Moral wiggle room reverted: Information avoidance is myopic

Homayoon Moradi*
Alexander Nesterov†

Job Market Paper
For the latest version, please visit
sites.google.com/site/moradiecon/

December 9, 2017

Abstract

We use a range of dictator game experiments to investigate why people avoid information. Dictators in our experiment know their own payoffs and can choose whether to learn the payoffs of the recipient. We vary whether dictators can learn the recipient's payoff before or after they are presented with their self-interested action. We find that dictators are more likely to avoid information when they do not yet know their self-interested action, and consequently act more selfishly in this case. These results go against two popular explanations of information avoidance: *self-image* and *default effects*. We study and test alternative explanations such as *wishful thinking, cognitive dissonance,* and *attention* and find support for the latter.
(JEL codes: C91, D64, D83, D01)
Keywords: Attention, Wishful Thinking, Self-Image, Default Effect, Information Avoidance, Moral Wiggle Room.

*Corresponding author, WZB Berlin Social Science Center, Reichpietschufer 50, 10785 Berlin, Germany; e-mail: moradi@wzb.eu.
†Higher School of Economics, Kantemirovskaya ul. 3, 194100 St.Petersburg, Russia; e-mail: asnesterov@hse.ru.

1 Introduction

Information avoidance is widespread and leads to reduced responsibility in settings such as corruption (Dana 2006; Simon 2005), the spread of disease (Sullivan et al. 2004), environmental pollution (Rayner 2012), and even atrocities (Cohen 2001). For instance, in corporate scandals, ranging from Bernie Madoff's Ponzi scheme to corruption at FIFA (the international governing body of soccer), it is difficult to imagine how so many people could have failed to notice the unethical behavior (Bazerman and Sezer 2016).

More structured evidence of information avoidance in a social context comes from the field of experimental economics. Participants in experiments on social preferences frequently sacrifice some of their own monetary gains when they know this action will help an anonymous recipient. But when given the choice to obtain the information about the consequences of their actions, they opt to avoid the information and choose the self-interested action (e.g., Dana et al. 2007; Ehrich and Irwin 2005). This *information avoidance* behavior is inconsistent with other-regarding distributional preferences.

Following the seminal experiment in Dana et al. (2007) (hereafter, DWK) most of the studies related to information avoidance in such pro-social settings associate the main reason behind information avoidance with self-image (described below). However, various motives can explain the information avoidance behavior:

Self-image. People are not necessarily inherently altruistic, yet they like to appear so, not only to others but also to themselves (Dana et al. 2007; Grossman and van der Weele 2017. See also Broberg et al. 2007; Dana et al. 2006; Lazear et al. 2012).

Default effect. People avoid information since there is a psychological cost for changing the default of not receiving information (Grossman 2014; Larson and Capra 2009).

Wishful thinking: People may use information avoidance to wishfully believe that their self-interested action is not so likely to have adverse consequences for others. Thus, their perceived probability of the adverse consequences of their self-interested actions is less than the actual probability, which leads to a higher expected payoff when avoiding the information instead of revealing the information. (Feiler 2014 based on Rabin 1995. See Brunnermeier et al. 2004 for the theoretical model of wishful thinking, and see Sharot 2011).

Attention. People are more myopic and concerned with the initial information that they see, which is the information regarding their own payoff, and are thus less concerned with respect to the payoff of others (Golman and Loewenstein 2016; Karlsson et al. 2009; Tasoff and Madarasz 2009).

Cognitive dissonance. People dislike being exposed to information that might conflict with their existing beliefs. Participants may decide to choose the self-

interested action when they are informed about it initially. Thus, they dislike receiving information that might conflict with their existing information (Matthey and Regner 2011; Konow 2000; Akerlof and Dickens 1982 for the theoretical model of cognitive dissonance).[1]

Empirical evidence of information avoidance has been extensively studied in the literature. However, while these studied are useful in providing evidence of information avoidance, they are less so when it comes to distinguishing motives behind this avoidance. The objective of this paper is to do precisely this: to put different motives to a test. To this end, we develop a model that combines these motives and we test predictions of motives with an experiment. Specifically, the two following questions help us to distinguish between different motives:

1. Are people strategic in information avoidance?
2. Does the level of uncertainty impact information avoidance?

We test different motives for information avoidance by using and slightly modifying the hidden-information dictator game of DWK. The *hidden-information dictator game* is a two-player game in which a dictator has a binary choice between two actions where one of them leads to a higher payoff than the other, but the consequences for the recipient are unknown. There are two possible–and equally likely–states of the world. In the conflicting case, the self-interested action leads to a lower payoff for the recipient, and in the non-conflicting case to a higher payoff. All the dictators need to do is click a button in order to find out if the case is conflicting or non-conflicting. DWK demonstrate that many dictators avoid the information and, as a result, the share of dictators who choose the self-interested action is significantly higher than in the baseline dictator game with full information.

To answer question 1, we modify the hidden-information dictator game by swapping the order of learning the self-interested action and the choice of whether to obtain information about the recipient's payoff or not. The dictator knows her own payoffs initially and has a chance to avoid information about the recipient payoff in both treatments. The fundamental experimental manipulation changes the decision-making timeline. In the original setting of DWK, the dictator is told about her self-interested action *before* she is presented with the information choice (**the Before treatment**). In contrast to this timeline, in the other setting she is told about her self-interested action *after* she has been presented with the information choice (**the After treatment**).

Table 1 displays the summary of the predictions of each motive. The self-image motive predicts that both the Before and After treatments will have the same rate of information avoidance. The reason is that the relevant information and the strategy set in both treatments are the same: the dictator initially knows her

[1] Although cognitive dissonance starting from Festinger (1957) covers a broad range of behavior which can include self-image as well, here we focus on the stated definition. See Abelson et al. (1968) for a wide-ranging volume taking stock of research on Festinger's theory.

payoffs and can remain ignorant about the recipient's payoff in both treatments. Therefore, she can protect her self-image by avoiding the information in both treatments. The default effect motive has the same predictions since in both treatments the default option is to avoid the information.

However, if either wishful thinking, attention, or cognitive dissonance are factors behind information avoidance, we should expect a difference between the Before and After treatments. For these three motives, prior knowledge of the self-interested action in the Before treatment makes the information avoidance choice more attractive than in the After treatment. This leads to lower rates of information avoidance and selfish choices in the After treatment than in the Before treatment. A wishful thinking dictator may underestimate the probability of the conflicting game and choose to avoid information. Underestimating the probability of the conflicting game increases the expected payoff of a wishful thinking dictator with other-regarding preferences. In this way, wishful thinking in information avoidance leads to a higher expected utility than obtaining information on the actual recipient's payoff with no wishful thinking.

A dictator may pay more attention to the initial information. She may focus more on her prior information about her self-interested action in the Before treatment and she may focus more on the prior information about the recipient's payoff in the After treatment. She is then more likely to avoid the information on the recipient's payoff in the Before treatment than in the After treatment.

A dictator with cognitive dissonance avoidance might not be keen to obtain the information that may conflict with her existing belief. Prior knowledge of the self-interested action in the Before treatment may prompt the dictator to choose the self-interested action. She may then prefer not to obtain the information on the recipient's payoff as this might conflict with her decision.

Our first hypothesis is that the rate of information avoidance and selfish choices do not differ significantly in the Before and After treatments. If the self-image or default effect is a factor behind information avoidance, we should not expect the first hypothesis to be rejected. If any of the remaining motives, wishful thinking, cognitive dissonance, or attention are factors behind information avoidance then we should expect a rejection of the first hypothesis.

Our results indicate that the information avoidance choice only leads to a significant increase in selfish choices in the Before treatment (where dictators know about their self-interested action when deciding whether to obtain information or not). In the Before treatment dictators avoid the information in over 34% of cases and choose the selfish option in over 58% of cases, almost double the rates in the After treatment (where dictators decide whether to obtain information or not without knowing their self-interested action): 16% of information avoidance choice and 34% of selfish choices. The latter 34% of selfish choices in the After treatment do not significantly differ from 29% in the dictator game with full information (**the Conflict treatment**).

Question 2 exposes the role of the uncertainty of adverse consequences in information avoidance. To further test different motives behind information avoid-

ance, the next experimental variation minimizes the uncertainty compared to 50% of cases with moral wiggle room, while still giving the chance for information avoidance. We replace the 50% probability of a conflicting case in the original setting of DWK with two extremes: a Before treatment with a 99% probability of conflicting payoffs (**Conflict-Before treatment**) and a Before treatment with a 1% conflicting payoff probability (**No-Conflict-Before treatment**).

As the probability of conflicting payoffs increases, different motives vary in their predictions of the changes in the rate of information avoidance choices. The self-image motive predicts the same rate of information avoidance for both Conflict-Before and No-Conflict-Before treatments. A dictator with self-image concern would like to not appear selfish, which is independent of changes in the underlying probabilities. The attention motive also predicts the same rate for both Conflict-Before and No-Conflict-Before treatments since it assumed to be entirely exogenous and do not depend on the probability of conflicting payoffs. The cognitive dissonance, on the other hand, foresees that the higher the probability of conflicting payoffs, the higher the rate of information avoidance. As the probability of conflicting payoffs increases, the chances that the dictator will face information that conflicts with existing information increases. This leads to a higher rate of information avoidance in the Conflict-Before treatment compare to the No-Conflict-Before treatment.

The default-effect and wishful thinking motive predict a lower rate of information avoidance in the Conflict-Before treatment compared to the No-Conflict-Before treatment. As the probability of conflict increases, the default costs decreases as the stakes become larger. Thus, the rate of information avoidance decreases. Wishful thinking, similarly, predicts that the higher the probability of conflict, the lower the rate of information avoidance. As the probability of conflicting payoffs increases, the expected cost of the required wishful thinking increases (as it becomes harder to turn a blind eye to the conflict), which leads to a lower rate of information avoidance. Therefore, wishful thinking predicts a higher rate of information avoidance in the No-Conflict-Before treatment than in the Conflict-Before treatment.

Our second hypothesis is that there is no significant difference between the Conflict-Before and No-Conflict-Before treatments. If the self-image or attention is a factor behind information avoidance, we should not expect a rejection of the second hypothesis. If any of the remaining motives of default-effect, wishful thinking, or cognitive dissonance, are factors behind information avoidance, we should expect a rejection of the second hypothesis.

We find that as the level of uncertainty changes, the rate of information avoidance does not significantly change, consistent with the attention and default effect model. For all probabilities of conflicting payoffs of 1%, 50%, and 99%, the rate of information avoidance does not differ significantly from approximately one-third (34%, 34%, and 38%, respectively).

We observe that the predictions of the attention motive is the closest to the ex-

Treatment	Selfish choices Before vs. After	Information avoidance choices in Before vs. After	No-Conflict-Before vs. Conflict-Before
Self-image	$=$	$=$	$=$
Default effect	$=$	$=$	$>$
Wishful thinking	$>$	$>$	$>$
Cognitive dissonance	$>$	$>$	$<$
Attention	$>$	$>$	$=$
Result	$>^{*}$	$>^{**}$	$=$
$p-value_{\chi^2(1)}$	0.04	0.008	0.7

Table 1: Summary of predictions of different motives of information avoidance and the results from the experiment.
Notes: For example, the third to the last row of the table means: "The attention motive predicts that the rate of selfish choices in the Before treatment is greater than in the After treatment."

periment results. This result suggests that information avoidance behavior may not be driven by self-image but rather by the attention. Instead of strategically protecting the self-image, they may be myopically reacting to the information that grabs their attention. Put differently, even if self-image concerns play a role in information avoidance, they can be neutralized by making the situation more complex. Subjects are not strategic in the protection of their self-image.

The paper is structured as follows: The next section presents the experimental design and spells out the hypotheses. Section 3 provides a theoretical framework with testable predictions for each motive to guide our understanding of information avoidance. Section 4 presents the experimental result, and section 5 concludes.

2 The experiment

2.1 Design

The experiment has five treatments, displayed in figure 2.1. The first two exactly replicate the DWK experiment, while three additional treatments feature variations of the timeline and the underlying probability. The After treatment tests the robustness of DWK's information avoidance result when dictators do not know their self-interested action by varying the decision-making timeline in the Before treatment. The Conflict-Before and No-Conflict-Before treatments are designed to further compare the predictions of the different motives for information avoidance by changing the underlying probability of conflicting payoffs to two extremes: highest and lowest probability.

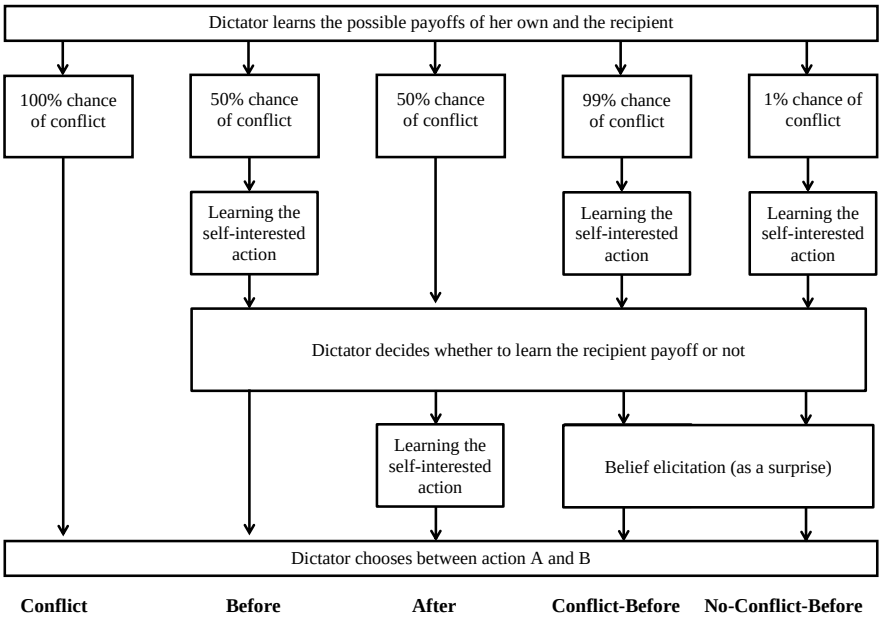

Figure 2.1: Overview of the Experimental Treatments

Player X's choices			
	A	X:6	Y:1
	B	X:5	Y:5

Figure 2.2: Conflict treatment's payoff table

1. *Conflict.* This treatment is an exact replication of the DWK's baseline treatment. The Conflict payoffs, as shown to the participants, are presented in figure 2.2. Dictators chose between A and B by clicking on two of the letters. If the dictator (Player X) choose option A, she receives €6 and the receiver (Player Y) receives €1. If the dictator choose option B, both players receive €5 In the Conflict treatment, the relationship between actions and outcomes is transparent. While the dictators were making their choices, recipients were asked to make the same decision hypothetically, as if they were a dictator.

Player X's choices			
	A	X:6	Y:?
	B	X:5	Y:?

Left			Right		
A	X:6	Y:1	A	X:6	Y:5
B	X:5	Y:5	B	X:5	Y:1

Figure 2.3: Before treatment's payoff table

2. *Before.* This treatment replicated the "hidden information" treatment of DWK. Participants were presented with two versions of the game as displayed in figure 2.3. In both games, if the dictator chose A, she would receive her highest payoff and if she chose B, she would receive a lower amount - action A is her self-interested action. The recipient's payoffs from these actions were uncertain. In game "Left" (with conflicting payoffs as in the Conflict treatment), the recipient would receive his lowest payoff if the dictator chose the action A, and a smaller amount if the dictator chose B. In the game "Right" (with non-conflicting payoffs), the recipient would get the opposite payoffs. Participants were told that the actual game they were playing was randomly selected with equal probability. The dictator could reveal which game was playing by clicking a "Reveal" button that did not entail any costs. Participants were informed that the dictator's decision as to whether to reveal would not be disclosed to the recipient. In the Before treatment, the dictator is presented with her self-interested action, action A, *before* she is presented with the choice of learning of the recipient's actual payoff.

Player X's choices		
	A	Y:?
	B	Y:?

Top Left		
A	X:6	Y:1
B	X:5	Y:5

Top Right		
A	X:6	Y:5
B	X:5	Y:1

Bottom Left		
A	X:5	Y:1
B	X:6	Y:5

Bottom Right		
A	X:5	Y:5
B	X:6	Y:1

Figure 2.4: After treatment's payoff table

3. *After.* This treatment differed from the Before treatment in that the order of observing the self-interested action and the information choice were reversed. The participants were presented with four versions of the game as displayed in the figure 2.4. Two games, called top games, are identical to the games in the Before treatment. In the top games, the self-interested action is action A. The other two games, called bottom games, only differ from top games in that the self-interested action is B. Participants were told that the actual game they were playing had been randomly selected with equal probability. Whether the game would be in the left or right column (i.e. what the recipient's payoff is) was never revealed publicly. But the dictator could either reveal this by clicking a "Reveal" button or could decide not to reveal it by clicking "Continue." After the dictator makes her revelation decision (and independent of her choice), she knows whether she is playing a game at the top or the bottom row i.e, she knows her own payoff from each action. Conditional on the revelation choice, participants observed the same payoff matrix in both treatments, as shown in the example in figure 2.5. In the After treatment, the dictator is presented with information regarding the self-interested action only *after* she has been

presented with the information choice.

(Revealed)			(Not revealed)	
A	X:6	Y:1	A	X:6
B	X:5	Y:5	B	X:5

Figure 2.5: Example of shown screen base on Revelation choice in all treatments

4. *Conflict-Before.* This treatment differed from the Before treatment in that the probability of the game (with conflicting payoffs as in the Conflict treatment) being played was 99 percent instead of 50 percent. After participants made the revelation choice, they were asked to elicit their beliefs using the binarized scoring rule on three matters to earn 1€ for each: 1) the likelihood of their actual game being the conflict game, 2) the likelihood of another dictator's actual game being the conflict game, 3) and the likelihood of another dictator choosing to reveal and choosing action B (the altruistic action).

5. *No-Conflict-Before.* This treatment differed from the Conflict-Before only in that the probability of conflicting payoffs was 1 percent instead of 50 percent.

Although we elicited participants' beliefs in treatments with extreme underlying probabilities, we did not do so in the Before and After treatments. Instead, we kept the design as close as possible to the moral wiggle room design of DWK for comparison reasons. In treatments with extreme underlying probabilities, however, we were only interested in information avoidance choices and these choices remained untouched by belief elicitation.

To maintain the possibility of subjects not knowing their self-interested action at the time of the information choice, we did the following. In the After treatment, dictators needed to click on a "Continue" button to choose information avoidance whereas in the Before treatment they did not need to click any button. However, Grossman (2014) found that such deviation does not produce significantly different results along several key measures.

The extreme probabilities in the Conflict-Before and No-Conflict-Before almost eradicate the uncertainty in the Before treatment while still keeping the possibility of information avoidance. While other studies have looked at some more intermediate level (van der Weele 2012 and Feiler 2014 look at an 80 percent probability of conflicting payoffs of 80 percent), the extreme probabilities allows us to have a pure test of impact of the level of uncertainty on information avoidance. Moreover, extreme probabilities help to test the wishful thinking and the self-image model in a more fitting environment. The wishful thinking model predicts a higher changes of information avoidance as the probability of conflicting payoffs decreases than when it increases. Feiler (2014) looks at the low probability of conflicting payoffs of 20 percent, but Grossman and van der Weele (2017) argue that the within-subject environment of Feiler (2014) with multiple periods deviates from the environment of the self-image model. We initially ran three sessions with a 25 percent probability of conflict in a one-shot

between-subjects setting and find no significant difference. Next, we decided to look at the extreme case to provide a fair chance for both the self-image model and also wishful thinking.

2.2 Procedures

The experiment took place at the Experimental Laboratory at the Technical University of Berlin from July 2016 to August 2017. Randomization across the three treatments occurred at the participant level using ORSEE (Online Recruitment System for Economic Experiments; Greiner (2015), which excludes those who had previously participated in an experiment related to charitable giving. The sessions were one-shot, between-subjects, gender balanced, with at least 16 participants present. Most of the participants were undergraduate and master's students from the Technical University of Berlin. The interface was programmed using the z-Tree software package (Fischbacher 2007). Experimental instructions are provided in the Appendix.

Participants were instructed that they would be playing a game with another person in the room with whom they had been randomly and anonymously matched. Upon arriving at the experiment, participants sat at computer terminals, and the instructions were read aloud. After participants had been told which role they had been assigned, they were allowed to make a (for the recipients, hypothetical) choice. Unless otherwise noted, 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 treatment-specific instructions in the replication treatments.

Participants completed a brief quiz to make sure they understood the instructions. To be certain that participants fully understood that the "Reveal" button depicted both players' payoffs and that the "Continue" button only showed the own payoff in the After treatment, the last two questions of the quiz focused on these two buttons. The quiz was administered just before the start of the experiment, so participants were unlikely to forget. The answer were read aloud; they were then asked whether they had any doubts or questions.

We conducted 33 sessions. A total of 736 students participated across the five treatments with exactly half (368) playing the role of a dictator (Player X). On average, participants earned €10.70, including a €5 show-up fee and incentive payment for the belief elicitation. Sessions lasted approximately 20 minutes.

3 Model

This is a model of decision-making that incorporates *wishful thinking*, *self-image, attention,* and *default effect*.[2] We are interested in the behavior of the dictator

[2]As mentioned in the introduction, attention is one of possible explanations for framing differences in Before and After treatments.

(hereinafter referred to as the agent) in a binary dictator game with incomplete information and voluntary revelation (Before, After, and Conflict treatments). The agent might be uncertain about the consequences of her actions on the receiver and might want to distort her prior beliefs (wishful thinking). At the same time the agent might care about the signal she is sending via her action: whether her action makes her look like a "bad" type or not (self-image). If the agent gets to see her payoffs first, she might be conditioned to pay more attention to her payoffs than to the receiver's payoffs (attention). In addition, one of the actions is a default action and choosing any other action is associated with a cost (default effect).

The agent is endowed with level of altruism β distributed according to cdf $F(\beta)$ on $[0,1]$ with a probability mass on zero $F(\beta = 0) = \varepsilon$. We call these $\beta = 0$ types **homo economicus** and assume complete rationality for them.

The agent faces a choice between two actions A and B. She knows her payoffs $X(A) = \overline{X}$ and $X(B) = \underline{X}$ ($\Delta X \equiv \overline{X} - \underline{X} > 0$) but does not know the recipient's payoffs $Y(A) \in \{\underline{Y}, \overline{Y}\}$ and $Y(B) = \{\underline{Y}, \overline{Y}\} \backslash Y(A)$ ($\Delta Y \equiv \overline{Y} - \underline{Y} > 0$).[3] There are two possible states of the world: the payoffs X and Y are either *conflicting* $Y(A) = \underline{Y}$ or *nonconflicting* $Y(A) = \overline{Y}$, the probability of conflict is $\Pr(Y(A) = \underline{Y} | X(A) = \overline{X}) = p$, $p \in [0,1]$.

Before choosing between A and B the agent chooses whether to reveal or not the recipient's payoffs (these actions are denoted by R for "reveal" and N for "do not reveal"). We will denote the agent's strategy as σ, for example, $\sigma_{N,X} = (N, \arg\max_{\{A,B\}} X)$ denotes that the agent first does not reveal and then chooses the action that maximizes her own payoff X, and $\sigma_{R,Y} = (R, \arg\max_{\{A,B\}} Y)$ denotes that the agent first reveals and then chooses the action that maximizes the recipient's payoff Y. The agent's type is her level of altruism denoted by β which is distributed according to cdf $F(\beta)$ on $[0,1]$

We assume that homo economicus type $\beta = 0$ does not have default effect and self-image concerns. We further assume that when she is indifferent between revealing a costless information or not then with probability $\mu \in [0,1]$ she will remain ignorant, and that μ is common knowledge.

The type's β utility under certainty is $U_\beta(\sigma) = X + (1+a)\beta Y - c_w(w) - c_s - c_d$, which has four components described below in detail:

- **Allocative utility** $X + \beta Y$, where X denotes the agent's payoff and Y denotes the recipient's payoff,
- **Attention parameter** a in $(1+a)\beta$ shifts the level of altruism β downwards;
- **Costs from wishful thinking** c_w – the costs from distorting the subjective probability about the state of the world;

[3] In the Conflict and in Before settings action A corresponds to the higher own payoff while action B corresponds to the lower own payoff, in the After setting each case is equally likely.

- **Self-image costs** c_s – the costs from pooling with the "bad" type (homo economicus);
- **Default effect costs** c_d – the costs of taking an action that is different from the default effect;

The allocative utility represents the standard distributional preferences as in Fehr and Schmidt (1999) and Bolton and Ockenfels (2000). The attention parameter $a \in [-1, 0]$ represents priming: if the agent is primed to be more selfish compared to the Conflict setting then $a < 0$, if there is no priming then $a = 0$. We assume that all types β are primed by the same treatment in the same way: for the Before treatments $a < 0$ (regardless the probability of conflict p), for Conflict and After treatments $a = 0$.

The costs from wishful thinking c_w depend on both the prior probability p and the distortion w and are given by the expression $c_w = c_1 p \log \frac{p}{p-w}$ for $w \geq 0$, where $c_1 \in [0, 1]$ is a constant. (We only consider $w \geq 0$ as only this will occur in the equilibrium; for $w < 0$ the costs can be specified symmetrically.) This functional form is chosen to represent these costs due to several reasons described below.

First, at the boundaries the costs satisfy sensible conditions: for $w = 0$ (zero distortion) the costs are zero and for $w = p$ (full distortion) the costs are infinite.[4] Second, the marginal costs $\frac{\partial c_w}{\partial w}$ also satisfy sensible conditions. First, the marginal costs are positive ($\frac{\partial c_w}{\partial w} > 0$) and increasing ($\frac{\partial^2 c_w}{\partial w^2} > 0$) and become infinite when the distortion w approaches the prior p. Second, the marginal costs monotonically decrease in p: ($\frac{\partial^2 c_w}{\partial w \partial p} = -\frac{w}{(p-w)^2} < 0$), which means that the costs of a 1% distortion are higher if the prior probability is low (e.g., distorting the probability from $p = 5\%$ to $(p - w) = 4\%$ is more expensive than distorting from $p = 50\%$ to $(p - w) = 49\%$). Third, the marginal cost at zero distortion $w = 0$ for each prior p is the same and equals $\frac{\partial c_w}{\partial w}|_{w=0} = c_1$. Thus, for each prior belief p it is equally costly to distort the belief by an infinitesimal amount of wishful thinking and these costs are captured by a constant c_1.

The self-image costs c_s of an agent i represent how "bad" the signal that i sends is, by playing strategy σ_i. We assume that the image costs are proportional to (1) $\int_{\sigma_0 = \sigma_i} dF$ – the mass of homo economicus agents that play the same strategy $\sigma_0 = \sigma_i$ and also to (2) $\int_{\sigma \neq \sigma_0} dF$ the mass of agents that play strategies σ different from what any homo economicus would play $\sigma \neq \sigma_0$. The reason behind assuming (1) is clear: the more homo economicus agents there are pooling with i, the worse the inference about i's type is. The reason behind assuming (2) is similar: the more agents there are that succeed in separating from the homo economicus agents (unlike agent i), again the worse the inference is about i's type, since her pooling with homo economicus becomes a more extreme type of behavior. For example, if all homo economicus agents play the same strategy as

[4] The costs of full distortion do not need to be infinite, it is enough that these costs are too high and thus $w = p$ never occurs in an equilibrium.

i does, while all other agents play a different strategy, then the image costs are maximal. Formally, we assume that c_s is proportional to a product of (1) and (2): $c_s(\sigma_i) = s \int_{\sigma_0=\sigma_i} dF \cdot \int_{\sigma\neq\sigma_0} dF$, where $s > 0$ is a non-negative constant.

The default effect costs c_d are the costs of choosing the option alternative to the default effect: not to reveal the information.[5]

In the Before setting the timing is as follows:

- t=0, the agent observes $\beta, a, F, X, \{\underline{Y}, \overline{Y}\}, p, c_d, c_c$ and functions c_w, c_s, and simultaneously chooses either to reveal Y or not, and, in the latter case, also chooses the amount of wishful thinking w,
- t=1, agents that revealed at t=0 observe Y and choose A or B; agents that did not reveal choose A or B based on the distorted probability $p - w$.

Equilibrium Now we show the existence of semi-separating equilibrium in which there is a threshold value β_1 that prescribes the strategy for each type: types below this threshold choose not to reveal (some of them will not engage in wishful thinking for it is too costly, for others their amount of wishful thinking will be increasing in their type) and choose option A, types above the threshold choose to reveal and, in case they discover conflicting payoffs, choose option B.

Proposition 1. *There exists a symmetric equilibrium in pure strategies $\sigma^*(\beta)$ characterized by a cutoff β_1 (such that $0 < \beta_1 \leq 1$): all types $\beta \in (0, \beta_1)$ choose $\sigma^*(\beta) = \sigma_{N,X} \equiv (N, w^*(\beta), A)$ with $w^*(\beta) \geq 0$ monotonically increasing in β, all types $\beta \in (\beta_1, 1]$ choose $\sigma^*(\beta) = \sigma_{R,Y} \equiv (R, \arg\max_{\{A,B\}} Y)$. The cutoff equals*

$$\beta_1 = \frac{1}{(1+a)\Delta Y - s\mu\varepsilon(1-\varepsilon)} \left(pc_1 e^{\frac{p\Delta X + c_d}{c_1 p^2} - 1} - s\mu\varepsilon(1-\varepsilon) \right). \tag{3.1}$$

The proofs of this and further propositions are presented in subsection 6.2 of the Appendix.

The equilibrium σ^* might not be unique since the game of minimizing the image costs c_s has a nature of coordination (or self-fulfilling equilibrium) and the agents might decide to coordinate on a different strategy. Specifically, if each agent with $\beta > 0$ that chooses action A expects that all other such agents will reveal (and thus pool with $(1 - \mu)$ of the homo economicus agents), then this might be an equilibrium. Indeed, in this case deviating from R to N causes a shift in image costs (since nobody with $\beta > 0$ pools with μ of homo economicus agents) that might outweigh the benefits of sticking with the default effect ($-c_d$) and benefits of wishful thinking.

[5] Alternatively, if as in Grossman (2014), the default effect would be to reveal, then the default effect costs denote the costs from not-revealing.

Yet, this can occur only in case no agent including the most altruistic reveals and in conflicting case chooses B, otherwise the equilibrium described above is generally unique.

Remark. If the most altruistic type $\beta = 1$chooses to reveal, then Proposition 1 defines the unique equilibrium.

Next, we determine the agent's behavior in the Conflict and After settings. For both settings we assume no priming and thus the attention shift disappears: $a = 0$.

In the After setting the timing is as follows:

- t=0, the agent observes $\beta, F, \{\underline{X}, \overline{X}\}, \{\underline{Y}, \overline{Y}\}, p$ and the function c_s, and simultaneously chooses either to reveal Y or not and,
- t=1, the agent observes X; agents that revealed at t=0 observe Y and chooses A or B; agents that did not reveal choose A or B based on the prior probability.

Proposition 2. *In the After setting the agent's strategy is determined by her type β:*

$$\begin{cases} \text{if } \beta = 0 & \text{agent chooses } \sigma_{R,X} \text{ or } \sigma_{N,X} \text{ at random }, \\ \text{if } 0 < \beta \leq \beta_2 & \text{agent chooses } \sigma_{N,X}, \\ \text{if } \beta > \beta_2 & \text{agent chooses } \sigma_{R,Y}, \end{cases}$$

where

$$\beta_2 = \frac{\frac{c_d}{p} + \Delta X - s\mu\varepsilon(1-\varepsilon)}{\Delta Y - s\mu\varepsilon(1-\varepsilon)}. \tag{3.2}$$

Next, proposition determines the agent's behavior in the Conflict setting.

Proposition 3. *In the dictator game with a complete information setting the strategy of the agent is determined by her type β:*

$$\begin{cases} \text{if } 0 \leq \beta \leq \beta_0 & \text{agent chooses } A, \\ \text{if } \beta > \beta_0 & \text{agent chooses } B, \end{cases}$$

where

$$\beta_0 = \frac{\Delta X - s\varepsilon(1-\varepsilon)}{\Delta Y - s\varepsilon(1-\varepsilon)}. \tag{3.3}$$

3.1 Model predictions

Based on the results above we formulate predictions regarding the behavior of agents in different treatments. These predictions help us to understand and give

Dominant motive	Assumptions for other motives	Conflict:β_0	Before:β_1	After:β_2
Attention, $a \in (-1,0)$	$w=0, s=0, c_d=0$	$\frac{\Delta X}{\Delta Y}$	$\frac{\Delta X}{(1+a)\Delta Y}$	$\frac{\Delta X}{\Delta Y}$
Wishful thinking	$a=0, s=0, c_d=0$	$\frac{\Delta X}{\Delta Y}$	$\frac{c_1 p}{\Delta Y} e^{\frac{\Delta X}{c_1 p}-1}$	$\frac{\Delta X}{\Delta Y}$
Self-image	$w=0, a=0, c_d=0$	$\frac{\Delta X - s\varepsilon(1-\varepsilon)}{\Delta Y - s\varepsilon(1-\varepsilon)}$	$\frac{\Delta X - \mu s\varepsilon(1-\varepsilon)}{\Delta Y - \mu s\varepsilon(1-\varepsilon)}$	$\frac{\Delta X - \mu s\varepsilon(1-\varepsilon)}{\Delta Y - \mu s\varepsilon(1-\varepsilon)}$
Default effect	$w=0, a=0, s=0,$	$\frac{\Delta X}{\Delta Y}$	$\frac{\Delta X + \frac{c_d}{p}}{\Delta Y}$	$\frac{\Delta X + \frac{c_d}{p}}{\Delta Y}$

Table 2: Summary of theoretical predictions of threshold levels of altruism for different dominant motives behind information avoidance
Notes: the last three columns of the table present the threshold values of the level of altruism for different treatments β_0 (Conflict), β_1 (Before), and β_2 (After) when the dominant motive behind information avoidance is either attention, wishful thinking, self-image, or default effect.

a theoretical support for the hypotheses presented in section 3.1 regarding which of the possible motives behind information avoidance is dominant. To single out the effect of each particular motive that is dominant, we shut down all other factors and compare the results for different treatments. The threshold values $\beta_0, \beta_1\beta_2$ for each of the four motives—attention, wishful thinking, self-image, default effect—are presented in Table 2.

1. Order of information First we look at the classic result in DWK by comparing thresholds β_0 and $\beta_1 (p = 0.5)$. All four factors give an unambiguous prediction of the result in DWK: there are more altruistic choices in the Conflict treatment than in the Before treatment ($\beta_0 < \beta_1$). In the case of attention and default effect the result is straightforward. In the case of self-image the result is given by the fact that $\Delta X < \Delta Y$. For wishful thinking the result comes from that $\frac{c_1 p}{\Delta Y} e^{\frac{\Delta X}{c_1 p}-1}$ is at least as high as $\frac{\Delta X}{\Delta Y}$ (since function $c_1 p e^{\frac{\Delta X}{c_1 p}-1}$ has a local minimum at $c_1 p = \Delta X$, where the difference is minimized to $\beta_1 = \beta_0$ and for all other values of $c_1 p$ the difference is positive).

Prediction 1.0 If dominant, each motive predicts that in the Before treatment the share of altruistic choices is lower than in the Conflict treatment.

Next we formulate the predictions regarding the role of the timing of revelation choice in Before and After by comparing the thresholds β_1 and β_2.

Only two factors – attention and wishful thinking – predict that $\beta_1 > \beta_2$, while self-image and default effect predict $\beta_1 = \beta_2$.

Prediction 1.1 If dominant, the attention and wishful thinking motives predict that in the After treatment the share of revealing subjects and the share of altruistic choices is higher than in the Before treatment.

However, the self-image and default effect motives being dominant predict that the share of revealing subjects and the share of altruistic choices in the After treatment is the same as in the Before treatment.

Finally, we formulate the model predictions regarding the behavior in the After and Conflict treatments as a result of Proposition 2 and Proposition 3. Attention and wishful thinking predict that the behavior in these treatments is the same $\beta_2 = \beta_0$, while the self-image and default effect predict that in After the behavior is more selfish: $\beta_2 > \beta_0$.

Prediction 1.2 If dominant, the attention and wishful thinking motives predict that in the After and the Conflict treatments the shares of altruistic choices are the same.

In contrast, if the self-image and default effect motives were dominant, then the share of altruistic choices is predicted to be lower in the After treatment than in the Conflict treatment.

2. Probability of conflicting payoffs Now we study the role of the probability of conflicting payoffs p by comparing the series of Before treatments for $p = .01$ (Conflict-Before), $p = .5$ (Before) and $p = .99$ (No-Conflict-Before) using the results of Proposition 1.

Attention and self-image predict that the probability of conflicting payoffs does not play a role. Wishful thinking and default effect predict that as the probability of conflicting payoffs p increases, the threshold $\beta_1(p)$ decreases.[6]

Prediction 2.1 If dominant, the wishful thinking and default effect motives predict that in the Conflict-Before treatment the share of revealing subjects is the same as in the No-Conflict-Before treatment.

If, however, the attention and self-image motives were dominant, then the share of revealing subjects is predicted to be higher in the Conflict-Before treatment than in the No-Conflict-Before treatment.

4 Experimental Results

Figure 4.1 displays the rates of information avoidance choices and selfish choices in all treatments. For the rate of information avoidance choices, we consider games with both conflicting and nonconflicting payoffs. For the rate of selfish choices, we only consider games with conflicting payoffs. In No-Conflict-Before treatment, we do not report the rate of selfish choices as there were not enough

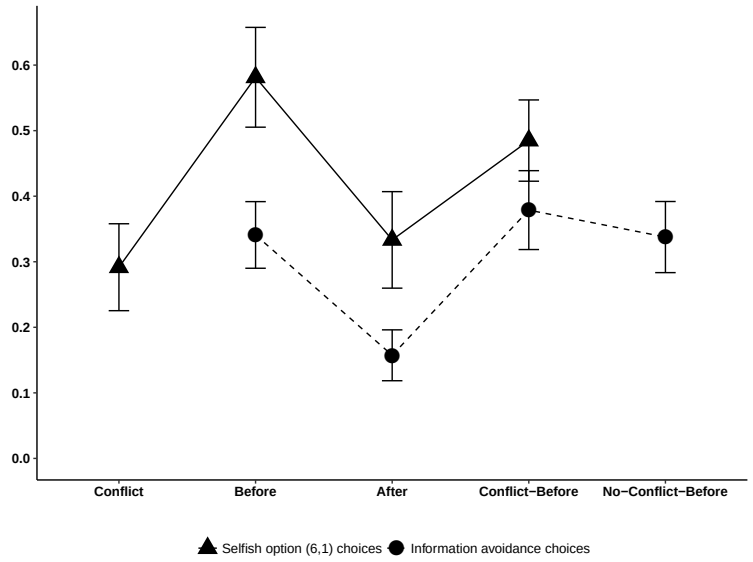

Figure 4.1: Information avoidance and selfish behavior by treatment. "Information avoidance choices" is the fraction of subjects choosing not to "Reveal." "Selfish option (6,1) choices" is the fraction of subjects who played the game with conflicting payoffs and chose the action with the payoff of (6,1) instead of (5,5). (Since conflicting payoffs are very rare in the No-Conflict-Before treatment, the result for this treatment is omitted.) The error bars represent $+/-1$ standard error.

observations (only one) of games with conflicting payoffs. We provide detailed results of the treatments in the Appendix.

Observation 1. Dictators are not strategic in information avoidance. In the Conflict treatment, 29% (14 out of 48) of dictators choose the selfish option. When the self-interested action is presented before the information choice, dictators are significantly more likely to choose the selfish option [$\chi^2(1) = 7$, p = 0.01], consistent with the DWK's result. They choose the selfish option in 58% (25 out of 43) of cases in the Before treatment. However, the rate of selfish choices drops to 33% (14 out of 42) of the cases in the After treatment [$\chi^2(1) = 4$, p = 0.04]. This does not differ significantly from that in the Conflict treatment [$\chi^2(1) = 0.04$, p = 0.8].

Information avoidance choices drive the treatment difference in the selfish behavior between the Before and the After treatment. The rate of dictators choosing information avoidance in the Before treatment is 34% (30 out of 88), which is significantly greater [$\chi^2(1) = 7$, p = 0.008] than the 16% (14 out of 89) rate in the After treatment. Most of those choosing information avoidance (90%) choose the selfish option and there is no significant difference between treatments. Dictators who choose to reveal the recipient's payoff in the Before and After treatment, choose the selfish option at the rate of 39% (11 out of 28) and 25% (9 out of 36), respectively, which does not differ significantly [$\chi^2(1) = 0.9$, p = 0.3]. Thus, there is no significant difference between dictators' selfish choices conditional on their information avoidance choice. Relative to the After treatment, the significantly higher of rate of information avoidance choices in the Before treatment leads to the significantly higher rate of selfish choices.

Hence, the increased selfish choices due to moral wiggle room found by DWK and replicated in the Before treatment depends on whether participants know their self-interested action. Relative to the Conflict treatment, the rate of selfish choices in the After treatment changes slightly, from 29% to 33%. The information avoidance choice only leads to a significant increase in selfish choices when it is presented after the announcement of the self-interested action.

Observation 2. As the probability of conflicting payoffs increases, the rate of information avoidance does not change significantly; it remains at approximately one-third. The rate of dictators choosing information avoidance is 38% (25 out of 66) in the Conflict-Before. In the No-Conflict-Before treatment, the rate is 34% (27 out of 77), which does not differ significantly [$\chi^2(1) = 0.1$, p = 0.7]. The rate of information avoidance choices in the Before treatment is 34% (30 out of 88). This rate does not differ significantly from the rates in both the Conflict-Before and the No-Conflict-Before treatments, [$\chi^2(1) = 0.1$, p = 0.8] and [$\chi^2(1) = 0$, p = 1], respectively.

Thus, for all 1 percent, 50 percent, and 99 percent probabilities of conflicting payoffs, the rate of information avoidance remains around one-third. This

[6] In case of wishful thinking for any small cost $c_1 \in (0, 1]$ the expression $\beta_1 = \frac{c_1 p}{\Delta Y} e^{\frac{\Delta X}{c_1 p} - 1}$ is decreasing in p on the entire domain $[0, 1]$.

result suggests that the probability of conflicting payoffs does not impact the information avoidance behavior.

Hence, we observe that the predictions of the attention model are the closest to the results of the experiment. This result suggests that DWK's avoid information result may be driven not be self-image but rather by the attention motive. Instead of strategically trying to protect self-image, they may be naively reacting to any relevant information that grabs their attention. Put differently, even if self-image concerns play a role for information avoidance, they can be neutralized by making the situation more complex. Subjects are not strategic in the protection of their self-image.

5 Conclusion

In the present experiments, we allow for the possibility of information avoidance but vary the probability of conflict and the timeline of knowing the self-interested action so that we can compare different model predictions. We are thus able to provide a test of the robustness of the moral wiggle room effect. It turns out that self-image does not seem to explain the manipulability of behavior in the dictator game.

Our paper has an implication for the design of information structures for cases with a conflict of interest. In various real-life situations, informed experts make a suggestion to uninformed customers. This is, for example, the case for highly specialized experts like doctors, financial advisers, head-hunters, and lawyers. When making this binding choice on behalf of the customer, the expert faces several incentives, which are not necessarily aligned: the benefit to the customer (the quality or fitness of the product) and the benefit to the expert (commission paid for a specific product compared to other products). As in our setting, the information about the quality of the product and the attached commission can be initially hidden and revealed by the expert (e.g., for a doctor prescribing a drug, both her patient's diagnosis and the drug producer's incentive program might be relevant and initially unknown). Our results suggest that if the information about the commission is revealed first, then it might have a detrimental effect on the expert's incentives to learn the information on the quality of the product and, as a consequence, may lead to some poor advice.

With free access to information, the order of receiving the information plays a significant role in cases with a conflict of interest. It is still an open question to what extent trivial changes in the order in which information is presented affect human behavior.

References

Abelson, R. P., Aronson, E., McGuire, W. J., Newcomb, T. M., Rosenberg, M. J., and Tannenbaum, P. H. (1968). Theories of cognitive consistency: a sourcebook.

Akerlof, G. and Dickens, W. (1982). The Economic Consequences of Cognitive Dissonance. *Science (New York, N.Y.)*, 72(3):308–319.

Bazerman, M. H. and Sezer, O. (2016). Bounded awareness: Implications for ethical decision making. *Organizational Behavior and Human Decision Processes*, 136:95–105.

Bolton, G. E. and Ockenfels, A. (2000). ERC: A theory of equity, reciprocity, and competition. *American Economic Review*, 90(1):166–193.

Broberg, T., Ellingsen, T., and Johannesson, M. (2007). Is generosity involuntary? *Economics Letters*, 94(1):32–37.

Brunnermeier, M. K., Parker, J. A., Alpert, M., Alpert, M., Raiffa, H., Raiffa, H., Weinstein, N. D., Weinstein, N. D., Buehler, R., and Buehler, R. (2004). Optimal Expectations. *American Economic Review*, 95(4):1092–1118.

Cohen, S. (2001). *States of denial : knowing about atrocities and suffering*. Polity.

Dana, J. (2006). Strategic Ignorance and Ethical Behavior in Organizations.

Dana, J., Cain, D. M., and Dawes, R. M. (2006). What you don't know won't hurt me: Costly (but quiet) exit in dictator games. *Organizational Behavior and Human Decision Processes*, 100(2):193–201.

Dana, J., Weber, R. A., and Kuang, J. X. (2007). Exploiting moral wiggle room: Experiments demonstrating an illusory preference for fairness. *Economic Theory*, 33(1):67–80.

Ehrich, K. R. and Irwin, J. R. (2005). Willful Ignorance in the Request for Product Attribute Information. *Journal of Marketing Research*, 42(3):266–277.

Fehr, E. and Schmidt, K. (1999). A Theory of Fairness,Competition and Cooperation. *Quarterly journal of Economics*, 114(August):817–868.

Feiler, L. (2014). Testing models of information avoidance with binary choice dictator games. *Journal of Economic Psychology*, 45:253–267.

Festinger, L. (1957). *A Theory of Cognitive Dissonance*.

Fischbacher, U. (2007). Z-Tree: Zurich toolbox for ready-made economic experiments. *Experimental Economics*, 10(2):171–178.

Golman, R. and Loewenstein, G. (2016). Information Gaps: A Theory of Preferences Regarding the Presence and Absence of Information. *Decision*.

Greiner, B. (2015). Subject pool recruitment procedures: organizing experiments with ORSEE. *Journal of the Economic Science Association*, 1(1):114–125.

Grossman, Z. (2014). Strategic Ignorance and the Robustness of Social Preferences. *Management Science*, 60(11):2659–2665.

Grossman, Z. and van der Weele, J. J. (2017). Self-Image and Willful Ignorance in Social Decisions. *Journal of the European Economic Association*, 15(1):173–217.

Karlsson, N., Loewenstein, G., and Seppi, D. (2009). The ostrich effect: Selective attention to information. *Journal of Risk and Uncertainty*, 38(2):95–115.

Konow, J. (2000). Fair Shares: Accountability and Cognitive Dissonance in Allocation Decisions. *American Economic Review*, 90(4):1072–1092.

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

Lazear, E. P., Malmendier, U., and Weber, R. A. (2012). Sorting in Experiments with Application to Social Preferences. *American Economic Journal: Applied Economics*, 4(1):136–163.

Matthey, A. and Regner, T. (2011). Do I Really Want to Know? A Cognitive Dissonance-Based Explanation of Other-Regarding Behavior. *Games*, 2(4):114–135.

Rabin, M. (1995). Moral Preferences, Moral Constraints, and Self-Serving Biases.

Rayner, S. (2012). Uncomfortable knowledge: the social construction of ignorance in science and environmental policy discourses. *Economy and Society*, 41(1):107–125.

Sharot, T. (2011). The optimism bias. *Current Biology*, 21(23).

Simon, W. H. (2005). Wrongs of Ignorance and Ambiguity: Lawyer Responsibility for Collective Misconduct. *Yale Journal on Regulation Yale Journal on Regulation Article*, 22(2).

Sullivan, P. S., Lansky, A., and Drake, A. (2004). Failure to return for HIV test results among persons at high risk for HIV infection. *Epidemiology and Social Science*, 35(5):511–518.

Tasoff, J. and Madarasz, K. (2009). A Model of Attention and Anticipation. *SSRN Electronic Journal*.

van der Weele, J. J. (2012). When ignorance is innocence: On information avoidance in moral dilemmas. *SSRN Electronic Journal*.

6 Appendix

Game trees with results

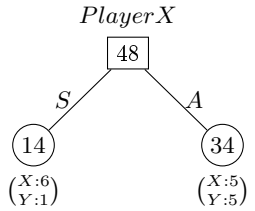

Figure 6.1: Conflict treatment
Notes: The number inside each node shows the number of subjects. The dictator chooses either Selfish (S), i.e., maximizing her own payoff, or Non-Selfish (NS).

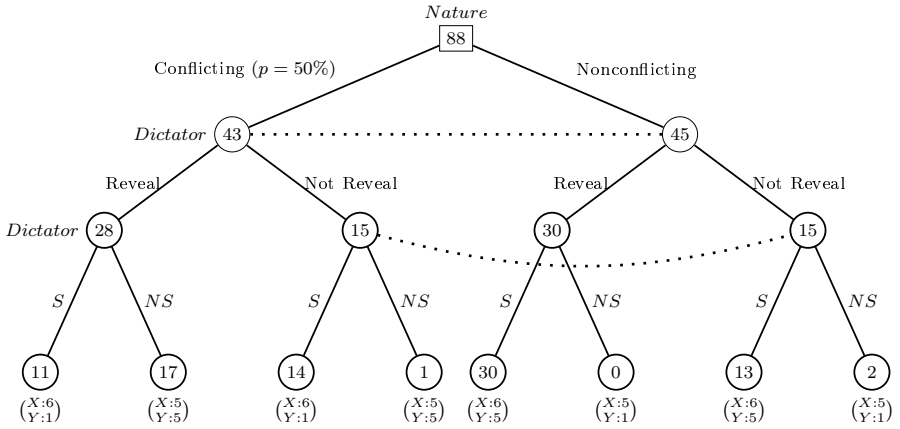

Figure 6.2: Before treatment
Notes: The number inside each node shows the number of subjects. The dictator chooses either Selfish (S), i.e., maximizing her own payoff, or Non-Selfish (NS).

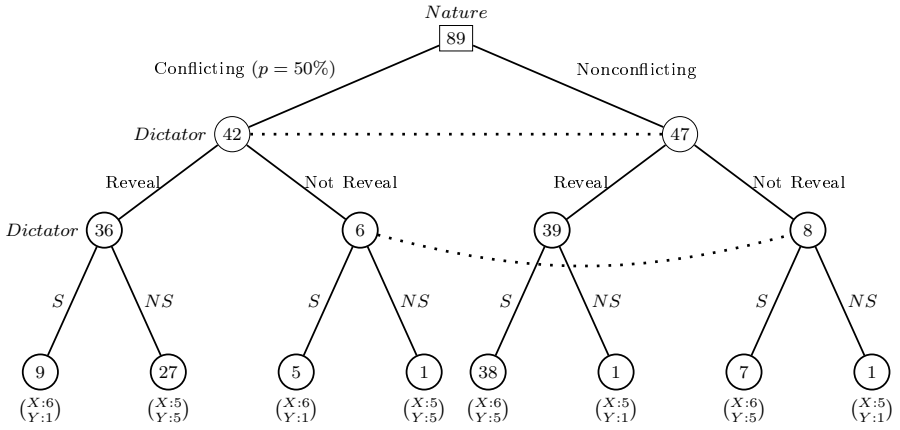

Figure 6.3: After treatment
Notes: The number inside each node shows the number of subjects. The dictator chooses either Selfish (S), i.e., maximizing her own payoff, or Non-Selfish (NS).

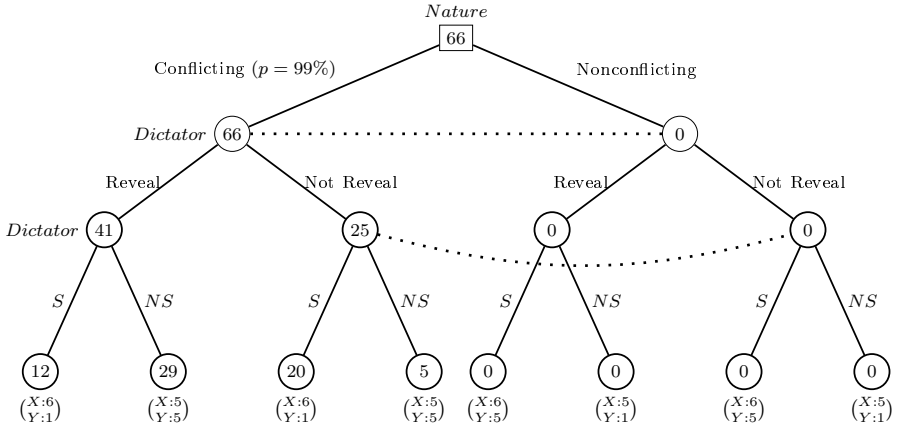

Figure 6.4: Conflict-Before treatment
Notes: The number inside each node shows the number of subjects. The dictator chooses either Selfish (S), i.e., maximizing her own payoff, or Non-Selfish (NS).

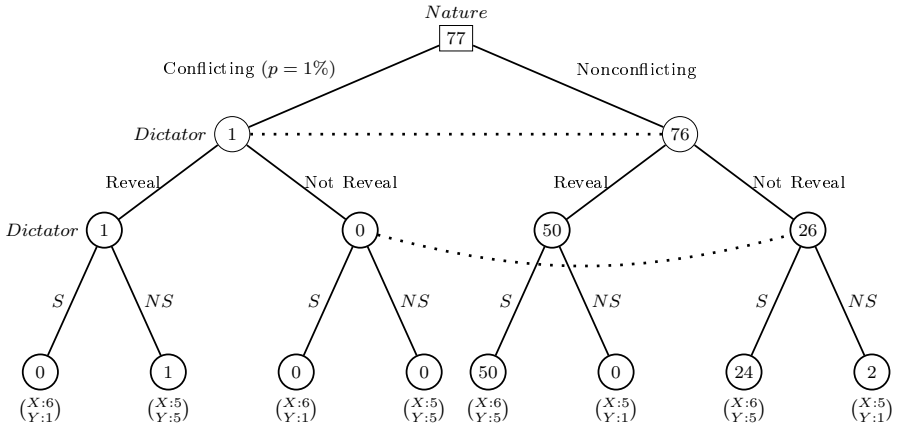

Figure 6.5: No-Conflict-Before treatment

Notes: The number inside each node shows the number of subjects. The dictator chooses either Selfish (S), i.e., maximizing her own payoff, or Non-Selfish (NS).

6.1 Instructions

All treatments

This is an experiment in the economics of decision-making. You will be paid for your participation in the experiment. The exact amount you will be paid will depend on your and/or others' decisions. Your payment will consist of the amount you accumulate plus a €5 participation bonus. You will be paid privately in cash at the conclusion of the experiment. If you have a question during the experiment, raise your hand and an experimenter will assist you. Please do not talk, exclaim, or try to communicate with other participants during the experiment. Please put away all outside materials (such as book bags, notebooks, etc.) before starting the experiment. Participants who violate the rules will be asked to leave the experiment and will not be paid.

In this experiment, each of you will play a game with one other person in the room. Before playing, we will randomly match people into pairs. The grouping will be anonymous, meaning that no one will ever know which person in the room they have played with. Each of you will be randomly assigned a role in this game. Your role will be player X or player Y. This role will also be kept anonymous. The difference between these roles will be described below. Thus, exactly one half of you will be a Player X and one half a Player Y. Also, each of you will be in a pair that includes exactly one of each of these types. The game your pair will play will be like the one pictured below. Player X will choose one of two options: "A" or "B." Player Y will not make any choice. Both players will receive payments based on the choice of Player X. The numbers in the table are the payments players receive. The payments in this table were chosen only to demonstrate how the game works. In the actual game, the payments will be different. For example, if player X chooses "B", then we should look at the right square for the earnings. Here, Player X receives 3 euros and Player Y receives 4 euros. Note that player X's payment is in the lower-left corner of the square, player Y's payment is in the upper-right corner.

Player X's choices	A	X:1	Y:2
	B	X:3	Y:4

At this point, to make sure that everyone understands the game, please answer the following questions:

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

Player X receives _ _

Player Y receives _ _

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

Player X receives _ _

Player Y receives _ _

<answers read aloud>

[Before treatment, No-Conflict-Before, Conflict-Before treatments]

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

		Left			Right	
Player X's choices	A	X:6	Y:1	A	X:6	Y:5
	B	X:5	Y:5	B	X:5	Y:1

You do not know which of the games you will be playing. However, note that for Player X, the payments will be identical. The only thing that differs is the payments for Player Y.

[*Before treatment:* The actual game you will play was determined by a coin flip before the experiment.]

[*Conflict-Before and No-Conflict-Before:* The actual game you will play was determined by a random draw using an urn with 100 balls consisting of [*Conflict-Before:* 1 Blue and 99 Black] [*No-Conflict-Before:* 99 blue and 1 black] balls. When the drawn ball is blue, the left game is played. If the drawn ball is black, the right game is played. For each pair of players the game will be determined in this way.]

However, we will not publicly reveal which game you are actually playing. Before playing, Player X can choose to find out which game is being played, if they want to do so, by clicking a button. This choice will be anonymous, thus Player Y will not know if X knows which game is being played. Player X is not required to find out and may choose not to do so. When the game ends, we will pay each player privately.

Player X's choices	A	X:6	Y:?	Reveal
	B	X:5	Y:?	

At this point, to make sure that everyone understands the game, please answer the following questions:

In both games, which action gives player X his or her highest payment of €6? _ _

If Player X chooses B, then Player Y receives _ _

1. €5
2. €1
3. either €5 or €1

[After TREATMENT]

The actual game you will play will be one of the four pictured below. Note that the TOP AND BOTTOM column games are the same except that Player X's payments are flipped between the two. Similarly, LEFT and RIGHT row games are the same except that Player Y's payments are flipped between the two.

Note that in games in TOP, Player X gets his or her highest payment of €6 by choosing A. In the TOP LEFT game, this gives Player Y's lowest payment of €1, and in the game TOP RIGHT, the highest payment of €5. Note that in games in TOP, if Player X chooses B, he or she gets a lower payment of €5. In the game TOP LEFT, this gives Player Y the highest payment of €5, and in the TOP RIGHT game, the lowest payment of €1.

Note that in games in BOTTOM, Player X gets his or her highest payment of €6 by choosing B. In the BOTTOM LEFT game, this gives Player Y's highest payment of €5, and in the game BOTTOM RIGHT, the lowest payment of €1. In games in BOTTOM, if Player X chooses A, he or she will get a lower payment of €5. In the game BOTTOM LEFT, this gives Player Y the lowest payment of €1, and in the BOTTOM RIGHT game, the highest payment of €5.

	Top Left			Top Right		
Player X's choices	A	X:6	Y:1	A	X:6	Y:5
	B	X:5	Y:5	B	X:5	Y:1

	Bottom Left			Bottom Right		
Player X's choices	A	X:5	Y:1	A	X:5	Y:5
	B	X:6	Y:5	B	X:6	Y:1

You do not know which of the games you will be playing. The actual game you will play was determined by two coin flips (one for TOP vs BOTTOM, and one for LEFT vs. RIGHT) before the experiment. However, we will not reveal publicly which game you are actually playing.

Before playing, Player X can choose to find out which games from LEFT and RIGHT are being played, if they want to do so, by clicking a "Reveal Player Y's Payoff" button. Note that for Player X, the payments will be identical. The only thing that will differ are the payments for Player Y. This choice will be anonymous; thus Player Y will not know if X knows which game is being played. Player X is not required to find out and may choose not to do so by clicking on the "Continue" button. After deciding to reveal or not, Player X will be informed which game(s) from TOP and BOTTOM is being played. This is independent of his or her actions. When the game ends, we will pay each player privately.

Player X's choices	A	Y:?	Reveal Y
	B	Y:?	

At this point, to make sure that everyone understands the game, please answer the following questions:

In TOP games, which action gives player X his or her highest payment of €6?
_ _
In TOP games , if Player X chooses B, then Player Y receives _ _

1. €5
2. €1
3. either €5 or €1

When player X clicks on "Reveal" button, his final payoff table will contain information about ...

1. Only player X's payoff
2. Only player Y's payoff
3. Both players' payoff

When player X does not click on "Reveal" button, his final payoff table will contain information about ...

1. Only player X's payoff
2. Only player Y's payoff
3. Both players' payoff

6.2 Proofs

Proof of Proposition 1.

Proof. First consider agents with $\beta = 0$. These agents will always choose option A at moment t=1, and are indifferent between revealing or not at moment t=0. By assumption, fraction μ of them does not reveal and fraction $(1-\mu)$ reveals. Therefore:

$$\int_{\sigma_0=(NR,A)} dF = \mu\varepsilon. \tag{6.1}$$

Consider type $\beta = \hat{\beta}_1$ that is indifferent between strategies $\sigma_{N,X}$ and $\sigma_{R,Y}$. If she chooses strategy $\sigma_{N,X}$ she guarantees a higher own payoff but incurs costs of self-image c_s since some of the homo economicus types choose the same strategy and also costs of wishful thinking c_w which occur with probability p:

$$EU_\beta(\sigma_{N,X}) = \overline{X} + (1-p+w)(1+a)\beta\overline{Y} + (p-w)(1+a)\beta\underline{Y} - pc_w(w) - (p-w)c_s. \tag{6.2}$$

If she chooses strategy $\sigma_{R,Y}$ she guarantees a high recipient's payoff but incurs the default effect costs c_d:

$$EU_\beta(\sigma_{R,Y}) = (1-p)\overline{X} + p\underline{X} + (1+a)\beta\overline{Y} - c_d. \tag{6.3}$$

Since she is indifferent, $EU(\sigma_{N,X}) = EU(\sigma_{R,Y})$ and thus

$$(p - w^*)((1+a)\beta\Delta Y + c_s) + pc_w(w^*) = p\Delta X + c_d, \tag{6.4}$$

where w^* denotes the optimal level of wishful thinking.

The RHS of the equation represents the costs from choosing $\sigma_{R,Y}$, these costs are constant for any type β. The LHS of the equation represents the costs from choosing $\sigma_{N,X}$ and we need to show that these costs monotonically increase in β. If this is the case, then for any type β that chooses $\sigma_{N,X}$, each type $\beta' < \beta$ will also choose this strategy. We need to show the following:

$$\frac{\partial[(p - w^*)((1+a)\beta\Delta Y + c_s) + pc_w(w^*)]}{\partial\beta} > 0. \tag{6.5}$$

.

Let us find the optimal level of wishful thinking w^* for each type $\beta \in (0, \beta_1)$. From the first order condition $\frac{\partial EU(\sigma_{N,X})}{\partial w} = 0$ we get

$$w^* = p(1 - \frac{c_1 p}{(1+a)\beta\Delta Y + c_s}). \tag{6.6}$$

If the optimal level is positive $w^* > 0$ then the costs of wishful thinking at w^* become $c_w(w^*) = c_1 p \log \frac{(1+a)\beta\Delta Y + c_s}{c_1 p}$.[7] We get

$$\frac{\partial[(p - w^*)((1+a)\beta\Delta Y + c_s) + pc_w(w^*)]}{\partial\beta} = \frac{\partial[c_1 p^2 + c_1 p^2 \log \frac{(1+a)\beta\Delta Y + c_s}{c_1 p}]}{\partial\beta} = \frac{(1+a)c_1 p^2 \Delta Y}{(1+a)\beta\Delta Y + c_s} > 0.$$

Otherwise, if the optimal level is zero $w^* = 0$ (in case $\frac{c_1 p}{(1+a)\beta\Delta Y + c_s} \geq 1$) then the costs of wishful thinking at $w^* = 0$ become $c_w = 0$. We get

$$\frac{\partial[(p - w^*)((1+a)\beta\Delta Y + c_s) + pc_w(w^*)]}{\partial\beta} = \frac{\partial[p((1+a)\beta\Delta Y + c_s)]}{\partial\beta} = (1+a)p\Delta Y > 0.$$

Therefore $\sigma^*$is an equilibrium.

Next, we find the cutoff $\hat{\beta}_1$ by plugging $w^* > 0$ from equation 6.6 into equation 6.4. We get:

$$c_1 p^2 + c_1 p^2 \log \frac{(1+a)\beta\Delta Y + c_s}{pc_1} = p(\Delta X + c_c) + c_d. \tag{6.7}$$

[7] Notice that c_s is a function of the threshold $\hat{\beta}_1$ and not of the type β and thus the derivative $\frac{\partial c_s}{\partial\beta} = 0$.

From the definition of c_s, using equation 6.1 and the fact that $\hat{\beta}_1$ is the cutoff we get that $c_s = s\mu\varepsilon \int_{\hat{\beta}_1}^{1} dF$. For simplicity, assume a linear cdf for $\beta > 0$: $F(\beta) = \varepsilon + \beta(1-\varepsilon)$ and thus:

$$c_s = s\mu\varepsilon(1-\varepsilon)(1-\beta). \tag{6.8}$$

Plugging these costs c_s in equation 6.7 we get equation 3.1:

$$\beta_1(w^* > 0) = \frac{1}{(1+a)\Delta Y - s\mu\varepsilon(1-\varepsilon)} \left(pc_1 e^{\frac{p\Delta X + c_d}{c_1 p^2} - 1} - s\mu\varepsilon(1-\varepsilon) \right).$$

Alternatively, if $w^* = 0$ by plugging c_s from 6.8 into equation 6.4 we get a different cutoff[8]

$$\beta_1(w^* = 0) = \frac{\Delta X + \frac{c_d}{p} - s\mu\varepsilon(1-\varepsilon)}{(1+a)\Delta Y - s\mu\varepsilon(1-\varepsilon)}.$$

□

Proof of the Remark.

Proof. Let $\hat{\beta}_1$ be the lowest type such that each type $\beta \in (\hat{\beta}_1, 1]$ chooses $\sigma_{R,Y}$ (by assumption we at least have one such type $\beta = 1$) and type $\hat{\beta}_1$ is at most indifferent between choosing $\sigma_{N,X}$ and $\sigma_{R,Y}$: $EU_{\hat{\beta}_1}(\sigma_{N,X}) \geq EU_{\hat{\beta}_1}(\sigma_{R,Y})$. Then we can use the monotonicity argument from the proof of Proposition 1 (equation 6.5): for each $\beta \in (0, \hat{\beta}_1)$ we get $EU_\beta(\sigma_{N,X}) > EU_\beta(\sigma_{R,Y})$. And thus the level $\hat{\beta}_1$ is the desired threshold level β_1 given by equation 3.1. □

Proof of Proposition 2.

Proof. Given the arguments used in the previous proof, the result is immediate. Since the agent does not distort the prior probability of conflicting payoffs in the After treatment, we get the following condition for the agent with $\beta > 0$ to be indifferent between playing $\sigma_{N,X}$ and $\sigma_{R,Y}$: $EU(\sigma_{N,X}) = EU(\sigma_{R,Y}) \iff \overline{X} + \beta\overline{Y} - p\beta\Delta Y = \overline{X} + \beta\overline{Y} - p\Delta X$ which gives us equation 3.2. Agents with $\beta = 0$ will choose the X-maximizing strategy and randomize w.r.t. the revelation choice. □

Proof of Proposition 3.

Proof. Following the arguments in the proofs above, the result is immediate. The threshold value β_0 is determined by equating the utility of action A (allocative utility minus the self-image costs) and the utility from action B: $U(A) = U(B) \iff \overline{X} + \beta\underline{Y} - s\varepsilon(1-\varepsilon)(1-\beta) = \underline{X} + \beta\overline{Y}$, which gives the equation 3.3. □

[8] This result is very similar to equation A.16 in Grossman and van der Weele (2017) that determines the cutoff type when self-image and the costs of revealing are present, but wishful thinking is shut down.