Self-image and willful ignorance in social decisions

Grossman, Z.; van der Weele, J.J.

DOI
10.2139/ssrn.2237496
10.1093/jeea/jvw001

Publication date
2017

Document Version
Final published version

Published in
Journal of the European Economic Association

SELF-IMAGE AND WILLFUL IGNORANCE IN SOCIAL DECISIONS

Zachary Grossman
Florida State University

Joël J. van der Weele
Department of Economics, University of Amsterdam

Abstract
Avoiding information about adverse welfare consequences of self-interested decisions, or *willful ignorance*, is an important source of socially harmful behavior. To understand this issue, 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 help 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. (JEL: D83, C72, C91)

"Living is easy with eyes closed."

The Beatles, Strawberry Fields Forever, 1967.

1. Introduction

Deliberate avoidance of evidence about the social impact of one's decisions, or "willful ignorance", has an important impact on prosocial behavior. 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)

The editor in charge of this paper was Nicola Gennaioli.
Acknowledgments: We are grateful to the editor, co-editor, and two anonymous referees for excellent comments. We would like to thank Roland Bénabou, Gary Charness, Aldo Rustichini, Jeroen van de Ven, Roberto Weber, Mark Le Quement, Joel Sobel, Heiner Schumacher, Elisabeth Schulte, Karine Nyborg, Leonie Gerhards, Tobias Broer, Andy Loo, and numerous seminar participants for useful comments.

E-mail: grossman@econ.ucsb.edu (Grossman); vdweele@uva.nl (van der Weele)

Journal of the European Economic Association January 2017 15(1):173–217 DOI: 10.1093/jeea/jvw001

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 hesitant to learn about or engage with the inconvenient facts of climate change (Norgaard 2006a; Stoll-Kleemann, O'Riordan, and Jaeger 2001).[1]

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. However, 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, Weber, and Kuang (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.[2]

These findings pose a puzzle for standard economic models of social behavior. The most prominent models assume preferences that are defined over distributions of outcomes, and 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. Models that rely on concerns for reputation or social image cannot explain the decision to ignore information either, as the experimental 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., Bénabou and Tirole 2006; Dana et al. 2007). Self-image is a source of motivation that has long been recognized by psychologists (e.g., Baumeister 1998; Bem 1972; Fiske 2013) and its relevance is echoed in the literature on climate change perceptions. For example, 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 analyze willful ignorance theoretically in the context of a Bayesian preference-signaling model and conduct a set of experiments designed to examine the implications of the model. Our theory builds on Bodner and Prelec (2003) and Bénabou and Tirole (2006), who incorporated self-image concerns

1. In the United States, 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.

2. In addition, several popular overviews of willful ignorance in a wide array of contexts 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).

in economics in the context of signaling model between two selves. 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. In our model, 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 insight of our model, 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 (altruistic) motivations of the decision-maker. Although the choice to avoid information 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 show that the ignorance equilibrium implies five distinct and testable behavioral patterns. First, signaling incentives to behave prosocially are weakened when there is initial uncertainty about the social benefit, 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 some people who act prosocially under full information choose not to obtain relevant information even if it is free. Such information avoidance may help explain 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. This has parallels with climate change perceptions, where Campbell and Kay (2014) demonstrate that ideological preferences determine how people deal with information on this topic. In their experiments, people who perceive higher economic or ideological costs of environmental regulations were 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 did so 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 altruistic preferences 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 *nonwillful ignorance* to be less prevalent than willful ignorance. Finally, inframarginal 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. In doing so, we extend or replicate several previous studies that provided evidence for the implied patterns in different experimental contexts. In Section 5, we use our theoretical results to organize the existing evidence and argue that other explanations cannot adequately account for the observed findings.

Taken together, our theoretical and experimental results provide evidence that self-signaling is an important driver of behavior in social situations. There have been many previous indications of the importance of self-image (Fischbacher and Föllmi-Heusi 2013; Gneezy et al. 2012; Mazar et al. 2009; Murnighan et al. 2001). These papers typically invoke self-signaling after more mainstream explanations fail, but do not explicitly test a particular model. In contrast, more direct tests of predictions about giving derived from signaling models have not found statistically significant effects (Grossman 2015; Van der Weele and Von Siemens 2014). Through the focus on the information acquisition choice, we are able to conduct theoretically informed empirical tests of Bayesian self-signaling in social decisions.

This paper 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 similar 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 nonsignaling 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. Although our approach has some similarities, our use of a self-signaling model endogenizes the self-image value of behavior. Furthermore, our model organizes a broad constellation of behavior and generates new, testable hypotheses for which we provide experimental evidence.

We argue that ignorance can serve to avoid large signaling investments in prosocial behavior that would be necessary to maintain a good self-image under full information. Others have explored ignorance as a commitment device in different contexts, for example, in the presence of time inconsistency (Bénabou and Tirole 2011; Carillo and Mariotti 2000). 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 (Eil and Rao 2011; Grossman and Owens 2012; Köszegi 1996; Möbius et al. 2011). 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 a 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 that 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, which 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 causes an uncertain impact on social welfare, W. She knows that $W = w$ with prior probability p and that $W = 0$ with complementary probability, where $w > c$. 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 σ_w denotes a "high signal" ($W = w$), σ_0 denotes a "low signal" ($W = 0$), and with some abuse of notation $\emptyset$ denotes the case in which no information is acquired.

Preferences Each agent is characterized by a preference type (θ, μ), 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 of the decision-maker self are represented by the following utility function:

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

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 θ, minus the cost of the action, and it is zero if she does not take the action. Thus, θ can be interpreted as the degree of altruism or prosocial motivation of the agent. The fact that types with low values of θ (low types) care less about welfare implies a single crossing-property that underlies the semiseparating 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 $E[\theta|\sigma, a]$ reflects the utility from (self-)image, given by the expectation of the agent's social-preference parameter, θ, 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 θ and μ. 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. Although these emotions are informative about the agents type, they are nonverifiable, and their intensity may be hard to imagine 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 θ and μ based upon the available information, namely, the agent's choices and the signal σ.[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 internal observer.[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-maker self also specifies behavior for counterfactual realizations of the signal.[5]

Our interpretation of the weighting parameter μ is that it captures the psychological benefits of a good self-image to the agent. Indeed, psychologists emphasize the purely psychological rewards of positive self-esteem (Fiske 2013). Alternatively, we could view μ as a reduced form representation of a model where the decision-maker self cares about the material benefits from having a good self-image, due, for example, to increased motivation or more success in social interactions.

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

1. With a probability ε that is bounded above 0, the agent is nonsocial agent or *homo economicus*, who only cares about her own material payoff, that is, $\theta = \mu = 0$. The existence of *homo economicus* is suggested by studies on prosocial behavior in, for example, the public goods game, in which a minority of subjects act like

3. Note that although both θ and μ are subject to the observer's inference, we assume that image utility depends only on beliefs about the agent's altruism θ. In doing so, we follow a large literature on the costly signaling of altruistic preferences. In our analysis, we will not explicitly discuss the simultaneous inferences the observer makes about μ. Although other conceptions of self-image concern are certainly possible, our objective is to examine whether existing, previously proposed models can explain behavior. An interesting, yet unexplored area for future research would be to examine the consequences of signaling over the image-concern parameter μ. One conjecture is that if behavior seemingly motivated by a less-than-pure desire to look good is heavily discounted, interventions aimed at leveraging image concerns to encourage good behavior will have limited effectiveness and may even backfire.

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.

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

free riders (e.g., Burlando and Guala 2004; Fischbacher, Gächter, and Fehr 2001; Kurzban and Houser 2005).

2. With probability $1 - \varepsilon$ the agent is a *social agent*, who cares about her image and about social welfare. For social agents, θ 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 μ to be homogeneous for social agents.[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 can be sustained as a pooling equilibrium.

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).

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, (θ, μ).
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 σ 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. Beliefs are formed by the application of Bayes' rule wherever possible, that is, $E[\theta|\sigma, a] = E[\theta|\sigma, a; s^*]$, where s^* is the equilibrium strategy profile. We assume the tie-breaking rule that all agents acquire information if they are indifferent between doing so or not doing so. Given that we want to explain willful ignorance, this represents the most conservative assumption and ensures our results are not driven by indifferent agents.

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

Although this tie-breaking assumption is not necessary for our main propositions, it simplifies the exposition.[7]

Though we consider potential pooling equilibria in Section 2.4, our main focus is on semiseparating equilibria, which are equilibria with a cutoff 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 θ^* 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). In this context, an equilibrium is unstable if a deviation from the threshold type implies that it is optimal for other types to also change their behavior.

Under what conditions can we rule out multiplicity and guarantee the stability of equilibrium? Consider the beliefs of the observer about θ. Upon seeing information acquisition and ignorance, these beliefs are $E[\theta|\theta \geq \theta^*, k]$ and $E[\theta|\theta < \theta^*, k]$, respectively, and we can define the image reward of information acquisition as $\delta(\theta^*, k)$: $=E[\theta|\theta \geq \theta^*, k] - E[\theta|\theta < \theta^*, k]$. In any ignorance equilibrium, the cutoff type must be indifferent between the expected cost of information $k - p(\theta^* w - c)$ and the 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 θ^* is positive. Moreover, a sufficient (but not necessary) condition for uniqueness of the threshold is that the total derivative is positive for all θ. This implies that the expected cost of information is always decreasing more steeply in θ than the image reward, which implies, in turn, that there is at most one θ for which they are equal. For the remainder of this paper, we will assume that this is the case, that is,

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

Here, the left-hand side depends on the density $f(\theta)$, and will not be far from zero if the distribution of types $f(\theta)$ is relatively flat, i.e. the mass is not concentrated on specific parts of the type space. In this case, and if μ is not too large, (2) is likely to be satisfied (see Bénabou and Tirole 2006, pp. 1667–1668 for a more extensive discussion). In Section 5.3, we show that the threshold is stable and unique in a numerical example featuring a (truncated) normal distribution of social types.

With respect to beliefs about out-of-equilibrium actions, we require that these beliefs satisfy a refinement we call "pD1", which 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. More formally, denote the

7. A previous version available at https://ideas.repec.org/p/cdl/ucsbec/qt0bp6z29t.html provided proofs of more general versions of the propositions featuring weaker assumptions about the tie-breaking rule. Notably, Propositions 1 (page 8 of the working paper) and 3 hold for any assumption about the probability that an indifferent agent will choose information and the main claim of Proposition 2 (page 11 of the working paper) holds as long as the fraction of indifferent agents choosing information is sufficiently high.

observer's beliefs by $\varphi(a, \sigma; \theta) \in [0, 1]$, and the equilibrium payoffs of the decision-maker with altruism level θ by $u^*(\theta)$. We define $D(\theta) := \{\varphi: u(\theta, a, I; \varphi) > u^*(\theta)\}$, that is, the set of observer beliefs for which a type with altruism θ would like to deviate to the off-equilibrium action. pD1 requires that $\varphi(a, \sigma; \theta')$ is positive only if $D(\theta')$ is maximal, that is, not a proper subset of any $D(\theta)$.[8]

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

1. *the homo economicus chooses $a = 0$ and acquires information if and only if $k \leq 0$,*
2. *all social types $\theta < \theta^*$ remain ignorant and choose $a = 0$, whereas 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 and obtain a high image 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 beliefs for this out-of-equilibrium action are specified to be 0 (if $k > 0$).[9] When choosing between these options, social agents

8. The traditional D1 refinement defines the set $D(\theta)$ over the set of *actions* 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.

9. These beliefs satisfy the pD1 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.

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 θ^* specifies that they behave prosocially 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 altruistic to behave prosocially.

Proposition 1 contains some sufficient conditions on the parameters: 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 the cost of information k is too high or too low, either no-one or everyone will acquire information.[10]

2.4. Multiplicity and Information-Seeking Equilibrium

Like many signaling games, this game has 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.

Which of these equilibria, if any, will be played is an empirical question that can only be settled by deriving behavioral predictions that distinguish between them. As we will show in detail below, 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.

10. 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 types would pool on information and sufficiently large k or c would lead to pooling on ignorance.

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 experimental literature on willful ignorance is the finding by DWK that although 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 the same subject is 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.

A first test of the self-signaling model is whether it can explain these results, which, as we will argue in Section 5.2, are inconsistent with theories of outcome-based preferences and with social signaling. To investigate this, we denote by Γ_C the information "choice game" defined in Section 2.1 (where subscript *C* stands for choice) and denote by Γ_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 agents who behave selfishly in Γ_I with the share that remains ignorant in game Γ_C, we first derive the equilibrium in game Γ_I. As we did in game Γ_C, we focus on a semiseparating equilibrium characterized by a unique cutoff 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 ζ be the image reward of being prosocial as a function of the threshold type $\hat{\theta}$ and the fraction of homo economicus, that is, $\zeta(\hat{\theta}, \varepsilon) := E[\theta|\theta > \hat{\theta}, \varepsilon] - E[\theta|\theta > \hat{\theta}, \varepsilon]$. The threshold type is such that the net cost of the prosocial action equals the image benefits, that is, $\hat{\theta}w - c + \mu\zeta(\hat{\theta}, \varepsilon) = 0$. A sufficient condition for uniqueness and stability is that the total derivative of this expression is positive, that is, $d\zeta(\theta)/d\theta > -w/\mu$ for all θ.

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

We compare this equilibrium with the ignorance equilibrium in the choice game Γ_C where, to be consistent with the experiments in the literature we assume $k = 0$, and obtain the following result.

PROPOSITION 2. *There exist* $\bar{\mu}$, $\bar{p} < 1$, *and* $\bar{\varepsilon} < 1/2$ *such that if* $\mu > \bar{\mu}$, $p > \bar{p}$ *and* $\varepsilon < \bar{\varepsilon}$, *then the fraction of all agents—including both social agents and homo economicus—who choose ignorance in the ignorance equilibrium of* Γ_C *is higher than the fraction of all agents who act selfishly in* Γ_I.

Proposition 2 shows that the signaling model can explain information avoidance under the right constellation of parameters. Information avoidance occurs because the social types want to avoid pooling with the *homo economicus*, which has different effects on their behavior in the two games. In game Γ_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 Γ_C, the *homo economicus* will choose information.[11] This *decreases* the signaling value of acquiring information and causes marginal social types in game Γ_C switch to ignorance, increasing the amount of selfish choices.

The conditions on the parameters are sufficient to ensure that the amount of social types who shift toward ignorance in game Γ_C outweighs the number of *homo economicus* who choose to reveal. First, ε should not be too large because when there are too many *homo economicus*, there will be little ignorance. Note that ε also needs to exceed 0, as we assumed earlier, in order to dilute the image associated with choosing information. Second, the importance of image concerns, as measured by μ, needs to be high enough to induce enough social types to shift toward ignorance. Third, the prior probability of there being a positive social benefit should be high enough to make ignorance attractive to the social types in game Γ_C. 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 Γ_I. In Section 5.3, we show a plausible numerical example in which these conditions are fulfilled and the model generates information avoidance.

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 nonsocial 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 differently

11. 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.

from those who obtained the same information passively. To investigate this idea formally, we compare the behavior of those who inform themselves in the ignorance-choice game Γ_C to the behavior of the players in the game with full information Γ_I (defined above).

PROPOSITION 3. *If $k > 0$, the fraction of agents that engages in prosocial behavior ($a = 1$) in the full information game Γ_I is lower than the corresponding fraction among agents who inform themselves in the ignorance equilibrium choice game Γ_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 Γ_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 prosocially, where the cutoff type depends on the parameters w, c, and μ. In comparison, in the choice game Γ_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 among 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 among 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 they 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. *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 $E[\theta|\sigma_w, 0] < E[\theta|\emptyset, 0] < E[\theta|\sigma_w, 1]$. 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 we could have chosen to know the actual consequences.

3.4. Willful versus Nonwillful 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, we would expect more people to choose information.

To establish this intuitive claim more formally, consider the modified game, Γ_N, which shares almost all the same features as the choice game Γ_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.[12]

PROPOSITION 4. *When $k \leq 0$, the fraction of agents choosing ignorance in Γ_N is lower than the corresponding fraction in the ignorance equilibrium of Γ_C.*

To see the logic behind this result, consider both games in turn. In game Γ_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 Γ_C. In contrast, in Γ_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.[13]

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, that

12. 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.

13. Note that when $k = 0$ our result depends on our tie-breaking rule, as the *homo economicus* is indifferent about information acquisition in both games, and all agents are indifferent in game Γ_N. We can relax our tie breaking rule, such that when a fraction α of indifferent agents reveals, the proposition holds when $\alpha > 1 - F(\theta^*)$.

is, $k < 0$, makes ignorance less attractive for all agents and causes θ^* 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 *homo economics* under full information, or pooling with the altruists at a cost of c.

COROLLARY 2. *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 and to demonstrate the robustness and replicability of the existing evidence, 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 nonignorance 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, whereas 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 was reproduced almost verbatim, as were the condition-specific instructions in the replication conditions.[14]

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. 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 (Table B.1) provides descriptive statistics of our data and instructions are provided in the Online Appendix. This can be found in Supplementary Data.

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 (Feiler 2014; Grossman 2014; Larson and Capra 2009), so our primary objective is not to establish the robustness of this result. Rather, we seek to (a) establish a baseline 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

14. 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.

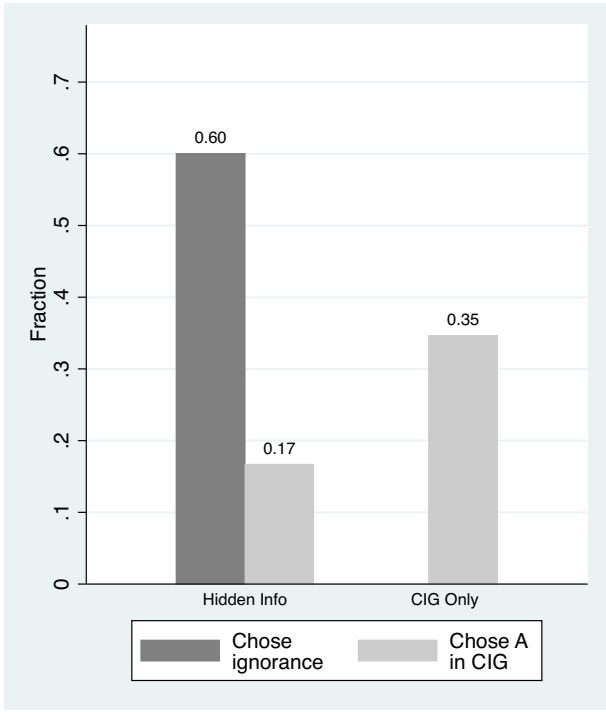


FIGURE 1. The dark bar represents the proportion of participants that chose ignorance in the *Hidden Information* treatment and the light bars represent the proportion that choose selfishly (*A*) conditional upon knowingly being in the *CIG*. Fractions of subjects choosing the relevant behavior are displayed 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 shows evidence consistent with sorting.

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, whereas the alternative hypothesis is that the ignorance rate is actually higher.

The results are shown in Figure 1.[15] First, although only 9 out of 26 (35%) dictators in the *CIG Only* condition chose *A*, in the *Hidden Information* condition 72 out of 120

15. 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). In 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.

(60%) chose ignorance. This difference is significant at 5% using a one (or two) sided Fisher's exact test ($p = 0.016$, FET).[16] 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, whereas 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 states. In the latter, participants can freely choose or avoid information, so those who know that they 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.[17]

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% of the dictators (4 of 24) knowingly facing the CIG choose selfishly, substantially fewer than the 35% (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.

Comparing Types across Behaviors. An alternative methodology to examine sorting is to provide an additional measurement of the types of the dictators, and 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

16. 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.

17. 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 the CIG. However, the small sample size (only eight dictators in the latter) means we cannot conclude much from this comparison.

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, Ackermann, and Handgraaf (2011). In the SVO task a subject is asked to choose an allocation of money between herself and another subject in six 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 prosociality for each subject, measured as an angle and determined by the mean allocations to the self and to the other person. The SVO score increases in the prosociality of the choices, with scores below 23 degrees indicating competitiveness and selfishness and higher scores indicating a more prosocial disposition. The SVO task is a standard instrument for the measurement of social preferences, both in social psychology and economics (see, e.g., Offerman et al. 1996).

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 score 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 scores 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.[18] 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 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 prosocially in the CIG.

18. Cumulative distribution functions for the SVO score and self-image rating are available in Appendix B.

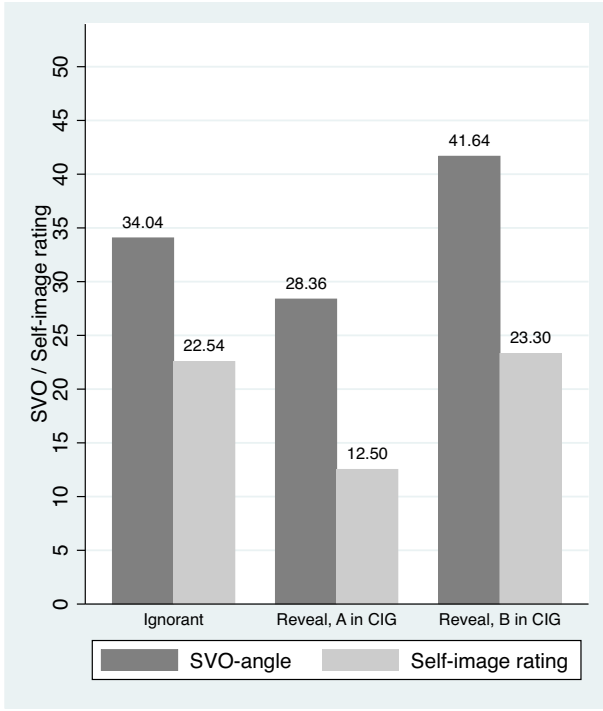


FIGURE 2. The dark bars represent the average score of the Social Value Orientation task among dictators. The light bars represent average importance-of-self-image ratings among dictators. Average ratings are displayed on top of the bars. The results of both measures are consistent with sorting in ignorance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

Figure 2 shows the mean SVO score and self-image rating for dictators, broken down by outcome in the *Hidden Information* game. The mean SVO score for the 46 ignorant dictators is 34.0, barely below the mean of all dictators. In contrast, the mean SVO score of the ten 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. Although the fact that only two dictators chose selfishly with knowledge of being in the CIG limits the ability to draw strong conclusions from the comparison, the mean SVO score 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 score 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 prosocially. 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.

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 chooses 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 nine 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 antisocial" to "very social". The recipients completed these ratings whereas 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 antisocial" 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

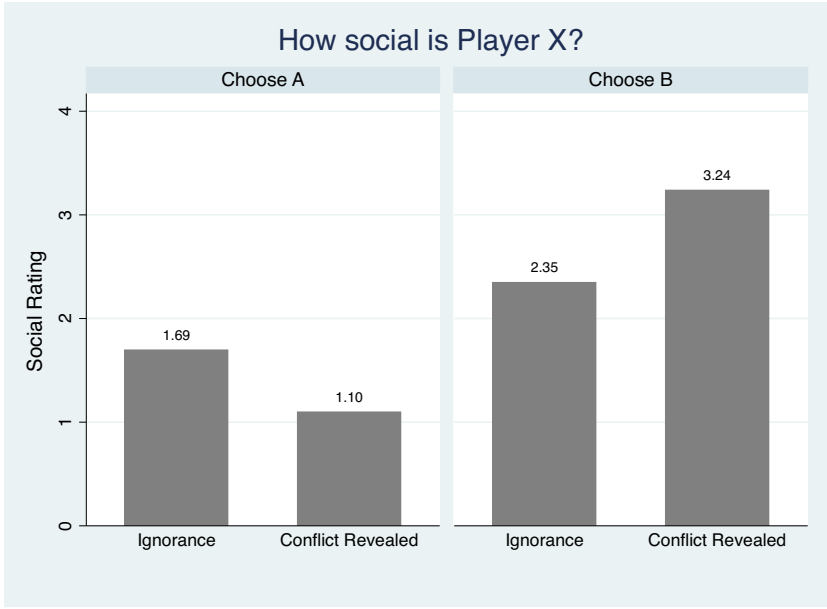


FIGURE 3. Answers to the question "How social do you rate Player X, based on each of the following actions ...", where actions included the joint decision to be informed or not and to choose A ($6) or B($5). Answers were given on a 5-point scale from "very antisocial" (0) to "very social" (4). Average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

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 "information seeking" equilibrium in which ignorance is associated with the lowest image.

4.4. Willful versus Nonwillful 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 afterward. To test this prediction, we contrast the *Hidden Information* treatment with a *Reveal Ex post* treatment, which is modeled on game Γ_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.[19]

Following Proposition 4, we hypothesize that subjects will be more likely to choose information in the *Reveal Ex post* treatment than in the *Hidden Information* treatment. In addition, the results of the *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 *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 *Reveal Ex post* treatment was 26%. This is significantly lower at the 1% level than the 60% rate in the *Hidden Information* condition, thus supporting our hypothesis that more agents reveal *after* than *before* the dictator decision.[20]

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 right bar in Figure 4 shows that the ignorance rate in the *Reveal Bonus* condition was 0.46, that is, much larger than zero.[21] Although the fraction of ignorance in the

19. This experiment partially replicates and overlaps with 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 reports the results of contemporaneous *Hidden Information* treatment sessions that are included in our data set, under the treatment name *Default NR*. However, the *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 *Hidden Information* treatment reported herein.

20. In the *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%. 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 the absence of image considerations.

21. 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.

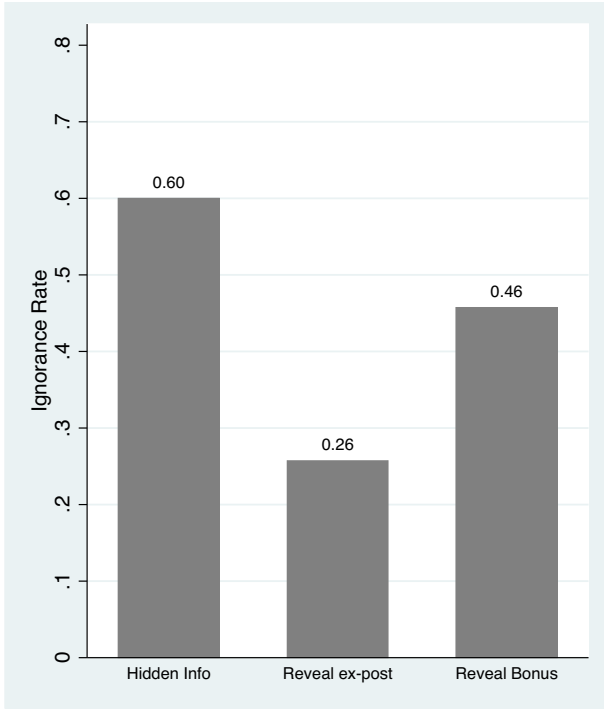


FIGURE 4. Ignorance rate by treatment, the fraction of subjects choosing ignorance is at the top of the bars. Comparing the *Hidden Information* treatment ignorance rate with that of the *Reveal Ex post* and *Reveal Bonus* treatments provides support for the predictions of ignorance equilibrium on nonwillful ignorance and paying for ignorance.

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 ($p = 0.096$, FET). Thus, although 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.[22]

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

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).

22. 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% (ten out of ten dictators), suggesting that the incentive was quite salient indeed.

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, we 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.

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.

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 statistically 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

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

	This study	Other studies	Consistent with		
			Material preferences	Ignorance equilibrium	Infomation equilibrium
Information avoidance	Replicate DWK in larger sample	*Find avoidance:* DWK, Larson and Capra (2009), Feiler (2014), Matthey and Regner (2011), Ehrich and Irwin (2005)	×	✓	×
Sorting	Find sorting by comparing both behavior across situations and type measures across behavior	*Find sorting:* Conrads and Irlenbusch (2013), Fong and Oberholzer-Gee (2011), Kajackaite (2015). *Do not find sorting:* Larson and Capra (2009)	✓	✓	×
Exculpation	Selfish dictators regarded "more social" when willfully ignorant	*Find exculpation:* Krupka and Weber (2013), Bartling, Engl, and Weber (2014), Conrads and Irlenbusch (2013)	×	✓	×
Nonwillful ignorance	Less ignorance when it cannot affect image	Grossman (2014) reports same finding with partially the same data	×	✓	×
WTP for ignorance	46% willing to pay for \$0.10 ignorance	Cain and Dana (2012): 23% willing to pay \$0.25 for ignorance	×	✓	×

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, Engl, and Weber (2014) find that given an unfair outcome, willful ignorance reduces the costly punishment meted out by a third party.

The nonwillful ignorance finding provides one of the biggest challenges for the 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 an artifact 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, whereas any ignorance in the *Reveal Ex post* condition cannot. To our knowledge, the comparison between willful and nonwillful ignorance has not been made before.

When it comes to paying for ignorance, Cain and Dana (2012) independently 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 is 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, that is, $(5, 5) \succsim (6, 1)$, 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 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 nonwillful 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 α 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 *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.

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. *Paying for ignorance* is also 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.

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 four out of five 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 information is costless. When we use the tie breaking rule that homo economicus chooses to inform himself, ignorance is zero.[23] Second, if there is no positive

23. Even if we used the opposite tie-breaking rule, ignorance is at most ε. Since at least a fraction ε do not take the social action in the full information game Γ_I, ignorance cannot be higher than selfishness in Γ_I.

information cost then there is also no *sorting*, as all agents choose information acquisition.[24] 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, *nonwillful 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 toward 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 does not know if it was from the choice of *B* in the *CIG* or *A* in the *AIG*, and 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.

24. If $k > 0$, *homo economicus* chooses ignorance. In that case, it is possible that a larger fraction of subjects chooses to be nice in game Γ_C (conditional on knowingly playing the CIG) than in Γ_I. However, since being selfish in the CIG does not involve pooling with *homo economicus*, selfishness in the CIG becomes more attractive relative to game Γ_I. If this effect is strong, it could even result in antisorting. Thus, the result is ambiguous for $k > 0$.

Apart from the recipient, the experimenter is another potential external audience for the dictator. We can think of audience concerns toward 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 judgment 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 toward 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, the results of the self-image ratings reported above are in line with the self-signaling interpretation, as are the questionnaire responses from dictators in a version of the *Hidden Information* treatment reported in Van der Weele (2014). On an open question asking them for a motivation behind their choices, four (out of eleven) 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 we think 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 (Spiekermann and Weiss 2016). 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 nonstandard 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 20 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, whereas 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 μ, ε, and the distribution of social types, as well as the utility valuations of monetary payoffs. In 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, we may wonder whether the equilibria discussed above are consistent with plausible parameter values, and conversely, whether there are plausible values that rule out willful ignorance.

In order to investigate whether the signaling model works for plausible parameter values, we conduct a simple calibration exercise using a simulated version of our model.[25] For the distribution of types, we assume that $\varepsilon = 0.1$ and that θ is distributed

25. Our simulations are conducted in Matlab; the program is available in the Online Appendix to this paper and this can be found in Supplementary Data.

according to a (truncated) normal distribution on [0, 1] with mean 0.5 and standard deviation 0.1. To match the probability in the experiment, we choose $p = 0.5$. To reflect the absence of monetary information costs in the experiment we choose $k = 0$, and we set $c = 1$ to capture the dollar cost of choosing B over A. We pick the values for the remaining parameters, μ and w, so as to generate equilibrium behavior that matches that observed in the experiment. Setting $w = 1.65$ and $\mu = 0.95$ we match exactly the rate of selfish choices in the baseline game Γ_I (35%) and the CIG (17%).[26] In addition, we find an ignorance rate of 43%, which is somewhat lower than in our experiment, but matches that of DWK.[27]

These outcomes show that we can generate both sorting and information avoidance in a simple example, while the model's restrictions $\mu < c < w$ are satisfied. Although this exercise does not constitute an empirical test, it illustrates that the model does not depend upon implausible assumptions in order to generate our main predictions. It also shows that not everything goes when it comes to the assumptions on the parameter values. For example, when we increase μ from 0.95 to more than 7 (thereby violating the assumption that $\mu < c$), we find that all social types would pool on acquiring information, and so would the *homo economicus* as long as $k \leq 0$. Thus, in situations where image concerns are extremely strong relative to the material payoffs, the model predicts we will not observe willful ignorance. Furthermore, in the calibration above, $\mu > 0.85$ is necessary to generate information avoidance, in line with the condition in Proposition 2.

When it comes to the stability of the ignorance equilibrium, setting $\mu > 2$ leads to existence of multiple equilibrium thresholds θ^*, some of which are unstable. A similar result occurs when the standard deviation of the type distribution becomes very small, implying a dense concentration of types around 0.5. These findings illustrate the remarks surrounding equation (2): multiple equilibria occur if a small change in the threshold causes a large shift in the image associated with different actions.

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 versus nonwillful ignorance." The predictions of "information avoidance" and "sorting" depend on the comparison of the equilibrium thresholds in different games. With multiple thresholds, it is not clear that these comparisons will hold for all pairs of thresholds without making additional assumptions.

26. Regarding w, note that the monetary benefit to the receiver in the experiment is 4. However, assuming $w = 4$ implies that the highest type values the other person's payoffs just like her own, which seems an implausibly high degree of altruism. The choice of $w = 1.65$ implies a more modest and probably more realistic level of altruism.

27. Note that we use two degrees of freedom to match three experimental outcomes. If we use more degrees of freedom and tinker with other parameters, the model can be made to fit the data better. If we allow all the parameters to vary, we can obtain similar outcomes with different constellations of parameters.

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. 2016). In pay-what-you-want (PWYW) schemes where some of the proceeds go to charity, Gneezy et al. (2012) find 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 (Broberg, Ellingsen, and Johannesson 2007; Dana, Cain, and Dawes 2006; Lazear, Malmendier, and Weber 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 (Broberg et al. 2007; Lazear et al. 2012) 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, μ. An agent with a low θ but high μ might be swayed by image-concern to act 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 diverse range of environments that reinforce our own experimental results. Together, these results give robust support to the broader theory of self-signaling and the usefulness of self-signaling models for understanding behavior.

From a policy perspective, our results indicate that making more information available about the social benefits of particular actions may not be enough. In the context of climate change, Osberghaus, Finkel, and Pohl (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. Our model suggests that the policies of providing information exogenously and raising (self-)image concerns may be more effective in tandem than in isolation.

Another avenue is subsidization of actions that benefit public goods like a stable 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, something that is also predicted by our model. More generally, there is a need to explore the design of choice environments that exploit self-signaling to advance the common good.

Appendix A: Proofs

A.1. 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 $\varphi_\sigma^a = E[\theta \mid a, \sigma; s]$ denote the expectation conditional on the equilibrium strategy profile, the chosen action a and information σ. Beliefs on the equilibrium path are as follows:

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

$$\varphi_0^0 = \varphi_0^0(\theta^*) \equiv E[\theta \mid 0, \sigma_0; \theta^*] = \int_{\theta^*}^1 \frac{(1-\varepsilon)\theta}{(1-\varepsilon)(1-F(\theta^*)) + \varepsilon} dF(\theta), \tag{A.2}$$

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

$$\varphi_w^0 = \varphi_w^0(\theta^*) \equiv E[\theta \mid 0, \emptyset; \theta^*] = 0. \tag{A.4}$$

With respect to beliefs off the equilibrium path, we assume that $\varphi_\emptyset^1 \leq \varphi_w^1$ and we make no assumption about φ_0^1.

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

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

It is immediate that in equilibrium all types $\theta \geq \bar{\theta}$ who observed $\sigma = \sigma_w$ take the prosocial action, and all $\theta < \bar{\theta}$ do not. Note that $\bar{\theta}$ is bounded above zero, because $\bar{\theta} \geq c - \mu/w$, while $c > \mu$.

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

$$\begin{aligned} u(a = 0 \mid \sigma = \sigma_0; \theta^*) &> u(a = 1 \mid \sigma = \sigma_0; \theta^*) \\ \times\, \mu\varphi_0^0 &> -c + \mu\varphi_0^1 \\ c &> \mu\big(\varphi_0^1 - \varphi_0^0\big), \end{aligned} \tag{A.6}$$

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

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

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

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

$$\begin{aligned} Eu(\text{acquire info}) &= Eu(\text{not acquire info}) \\ (1-p)\mu\varphi_0^0 + p\left(\theta^* w - c + \mu\varphi_w^1\right) - k &= \mu\varphi_\emptyset^0 \\ \phi &= \frac{pc + k - \mu\left(p\varphi_w^1 + (1-p)\varphi_0^0 - \varphi_\emptyset^0\right)}{pw}. \end{aligned} \tag{A.8}$$

It is straightforward that all types $\theta < \theta^*$ remain ignorant and all types $\theta \geq \theta^*$ acquire information. Existence of θ^* follows from the continuity of both sides of (A.8).

For future reference, we now establish the conditions under which θ^* is bounded above zero. Plugging $\theta^* = 0$ into (A.8) it follows that this is the case iff k is bounded above $\mu\left(pE_F\theta + (1-p)\varphi_0^0\right) - pc$. The fact that ε is bounded above zero implies that φ_0^0 is bounded below $E_F\theta$, so $\mu\left(pE_F\theta + (1-p)\varphi_0^0\right) - pc$ is bounded below $\mu E_F\theta - pc$. Thus, if

$$p > \frac{\mu E_F\theta}{c} \equiv p_0, \tag{A.9}$$

then we can find a k_0 bounded below 0 such that θ^* is bounded away from 0 if $p > p_0$ and $k > k_0$. Note that since we assumed that $f(\theta)$ has full support and $\mu < c$, both $E_F\theta$ and p_0 are bounded away from 1.

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

Some algebra shows that if $k \leq 0$, then $\theta^* < \tilde{\theta}$ iff

$$\begin{aligned} (1-p)c &> \mu\left(\varphi_\emptyset^1 - p\varphi_w^1 - (1-p)\varphi_0^0\right) \\ (1-p)\left[c - \mu\left(\varphi_w^1 - \varphi_0^0\right)\right] &> \mu\left(\varphi_\emptyset^1 - \varphi_w^1\right). \end{aligned} \tag{A.10}$$

This inequality is satisfied: the LHS is positive (since $c > \mu$) and the RHS is negative (as we assumed $\varphi_\emptyset^1 \leq \varphi_w^1$).

Similar algebra comparing (7) and (A.8) yields that $\bar{\theta} < \theta^*$ if and only if

$$k > \mu\left((1-p)\varphi_0^0 - \varphi_\emptyset^0\right) \equiv k_1. \tag{A.11}$$

We require that k_1 is bounded below 0. A necessary condition is that $\varphi_\emptyset^0$ is bounded above 0, which is the case if θ^* is bounded above 0 (i.e., if $p > p_0$ and $k > k_0$ as we established above). This ensures that there exists a $p_1 > 1 - \varphi_\emptyset^0/\varphi_0^0$ and bounded below 1 such that k_1 is bounded below 0 if $p > p_1$.

It remains to check that $\theta^* < 1$. Note that if the threshold type is $\theta^* = 1$, we have $\varphi_w^1 = 1$, $\varphi_0^0 = 0$, and $\varphi_\emptyset^0 = E_F\theta$. Plugging $\theta^* = 1$ into (A.8), it is straightforward to show that a sufficient condition is $p > \mu E_F\theta/(w - c + \mu) \equiv p_2$.

Combining arguments, an interior equilibrium exists if $k > \max\{k_0, k_1\} \equiv \underline{k}$ and $p > \max\{p_0, p_1, p_2\} \equiv \tilde{p}$, where $\underline{k}$ and $\tilde{p}$ are bounded below 0 and 1, respectively.

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

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

$$\varphi_w^1 = \varphi_0^0 = \varphi_0^0(\theta^*) \equiv E[\theta \mid 0, \sigma_0; \theta^*] = \int_{\theta^*}^1 \frac{\theta dF(\theta)}{1 - F(\theta^*)}. \quad \text{(A.13)}$$

With respect to beliefs off the equilibrium path, we assume that $\varphi_w^0 = 0$ and $\varphi_\emptyset^1 \leq \varphi_w^1$.

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

$$\bar{\theta} = \frac{c - \mu\varphi_w^1}{w}, \quad \text{(A.14)}$$

$$\tilde{\theta} = \frac{c - \mu(\varphi_\emptyset^1 - \varphi_\emptyset^0)}{pw}, \quad \text{(A.15)}$$

$$\theta^* = \frac{pc + k - \mu\left(\varphi_w^1 - \varphi_\emptyset^0\right)}{pw}. \quad \text{(A.16)}$$

First, we establish a sufficient condition for θ^* being bounded below 1 and above 0. By substituting $\theta^* = 1$ into (A.16), we see that $\theta^* < 1$ iff $k < p(w - c) + \mu(1 - \varphi_\emptyset^0)$. Since $\varphi_\emptyset^0 < E_F\theta$, a sufficient condition is

$$k < p(w - c) + \mu(1 - E_F\theta). \quad \text{(A.17)}$$

Since the RHS of this expression is strictly positive and bounded above 0, we can find a k_4 that is strictly positive and bounded away from 0, such that θ^* is bounded below 1 if $k < k_4$.

Furthermore, by substituting $\theta^* = 0$ into (A.16), we see that $\theta^* > 0$ iff $k > \mu E_F\theta - pc$. Since $\mu E_F\theta$ is bounded below 1, we can find a p_4 bounded below 1 such that this condition holds for any $k > 0$ if $p > p_4$.

Next we establish sufficient conditions for the existence of a $\theta^* \in (0, 1)$ such that $\bar{\theta} < \theta^* < \tilde{\theta}$. First, $\theta^* < \tilde{\theta}$ iff

$$k < c(1 - p) + \mu\left(\varphi_w^1 - \varphi_\emptyset^1\right) \equiv k_5. \quad \text{(A.18)}$$

The fact that θ^* is bounded below 1 implies that $\varphi_w^1 - \varphi_\emptyset^1$ is bounded above 0, so $k < k_4$ implies that k_5 is bounded above 0.

Next, some algebra yields that $\bar{\theta} < \theta^*$ iff $k > \mu(\varphi_w^1(1-p) - \varphi_\emptyset^0)$, which is satisfied for all $k > 0$ if $p \geq 1 - \varphi_\emptyset^0/\varphi_w^1 \equiv p_5$. We know θ^* (and therefore $\varphi_\emptyset^0$) is bounded above 0 if $p > p_4$, implying that p_5 is bounded below 1.

Thus, we have established that there exists a $\bar{k} \equiv \min\{k_4, k_5\}$ bounded above 0 and $\hat{p} \equiv \max\{p_4, p_5\}$ bounded below 1 such that there exists a $\bar{\theta} < \theta^* < \tilde{\theta}$ if $k < \bar{k}$ and $p > \hat{p}$.

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. $\varphi_w^0 = 0$ when $k > 0$.
   Types $\theta < \theta^*$ would deviate if the deviation payoffs are bigger than equilibrium payoffs

$$(1-p)\mu\varphi_0^0 + p\mu\varphi_w^0 - k > \mu\varphi_\emptyset^0\varphi_w^0 > \frac{\varphi_\emptyset^0 - (1-p)\varphi_0^0 + k/\mu}{p}. \tag{A.19}$$

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

$$(1-p)\mu\varphi_0^0 + p\mu\varphi_w^0 - k > (1-p)\mu\varphi_0^0 + p(\theta w - c + \mu\varphi_w^1) - k$$
$$\varphi_w^0 > \varphi_w^1 + \frac{\theta w - c}{\mu}. \tag{A.20}$$

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

$$\varphi_w^1 + \frac{\theta w - c}{\mu} \geq \frac{\varphi_\emptyset^0 - (1-p)\varphi_0^0 + k/\mu}{p}\theta \geq \frac{pc + k - \mu(\varphi_w^1 + (1-p)\varphi_0^0) - \varphi_\emptyset^0)}{pw}. \tag{A.21}$$

   Since the RHS is exactly equal to θ^*, this expression is satisfied in equilibrium for all θ .

   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 $\varphi_w^0 = 0$.
2. $\varphi_\emptyset^1 \leq \varphi_w^1$.
   Types $\theta \geq \theta^*$ deviate if deviation payoffs are bigger than equilibrium payoffs

$$\mu\varphi_\emptyset^1 - c + \theta pw > (1-p)\mu\varphi_0^0 + p(\theta w - c + \mu\varphi_w^1) - k$$
$$\varphi_\emptyset^1 > \frac{c(1-p) - k}{\mu} + (1-p)\varphi_0^0 + p\varphi_w^1. \tag{A.22}$$

This expression does not depend on θ, so pD1 does not distinguish between types $\theta \geq \theta^*$. It is thus reasonable to assume $\varphi^1_{\emptyset} = \theta^*$, which is lower than φ^1_w. (One can show that types $\theta < \theta^*$ will deviate for a smaller set of off-equilibrium beliefs, so θ^* is the lowest belief that can be supported by pD1).

A.2. Proof of Lemma 1

First, consider the equilibrium in game Γ_I. Define

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

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

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

$$\begin{aligned} Eu(a=1) &= Eu(a=0), \\ \hat{\theta}w - c + \mu\hat{\varphi}^1_w &= \mu\hat{\varphi}^0_w, \\ \hat{\theta} &= \frac{c - \mu\left(\hat{\varphi}^1_w - \hat{\varphi}^0_w\right)}{w}. \end{aligned} \tag{A.25}$$

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).

A.3. Proof of Proposition 2

We can denote the fraction of people who choose prosocially Γ_I by $(1-\varepsilon)(1 - F(\hat{\theta}(\varepsilon)))$. The fraction of people who choose to reveal information in game Γ_C is $\varepsilon + (1-\varepsilon)(1 - F(\theta^*(\varepsilon)))$. Thus, we want to prove that

$$\begin{aligned} (1-\varepsilon)(1 - F(\hat{\theta}(\varepsilon))) &> \varepsilon + (1-\varepsilon)(1 - F\left(\theta^*(\varepsilon)\right)), \\ F\left(\theta^*(\varepsilon)\right) - F\left(\hat{\theta}(\varepsilon)\right) &> \frac{\varepsilon}{1-\varepsilon}. \end{aligned} \tag{A.26}$$

Let us first consider the RHS of (28). The RHS is increasing in ε and is 1 if $\varepsilon = 1/2$. Since $F\left(\theta^*(\varepsilon)\right) - F\left(\hat{\theta}(\varepsilon)\right) \leq 1$, $\varepsilon < 1/2$ is a necessary condition for (28) to be fulfilled.

Let us now consider the LHS of (28). To ease notation, define $\Delta(\varepsilon, p) \equiv \theta^*(\varepsilon, p) - \hat{\theta}(\varepsilon)$. Note that the LHS of (28) is increasing in Δ and approaches 1 if Δ approaches 1. Substituting in the expressions for θ^* from (A.8) and $\hat{\theta}$ from (A.25)

and setting $k = 0$ we find

$$\Delta(\varepsilon, p) = \frac{pc - \mu\left(p\varphi_w^1 + (1-p)\varphi_0^0(\varepsilon) - \varphi_\emptyset^0\right)}{pw} - \frac{c - \mu\left(\hat{\varphi}_w^1 - \hat{\varphi}_w^0(\varepsilon)\right)}{w}$$
$$= \frac{\mu\left(p(\hat{\varphi}_w^1 - \hat{\varphi}_w^0(\varepsilon)) - (p\varphi_w^1 + (1-p)\varphi_0^0(\varepsilon) - \varphi_\emptyset^0)\right)}{pw}. \tag{A.27}$$

Whether or not $\Delta(\varepsilon, p) > 0$ depends on ε and p. First note that if $\varepsilon = 0$ and $p = 1$, the thresholds θ^* and $\hat{\theta}$ coincide, so $\Delta(0, 1) = 0$. Second, by implicit differentiation of equations (A.8) and (A.25), we can derive

$$\frac{d\Delta(\varepsilon, p)}{d\varepsilon} = \frac{d\theta^*}{d\varepsilon} - \frac{d\hat{\theta}}{d\varepsilon} = \left(\frac{-\mu(1-p)\frac{d\varphi_0^0}{d\varepsilon}}{pw + \mu\left(p\frac{d\varphi_w^1}{d\theta^*} + (1-p)\frac{d\varphi_0^0}{d\theta^*} - \frac{d\varphi_\emptyset^0}{d\theta^*}\right)}\right)$$
$$- \left(\frac{\mu\frac{d\hat{\varphi}_w^0}{d\varepsilon}}{w + \mu\left(\frac{d\hat{\varphi}_w^1}{d\theta^*} - \frac{d\hat{\varphi}_w^0}{d\theta^*}\right)}\right) > 0 \tag{A.28}$$

since the denominators of both terms on the RHS are positive, as is guaranteed by the uniqueness conditions for θ^* and $\hat{\theta}$, and $d\varphi_0^0/d\varepsilon$, $d\hat{\varphi}_w^0/d\varepsilon < 0$. Since $\Delta(0, 1) = 0$, $d\Delta(\varepsilon, p)/d\varepsilon > 0$ implies that $\Delta(\varepsilon, p) > 0$ for $\varepsilon > 0$. Moreover, since ε is bounded away from 0, and Δ is continuous in p, we can find a $\bar{p}$ bounded below 1 such that $\Delta(\varepsilon, p) > 0$ for $p > \bar{p}$.

Thus, if $\varepsilon > 0$ and p is high enough, then $\Delta(\varepsilon, p) > 0$. From (A.27) and the fact that $p, w, \mu > 0$, it follows that $\Delta(\varepsilon, p) > 0$ implies that $d\Delta/d\mu > 0$. (Note that for this conclusion, the stability conditions for the thresholds θ^* and $\hat{\theta}$, given by equation (2) and discussed for $\hat{\theta}$ in Section 3.1, respectively, need to hold.) Thus, $p > \bar{p}$ implies that $\Delta(\varepsilon, p) > 0$ and $\Delta_\mu > 0$, which proves our proposition.

A.4. Proof of Proposition 3

With respect to game Γ_I, we know from the first part of the proof of Proposition 1 that there is a unique cutoff 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 Γ_C, if $k > 0$, all *homo economicus* chooses to be ignorant and only altruistic agents select information in the ignorance equilibrium of game Γ_C. This implies that the fraction behaving prosocially conditional on observing σ_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 ε is small enough.

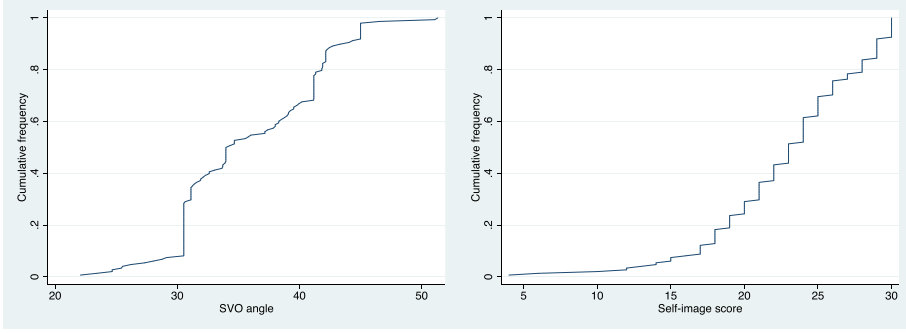
A.5. Proof of Proposition 4

Please see the argument in the main text.

Appendix B: Descriptive Statistics

TABLE B.1. Percent of dictators choosing ignorance and percent of selfish (*A*) choices by information state (sample size in parentheses).

Treatment	Chose ignorance	Ignorant	Chose A AIG	CIG
Hidden Information	60 (120)	85 (72)	96 (24)	17 (24)
CIG Only	–	–	–	35 (26)
Reveal Ex post	26 (35)	–	91 (35)	54 (35)
Reveal Bonus	46 (35)	88 (16)	100 (11)	25 (8)



(a) SVO score (b) Self-image rating

FIGURE B.1. Cumulative distribution functions for the two attribute measures used to investigate sorting.

References

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

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

Andreoni, James, Justin M. Rao and Hannah Trachtman (2016). "Avoiding the Ask: A Field Experiment on Altruism, Empathy, and Charitable Giving". *Journal of Political Economy*, in press.

Aquino, Karl and Americus Reed (2002). "The Self-Importance of Moral Identity". *Journal of Personality and Social Psychology*, 83, 1423–1440.

Bandura, Albert, Claudio Barbaranelli, Gian Vittorio Caprara and Concetta Pastorelli (1996). "Mechanisms of Moral Disengagement in the Exercise of Moral Agency". *Journal of Personality and Social Psychology*, 71, 364–374.

Banks, Jeffrey S. and Joel Sobel (1987). "Equilibrium Selection in Signaling Games". *Econometrica*, 55, 647–661.

Bartling, Björn, Florian Engl and Roberto A. Weber (2014). "Does Willful Ignorance Deflect Punishment? An Experimental Study". *European Economic Review*, 70, 512–524.

Baumeister, Roy (1998). "The Self." In *The Handbook of Social Psychology*, edited by D. Gilbert, S. Fiske and G. Lindzsey McGraw-Hill.

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

Bénabou, Roland (2013). "Groupthink: Collective Delusions in Organizations and Markets". *The Review of Economic Studies*, 80, 429–462.

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

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

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

Broberg, Tomas, Tore Ellingsen and Magnus Johannesson (2007). "Is Generosity Involuntary?" *Economic Letters*, 94, 32–37.

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

Cain, Daylian and Jason Dana (2012). "Paying People to Look at theConsequences of Their Actions." Working Paper.

Campbell, Troy H. and Aaron C. Kay (2014). "Solution Aversion: On the Relation between Ideology and Motivated Disbelief." *Journal of Personality and Social Pscyhology*, 107, 809–824.

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

Charness, Gary and Matthew Rabin (2002). "Understanding Social Preferences with Simple Tests." *Quarterly Journalof Economics*, 117(3), 817–869.

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

Conrads, Julian and Bernd Irlenbusch (2013). "Strategic Ignorance in Ultimatum Bargaining". *Journal of Economic Behavior & Organization*, 92, 104–115.

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

Dana, Jason, Daylian M. Cain and Robyn M. Dawes (2006). "What You Dont Know Wont Hurt Me: Costly (B Quiet) Exit in Dictator Games". *Organizational Behavior and Human Decision Processes*, 100, 193–201.

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

Domingues-Martinez, Silvia, Randolph Sloof and Ferdinand von Siemens (2014). "Monitoring Your Friends, Not your Foes: Strategic Ignorance and the Delegation of Real Authority". *Games and Economic Behavior*, 85, 289–305.

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

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

Ellingsen, Tore and Magnus Johannesson (2008). "Pride and Prejudice: The Human Side of Incentive Theory." *American Economic Review*, 98(3), 990–1008.

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

Fehr, Ernst and Klaus M. Schmidt (1999). "A Theory of Fairness, Competition, and Cooperation". *The Quarterly Journal of Economics*, 114, 817–868.

Feiler, Lauren (2014). "Testing Models of Information Avoidance with Binary Choice Dictator Games". *Journal of Economic Psychology*, 45(12), 253–267.

Fischbacher, Urs (2007). "z-Tree: Zurich Toolbox for Ready-made Economic Experiments". *Experimental Economics*, 10, 171–178.
Fischbacher, Urs, Franziska and Föllmi-Heusi (2013). "Lies in Disguise—An experimental study on cheating". *Journal of the European Economic Association*, 11, 525–547.
Fischbacher, Urs, Simon, Gächter and Ernst Fehr (2001). "Are People Conditionally Cooperative". *Economics Letters*, 71, 397–404.
Fiske, Susan (2013). *Social Beings: A Core Motives Approach to Social Psychology*. Wiley, New York.
Fong, Christina, Felix and Oberholzer-Gee (2011). "Truth in Giving: Experimental Evidence on the Welfare Effects of Informed Giving to the Poor". *Journal of Public Economics*, 95, 436–444.
Gneezy, Ayelet, Uri Gneezy, Gerhard Riener and Leif D. Nelson (2012). "Pay-what-you-want, identity, and self-signaling in markets". *Proceedings of the National Academy of Sciences USA*, 109, 7236–7240.
Greiner, Ben (2003). "An Online Recruitment System for Economic Experiments". *Forschung und wissenschaftliches Rechnen*, 63, 79–93.
Grossman, Zachary (2014). "Strategic Ignorance and the Robustness of Social Preferences". *Management Science*, 60, 2659–2665.
Grossman, Zachary (2015). "Self-Signaling and Social-Signaling in Giving". *Journal of Economic Behavior & Organization*, 117, 26–39.
Grossman, Zachary and David Owens (2012). "An Unlucky Feeling: Overconfidence and Noisy Feedback". *Journal of Economic Behavior & Organization*, 84, 510–524.
Heffernan, Margaret (2012). *Willful Blindness: Why We Ignore the Obvious at Our Peril*. Walker Books.
Hobson, Kersty and Simon Niemeyer (2013). "What Sceptics Believe: The Effects of Information and Deliberation on Climate Change Scepticism." *Public Understanding of Science*, 22, 396–412.
IPCC, (2007). "Climate Change 2007: Synthesis Report. Contribution of Working Groups I, II and III to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change." Tech. rep., edited by R.K. Pachauri and A. Reisinger. IPCC, Geneva.
Kajackaite, Agne (2015). "If I Close My Eyes, Nobody will Get Hurt. The Effect of Ignorance on Performance in a Real Effort Experiment". *Journal of Economic Behavior and Organization*, 116, 518–524.
Kolpin, Van (1992). "Equilibrium Refinement in Psychological Games". *Games and Economic Behavior*, 4, 218–231.
Konow, James (2000). "Fair Shares: Accountability and Cognitive Dissonance in Allocation Decisions". *American Economic Review*, 90(4), 1072–1091.
Köszegi, Botond (1996). "Ego Utility, Overconfidence, and Task Choice". *Journal of the European Economic Association*, 4, 673–707.
Krupka, Erin L. and Roberto A. Weber (2013). "Identifying Social Norms Using Coordination Games. Why Does Dictator Game Sharing Vary?" *Journal of the European Economic Association*, 11, 495–524.
Kunda, Ziva (1990). "The Case for Motivated Reasoning". *Psychological Bulletin*, 108, 480–498.
Kurzban, Robert and Daniel Houser (2005). "Experiments Investigating Cooperative Types in Humans: A Complement to Evolutionary Theory and Simulations". *Proceedings of the National Academy of SciencesUSA*, 102, 1803–1807.
Larson, Tara and Monica C. Capra (2009). "Exploiting Moral Wiggle Room: Illusory Preference for Fairness? A Comment". *Judgment and Decision Making*, 4, 467–474.
Lazear, Edward P., Ulrike Malmendier and Roberto A. Weber (2012). "Sorting in Experiments with Application to Social Preferences". *American Economic Journal: Applied Economics*, 4, 136–163.
Matthey, Astrid and Tobias Regner (2011). "Do I Really Want to Know? A Cognitive Dissonance-Based Explanation of Other-Regarding Behavior". *Games*, 2, 114–135.
Mazar, Nina, On Amir and Dan Ariely (2009). "The Dishonesty of Honest People: A Theory of Self-Concept Maintenance". *Journal of Marketing Research*, XLV, 633–644.

McGoey, Linsey (2012). "Strategic Unknowns: Towards a Sociology of Ignorance". *Economy and Society*, 41, 1–16.
Möbius, Markus, Muriel Niederle, Paul Niehaus and Tanya Rosenblat (2011). "Managing Self-Confidence: Theory and Experimental Evidence." Mimeo. Stanford.
Murnighan, Keith, John M. Oesch and Madan Pillutla (2001). "Player Types and Self-Impression Management in Dictatorship Games: Two Experiments". *Games and Economic Behavior*, 37, 388–414.
Murphy, Ryan O., Kurt A. Ackermann and Michel J.J. Handgraaf (2011). "Measuring Social Value Orientation". *Judgment and Decision Making*, 6, 771–781.
Niehaus, Paul (2014). "A Theory of Good Intentions." Mimeo. UC, San Diego.
Norgaard, Kari Marie (2006a). "People to Protect Themselves a Little Bit: Emotions, Denial, and Social Movement Nonparticipation." *Sociological Inquiry*, 76, 372–396.
Norgaard, Kari Marie (2006b). "We Don't Really Want to Know. Environmental Justice and Socially Organized Denial of Global Warming in Norway." *Organization and Environment*, 19, 347–370.
Nyborg, Karine (2011). "I Don't Want to Hear About It: Rational Ignorance among Duty-Oriented Consumers". *Journal of Economic Behavior and Organization*, 79, 263–274.
Offerman, Theo, Joep Sonnemans and Arthur Schram (1996). "Value Orientations, Expectations and Voluntary Contributions in Public Goods". *The Economic Journal*, 106, 817–845.
Osberghaus, Daniel, Elyssa Finkel and Max Pohl (2010). "Individual Adaptation to Climate Change: The Role of Informationand Perceived Risk." ZEW - Centre for European Economic Research Discussion Paper No. 10-061. Available at SSRN: https://ssrn.com/abstract=1674840 or http://dx.doi.org/10.2139/ssrn.1674840.
Posner, Eric A. (2000). *Law and Social Norms*. Harvard University Press, Cambridge, MA.
Schwartz, Stephan A. (2012). "Climate Change and Willful Ignorance". *Explore*, 8, 268–70.
Spiekermann, Kai and Arne Weiss (2016). "Objective and Subjective Compliance: A Norm-Based Explanation of 'Moral Wiggle Room"'. *Games and Economic Behavior*, 96(3), 170–183.
Stoll-Kleemann, Susanne, Tim, O'Riordan and Carlo C. Jaeger (2001). "The Psychology of Denial Concerning Climate Mitigation Measures: Evidence from Swiss Focus Groups". *Global Environmental Change*, 11, 107–117.
Sweeny, Kate, Darya Melnik, Wendi Miller and James Shepperd (2010). "Information Avoidance: Who, What, When and Why". *Review of General Psychology*, 14, 340–353.
Tadelis, Steve (2011). "The Power of Shame and the Rationality of Trust." Mimeo. Haas School of Business, Berkeley.
Van der Weele, Joël J. (2012). "When Ignorance is Innocence: On Information Avoidance in Moral Dilemmas." Mimeo, Goethe University Frankfurt.
Van der Weele, Joël J. (2014). "Inconvenient Truths: Determinants of Strategic Ignorance in Moral Dilemmas." Mimeo, University of Amsterdam, available at SSRN: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2247288.
Van der Weele, Joël J. and Ferdinand Von Siemens (2014). "Bracelets of Pride and Guilt? An Experimental Test of Self-signaling in Charitable Giving." CESIfo Working Paper 4674. Munich.
Zerubavel, Eviatar (2007). *The Elephant in the Room: Silence and Denial in Everyday Life*. Oxford University Press.

Supplementary data

Supplementary data are available at *JEEA* online.