamma_C$, the homo economicus will choose information. This decreases the signaling value of acquiring information and causes marginal social types in game $\Gamma_C$ switch to ignorance, increasing the amount of selfish choices.

12 The tie-breaking assumption that all indifferent agents inform themselves is not necessary for the result. What is necessary is that the share of indifferent agents who choose to inform themselves is large enough.
The conditions on the parameters are sufficient to ensure that the amount of social types who shift towards ignorance in game $\Gamma_C$ outweighs the number *homo economicus* who choose to reveal. First, $\varepsilon$ should neither be too large nor too small; when there are too many *homo economicus* there will be little ignorance, and when there are too few they will not sufficiently dilute the image associated with choosing information. Second, the importance of image concerns, as measured by $\mu$, needs to be high enough to induce enough social types to shift towards ignorance. If these conditions are not satisfied, willful ignorance may still occur, but the fraction of ignorant agents may not outweigh those who are selfish in game $\Gamma_I$.

In sum, 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 choosing information means pooling with some of the ‘worst’ types. This dilution of signaling incentives can explain why willful ignorance can exceed selfish behavior in environments without uncertainty.

### 3.2. Sorting

In the ignorance equilibrium certain types of people choose information. As a result, people who actively chose to obtain information will on average behave more prosocially than those who obtained the same information passively. To investigate this idea formally, we compare the the behavior of those who inform themselves in the ignorance-choice game $\Gamma_C$ to the behavior of the players in the game with full information $\Gamma_I$ (defined above).

**Proposition 3 (Sorting).** If $k > 0$, the fraction of agents that engages in prosocial behavior ($a = 1$) in the full information game $\Gamma_I$ is lower than the corresponding fraction amongst agents who inform themselves in the ignorance equilibrium choice game $\Gamma_C$ and received signal $\sigma = \sigma_w$. If $k \leq 0$, there exists an $0 < \tilde{\varepsilon} < 1$, such that the same result holds when $\varepsilon < \tilde{\varepsilon}$.

The intuition behind this result is that average prosocial behavior in the full information game $\Gamma_I$ is taken over all agents in the population, where average altruism is equal to $E[\theta]$. Only the most altruistic types in this sample will choose to behave pro-socially, where the cutoff type depends on the parameters $w$, $c$, and $\mu$. In comparison, in the choice game $\Gamma_C$ it is mostly the altruistic types who will choose to inform themselves, so conditional on information acquisition, altruism is higher than $E[\theta]$ and prosocial behavior is higher than in the full-information game.

Note that if $k > 0$, the *homo economicus* chooses ignorance, so all agents who acquire information will be prosocial if they see signal $\sigma = \sigma_w$. If $k \leq 0$, then the *homo economicus* chooses information, which lowers the relative incidence of prosocial behavior amongst those who are informed. In that case, the condition $\varepsilon < \tilde{\varepsilon}$ is necessary and sufficient for the result. If the share of *homo economicus* becomes very large, then prosocial behavior may actually be lower amongst those who choose to inform themselves.
3.3. Exculpation

Sorting of types into different actions implies that in equilibrium, where the image associated with an action corresponds to the expected type choosing that action, the image associated with being selfish depends on the preceding information-acquisition choice. Specifically, in the ignorance equilibrium, agents who do not take the social action are perceived as more selfish when they made their choice with full information of the adverse social consequences compared to when the were uninformed. It is this feature of the equilibrium that drives the social types $\theta < \theta^*$ to strictly prefer ignorance. The following corollary makes this statement more precise.

**Corollary 1 (Exculpation).** Agents who choose $a = 0$ have a higher image if they do so after choosing not to see a signal than after seeing $\sigma = \sigma_w$.

This result follows directly from the order of beliefs in the ignorance equilibrium, which is $E[\theta | w, 0] < E[\theta | \emptyset, 0] < E[\theta | w, 1]$. Thus, the image value associated with different actions is lowest for people who engage in harmful acts with knowledge of the harmful consequences, intermediate for ignorant agents, and highest for those who abstain from harmful acts after having acquired knowledge of the harmful consequences. This means that willful ignorance can indeed be used as an ‘excuse’ to oneself or others, even if it is clear that one could have chosen to know the actual consequences.

3.4. Willful vs. non-willful ignorance

The value of ignorance in equilibrium comes from avoiding the tradeoff between taking a costly prosocial action or being revealed as a selfish individual. Consider, however, a situation where the agent has already revealed her type in some way or another, and faces the decision to obtain information about the state only after this revelation. In this case, the decision maker can no longer use ignorance to hide the choices she would make under full information. Therefore, one would expect more people to choose information.

To establish this intuitive claim more formally, consider the modified game, $\Gamma_N$, which shares almost all the same features as the choice game $\Gamma_C$ described in Section 2.1. However, the action choice takes the form of a contingent strategy, formally defined as a vector, $a = (a_0, a_w)$, with each component indicating whether or not the dictator chooses to take the prosocial action in each of the respective states of $W$. Moreover, the timing of the information choice and action choice are reversed, so that the action choice is made before the signal about $W$ is chosen. Thus, the agent is forced to reveal her behavior, and therefore her type, in the case where there is a trade-off between her own welfare and that of others.\(^{13}\)

**Proposition 4 (Non-willful ignorance).** When $k \leq 0$, the fraction of agents choosing ignorance in $\Gamma_N$ is lower than the corresponding fraction in the ignorance equilibrium of $\Gamma_C$.

\(^{13}\)Because it is the timing of the observation of the signal that is important, the information choice itself may actually be made simultaneously with the action choice.
To see the logic behind this result, consider both games in turn. In game $\Gamma_C$, when $k \leq 0$, a fraction $(1 - \varepsilon)F(\theta^*)$ strictly prefers not to have information (i.e. the social types who care mainly about image). These agents will forgo information’s instrumental value in exchange for ignorance’s image value if they can do so before the outcome decision. Therefore, under the tie breaking rule that indifferent agents reveal, a fraction $1 - (1 - \varepsilon)F(\theta^*)$ will reveal in game $\Gamma_C$. In contrast, in $\Gamma_N$, neither information nor ignorance has any value after the decisions about $a$ have been made and the agent has revealed her type, so all agents will reveal.\footnote{Note that when $k = 0$ our result depends on our tie-breaking rule, as the \textit{homo economicus} is indifferent about information acquisition in both games, and all agents are indifferent in game $\Gamma_N$. We can relax our tie breaking rule, such that when a fraction $\alpha$ of indifferent agents reveals, the proposition holds for $\alpha > 1 - F(\theta^*)$.}

### 3.5. Paying for ignorance

The ignorance equilibrium can be used to derive comparative statics on the price of information, $k$. All other things being equal, a positive price of ignorance, i.e. $k < 0$, makes ignorance less attractive for all agents and causes $\theta^*$ to be lower than it would be if ignorance is costless. Nevertheless, the social types $\theta < \theta^*$ are strictly better off paying for ignorance, as it beats the alternatives, namely pooling with the \textit{homo economics} under full information, or pooling with the altruists at a cost of $c$.

**Corollary 2 (Paying for ignorance)**. A positive share of agents in ignorance equilibrium is willing to pay to remain ignorant.

### 4. Experimental tests

Previous experiments separately provide evidence relating to each of the implications derived above, but no single experiment covers all five patterns. In order to provide a comprehensive set of results on all five implications in a unified framework, we conducted our own experiments with treatments designed specifically to examine each behavioral prediction. After presenting the experiments and results in this section, we discuss in Section 5.1 their relationship with existing experimental results.

Note that our predictions and thus the alternative hypotheses in our experimental tests, represent behaviors and outcomes consistent with ignorance equilibrium and inconsistent with other models. In contrast, our null hypotheses reflect behavior potentially consistent with multiple models, including the non-ignorance equilibrium in the self-image model. Thus, our statistical tests are not designed to reject ignorance equilibrium in the self-image model, rather, to reject reasonable alternatives. In Section 5.2 we consider alternative explanations for our results in detail.

Our experiments are based on the design of DWK and implement some subtle variations of their original game. 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 yielded 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’.

Before participants were told which role had been assigned to them 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.\(^\text{15}\)

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 was programmed using the Z-Tree software package (Fischbacher, 2007) and subjects were recruited using the ORSEE system (Greiner, 2003). The experiment was carried out at the Experimental and Behavioral Economics Laboratory (EBEL) at the University of California, Santa Barbara.

The dictator made her decision anonymously and 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. Appendix B provides descriptive statistics of our data and instructions are provided in Appendix C.

4.1. Information Avoidance

We begin by replicating the avoidance result of DWK, which plays a central role in the literature on willful ignorance and is one of the most important inspirations for this paper. The result has seen multiple replications (Larson and Capra, 2009; Feiler, 2014; Grossman, 2014), so our primary objective is not to establish the robustness of this result. Rather, we seek to a) establish a baseline

\(^{15}\)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.
of comparability of our methods so as to eliminate doubts about our other results; b) provide our own evidence on all five theoretical implications of ignorance equilibrium in a unified framework; and c) provide our own results for DWK’s standard Hidden Information treatment, to be compared with the results of other treatments as we examine the other behavioral implications. In so doing, we do indeed replicate DWK’s result with a much larger sample.

Toward these ends, we run two treatments. First, the CIG Only treatment exactly replicated the DWK baseline treatment. Dictators played the CIG game with certainty, so the link between actions and outcomes was transparent. Second, the Hidden Information treatment exactly replicated the treatment of DWK with the same name. 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”. The null hypothesis is that the rate at which participants choose selfishly in the CIG Only treatment is at least as high as the ignorance rate in the Hidden Information treatment, while the alternative hypothesis is that the ignorance rate is actually higher.

The results are shown in Figure 1. First, while only 9 out of 26 (35%) dictators in the CIG Only condition chose A, in the Hidden Information condition 72 out of 120 (60%) chose ignorance. This difference is significant at 5% at a one (or two) sided Fisher’s exact test or FET ($p = 0.016$, FET). Thus, the main result of DWK is replicated.

4.2. Sorting

We examine sorting of types into different actions in two separate ways: a) by comparing the behavior of people with the same information—knowledge that they are in the CIG—but obtained either exogenously, in the CIG Only treatment, or by choice, in the Hidden Information treatment; and b) by providing separate measures of social type and self-image type and comparing these measures across participants exhibiting different behaviors in the Hidden Information treatment. The former approach is based on the design of DWK, while the latter is entirely original. Our results complement other evidence suggestive of ignorance and sorting from previous experiments, discussed in Section 5.1, with different designs.

Comparing behavior across sorting environments. The CIG Only and Hidden Information treatments described above represent two different sorting environments. In the former, the information is given exogenously to all participants so there is no opportunity to sort into different information

\[^{16}\text{Note that the data of the sessions for the Hidden Information treatment were collected at different points in time. The earliest sessions were run in 2009, simultaneously with the CIG only and Reveal Ex-post treatment (see below). Early 2015 we conducted additional sessions, to compare them with the simultaneously run Reveal Bonus treatment. Results are similar and statistically indistinguishable across these two time periods, so we pool the data. All our results hold when we only compare results from contemporaneously run sessions.}\]

\[^{17}\text{Unless otherwise indicated, all results reported hold for a one-sided Fisher’s exact test, and are often stronger for the more powerful one-sided exact z-test for equal proportions.}\]
Figure 1: The dark bar represents the proportion of participants that chose ignorance in the Hidden Information treatment and the light bars represent the proportion that choose selfishly (A) conditional upon knowingly being in the CIG. Fractions of subjects choosing the relevant behavior are displayed on top of the bars. Comparing ignorance in the Hidden Information treatment with selfish behavior in the CIG Only treatment provides evidence of information avoidance. Comparing the latter with selfish behavior in the revealed CIG in the Hidden Information treatment is consistent with sorting.

states. In the latter, participants can freely choose or avoid information, so those who know that are in the CIG are self-selected. To test the hypothesis derived in Proposition 3, we compare the frequency of prosocial behavior in the CIG Only game with the frequency observed among those who revealed and found themselves in the CIG in the Hidden Information treatment.\textsuperscript{18}

The light bars in Figure 1 indicate the percentage of self-interested choices in the CIG under full information. When information is endogenous in the Hidden Information treatment, 17 percent of the dictators (4 of 24) knowingly facing the CIG choose selfishly, substantially fewer than the 35 percent (9 of 26) who do so when information is imposed exogenously in the CIG Only treatment ($p = 0.13$, FET, and $p = 0.084$ on a one-sided $z$-test.) Thus, comparing behavior across sorting environments provides some evidence consistent with ignorance equilibrium behavior, though it is not very strong statistically.

\textsuperscript{18}In DWK, the share of participants who make a fair choice in the baseline treatment is 74%, similar to the 75% of participants who chose information and chose a fair outcome when confronted with a social dilemma. However, the small sample size (only 8 dictators in the latter) means we cannot conclude much from this comparison.
Comparing types across behaviors. An alternative methodology to examine sorting is to separately assess the types of the dictators to investigate if the types that choose different actions correspond to those predicted by the ignorance equilibrium. To this end, in nine of the Hidden Information sessions, we elicited two individual attribute measures from participants after they completed the hidden information game and had received feedback about their earnings from that game.

The first task assessed social-value orientation (SVO), also known as ‘the ring test’, implemented using the slider method of Murphy et al. (2011). In the SVO task a subject is asked to choose an allocation of money between herself and another subject in 6 different situations, where in each situation the trade-offs between the own and other’s payoffs are changed. From these six choices, Murphy et al. (2011) show how one can back out a measure of pro-sociality for each subject. The resulting measure is an angle determined by the mean allocations to the self and to the other person. The SVO angle increases in the prosociality of the choices, with angles below 23 degrees indicating competitiveness and selfishness and higher angles indicating a more prosocial disposition. The SVO task is a standard instrument for the measurement of social preferences, both in social psychology and economics.\(^\text{19}\)

Second, we assessed the importance of self-image through a questionnaire based on Aquino and Reed (2002). In this questionnaire the subjects were prompted to consider the attributes of being fair, generous, and kind and then asked to indicate agreement or disagreement with six statements about the importance of those attributes to their sense of self on a six-point Likert scale. We scored each item on a scale from 0 to 5, with ‘strongly disagree’ scored as 0 and ‘strongly agree’ scored as 5, and generated a measure of the importance of self-image by summing these six scores.

We collected these measures from 148 participants, including 74 dictators, with subjects earning between $0.60 and $4.00 in additional earnings from the SVO task. The mean SVO angle was 35.9 degrees and the median was 34.1, both in the range described as prosocial by Murphy et al. (2011). Restricting attention to dictators, the mean and median SVO angles are 35.9 and 34.5 degrees, indicating successful randomization. The mean and median self-image rating were 22.9 and 23, respectively, for the full sample, and 22.6 and 23, respectively, for the dictators.\(^\text{20}\) To give these numbers some meaning, a participant who selects ‘slightly agree’ in response to all six statements about the importance of being kind, generous, and fair to her sense of self would obtain a self-image score of 18, indicating ‘agree’ for all six statements would yield 24, and indicating ‘strongly agree’ would yield 30. Thus, the average participant agrees with the statements.

Proposition 1 gives predictions about which types should choose the different available actions. First, dictators who choose ignorance are predicted to be more prosocial than those who reveal and then choose selfishly in the CIG and are less prosocial than those who choose to reveal and

---

\(^{19}\)See, for example, Offerman et al. (1996).

\(^{20}\)Cumulative distribution functions for the SVO angle and self-image rating are available in Appendix B.
choose to share in the CIG. Second, dictators who reveal and then choose selfishly in the CIG are predicted to care less about their self-image than do those who choose ignorance or who choose reveal and choose pro-socially in the CIG.

Figure 2: The dark bars represent the average score of the Social Value Orientation task amongst dictators. The light bars represent average importance-of-self-image ratings amongst dictators. Average ratings are displayed on top of the bars. The results of both measures are consistent with sorting in ignorance equilibrium.

Figure 2 shows the mean SVO angle and self-image rating for dictators, broken down by outcome in the *Hidden Information* game. The mean SVO angle for the 46 ignorant dictators is 34.0, barely below the mean of all dictators. In contrast, the mean SVO angle of the 10 dictators who revealed and chose B in the CIG is 41.6, significantly higher (*p* = 0.001, one-sided Mann-Whitney-U test) than that of the ignorant dictators. While the fact that only 2 dictators chose selfishly with knowledge of being in the CIG limits the ability to draw strong conclusions from the comparison, the mean SVO angle of these dictators is 28.36, which is significantly lower (*p* = 0.04, one-sided Mann-Whitney-U test) than that of the ignorant dictators. Thus, the SVO angle measure supports the sorting predictions of ignorance equilibrium.

The mean self-image rating of ignorant dictators is 22.54 and that of those who chose B in the CIG is a statistically indistinguishable 23.30. In contrast, the mean self-image rating of the two dictators who were selfish in the CIG is 12.50, which is significantly lower than that of ignorant dictators (*p* = 0.07, one-sided Mann-Whitney-U test). Again, the small sample limits the strength of our conclusions, but the large differences suggest that dictators in this position report lower self-image concerns than those who sorted into ignorance or chose pro-socially. Thus, the self-image
measure offers some support for the sorting predictions of ignorance equilibrium.

In summary, we find support for sorting as predicted by the self-signaling model both by comparing behavior across sorting environments as well as by comparing independent measures of types across behaviors.

4.3. Exculpation

To test whether ignorance functions as an excuse, as hypothesized in Section 3.3, we asked recipients to evaluate the character of the dictator conditional on playing different strategies. This approach was new and did not directly replicate previous work, though it bears similarities to the approach taken by Krupka and Weber (2013), which we discuss further in Section 5. Our results complement other measures of blameworthiness used in previous experiments that we discuss in Section 5.1 below.

As our focus is on self-image, our preferred measure of exculpation would be of one’s opinions of oneself. However, dictator’s self-reported opinions of their own behavior might interact with their choices and a person who choose an action for reasons entirely divorced from self-image concern might report higher self-image for the action she chose, merely to appear consistent ex post. Lacking a methodologically acceptable way to elicit such self evaluations, we instead use recipients' opinions of the dictator as a proxy. Because the model’s predictions depend on the information available to the observer and not the observer’s identity, substituting others’ opinions of the decision-maker for her own opinion of herself is theoretically justified as long as the evaluators have the same information the decision-maker would have. Our procedure, described below, maintains this equivalence by eliciting responses separately for each informational and behavioral contingency.

In 9 sessions of the Hidden Information treatment, we listed six possible dictator behaviors: choosing A or B without revealing, revealing and choosing A or B in the CIG, and revealing and choosing A or B in the AIG. For each outcome recipients answered the question “How social (as opposed to antisocial) do you view Player X if he or she chooses the following action?” by selecting a point on a 5-point scale ranging from “very anti-social” to “very social”. The recipients completed these ratings while the dictators were making their decisions.

Each recipient’s response for each outcome was coded as an integer from 0 to 4, with 0 corresponding to “very anti-social” and 4 corresponding to “very social”. Figure 3 shows the mean ratings across 72 recipients for dictators who chose A versus B, both under chosen ignorance or knowingly in the CIG. Recipients ascribe to Player X a mean social rating of 1.69 when she chooses A under self-imposed ignorance, which is significantly higher or more social than the mean rating of 1.10 assigned for a dictator who knowingly chooses A in the CIG. Conversely, Player X is judged to be more social when she chooses B in the CIG, with a mean social rating of 3.24, significantly lower rather than the 2.35 mean rating under ignorance. A Mann-Whitney test shows that the distributions of responses in both cases differ significantly at the 1% level.
Summarizing, willful ignorance indeed functions as an excuse. As we discuss in more detail below, this finding is consistent with the ignorance equilibrium described in Proposition 1 but runs exactly counter to the alternative ‘info-seeking’ equilibrium in which ignorance is associated with the lowest image.

Figure 3: Answers to the question “How social do you rate Player X, based on each of the following actions ...”, where actions included the joint decision to be informed or not and to choose $6 or $5. Answers were given on a 5-point scale from “very anti-social” (0) to “very social” (4). Average ratings are based on 72 participants, and are displayed on top of the bars.

4.4. Willful vs. non-willful ignorance

In Section 3.4, we showed that under plausible conditions, subjects may be more interested in information about the consequences of their actions after taking an action than before they take it, even though the information is ‘useless’ afterwards. To test this prediction, we contrast the Hidden information treatment with a Reveal Ex-post treatment, which is modeled on game $\Gamma_N$ (defined in Section 3.4). This condition differed from the Hidden information condition only in that the dictator entered her outcome choice using the strategy method, specifying her 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 Hidden information condition, the dictator could reveal the payoffs by clicking a button on the same screen, but this information would only be revealed ex post, after the conditional outcome choice was entered and the outcome realized.21

---

21This experiment partially replicates and overlaps with the the experiment reported in Grossman (2014). The results of the Reveal Ex-post treatment are reported in that paper as the Strategy Method treatment. A working paper version of that paper also report the results of contemporaneous Hidden Information treatment sessions that
Following Proposition 4, we hypothesize that subjects will be more likely to choose information in the \textit{Reveal Ex-post} treatment than in the \textit{Hidden Information} treatment. In addition, the results of the \textit{Reveal Ex-post} treatment can be used to test two maintained assumptions. First, we can see if people are willing to overcome the default of no-information, countering the criticism that ignorance occurs ‘merely’ because it is the default choice. Second, because information has no instrumental value \textit{Reveal Ex-post} treatment, a low ignorance rate would provide evidence for our tie-breaking rule that most indifferent agents reveal.

The results are depicted in Figure 4. Comparing the two treatments, the overall ignorance rate in the \textit{Reveal Ex-post} treatment was 26%. This is significantly lower at the 1% level than the 60% rate in the \textit{Hidden Information} condition, thus confirming our hypothesis that more agents reveal \textit{after} than \textit{before} the dictator decision.\footnote{In the \textit{Reveal Ex-post} 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\%. 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 image considerations.}

![Figure 4: Ignorance rate by treatment, the fraction of subjects choosing ignorance is at the top of the bars. Comparing the \textit{Hidden Information} treatment ignorance rate with that of the \textit{Reveal Ex-post} and \textit{Reveal Bonus} treatments provides support for the predictions of ignorance equilibrium on non-willful ignorance and paying for ignorance.](image.png)

\textit{Default NR} are included in our data set, under the treatment name \textit{Default NR}. However, the \textit{Default NR} treatment reported in the published paper was modified so as to be more comparable to the other default treatments and, though similar, is not identical to the \textit{Hidden Information} treatment reported herein.
4.5. Paying for ignorance

In Section 3.5 we derived the prediction that all social types with \( \theta < \theta^* \) are willing to pay for ignorance. In addition, the comparative statics in the ignorance equilibrium imply that ignorance is increasing in the cost of information \( k \). Here we test these two predictions by introducing the Reveal Bonus treatment. In this treatment, subjects could earn an additional $0.10 if they chose to reveal. They were made aware of this bonus in the instructions, and next to the button that they should click to reveal the game it said “Reveal game + $0.10”. Since this makes information acquisition more attractive, we hypothesized that ignorance should go down. However, consistent with corollary 2 we still expected a substantial ignorance rate in this condition. This experiment is original and was conducted contemporaneously with a similar experiment in Cain and Dana (2012), which we discuss further in Section 5.

The last bar in Figure 4 shows that the ignorance rate in the Reveal Bonus condition was 0.46, i.e. much larger than zero.\(^{23}\) While the fraction of ignorance in the Reveal Bonus condition is lower than in the Hidden Information condition, a one-sided test of the difference is significant only at the 10\% level (\( p = 0.096, \text{FET} \)). Thus, while we find some evidence that ignorance decreases when it is costly, we find that almost half of the subjects is willing to pay a small cost not to know the outcomes for the receiver.\(^{24}\)

5. Discussion

In this section we discuss self-image theory in relation to our experimental results and evidence from previous studies. We then contrast the self-image interpretation with alternative theories. Next, we examine the plausibility of the assumptions and conditions that are necessary for the existence of an ignorance equilibrium. We demonstrate with a simple example that the existence of a unique ignorance equilibrium featuring behavior similar to that which we observe in the experiment does not require implausible assumptions about parameter values and type distributions. Finally, discuss how our model relates to other situations in which people have the opportunity to avoid situations where they are confronted with a giving decision.

\(^{23}\)Although this seems obvious that this fraction is different from zero, it is hard to test statistically as a binomial test with the null-hypothesis that it is equal to zero would reject for any positive fraction. However, when we compare this result to a fictitious sample of the same size with a zero ignorance rate, we reject the null hypothesis of equal fractions with \( p < 0.001 \) (FET).

\(^{24}\)The small difference between the Hidden Information and Reveal Bonus treatments may lead one to think that perhaps the cost of ignorance was not clear or salient to the subjects. To see whether this was the case, we ran a session where we reversed the cost. The only difference was that the subjects now saw a −$0.10 charge for revealing, instead of a +$0.10 bonus. We found an ignorance rate of 100\% (10 out of 10 dictators), suggesting that the incentive was quite salient indeed.
5.1. Other evidence of ignorance equilibrium

This paper comes at the heels of a wave of experimental investigations into the phenomenon of willful ignorance. Here we discuss all experimental papers in the economics literature on willful ignorance that we are aware of, and explain how our results and the self-image theory relate to this growing literature. Column 1 of Table 1 shows the different empirical implications of our theory that we derived in Section 3. Column 2 presents our main results, and column 3 summarizes the result of other studies. The third and fourth columns indicate whether (√) or not (×) the accumulated evidence is consistent with models of purely distributional or outcome-based preferences or self-signaling, respectively. In column 5, we present the predictions of our alternative information-seeking equilibrium, discussed in Section 2.4, in the case that information is costless.

<table>
<thead>
<tr>
<th></th>
<th>This study</th>
<th>Other studies</th>
<th>Consistent with</th>
</tr>
</thead>
<tbody>
<tr>
<td>Information avoidance</td>
<td></td>
<td></td>
<td>×</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Find sorting: Conrads and Irlenbusch (2013); Fong and Oberholzer-Gee (2011); Kajackaite (2015)</td>
<td>✓</td>
</tr>
<tr>
<td>Sorting</td>
<td>Find sorting by comparing both behavior across situations and type measures across behavior</td>
<td>Don’t find sorting: Larson and Capra (2009)</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>×</td>
</tr>
<tr>
<td>Exculpation</td>
<td>Selfish dictators regarded “more social” when willfully ignorant</td>
<td>Find exculpation: Krupka and Weber (2013); Bartling et al. (2014); Conrads and Irlenbusch (2013)</td>
<td>×</td>
</tr>
<tr>
<td>Non-willful ignorance</td>
<td>Less ignorance when it cannot affect image</td>
<td>Grossman (2014) reports same finding with partially the same data</td>
<td>×</td>
</tr>
<tr>
<td>WTP for ignorance</td>
<td>46 % willing to pay for $0.10 ignorance</td>
<td>Cain and Dana (2012): 23% willing to pay $0.25 for ignorance.</td>
<td>×</td>
</tr>
</tbody>
</table>

Table 1: Evidence for testable implications of the ignorance equilibrium in Proposition 1.

With respect to ‘avoidance’, we replicate the result of DWK with a much larger sample. Feiler (2014) and Larson and Capra (2009) also find avoidance in a similar context. Matthey and Regner (2011) provide strong evidence for avoidance in a slightly modified game. In a more applied experimental setting, Ehrich and Irwin (2005) let participants request information about different product attributes and find that those who indicate that they value sustainability are especially reluctant to request information on this dimension.

We also find clear evidence of ‘sorting’ by comparing behavior when information is chosen versus imposed and by comparing individual attributes across behaviors. These results add to
several earlier studies that compare conditions with endogenous information acquisition about the recipient to situations where information is exogenously given. First, Fong and Oberholzer-Gee (2011) demonstrate this in a dictator game where it was unknown whether the recipient was either ‘worthy’ (a mother living in poverty) or not (a drug addict). Subjects who paid to see information that the recipient was worthy were more generous on average than those that were given the information exogenously.

Second, Kajackaite (2015) studies a real effort task that potentially yields benefits to a ‘bad charity’ (the National Rifle Association or NRA). She finds that agents who were exogenously informed that their effort generated earnings for the NRA exerted more effort and earned more for the NRA than agents who chose to inform themselves and learned this fact, although this difference is not significant. Conrads and Irlenbusch (2013) investigate an ultimatum game where the proposer’s information about the offer for the responder is varied in different experimental conditions. They find that 91% of proposers who reveal initially hidden information make ‘nice offers’, compared to only 44% for proposers who got the information exogenously. Standing in contrast to these results, Larson and Capra (2009) find little evidence of sorting in a direct replication of DWK, but their sample is small, lacking statistical power.

When it comes to ‘exculpation’, the evidence provided by our recipients’ normative ratings complements that provided by several studies that have looked at other measurements of blameworthiness and found evidence for the exculpatory effects of willful ignorance. Krupka and Weber (2013) use an incentive compatible elicitation procedure in the context of the same hidden-information game to 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 (2013) find that unequal proposals in an ultimatum game are rejected less often if the proposer chose to be ignorant of the payoffs for the responder. Bartling et al. (2014) find that given an unfair outcome, willful ignorance reduces the costly punishment meted out by a third party.

The non-willful ignorance finding provides one of the biggest challenges for alternative theories that we discuss below. Grossman (2014) reports a similar result, based on partly the same data, and uses it to argue that, despite major default effects, willful ignorance is not merely a artefact of status quo bias. Our self-image model provides a theoretical backing for this idea by establishing that the ignorance in the Hidden Information condition can be driven by image considerations, while any ignorance in the Reveal Ex-post condition cannot. To our knowledge, the comparison between willful and non-willful ignorance has not been made before.

Concurrent with this paper, Cain and Dana (2012) also investigated a treatment where ignorance was costly. Interestingly, their bonus for information acquisition ($0.25) was slightly higher than ours, but their results show that even with this amount, 23% of subjects are willing to forego information. Together, these results build a robust case that subjects in these experiments are
resisting becoming informed about the social consequences of their behavior.

5.2. Alternative explanations for willful ignorance?

The previous section shows that self-signaling theory can explain a number of apparently puzzling results related to willful ignorance. An important question is whether the results can be explained by other, perhaps simpler theories.

Models of distributional preferences. A first stop are models of distributional or ‘social’ preferences, which can be considered the ‘first generation’ of models dedicated to understanding prosocial behavior. In these models, subject’s preferences are defined over the material outcomes of all players in the game. A general utility function therefore looks like \( u_i = u_i(x_i, x_{-i}) \), where \( x_i \) are the material outcomes of player \( x_i \), and \( x_{-i} \) are the outcomes of the other players. Our model implements a simple version of such redistribution preferences in the special case where \( \mu = 0 \) and \( W = x_{-i} \). Other well-known instances are the inequality aversion model of Fehr and Schmidt (1999) and Charness and Rabin (2002).

To see if these models can explain the stylized facts surrounding willful ignorance, first consider information avoidance (see also Feiler, 2014). If a person prefers the fair distribution in the DWK game with full certainty, i.e. \( (5, 5) \succsim (6, 1) \), then the independence axiom implies that \( p(5, 5) + (1 - p)(6, 5) \succsim p(6, 1) + (1 - p)(6, 5) \), which reflects the lottery resulting from becoming informed on the left-hand side, and the lottery under ignorance on the right hand side. Thus, information avoidance cannot be explained by the combination of social preferences and expected utility.

The same is true for the comparison between willful and non-willful ignorance. 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 the same action in both states and will always be indifferent about acquiring information. Suppose a fraction \( \alpha \) of those types acquires information. Such a model predicts that the ignorance rate will be \( (1 - \beta)(1 - \alpha) \) in the Reveal Before treatment, and \( 1 - \alpha \) in the Reveal Ex-post 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.

Paying for ignorance is obviously inconsistent with social preference models. Since the distribution of outcomes cannot be improved by having less information, no agent should be willing to pay for it. Exculpation, the fact that ignorance deflects punishment, also poses a puzzle for social preference models. In these models, punishment only serves to alter the relative payoffs of different parties, for example to reduce the gains of someone who came out ahead of others. Without adding a theory of intentions or image, outcome based theories cannot explain why punishment differs between identical outcomes, depending only on the informational choice of the person being punished.
The phenomenon of *sorting* is an exception in this list. Sorting can be explained by social preferences, as these models predict that “fair types” will be willing to pay for information and “selfish” types will not. Nevertheless, the fact that outcome-based models fail to explain 4 out of 5 stylized facts relating to willful ignorance seems to us enough reason to reject them as an explanation. In doing so, we do not want to dismiss models of fairness, as they may provide a tractable and suitable analytic tool in many contexts. Rather, we argue that these models are not rich enough to adequately explain the role of uncertainty in social decisions.

*Information-seeking equilibrium.* The information-seeking equilibrium we discussed in Section 2.4 predicts none of the five behavioral patterns we identified as consistent with ignorance equilibrium. First, there is no *information avoidance* when when information is costless. When we use the tie breaking rule that homo-economicus chooses to inform himself, ignorance is zero.\(^{25}\) Second, if there is no positive information cost then there is also no *sorting*, as all agents choose information acquisition.\(^{26}\) Third, there is no *exculpation* from ignorance, as it is associated with the worst image. Fourth, since in this equilibrium there is no willful ignorance, *non-willful ignorance* by definition occurs no less frequently than willful ignorance. Finally, there is no *willingness to pay for ignorance*, as no type strictly prefers ignorance when there is no image benefit. Thus, at least in this experimental paradigm there is unequivocal support for a social norm that is lenient towards the ‘hypocritical’ action of willful ignorance, even though theoretically there is an equally attractive alternative norm that does not display such leniency.

*Social-image concerns.* Although we have stressed the importance of self-image, the model is consistent with a social-image interpretation, as the predictions do not depend upon the identity of the observer. The most plausible interpretation in any given application should therefore be judged by the appropriateness of the assumptions on the information structure. As we argued above, it is often implausible to assume that an external audience observes the decision-maker’s information choice, signal, and outcome choice. However, for an observer self, *not* having access to this information seems to require some additional layer of self-deception.

These considerations extend to our experimental environment. Suppose that a dictator in the *Hidden Information* treatment cares what her recipient thinks about her. In both the DWK experiment and in our experiments, the recipient observes neither the information acquisition choice, the signal, nor the dictator’s outcome choice. Instead he observes his own payoff. Thus, if he receives a high payoff, he doesn’t know if it was from the choice of \(B\) in the CIG or \(A\) in the AIG, and

\[\text{25}^{\text{Even if we used the opposite tie-breaking rule, ignorance is at most } \varepsilon. \text{ Since at least a fraction } \varepsilon \text{ do not take the social action in the full information game } \Gamma_F, \text{ ignorance cannot be higher than selfishness in } \Gamma_I.}\]

\[\text{26}^{\text{If } k > 0, \text{ homo economicus chooses ignorance. In that case, it is possible that a larger fraction of subjects chooses to be nice in game } \Gamma_C \text{ (conditional on knowingly playing the the CIG) than in } \Gamma_I. \text{ However, since being selfish in the CIG does not involve pooling with homo economicus, selfishness in the CIG becomes more attractive relative to game } \Gamma_I. \text{ If this effect is strong, it could even result in anti-sorting. Thus, the result is ambiguous for } k > 0.}\]
whether the dictator knew or was ignorant of the state. Because the choice of ignorance cannot influence the observer’s beliefs, there is no strategic reason to choose it. Such social signaling can therefore not explain the phenomenon of information avoidance, exculpation or the willingness to pay for ignorance.

The results of the Reveal Ex-post treatment further discredit the idea that dictators signal to the recipient. The use of the strategy method in this treatment forces the dictator to express binding choices for all contingencies. The recipient does not observe these choices and the instructions and information provided to the recipient in the Reveal Ex-post treatment is identical to that in the Hidden Information treatment. So if the recipient’s inferences were central to the decision maker, this experimental manipulation should not affect information acquisition. The fact that it actually does so strengthens the self-signaling interpretation, namely that ignorance loses its value if the dictator can no longer tell herself that she would have acted differently with full information.

Apart from the recipient, the experimenter is another potential external audience for the dictator. We can think of audience concerns towards the experimenter in two ways. First, subjects could fear direct identification in the lab, which could occur during the payment phase. Since the participant’s role was not apparent to the experimenter during the payment phase, it would have been hard to infer choices from the paid amount. Also, the final payoff did not reveal the information choice or signal content. Thus, the experimenter’s information is similar to that of the recipient, and so are the behavioral consequences of the two kinds of signaling. For the same reason we can dismiss signaling to the recipient, we can also dismiss this kind of signaling to the experimenter.

Second, subjects could be influenced by the prospect of future judgement by the experimenter. When analyzing the data, the experimenter has the exact same information as a self-signaler, so the model would predict the same outcome. However, it strains credibility that researchers conducting studies with hundreds of participants might be able to form social judgments in a way that can be linked back to participants individually. The anonymity protocols and the physical environment of the EBEL at UC Santa Barbara make it very difficult to directly monitor the choices in the lab and link them to the person making them in real time. Moreover, this is readily apparent to participants.

Perhaps the experimenter’s retroactive assessment matters to subjects in a more abstract way, or the lab environment may trigger feelings of “being watched”. We believe that such effects are quite different from signaling towards peers, other experimental participants or even directly to the experimenter. Like concerns about heavenly judgment, or thoughts like “what would my mother think?”, there is no behavioral test that can distinguish such influences from those of an internal observer. We therefore see no objections against treating the behavioral effects of such abstract audiences as instances of self-signaling.

Finally, Van der Weele (2014) reports questionnaire responses from dictators in a version of the
Hidden Information treatment. Here, 4 (out of 11) subjects who chose to remain ignorant reported that they did so in order not to have “a bad conscience”, indicating that subjects felt an internal obligation to contribute in the CIG.

Cognitive dissonance, social norms, and other theories. Other theories that are less well studied by economists may help explain willful ignorance. For example, some have argued that willful ignorance relates to cognitive dissonance, the psychological cost that arises from the knowledge of having acted contrary to what one thinks is morally right (see Konow, 2000; Matthey and Regner, 2011). We think this explanation is compatible with self-image concerns, as a low self-image resulting from transgressive behavior can naturally be viewed as a form of dissonance. Without the need for any specific assumptions, the model thus naturally explains how ignorance reduces dissonance, by raising self-image relative to the situation of acting selfishly with full knowledge.

Another idea is that social behavior is governed by different social norms than information acquisition (Spiekerman and Weiss, 2013). Again, we see self-signaling as compatible with such an explanation. To the extent that norms can be interpreted as Nash equilibria in signaling games (see e.g. Posner, 2000), we show how differential norms for social behavior under uncertainty and under full information arise endogenously from variations in the strength of signaling concerns.

We do not rule out that other theories that make additional assumptions will be able to explain the results. For example, a model that assumes that disutility from emotions such as guilt or pity is convex in the (perceived) probability of harming another person may generate information aversion. However, an important advantage of the Bayesian image-based theory presented here is that it does not necessitate any non-standard assumptions about the relation between probabilities and preferences.

Finally, a challenge to the self-image model as an explanation for willful ignorance comes from Feiler (2014), who manipulates the probability of the CIG in an experiment sharing the basic design of DWK’s hidden information game. She argues informally that self-signaling predicts lower rates of ignorance when the probability of the CIG decreases, and she finds evidence for the opposite. Van der Weele (2012), who analyzed an earlier version of the model of this paper, shows that the validity of this prediction depends upon parameter values and finds experimental evidence that does not support Feiler’s theoretical claim. Furthermore, Feiler’s experiment relies on a within-subjects design, with participants making decisions in twenty closely related variations of the hidden information game. It is unclear how the predictions of the signaling model applied to a one-shot decision would change in an environment with multiple signaling opportunities. Thus, Feiler’s results do not contradict the self-image model’s predictions, while they do offer additional evidence of information avoidance and sorting.
5.3. Parameter values

Some parameters of the self-signaling model are not directly observable. These include \( \mu, \epsilon, \) and the distribution of social types, but also the utility valuations of monetary payoffs. By contrast, the parameters \( p \) and \( k \) are observable, at least in experimental conditions. Other auxiliary assumptions such as the stability condition (2) or the tie-breaking assumption may be testable in principle, but it is not obvious how to do so. With so many degrees of freedom, one may wonder whether the ignorance equilibrium is consistent with plausible parameter values, and conversely, whether there are plausible values where we would not expect this equilibrium to occur. We address these questions in turn.

**Ignorance equilibrium: an example.** Consider the following example of an environment that conforms to our general model with simple and plausible specifications for parameter values and the distribution of types. We begin with basic assumptions about the distribution of types that are chosen to aid tractability: let \( \theta \) be distributed uniformly on \([0, 1]\) and \( \epsilon = 0.1 \). To match the probability in the experiment, we choose \( p = \frac{1}{2} \) and to reflect the absence of monetary information costs in the experiment and the monetary cost of choosing \( B \), we choose \( k = 0 \) and \( c = 1 \), respectively, though there may have been other, unobservable costs in the experiment. Finally, though we cannot observe \( \mu \) directly, a value of \( \mu = 0.25 \) does not seem unreasonable and satisfies the assumption \( \mu < c \). The remaining free parameter is the social benefit \( w \). While the monetary benefit to the other person is 4, assuming \( w = 4 \) requires that the highest type values the other person’s payoffs just like her own, which seems an implausibly high degree of altruism. If we choose \( w \) to be 1.27, which implies a more modest level of altruism, we obtain an ignorance equilibrium threshold of \( \theta^* = 0.63 \).

With the faction \( \epsilon \) of the population being homo economicus and choosing information, this leaves a fraction \( (1 - \epsilon)\theta^* = 0.57 \) of the population remaining ignorant, which is close to the 60% ignorance rate we observed in the Hidden Information treatment. Furthermore, among the mix of social types and homo economicus that reveal and receive the signal that \( W = w \), a fraction \( \frac{\epsilon}{\epsilon + (1-\epsilon)(1-\theta^*)} \) or 23% of the population decline to take the prosocial action, which is also close to the 17% that chose \( A \) in the experiment. Thus, the ignorance equilibrium implied by these values yields behavior that resembles the ignorance level and prosocial behavior observed in the Hidden Information treatment of the experiment. Furthermore, a comparison of the image-reward and material cost of information functions illustrates that this equilibrium is stable and unique among interior ignorance equilibria.

This simple calibration exercise does not constitute a test of the model, but it shows that the existence of the equilibrium described in Proposition 1 does not depend upon implausible assumptions. The experiment pinned down only some of the model’s parameters, so there are other degrees of freedom that the modeler may use, such as the distribution of types and the tie-
breaking rule. Since these elements are not observable, it is not possible to make reliable inferences on the parameter values of the model from the current results. At the same time, variations in these elements offer the potential to fit the results of the experiment even more closely.

**Other parameter values.** The additional degrees of freedom allow the self-signaling model to explain novel phenomena, but a potential weakness is that the model may be hard to discipline. While the above example demonstrates that one can find plausible parameter values such that equilibrium is unique and stable, it can also demonstrate when an ignorance equilibrium may not exist. For example, changing $\mu$ from 0.25 to 1.25, thereby violating the assumption that $\mu < c$, implies that there is no equilibrium threshold $\theta^*$ in $[0, 1]$. For such high values of $\mu$, all social types would pool on acquiring information, and so would the *homo economicus* as long as $k \leq 0$. Thus, in situations where (self-)image concerns are extremely strong relative to the material payoffs, we might not observe willful ignorance.

When it comes to stability, alternate values of $\mu = 0.5$ and $w = 0.94$ would also yield $\theta^* = 0.63$ as the cutoff value for a stable equilibrium, with all the same behavioral outcomes. However, it would violate the stability condition, and yield another equilibrium cutoff at $\theta^* = 0.98$, this one unstable. Note that even if the uniqueness or stability conditions fail to hold, the existence result in Proposition 1 is still valid. The same is true for the predictions that are derived directly from this equilibrium, namely “exculpation”, “willingness to pay for ignorance”, and “willful vs. non-willful ignorance”. The predictions of “information avoidance” and “sorting” depend on the comparison of the equilibrium thresholds in different games. Under multiple thresholds, it is not clear that these comparisons will hold for all pairs of thresholds without making additional assumptions.

### 5.4. Similarities to other “opt-out” situations

Willful ignorance bears some similarities to other situations where people can choose “outside options” in order to avoid a giving situation. For instance, people will sometimes go out of their way to avoid being asked for a donation (Andreoni *et al.*, 2011). In pay-what-you-want (PWYW) schemes where some of the proceeds go to charity, Gneezy *et al.* (2012) finds that purchases drop compared to a scheme with a fixed price, even though consumers under PWYW could have chosen any price.

In laboratory dictator games, a substantial minority of people are willing to pay not to implement their decision but instead to opt out of the game (Dana *et al.*, 2006; Broberg *et al.*, 2007; Lazear *et al.*, 2012). These “reluctant sharers” (Lazear *et al.*, 2012) resemble the types who avoid information in the ignorance equilibrium but would have given in the counterfactual situation where they had been informed. We thus conjecture that the people who opt out of dictator games may have initially contributed mainly because of image concerns.

The fact that those who share more are more willing to drop out (Lazear *et al.*, 2012; Broberg *et al.*, 2007), is at least superficially in line with our finding that some subjects who remain ignorant
would behave prosocially under full information, but is not fully explained by our model. This may reveal a limitation of assuming that social types all have the same intensity of image concern, \( \mu \). An agent with a low \( \theta \) but high \( \mu \) might be swayed by image-concern to act very generously in a dictator game, but strongly desire to avoid being placed in such a giving environment. Apart from this, several differences between the information acquisition environment that we study and these “exit” games make it difficult to apply our model directly. These include the absence of uncertainty, the timing of the game (which reverses the prosocial choice and the opt-out decision) and the continuous action spaces in the dictator game. Extensions of our framework with continuous action spaces are not straightforward, but exploring the theoretical underpinnings of avoidance in settings without uncertainty is an interesting goal for future research.

6. Conclusion

We have shown that Bayesian self-signaling theory can account for the phenomenon of willful ignorance in social decisions. An equilibrium exists in which ignorance is chosen by less-than-completely-selfish types who fear the trade-off between self-image and material rewards that occur in a more transparent environment. Because of such sorting, ignorance provides at least a partial excuse for selfish behavior. Our study thus explains why “looking the other way” is such an important source of socially harmful behavior. Seemingly obvious evasions of responsibility like willful ignorance can, at least partially, neutralize the demands of conscience and provide us with the excuses we need to behave selfishly.

We present a body of evidence from a unified experimental framework that offers consistent support for the predictions associated with this equilibrium. Our theory also ties together a large number of previous studies from a diversity of environments that reinforce our results. The end result is a robust case that supports the broader theory of self-signaling and the usefulness of self-signaling models for understanding behavior.

From a policy perspective, our results indicate that it may not be enough to simply make more information available about the social benefits of some actions. In the context of climate change, Osberghaus et al. (2010) did not find that providing information about climate change increased self-reported risk perceptions. Stoll-Kleemann et al. (2001) advocate to increase self-image and intrinsic concerns for good behavior, for instance by providing appropriate narratives about climate change in schools and by community leaders. Another avenue is subsidization of actions that benefit common goods like the climate. Van der Weele (2014) shows in the context of the DWK game that when prosocial actions become cheaper, experimental participants are more prone to acquire information.

More generally, future research may address how authorities and other social actors can design choice environments that exploit self-signaling to advance the common good, for example...
through increased salience of self-image concerns such as honor codes and moral appeals. Another promising direction is to extend the idea that people avoid high signaling incentives to other choice environments that feature “outside options”, as described in Section 5.4.

References


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 \( \phi_a^\theta = E[\theta | a, \sigma; s] \) denote the expectation conditional on the equilibrium strategy profile, the chosen action \( a \) and information \( \sigma \). Beliefs on the equilibrium path are as follows:

\[
\phi_0^0 = \phi_0^0(\theta^*) \equiv E[\theta | 0, 0; \theta^*] = \int_0^{\theta^*} \frac{\theta dF(\theta)}{F(\theta^*)} \tag{A.1}
\]

\[
\phi_0^0 = \phi_0^0(\theta^*) \equiv E[\theta | 0, \sigma_0; \theta^*] = \int_{\theta^*}^1 \frac{(1-\varepsilon)\theta}{\varepsilon (1-\varepsilon) (1-F(\theta^*)) + \varepsilon} dF(\theta) \tag{A.2}
\]

\[
\phi_1^w = \phi_1^w(\theta^*) \equiv E[\theta | 1, \sigma_w; \theta^*] = \int_{\theta^*}^1 \frac{\theta dF(\theta)}{1-F(\theta^*)} \tag{A.3}
\]

\[
\phi_1^w = \phi_1^w(\theta^*) \equiv E[\theta | 0, \theta^*] = 0 \tag{A.4}
\]

With respect to beliefs off the equilibrium path, we assume that \( \phi_1^w \leq \phi_1^w \) and we make no assumption about \( \phi_0^0 \).

**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 $\theta$ will take the prosocial action iff

\begin{align*}
u(a = 1 | \sigma = \sigma_w; \theta^*) & \geq \nu(a = 0 | \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 \tilde{\theta}. \quad (A.5)
\end{align*}

It is immediate that in equilibrium all types $\theta \geq \tilde{\theta}$ who observed $\sigma = \sigma_w$ take the prosocial action, and all $\theta < \tilde{\theta}$ do not. Note that $\tilde{\theta} > 0$, 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{align*}
u(a = 0 | \sigma = \sigma_0; \theta^*) & > \nu(a = 1 | \sigma = \sigma_0; \theta^*) \\
\mu \phi_0^0 & > -c + \mu \phi_1^0 \\
c & > \mu (\phi_1^0 - \phi_0^0). \quad (A.6)
\end{align*}

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{align*}
u(a = 0 | \sigma = \emptyset; \theta^*) & > \nu(a = 1 | \sigma = \emptyset; \theta^*) \\
\mu \phi_0^0 - p \theta w & > -c + \mu \phi_1^0 \\
\theta & < \frac{c - \mu (\phi_1^0 - \phi_0^0)}{pw} \equiv \hat{\theta}. \quad (A.7)
\end{align*}

Step 2. We now check which type will acquire information. Since the homo economicus cares only about her 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 $\theta^*$, we can derive that $\theta^*$ is given implicitly by the fixed point equation

\begin{align*}Eu(\text{acquire info}) = Eu(\text{not acquire info}) \\
(1 - p) \mu \phi_0^0 + p (\theta^* w - c + \mu \phi_1^w) - k = \mu \phi_0^0 \\
\theta^* = \frac{pc + k - \mu (p \phi_1^w + (1 - p) \phi_0^0 - \phi_0^w)}{pw}. \quad (A.8)
\end{align*}

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

Only if $\theta^* < \hat{\theta}$ do all ignorant types take the self-interested action and only if $\tilde{\theta} < \theta^*$ 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 $\tilde{\theta} < \theta^* < \hat{\theta}$.

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

\begin{align*}k & < (1 - p) c - \mu (p \phi_1^w + (1 - p) \phi_0^0 - \phi_0^w). \quad (A.9)
\end{align*}
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 $\phi^1_0 \leq \phi^1_w$, which we assumed above.

Similar algebra comparing (A.5) and (A.8) yields that $\bar{\theta} < \theta^*$ if and only if
\[
k > \mu \left( (1 - p) \phi^0_0 - \phi^0_\emptyset \right) \equiv \bar{k}.
\] (A.10)

A necessary and sufficient condition for $\bar{k} < 0$ is that $p > 1 - \frac{\phi^0_\emptyset}{\phi^0_0}$.

It remains to check that $\theta^*$ exists and is in the interior. We have already verified that $\theta^* > 0$, since $\bar{\theta} > 0$. To check that $\theta^* < 1$, note that if the threshold type is $\theta^* = 1$, we have $\phi^1_w = 1$, $\phi^0_0 = 0$, and $\phi^0_\emptyset = E_F \theta$. 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 $\bar{k} < k \leq 0$ and
\[
p > \max \left\{ 1 - \frac{\phi^0_\emptyset}{\phi^0_0}, \frac{\mu E_F \theta}{w - c + \mu} \right\}.
\] (A.11)

Since $\frac{\mu E_F \theta}{w - c + \mu} < 1$ and $\frac{\phi^0_\emptyset}{\phi^0_0}$ is positive for any interior value of $\theta^*$ (i.e. whatever the value of $p$), there exists a $\hat{p} < 1$ such that this condition will be satisfied for some $p > \hat{p}$.

Comparative statics with respect to ignorance levels can be obtained by implicit differentiation of (A.8) with respect to $\theta^*$ 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:

\[
\phi^0_0 = \phi^0_0(\theta^*) = E[\theta \mid 0, \emptyset; \theta^*] = \int_0^{\theta^*} \frac{(1 - \varepsilon) \theta}{(1 - \varepsilon) F(\theta^*) + \varepsilon} dF(\theta),
\] (A.12)
\[
\phi^1_w = \phi^1_0 = \phi^0_0(\theta^*) = E[\theta \mid 0, \sigma_0; \theta^*] = \int_{\theta^*}^{1} \frac{\theta dF(\theta)}{1 - F(\theta^*)}, \quad (A.13)
\]

With respect to beliefs off the equilibrium path, we assume that $\phi^0_0 = 0$ and $\phi^1_\emptyset \leq \phi^1_w$.

The analysis proceeds like before, and is not reconstructed here in detail for reasons of space. We obtain
\[
\bar{\theta} = \frac{c - \mu \phi^1_w}{w}, \quad (A.14)
\]
\[
\tilde{\theta} = \frac{c - \mu (\phi^1_0 - \phi^0_0)}{pw}, \quad (A.15)
\]
\[
\theta^* = \frac{pc + k - \mu (\phi^1_w - \phi^0_0)}{pw}. \quad (A.16)
\]

Some algebra yields that $\bar{\theta} \leq \theta^*$ iff $k \geq \mu (\phi^1_w (1 - p) - \phi^0_0)$, which is satisfied if $p \geq 1 - \frac{\phi^0_0}{\phi^1_w}$. Furthermore, $\theta^* < \tilde{\theta}$ iff
\[
k < c (1 - p) + \mu (\phi^1_w - \phi^0_0) \equiv \bar{k}.
\] (A.17)

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

42
It remains to check that $\theta^* < 1$. By substituting $\theta^* = 1$ into (A.16), we can find the sufficient condition $p > \frac{k-\mu(1-E\theta)}{w-c}$. Using (A.17) we obtain $p > \frac{c-\mu(1-\phi^1_w + \phi_0^1 - E\theta)}{w}$.

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

$$p > \max \left\{ 1 - \frac{\phi_0^0}{\phi_w^0}, \frac{c - \mu(1 - \phi_w^0 + \phi_0^0 - E\theta)}{w} \right\}. \quad (A.18)$$

Since $c < w$ and $\phi_0^0 < \phi_w^1$ for any value of $\theta^*$, 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 using our refinement $pD1$, defined in the main text. We check the following assumptions on off-equilibrium beliefs that were made above:

1. $\phi_0^0 = 0$ when $k > 0$.

Types $\theta < \theta^*$ would deviate if the deviation payoffs are bigger than equilibrium payoffs

$$(1-p)\mu\phi_0^0 + p\mu\phi_w^0 - k > \mu\phi_0^0$$

$$\phi_0^0 > \frac{(1-p)\phi_0^0 + k/\mu}{p} \quad (A.19)$$

Types $\theta \geq \theta^*$ deviate if the deviation payoffs are bigger than equilibrium payoffs

$$(1-p)\mu\phi_0^0 + p\mu\phi_w^0 - k > (1-p)\mu\phi_0^0 + p(\theta w - c + \mu\phi_1^1) - k$$

$$\phi_0^0 > \phi_1^1 + \frac{\theta w - c}{\mu} \quad (A.20)$$

Combining the RHS of (A.19) and (A.20), we see that the latter is higher if

$$\phi_1^1 + \frac{\theta w - c}{\mu} \geq \frac{(1-p)\phi_0^0 + k/\mu}{p}$$

$$\theta \geq \frac{pc + k - \mu(\phi_1^1 + (1-p)\phi_0^0) - \phi_0^0}{\phi_0^0 w} \quad (A.21)$$

Since the RHS is exactly equal to $\theta^*$, this expression is satisfied in equilibrium for all $\theta$. Thus, types $\theta < \theta^*$ deviate for a larger set of off-equilibrium beliefs. Since $pD1$ does not distinguish further between types $\theta < \theta^*$, it is reasonable to assume $\phi_0^0 = 0$.

2. $\phi_0^1 \leq \phi_w^1$.

Types $\theta \geq \theta^*$ deviate if deviation payoffs are bigger than equilibrium payoffs

$$\mu\phi_0^1 - c + \theta pw > (1-p)\mu\phi_0^0 + p(\theta w - c + \mu\phi_1^1) - k$$

$$\phi_0^1 > \frac{c(1-p) - k}{\mu} + (1-p)\phi_0^0 + p\phi_1^1 \quad (A.22)$$

This expression does not depend on $\theta$, so $pD1$ does not distinguish between types $\theta \geq \theta^*$. It is thus reasonable to assume $\phi_0^1 = \theta^*$, which is lower than $\phi_w^1$. (One can show that types $\theta < \theta^*$ will deviate for a smaller set of off-equilibrium beliefs, so $\theta^*$ is the lowest belief that can be supported by $pD1$).
Proof of Lemma 1. First, consider the equilibrium in game $\Gamma_I$. Define

$$\hat{\phi}_w^0 = \hat{\phi}_w^0(\hat{\theta}) \equiv E[\theta \mid 0; \hat{\theta}] = \int_0^{\hat{\theta}} \frac{(1 - \varepsilon)\theta}{(1 - \varepsilon)F(\theta) + \varepsilon} dF(\theta). \quad (A.23)$$

$$\hat{\phi}_w^1 = \hat{\phi}_w^1(\hat{\theta}) \equiv E[\theta \mid 0; \hat{\theta}] = \int_{\hat{\theta}}^1 \frac{1 - \varepsilon}{1 - F(\theta)} dF(\theta). \quad (A.24)$$

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

$$Eu(a = 1) = Eu(a = 0) \quad \hat{\theta}w - c + \mu \hat{\phi}_w^1 \quad \hat{\theta} = \frac{c - \mu(\hat{\phi}_w^1 - \hat{\phi}_w^0)}{w}. \quad (A.25)$$

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.25).

Proof of Proposition 2. We can denote the fraction of people who choose pro-socially $\Gamma_I$ by $(1 - \varepsilon)(1 - F(\hat{\theta}(\varepsilon)))$. The fraction of people who choose to reveal information in game $\Gamma_C$ is $\varepsilon + (1 - \varepsilon)(1 - F(\theta^*(\varepsilon)))$. Thus, we want to prove that

$$(1 - \varepsilon)(1 - F(\hat{\theta}(\varepsilon))) > \varepsilon + (1 - \varepsilon)(1 - F(\theta^*(\varepsilon)))$$

$$F(\theta^*(\varepsilon)) - F(\hat{\theta}(\varepsilon)) > \frac{\varepsilon}{1 - \varepsilon}. \quad (A.26)$$

Let us first consider the LHS of (A.26). To ease notation, define $\Delta(\varepsilon) \equiv \theta^*(\varepsilon) - \hat{\theta}(\varepsilon)$. Substituting in the expressions for $\theta^*$ from (A.8) and $\hat{\theta}$ from (A.25) and setting $k = 0$ we find

$$\Delta(\varepsilon) = \frac{pc - \mu(p\hat{\phi}_w^1 + (1 - p)\phi_0^0(\varepsilon) - \phi_0^0(\varepsilon)) - c - \mu(\hat{\phi}_w^1 - \hat{\phi}_w^0(\varepsilon))}{pw}$$

$$= \frac{p(\hat{\phi}_w^1 - \hat{\phi}_w^0(\varepsilon)) - (p\hat{\phi}_w^1 + (1 - p)\phi_0^0(\varepsilon) - \phi_0^0(\varepsilon))}{pw}, \quad (A.27)$$

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 $\varepsilon$ is large enough. First note that $\Delta(0) < 0$, because $\phi_0^0(0) = \phi_0^1$, and so

$$\Delta(0) = \frac{\mu(p(\hat{\phi}_w^1 - \hat{\phi}_w^0(0)) - (\hat{\phi}_w^1 - \phi_0^0(0)))}{pw}. \quad (A.28)$$
Consider now the case where \( \varepsilon \) is close to 1. Suppose that \( \Delta(\varepsilon) \leq 0 \), i.e. \( \hat{\theta} \geq \theta^* \). This implies that 
\[
\hat{\phi}_w^1 \geq \phi_w^0, \text{ so it must be that }
\]
\[
p\phi_w^0 + (1 - p)\phi_w^0(\varepsilon) \geq \phi_w^0(\varepsilon)
\]
When \( \varepsilon \) gets large, the LHS (A.29) approaches 0, while the RHS is positive. 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.25), we can derive
\[
d\frac{d\Delta(\varepsilon)}{d\varepsilon} = d\frac{\theta^*}{d\varepsilon} - \frac{\hat{\theta}}{d\varepsilon} \left( \frac{\partial \theta^*}{\partial \varepsilon} - (1 - p)\frac{\partial \phi_w^0}{\partial \varepsilon} \right) - \left( \frac{\partial \phi_w^0}{\partial \theta^*} - \frac{\partial \phi_w^0}{\partial \hat{\theta}} \right)
\]
It is easy to show that \( \frac{\partial \phi_w^0}{\partial \varepsilon}, \frac{\partial \phi_w^0}{\partial \hat{\theta}} < 0 \). We can guarantee that the denominators of both terms on the RHS are positive by invoking the uniqueness condition (2) for \( \theta^* \) as well as an equivalent condition for \( \hat{\theta} \)
\[
\frac{d}{d\theta} \left( \hat{\phi}_w^1(\hat{\theta}) - \phi_w^0(\hat{\theta}) \right) > -\frac{w}{\mu}
\]
Therefore, expression A.30 is positive if \( (1 - p)\frac{\partial \phi_w^0}{\partial \varepsilon} < \frac{\partial \phi_w^0}{\partial \hat{\theta}} \), which is satisfied since \( \frac{\partial \phi_w^0}{\partial \varepsilon} = 0 \), while \( \frac{\partial \phi_w^0}{\partial \hat{\theta}} \) is negative.

Summarizing, we have established that \( \exists \hat{\varepsilon} < 1 \), such that if \( \varepsilon > \hat{\varepsilon} \), then the LHS of (A.26) 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.25), we find that
\[
d\frac{d\Delta(\varepsilon)}{d\mu} = d\frac{\theta^*}{d\mu} - \frac{\hat{\theta}}{d\mu} \left( \frac{\partial \theta^*}{\partial \mu} - (1 - p)\frac{\partial \phi_w^0}{\partial \mu} \right) - \left( \frac{\partial \phi_w^0}{\partial \theta^*} - \frac{\partial \phi_w^0}{\partial \hat{\theta}} \right)
\]
Suppose that \( \varepsilon \) is large, so that \( \frac{\partial \phi_w^0}{\partial \theta^*}, \frac{\partial \phi_w^0}{\partial \hat{\theta}} \) and \( \frac{\partial \phi_w^0}{\partial \varepsilon} \) are small and can be ignored. Stability and uniqueness require that \( f(\cdot) \) is relatively flat, so that \( \frac{\partial \phi_w^0}{\partial \hat{\theta}} \approx \frac{\partial \phi_w^0}{\partial \varepsilon} \). Using these findings, we obtain that (A.32) is positive when
\[
p(\hat{\phi}_w^1 - \phi_w^0(\varepsilon)) - (p\phi_w^1 + (1 - p)\phi_w^0(\varepsilon) - \phi_w^0(\varepsilon)) > 0,
\]
i.e. if and only if \( \Delta(\varepsilon) > 0 \). Above, we have shown that this is the case for \( \varepsilon \) sufficiently large.
Let us now consider the RHS of (A.26). Since \( \lim_{\varepsilon \to 1} \frac{\varepsilon}{1 - \varepsilon} = \infty \), \( \exists \bar{\varepsilon} \) such that if \( \varepsilon > \bar{\varepsilon} \), then (A.26) does not hold. Since we have shown that the LHS of (A.26) is positive for \( \varepsilon \) sufficiently large, and increasing in \( \mu \), we obtain the statement in Proposition 2.

Proof of Proposition 3. With respect to game \( \Gamma_I \), we know from the first part of the proof of Proposition 1 that there is a unique cut-off equilibrium with threshold \( \hat{\theta} \), where \( 0 < \hat{\theta} < 1 \). Thus the fraction of prosocial behavior is somewhere in \( (0, 1) \).

With respect to game \( \Gamma_C \), if \( k > 0 \), all *homo economicus* chooses to be ignorant and only altruistic agents select information in the ignorance equilibrium of game \( \Gamma_C \). This implies that the fraction behaving prosocially conditional on observing \( \sigma_w \) is 1. Thus, prosocial behavior is higher under endogenous information.

If \( k \leq 0 \), then the *homo economicus* chooses information. In this case, the fraction of prosocial behavior in the CIG is \( 1 - \varepsilon \). Thus prosocial behavior is higher under endogenous information if \( \varepsilon \) is small enough.

Proof of Proposition 4. Please see the argument in the main text.

Appendix B: Descriptive Statistics

<table>
<thead>
<tr>
<th>Treatment</th>
<th>Chose ignorance</th>
<th>Chose A Ignorant</th>
<th>Chose A AIG</th>
<th>Chose A CIG</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hidden Information</td>
<td>60 (120)</td>
<td>85 (72)</td>
<td>96 (24)</td>
<td>17 (24)</td>
</tr>
<tr>
<td>CIG Only</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>35 (26)</td>
</tr>
<tr>
<td>Reveal Ex-post</td>
<td>26 (35)</td>
<td>-</td>
<td>91 (35)</td>
<td>54 (35)</td>
</tr>
<tr>
<td>Reveal Bonus</td>
<td>46 (35)</td>
<td>88 (16)</td>
<td>100 (11)</td>
<td>25 (8)</td>
</tr>
</tbody>
</table>

Table 2: Percent of dictators choosing ignorance and percent of selfish (A) choices by information state (sample size in parentheses)

Figure 5: Cumulative distribution functions for the two attribute measures used to investigate sorting
Appendix C: Instructions and screenshots [NOT FOR PUBLICATION]

Below we reproduce the instructions for the experiments in the form of screen shots.
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.
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.

<table>
<thead>
<tr>
<th>Player X chooses</th>
<th>Player X receives</th>
<th>Player Y receives</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>B</td>
<td>3</td>
<td>4</td>
</tr>
</tbody>
</table>
At this point, to make sure that everyone understands the game, please answer the two questions below.

<table>
<thead>
<tr>
<th>Player X chooses</th>
<th>Player X receives</th>
<th>Player Y receives</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>B</td>
<td>3</td>
<td>4</td>
</tr>
</tbody>
</table>

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

I choose

<table>
<thead>
<tr>
<th>Player X chooses</th>
<th>Player X receives</th>
<th>Player Y receives</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>6</td>
<td>1</td>
</tr>
<tr>
<td>B</td>
<td>5</td>
<td>5</td>
</tr>
</tbody>
</table>
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.

<table>
<thead>
<tr>
<th>Player X chooses</th>
<th>Player X receives</th>
<th>Player Y receives</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>6</td>
<td>1</td>
</tr>
<tr>
<td>B</td>
<td>5</td>
<td>5</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Player X chooses</th>
<th>Player X receives</th>
<th>Player Y receives</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>6</td>
<td>5</td>
</tr>
<tr>
<td>B</td>
<td>5</td>
<td>1</td>
</tr>
</tbody>
</table>
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
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.

<table>
<thead>
<tr>
<th>Player X chooses</th>
<th>Player X receives</th>
<th>Player Y receives</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>6</td>
<td>1</td>
</tr>
<tr>
<td>B</td>
<td>5</td>
<td>5</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Player X chooses</th>
<th>Player X receives</th>
<th>Player Y receives</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>6</td>
<td>5</td>
</tr>
<tr>
<td>B</td>
<td>5</td>
<td>1</td>
</tr>
</tbody>
</table>
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

<table>
<thead>
<tr>
<th>Player X chooses</th>
<th>Player X receives</th>
<th>Player Y receives</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>6</td>
<td>?</td>
</tr>
<tr>
<td>B</td>
<td>5</td>
<td>?</td>
</tr>
</tbody>
</table>
Reveal Ex-post

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 make 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

<table>
<thead>
<tr>
<th>Player X chooses</th>
<th>Player X receives</th>
<th>Player Y receives</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>6</td>
<td>?</td>
</tr>
<tr>
<td>B</td>
<td>5</td>
<td>?</td>
</tr>
</tbody>
</table>
In this part of the session you have again been paired with another participant in this session, whom we will refer to as the other. You were paired randomly, so this other person is not necessarily the same person you were matched with before. You and this other person will remain mutually anonymous. All of your choices are completely confidential. You will be making a series of decisions about allocating money between you and this other person. For each decision, please indicate the distribution you prefer.

There are no right or wrong answers, this is all about personal preferences. After you have made your decisions, the computer will randomly select ONE of the decisions made either by your or by the other person and this decision will determine the earnings for the two of you from this part of the session. If one of your decisions is selected, then your earnings will be the amount that you allocated to yourself for that decision. If one of the other person's decisions is selected, your earnings will be the amount that they allocated to you for that decision.

Recall that your final earnings for the session are determined by combining your earnings from each of the two parts.
For each of the following decisions, please indicate the distribution you prefer most by marking the respective position along the midline. The numbers indicate how many CENTS are to be allocated to each person.

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Other receives</td>
<td>340</td>
<td>305</td>
<td>270</td>
<td>235</td>
<td>200</td>
<td>185</td>
</tr>
</tbody>
</table>

You receive (in cents) | 0
Other receives (in cents) | 0
For each of the following decisions, please indicate the distribution you prefer most by marking the respective position along the midline. The numbers indicate how many CENTS are to be allocated to each person.

<table>
<thead>
<tr>
<th></th>
<th>340</th>
<th>348</th>
<th>355</th>
<th>363</th>
<th>370</th>
<th>378</th>
<th>385</th>
<th>393</th>
<th>400</th>
</tr>
</thead>
<tbody>
<tr>
<td>You receive</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Other receives</td>
<td>60</td>
<td>78</td>
<td>95</td>
<td>113</td>
<td>130</td>
<td>148</td>
<td>165</td>
<td>183</td>
<td>200</td>
</tr>
</tbody>
</table>

You receive (in cents) 0
Other receives (in cents) 0
For each of the following decisions, please indicate the distribution you prefer most by marking the respective position along the midline. The numbers indicate how many CENTS are to be allocated to each person.

<table>
<thead>
<tr>
<th>You receive</th>
<th>200</th>
<th>218</th>
<th>235</th>
<th>253</th>
<th>270</th>
<th>288</th>
<th>305</th>
<th>323</th>
<th>340</th>
</tr>
</thead>
<tbody>
<tr>
<td>Other receives</td>
<td>400</td>
<td>393</td>
<td>385</td>
<td>378</td>
<td>370</td>
<td>363</td>
<td>355</td>
<td>348</td>
<td>340</td>
</tr>
</tbody>
</table>

You receive (in cents) 0
Other receives (in cents) 0
For each of the following decisions, please indicate the distribution you prefer most by marking the respective position along the midline. The numbers indicate how many CENTS are to be allocated to each person.

<table>
<thead>
<tr>
<th>You receive</th>
<th>200</th>
<th>218</th>
<th>235</th>
<th>253</th>
<th>270</th>
<th>288</th>
<th>305</th>
<th>323</th>
<th>340</th>
</tr>
</thead>
<tbody>
<tr>
<td>Other receives</td>
<td>400</td>
<td>358</td>
<td>315</td>
<td>273</td>
<td>230</td>
<td>188</td>
<td>145</td>
<td>103</td>
<td>60</td>
</tr>
</tbody>
</table>

You receive (in cents) 0
Other receives (in cents) 0
For each of the following decisions, please indicate the distribution you prefer most by marking the respective position along the midline. The numbers indicate how many CENTS are to be allocated to each person.

<table>
<thead>
<tr>
<th>You receive</th>
<th>400</th>
<th>375</th>
<th>350</th>
<th>325</th>
<th>300</th>
<th>275</th>
<th>250</th>
<th>225</th>
<th>200</th>
</tr>
</thead>
<tbody>
<tr>
<td>Other receives</td>
<td>200</td>
<td>225</td>
<td>250</td>
<td>275</td>
<td>300</td>
<td>325</td>
<td>350</td>
<td>375</td>
<td>400</td>
</tr>
</tbody>
</table>

You receive (in cents): 0
Other receives (in cents): 0

OK
For each of the following decisions, please indicate the distribution you prefer most by marking the respective position along the midline. The numbers indicate how many CENTS are to be allocated to each person.

<table>
<thead>
<tr>
<th></th>
<th>You receive</th>
<th>Other receives</th>
<th>You receive (in cents)</th>
<th>Other receives (in cents)</th>
</tr>
</thead>
<tbody>
<tr>
<td>6 of 5</td>
<td>400 393 385 378 370 363 355 348 340</td>
<td>200 218 235 253 270 288 305 323 340</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
Attribute measures: self-image rating

While we are calculating your payment, please answer a few more questions after reading the instructions below.

Here are some characteristics that could describe a person: fair, generous, kind. The person with these characteristics could be you or it could be someone else. For a moment, visualize in your mind the kind of person who has these characteristics. Imagine how that person would think, feel, and act.

When you have a clear image of what this person would be like, please respond to the statements below by indicating how strongly you agree or disagree. For each statement, the responses are (from left to right):

- strongly disagree
- disagree
- disagree slightly
- agree slightly
- agree
- strongly agree

Choose the response that best expresses your feelings.

1. It would make me feel good to be a person who has these characteristics.
2. Being someone who has these characteristics is an important part of who I am.
3. A big part of my emotional well-being is tied up in having these characteristics.
4. Having these characteristics is an important part of my sense of self.
5. I strongly desire to have these characteristics.
6. The types of things I do in my spare time (e.g., hobbies) clearly identify me as having these characteristics.
One of the other person's choices was selected for payment.

It was the following:
- Choice: 6
- Another person assigned to you: 200
- Another person assigned to self: 400

Your total payment for this part (in dollars): 2.00
One of your decisions was selected for payment.

It was the following: Choice 3
You assigned to yourself 200
You assigned to other 400

Your total payment for this part (in dollars): 2.00
You are Player Y. You will not make a choice. Instead, we ask that you evaluate Player X on the basis of his or her possible decisions. Assume that Player X can choose between finding out or not finding out the actual game being played without any consequences his or her payment, i.e. there is no cost.

Please note that your answers will not affect your payment or that of Player X.

How social (as opposed to antisocial) do you view Player X if he or she chooses the following action?

- Does not find out the game and chooses A ($6, $5).
- Does not find out the game and chooses B ($5, $5).
- Does find out the game and has a choice between A ($6, $1) and B ($5, $5). Chooses A ($6, $1).
- Does find out the game and has a choice between A ($6, $1) and B ($5, $5). Chooses B ($5, $5).
- Does find out the game and has a choice between A ($6, $5) and B ($5, $1). Chooses A ($6, $5).
- Does find out the game and has a choice between A ($6, $5) and B ($5, $1). Chooses B ($5, $1).
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 + $0.10". This will add $0.10 to your earnings.

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

I choose:

- □ A
- □ B
- □ Reveal Game + $0.10

<table>
<thead>
<tr>
<th>Player X chooses</th>
<th>Player X receives</th>
<th>Player Y receives</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>6</td>
<td>?</td>
</tr>
<tr>
<td>B</td>
<td>5</td>
<td>?</td>
</tr>
</tbody>
</table>