WARWICK
ECONOMICS

The Effect of Self-Awareness on Dishonesty

Ceren Bengu Cibik & Daniel Sgroi

October 2020 No: 1307

Warwick Economics Research Papers

ISSN 2059-4283 (online)
ISSN 0083-7350 (print)

The Effect of Self-Awareness on Dishonesty*

Ceren Bengü Çıbık & Daniel Sgroi

October 16, 2020

Department of Economics,

University of Warwick

Abstract

We investigate the relationship between self-awareness and dishonesty in a pre-registered experiment with 1,260 subjects. In a first experiment, we vary the level of awareness of subjects' own past dishonesty and explore the impact on behaviour in tasks that include the scope to lie. We find that in single-person non-interactive tasks, self-awareness of dishonesty helps to lower dishonesty in the future. However, in tasks that are competitive in nature becoming more aware of past dishonesty raises the likelihood of dishonesty. We argue that this behaviour is consistent with cognitive dissonance. In a second experiment we vary the degree of competitiveness in one of our core tasks to further explore the interactions between self-awareness, (dis)honesty and competition. Our results show when and why pointing out those who have been (dis)honest in the past can be an effective way to induce honesty in the future and when it might back-fire badly, and perhaps also shed some light on perceived increases in dishonesty in politics, the media and everyday life.

I Introduction

We now have a sound grasp of how motivated reasoning and self-deception are both ubiquitous and potentially very damaging for society as a whole (Benabou & Tirole (2016), Gino et al. (2016)). What is less well understood is what happens when self-deception fails and individuals realize they are not the upstanding morally unambiguous individuals they might wish to be. In this paper we consider what happens when individuals become more aware of their own dishonesty. Will this push them towards greater honesty or will this realization only increase their propensity for dishonesty in the future? The first, more positive, reaction to increased self-awareness is normally referred to as "moral balancing": an effort to correct the sins of the past by leaning towards greater morality in the future. The second, more negative response reflects the psychological costs associated with cognitive dissonance, the

*Ethical approval reference ECONPGR 05/18. AEA RCT Registrations AEARCTR-0005142 (wave 1) and AEARCTR-0005955 (wave 2).

high cost of attempting to carry two opposing viewpoints at once, and might be expected to lead to acceptance and greater levels of dishonesty. Which force is more prevalent is an open and topical question in a world increasingly characterized by dishonesty at all levels of society whether it be in our leaders, our media (especially in an era of "fake news") and in social media-based interactions in which lying can seem relatively easy, but where revelations about dishonest behaviour in the past occur on a daily basis. It would be wonderful to think that self-awareness of dishonesty would result in reduced incidence of dishonest behaviour in the future. Unfortunately our results suggest that this may not be the case, but that on the contrary, the constant stream of ineffectual recriminations about previous lying may actually result in higher levels of dishonesty in the future.

In wave 1 of our study we ran a pre-registered experiment involving 892 subjects. We first allocated them to one of several treatments in which they were asked to write about incidents in their lives which involved dishonesty of various types. Alternatively, they were allocated to a control group where they were not asked to engage in any writing. We then ask them to undertake two incentivized tasks, each of which included the scope to be dishonest. In one task, the "matrix puzzle" taken from Ariely (2012), subjects were asked to identify the number of cells in a matrix which sum to 10 and then report their answer earning higher payments for higher reported numbers. We compared both the absolute number reported in the treatment(s) vs the control as one measure of dishonesty but were also able to consider subjects who reported numbers higher than what was possible, and so must have lied. In another task, a sender-receiver game taken from Gneezy (2005), we asked them to send a message to a partner, the message being either true or false with the likelihood being that the false message would result in higher rewards. In both tasks they faced incentivized temptation to lie though the contexts were very different. We also collected responses to standard demographic questions as well as answers to various psychometric and personality questions. In wave 2 we repeated the basic setup with a further 368 subjects, but asked subjects to undertake a single task: a variant of the matrix puzzle task where we modified the structure to raise the competitive nature of the task by making rewards partly dependent upon reporting a number that placed subjects inside the top 50% of the distribution. Our experiment gave us the ability to compare our treatment and control groups to test whether self-awareness had an impact on behaviour (it did), compare the two tasks to see if competition and strategic interaction mattered (it did) and to compare the matrix task in wave 1 with the more competitive variant of the task in wave 2 to see if raising the level of competition mattered (it did). These three core results link to pre-registered hypotheses and conjectures driven by a model that draws on a rational choice framework developed in Rabin (1994). The model explains how higher levels of competition can exhibit a form of crowding out, allowing individuals to see dishonest behaviour as acceptable in the context of a competitive environment, which in turn interacts with self-awareness to drive up dishonesty: we see this when comparing tasks in wave 1, and it is especially clear when we compare the matrix task between waves 1 and 2. Whenever the competitive nature of the task increases, dishonesty rises in the face of self-awareness, while in a single-person task (the basic matrix puzzle from wave 1) self-awareness drives down dishonesty. We also see that incentives matter in the sender-receiver game: as the incentives to lie increase lying goes up, but this does not apply to the matrix puzzle.

Our work lends itself to several immediate policy conclusions. Becoming aware of your

own dishonesty can help boost honesty but only in non-competitive settings. Where there are issues of rank or of bettering others, then if anything self-awareness will boost dishonesty further. This makes it important to be careful when challenging the lies of others: in some settings this will produce beneficial outcomes, but in others it will only make things worse. We can also see why in very competitive settings, whether it be attempting to win votes or sell newspapers, becoming more aware of your own dishonesty will only lead to yet more lies.

II Literature Review

The study of dishonesty, whether lying or cheating, has been a popular topic within academia over the past few decades. In this section, we will review the studies that are close to our research question and experimental design. Various researchers have employed a version of a cheap talk sender-receiver game to study dishonesty (Gneezy 2005, Sutter 2009, Erat & Gneezy 2012). For example, Gneezy (2005) investigated the consequences of lying on (dishonest) behaviour. In this task as in our experiment, subjects in the sender role were made aware of the respective payoff allocations of two options (A and B). Senders could send either an honest or dishonest message to the receiver as a form of communication but receivers were free to ignore the message. The results suggest that on average 36% of senders lied when the cost and benefit of the lie amounted to $1, 17% with a cost of $10 and benefit of $1, and 52% in with a cost and benefit of $10. In our own experiment we employ a variant of the method used in Gneezy (2005).

The second incentivized task employed in our experiment to elicit dishonesty is the matrix puzzle (Ariely 2009) which has been widely-used in the literature (Mazar et al. 2008, Mead et al. 2009, Shu et al. 2011, Ariely 2012). In Mazar et al. (2008), participants were given 20 matrices and asked to find two numbers that added up to 10 from among those numbers in 5 minutes in return for $0.50 per correct answer. Participants in their control group (where cheating was not possible) solved 3.4 matrices on average, while those in a group where there was an option to cheat reportedly solved 6.1 matrices on average, with no significant difference when the incentive for cheating was increased to $2 per correct answer.

Another popular experimental tool used to measure dishonesty is a simple coin-flip where participants are paid on the basis of the number of heads (or tails) reported (Bucciol & Piovesan 2011, Houser et al. 2012, Abeler et al. 2014, Cohn et al. 2015). Notably, Abeler et al. (2014) found that 55% of the subjects reported an outcome that provided no material payoff even though there was no incentive to behave honestly in their experiment. Field experiments are also widely-used in the dishonesty literature, for instance Yezer et al. (1996), Franzen & Pointner (2013), Stoop (2014) (lost/misdirected letters), West (2005), Cohn et al. (2019) (wallets found on the street where the return rate of the lost items determines the level of honesty among the sample and Pruckner & Sausgruber (2013) (newspaper sales on the street). Rosenbaum et al. (2014) provides a review of the methodology and the results of 63 economic and psychology experiments on dishonesty. Their review highlights one of the most robust findings across the literature: the existence of unconditional cheaters, never-cheaters and partial cheaters who are sensitive to monitoring costs and the intrinsic costs of being dishonest.

There have been numerous attempts to explain the phenomena of "partial cheaters".

For example, Levit (2006) argues that the existence of partial liars who do not maximize their material payoff by lying could be explained by a model in which people receive internal rewards from being honest. Moreover, Ariely (2012) explains this phenomena as a form of what he calls "Fudge Factor Theory". Fudge Factor Theory argues that there are two conflicting motivations for people while they are deciding to behave honestly or not. The first is the economic motivation which corresponds to the material benefit of cheating. The second is the psychological motivation or the avoidance of cognitive dissonance which drives people to behave honesty in an effort to see themselves as moral. The result of this conflict, which could generate cognitive dissonance determines the action which is susceptible to one's self-control or contextual framework such as the intrinsic cost of lying or the existence of moral wiggle room. Similarly, Fischbacher & Föllmi-Heusi (2013) argues that partial liars exist because people care not only about the material gains from lying, but also about maintaining their morally upright self-image.

A second theory that can explain partial lying is "moral balancing", which relates to an individual's attempts to maintain an adequate level of morality in the light of past misdeeds. If a person's moral self-image drops below a personal standard (that she determines through her own beliefs about her morality and her beliefs about how other's perceive her morality) she would react by engaging in moral compensation to reduce cognitive dissonance. This also suggests that if her moral self-image rises above her personal standards, then she might become prone to behave immorally (Ploner & Regner 2013). In line with Ploner & Regner (2013)'s argument, in their earlier work Mazar & Zhong (2010) observed that participants who reported buying environment-friendly products in the past were more likely to over-report their performance in a simple math task and steal money during the experiment.

Within the psychology literature, Diener & Wallbom (1976) provides perhaps the closest experiment to our own, albeit without monetary incentives and quite modest in size. In their experiment, one group of participants were seated in front of a mirror and listened to a self-recorded tape about their physical characteristics and occupation (a method of inducing self-awareness), while a second group listened to a recording made by a stranger. Participants were then given the opportunity to cheat on an anagrams test. They found that significantly fewer people cheated when they were made more self-aware. While not directly comparable (since we use a different self-awareness procedure and honesty tasks) our results are somewhat different: we find that self-awareness can both raise and lower the level of dishonesty depending upon context. Our divergence from the findings in Diener & Wallbom (1976) are likely because of the nature of our tasks which vary the degree of interaction with others and material incentives (neither of which feature in Diener & Wallbom (1976)) and the significant difference in power. We also chose to adopt Fenigstein & Levine (1984)'s priming technique by asking participants to write a short story about themselves which is related to the moral code of being honest to impose awareness of one's ideal-self. They hypothesized that priming in the context of a task which requires people to make judgments about their causal effect on certain outcomes activates self-related cognition and would make the self more accessible as a causal agent for subsequent events which in turn contributes to an increase in self-awareness.

III Theoretical Framework

Our hypotheses are based on a simple rational-choice model which follows Rabin (1994). In this model, there are three important variables that affect people's choice to behave dishonestly. These factors are the material benefit obtained from dishonesty, the psychological cost of dishonesty (which derives from cognitive dissonance) and the cost of developing beliefs that are not consistent with the honest set of beliefs about the morality of dishonesty.

Let $X \in [0, \infty)$ be the chosen level of dishonesty with corresponding utility from the dishonest activity $U(X)$ where $U'(X) > 0$ and $U''(X) < 0$ for all X. Absent any cost of lying $X = \infty$. However, in our model lying does indeed generate cost that arises through the existence of cognitive dissonance. This can be described as the conflict between the benefits of believing in one's ideal self and the costly realization that the true self does not meet this standard. We assume the existence of a level of dishonesty, Y, above which cognitive dissonance takes hold. For simplicity we assume that if $X < Y$, there is no cognitive dissonance and hence no psychological cost. We then measure the cost (cognitive dissonance) suffered as $D(X - Y)$ where $D'(X - Y) > 0$, $D''(X - Y) > 0$ for all values of $X - Y > 0$, and $D(X - Y) = 0$ if $X \leq Y$.

Individuals may change their beliefs about the degree of dishonesty associated with any particular behaviour in an attempt to avoid high levels of cognitive dissonance: for instance convincing yourself that lying is acceptable in a given context. However, changing beliefs about the morality associated with certain (dishonest) actions comes with a cost. The cost of developing beliefs which are different from the natural, true set of belief about the morality of dishonesty is represented by the function $C(Y)$ where $C(0) = 0$ and $C'(Y) > 0$, $C''(Y) > 0$ for all Y. In this case people can adjust Y by incurring this cost. We normalize the initial level of the threshold to be $Y = 0$ which allows us to consider Y to also represent the distance between a person's beliefs and the true moral level of activity. Therefore, we can summarize the maximization problem as follows.

$$\begin{aligned} \max_{X,Y} L(X,Y) = \quad & U(X) - D(X - Y) - C(Y) \\ \textit{where} \quad & X, Y \geq 0; \quad U'(X), D'(X - Y), C'(Y) > 0; \\ \textit{and} \quad & U''(X) < 0, D''(X - Y), C''(Y) > 0. \end{aligned} \tag{1}$$

In the original version of this model, Rabin suggests that if a person receives lower material utility from engaging in an activity, or it becomes more costly to maintain modified beliefs about the morality of behaviour, or there is greater distaste for cognitive dissonance, then that individual would reduce the amount of the perceived immoral activity.

In our experiment, we vary the material benefits obtained from dishonesty and the psychological cost of dishonesty (or cognitive dissonance)[1] We vary the material incentive for each dishonesty task to observe the effect of the function $U(.)$ in the model. To observe the effect of cognitive dissonance we employ two methods. The first method comes through the use of positive or negative self-image induction methods that are reflected in the function $D(.)$ in the model. In the dishonesty treatment groups, subjects are asked to write about recent experiences of (dis)honest behaviour: ranging from behaviour that caused harm to

[1]Since in our model the probability of getting caught is zero, we eliminate the material cost of lying.

others through behaviour with no ramification and finally honest behaviour. In this way we vary the initial level of cognitive dissonance, increasing the gap between X and Y pushing up the psychological cost of dishonesty. The second method is to vary the context under which decisions are made which might shift beliefs about the true underlying morality of dishonesty. We argue that under some contexts, dishonest behaviour could be seen more acceptable which leads to a higher level of Y and a lower level of $D(.)$ for a given X.

To explore this further we categorize our tasks across four characteristics that might help to determine the level of cognitive dissonance or psychological cost that comes about through dishonest behaviour: the degree of competition in the game, whether the final decision is only the responsibility of one person, whether dishonesty is salient, and whether the task is ego-relevant or not. The impact and even direction of the effect of changes in self-awareness on the level of dishonesty depend on these characteristics. Firstly, behaviour that might seem to be unambiguously dishonest in a single-person context, might seem more acceptable during a competitive interaction when "defeating" your rival is the ultimate aim and so damaging the other player's payoff is a key component of the game, bringing to mind the proverb "all is fair in love and war". Secondly, if a game includes lying as a possible action in the list of actions, it might be easier to mentally categorize this action as selecting a strategy presented to you by the experimenter, rather than undertaking a dishonest action, and so it is easier to overcome cognitive dissonance: we see this as a form of moral "wiggle room" which acts to reduce the salience (and cost) of lying. Thirdly, if the payoffs in the game depend not only the player's action itself, but also on the other player's action, it might become easier to feel a lower level of responsibility for any dishonest action. Consider for instance the situation when the ultimate decision of whether to believe and act upon a lie is in the hands of another person (as in the case in the sender-receiver game in which the sender may send a dishonest message but the receiver chooses whether to act upon this lie). Finally, if a game is ego-relevant, the cost of lying might be higher since it includes elements of self-deception or the formation of motivated beliefs which come with a cost (Benabou & Tirole 2016), as well as the deception of others.

IV Experimental Design

The experiment itself consisted of three stages and two waves. The first stage included a questionnaire containing basic demographic questions, together with questions designed to elicit preferences for fairness, risk, integrity and their ethical stance, and a brief version of the Big Five Inventory designed to detect personality traits (Rammstedt & John 2007). In the second stage of the experiment, subjects were randomly assigned to one of four different groups: a control group, honesty treatment group, low dishonesty treatment group or high dishonesty treatment group. In the three treatment groups, subjects were asked to write about a real-life event that took place in the 12 months before the experiment, during which they were completely honest (the honest treatment), dishonest but with no negative effect on others (the low dishonesty treatment) or dishonest with a negative effect on others (the high dishonesty treatment). In the control group subjects were not asked to write anything but instead progressed directly to the third stage of the experiment. The aim of the second stage was to generate between-subject variation in self-awareness. In particular forcing subjects

to recall an honest event should not impose any prior level of cognitive dissonance while recalling events featuring dishonesty might generate an initial level of cognitive dissonance which could be even higher if their dishonesty ended up harming another person. The third stage of the experiment contained incentivized tasks, undertaken in a random order, which included the option to be dishonest which we will call "dishonesty tasks" for simplicity.[2] The two waves differed only with respect to the dishonesty tasks used.

The first wave of the experiment included two dishonesty tasks: the matrix puzzle game (Ariely 2012) and the cheap talk sender-receiver game (Gneezy 2005). In the matrix game, participants were presented with an image containing 20 different matrices and asked to find as many pairs of numbers that sum to 10 in these matrices as possible within 5 minutes. Once the time was over, subjects were directed to the next page to report the number of pairs they found. The potential bonus they made depended on the number of matrices they reported solving. The experiment included two variations where the incentives were \$0.10 (low incentive) and \$0.30 (high incentive) per correct answer. Since the subjects' answers could not be confirmed, they were free to report any number they wished, and it was not possible to detect dishonesty at an individual level *unless* they reported an infeasibly large number. In general, the number of reported answers is normally assumed to be increasing in the level of dishonesty of the subject. As a special property of our design, the maximum number of correct answers was 10, and so any reported number above 10 was considered to be infeasible.

In the sender-receiver game, subjects were asked to imagine that they were matched with another anonymous MTurk worker, and played the role of the sender. The other worker (the receiver) made the final decision about which of two possible options, 'Option A' or 'Option B', would be selected, which determined the payoffs of both players. In our design, two different allocations related to Option A and Option B were used but in both of these two cases, Option B always gave a higher payoff to the sender (respectively a lower payoff to the receiver) than Option A. However, the payoffs associated with these two options were visible only to the sender, not to the receiver. Instead, the subject, who did see the payoffs associated with the two options, had the opportunity to send a message to the receiver to help guide their choice. In our design the payoff allocations were set as \$1 for the sender (subject) and \$X for the receiver under option A and \$X for the sender (subject) and \$1 for the receiver under option B, where X was set equal to \$1.20 in the low incentive setting, and \$3 in the high incentive setting. The subject was given two possible messages to send: "Option B will earn you more money than Option A" (which was not true and was hence classified as a dishonest message) or "Option A will earn you more money than Option B" (which was true and was hence classified as an honest message).

The second wave of the experiment differed from the first wave in only one respect: it included only one dishonesty task, a modified version of the matrix puzzle game. As in the wave 1 version of the task, subjects were asked to report the number of pairs that they found in the matrix puzzle which add up to 10. The only difference from the wave 1 task was the payment scheme. While in the wave 1, participants were paid for each matrix they reportedly solved, in the modified version, they were paid a lump-sum amount only if they

[2]Subjects were paid a bonus based on their actions in one of these dishonesty tasks chosen at random at the end of the experiment.

were in the top 50% of the distribution. The lump-sum amount was decided based on the difference between the average payment that participants in the top 50% of the distribution received and the average payment that participants in the bottom 50% received in wave 1 of the experiment. This added a competitive element to the matrix puzzle game but maintained the same average payment in expectation by applying a mean-preserving spread by awarding a bonus only to those in the top half of the distribution. As in the wave 1 task, the experiment included two versions of this task with low or high material incentives allowing us to check whether our results were responsive to a change in the incentive payments. The lump-sum payments were set to be \$0.72 in the low incentive task and \$2.13 in the high incentive task, which were calculated to yield the same average payment as in the wave 1 variant of the game.

The dishonesty tasks were followed by two dictator games designed to elicit subjects' preferences for altruism. They were asked to indicate what percentage of their actual bonus from this experiment they would like to donate to Macmillan Cancer Support and/or to the researchers of this experiment to be used for research purposes. By employing dictator games, we aimed to test whether the moral balancing argument held in our sample. These games gave subjects the opportunity to balance their earlier dishonesty by donating either to the charity or to the researchers, and so we could directly compare actions in the dishonesty tasks with donation levels. The last task in the final stage of the experiment aimed to provide data to enable us to check for possible experimenter demand effects. First, subjects were asked to report how likely is might be for a person who did something dishonest to behave more honestly in the future. Then, they were asked to indicate the percentage chance that the experimenter expected them to behave honestly in the various tasks.

The experiment was conducted in 2020 on Amazon M-Turk, the first wave in February and the second wave in July.[3] Subjects earned a show-up fee of \$2 plus a performance related bonus payment. The experiment took approximately 25 minutes on average. Out of the 892 subjects who took part in wave 1, 284 were randomly allocated to the control group, 205 to the honesty treatment group, 208 to the low dishonesty treatment group and 195 to the high dishonesty treatment group. Out of the 368 subjects who took part in wave 2, 101 were randomly allocated to the control group, 76 to the honesty treatment group, 104 to the low dishonesty treatment group and 87 to the high dishonesty treatments group.[4]

Full experimental instructions can be found in the Appendix, together with a simplified timeline of events, which is presented in Table 5.

[3]Note that in most cases in the results to follow we will compare control and treatment groups *within* waves and so the temporal gap *between* waves is not relevant. In the cases where we *do* compare between waves, we will always include a full set of demographic variables to control for any demographic changes that might have taken place between the two waves.

[4]Initially, we collected 1110 observations in wave 1 and 705 in wave 2. Before starting to analyze our data, we imposed a relevance filter on the text (described in the RCT pre-registration analysis plan) to allow us to remove subjects who did not take our self-awareness task seriously, for example by entering irrelevant text. In total, we eliminated 218 subjects in wave 1 and 337 subjects in wave 2 who wrote irrelevant text in the second stage of the experiment.

V Hypotheses

Having outlined the design, we are now in a position to summarize much of the discussion in section III in six testable hypotheses. These are also referenced in our pre-registered RCT entry and analysis plan. The first two of our hypotheses relate directly to our discussion of the role of self-awareness and cognitive dissonance:

Hypothesis 1: Does self-awareness matter? Self-awareness affects the level of dishonesty. However, the direction is determined by the context under which people make their decision to behave dishonestly or honestly.

Our next hypothesis addresses the issue of context. As we argue in section III (and in our pre-registered analysis plan), if dishonest behaviour in a task is expected to create a higher (lower) level of cognitive dissonance, regardless of the positive or negative self-awareness imposed, the level of dishonesty will be lower (higher) in this task than in a task where the dishonest behaviour is associated with lower (higher) levels of cognitive dissonance. We can refine this logic further so that applies directly to the specific form of games described in section IV:

Hypothesis 2: Self-awareness stemming from any of our treatments should result in a decrease in levels of dishonesty in the matrix puzzle game but an increase in levels of dishonesty in the sender-receiver game.

We argued earlier that the wave 2 variant of the matrix puzzle game induces lower levels of cognitive dissonance and so should also be associated with higher levels of dishonesty which yields our next hypothesis:

Hypothesis 3: We expect participants to incur lower levels of cognitive dissonance from dishonest behaviour in the wave 2 version of the matrix puzzle than in the wave 1 version, therefore they should behave more dishonestly in wave 2.

As a related supplementary hypothesis we can also consider the extent to which material incentives influence behaviour and change the nature of the relationship between self-awareness and dishonesty. To that end we have our next testable hypothesis:

Hypothesis 4: Material incentives play a role in determining the relationship between self-awareness and dishonesty.

As an additional exercise we might also consider the characteristics of those who lie in a regular and detectable way. We will do so in the results section to follow and can here identify two testable hypotheses relating to consistency and moral balancing:

Hypothesis 5: (Consistency) Those who lie more in one task are likely to lie more in the other.

Hypothesis 6: (Moral balancing) Lying more should result in higher donations to charity and/or the researcher.

VI Results

In what follows we will structure our results first into an attempt to investigate the core hypotheses relating to the link between self-awareness and dishonesty (identified above as hypotheses 1 to 4) via comparisons between the treatment and control groups before moving on to a discussion of the characteristics of lying that are detectable at the individual level (relating to hypotheses 5 and 6). Note that our results are designed to link easily to the corresponding hypothesis, so result 1 refers to hypothesis 1, and so on.

A Group-level Results

We can first start with a mean value comparison test between the treatment groups and the control group for the dishonesty variables obtained from the different dishonesty tasks which will provide us with one useful insight. Table 1 reports the mean values for all the dishonesty tasks conducted in the first and second waves of the experiment for the treatment groups and the control group separately. The findings in this table allow us to make a useful simplification driven by the data: different types of treatment groups seem to behave in a similar way to each other when compared with the control group. In other words, what seems to matter is making individuals self-aware of past honesty/dishonesty rather than the way in which this self-awareness comes about. As a result from now on, we will combine the treatment groups to conduct our analysis and perform our analysis as treatment vs control.[5].

Asking subjects to recall recent experiences with honesty or dishonesty has a significant impact on their future behaviour in the experiment. We can see from table 1, the treatment group significantly differs from the control group for all dishonesty variables. This result provides immediate support for Hypothesis 1 which states that self-awareness affects the level of dishonesty:

Result 1: Self-awareness matters: self-awareness affects the level of dishonesty in the future. Moreover, this impact is largely neutral to the type of self-awareness.

TABLE 1: Mean Value Comparisons of Various Dishonesty Tasks

	Wave 1				**Wave 2**	
	Matrix Puzzle		**CT Sender Receiver Game**		**Modified Matrix Puzzle**	
	No of matrix reported to be solved		% of people who sent a dishonest message		No of matrix reported to be solved	
	Low Incentive	*High Incentive*	*Low Incentive*	*High Incentive*	*Low Incentive*	*High Incentive*
Mean Values						
Control Group	5.746	5.799	0.394	0.500	7.881	7.059
Honesty Treatment	4.766	4.868	0.576	0.634	5.408	6.053
Low Dishonesty Tr.	4.822	4.827	0.514	0.543	5.712	5.356
High Dishonesty Tr.	5.528	5.487	0.497	0.595	5.138	4.885
Treatment Groups	5.03	5.05	0.530	0.590	5.438	5.401
T-test[1]						
Honesty vs Low Dishonesty	0.9078	0.9302	0.2128	0.0608*	0.6957	0.3547
Honesty vs High Dishonesty	0.1303	0.2248	0.1176	0.421	0.714	0.1201
Low Dishonesty vs High Dishonesty	0.1722	0.1972	0.734	0.2971	0.4261	0.5029
Control vs Treatment	0.053*	0.042**	0.0002***	0.0112**	0.0001***	0.0055***

[1] p-values from a two-tailed t-test are reported. * p<0.10, ** p<0.05 , *** p<0.01

[5]This also helps minimize the need for p-value adjustments from multiple testing.

To investigate whether the direction of the effect is determined by the context under which people make their decision, we compare the effect of self-awareness on dishonesty differentiating by dishonesty task. We can already see that in line with Hypothesis 2, self-awareness decreases the level of dishonesty in the matrix puzzle game whereas it increases the level of dishonesty in the sender-receiver game in Wave 1. Table 1 shows that the mean number of matrices reportedly solved by subjects in the control group is 5.746 (5.799) whereas it is 5.03 (5.05) in the treatment group for the low (high) incentive task. This suggests that for subjects playing the matrix puzzle game, self-awareness significantly lowers the level of dishonesty. On the other hand, the proportion of people who sent a dishonest message in the control group in the low (high) incentive sender-receiver game is 39.4% (50%) whereas it is 53% (59%) in the treatment group. We see a significant increase in dishonesty after self-awareness has been induced in the context of the sender-receiver game. We argued in section III (and in our pre-registration) that dishonest behaviour is psychologically more costly in the matrix puzzle game than the sender-receiver game and so this result also supports Hypothesis 2. We can summarize all of this in our next result:

Result 2: Context matters: self-awareness (stemming from any of the treatments) leads to a decrease in dishonesty in the matrix puzzle game but also leads to an increase in dishonesty in the sender-receiver game.

TABLE 2: Regression Analysis

	Matrix Puzzle		**Sender Receiver Game**	
	(1)	(2)	(3)	(4)
VARIABLES	Model 1	Model 2	Model 1	Model 2
Treatment	-2.051***	-1.932***	0.113***	0.108***
	[0.601]	[0.609]	[0.0304]	[0.0306]
Wave 1	-1.840***	-1.558**	-	-
	[0.626]	[0.628]		
Treatment x Wave 1	1.319*	1.279*	-	-
	[0.700]	[0.703]		
High Incentive	-0.253	-0.256	0.0758***	0.0773***
	[0.196]	[0.197]	[0.0172]	[0.0174]
High Incentive x Wave 1	0.285	0.293	-	-
	[0.223]	[0.224]		
Constant	7.597***	8.003***	-	-
	[0.545]	[1.641]		
Observations	2,520	2,512	1,784	1,778
R-squared	0.016	0.040	0.0121	0.0252
Control Variables	✗	✓	✗	✓

Standard errors are clustered at the individual level and shown in brackets. Column 1 and 2 represent the linear regression results on the number of matrices reported to be solved in the matrix puzzle game. Column 3 and 4 represent the probit regression results on a variable which takes 1 if a dishonest message is sent in the sender-receiver game and 0 if an honest message is sent. Control variables include age, age square, being married, having at least a college degree and being American. Stars indicate statistical significance as follows: * p<0.10, ** p<0.05 and *** p<0.01.

In order to test Hypothesis 3, we conducted a second wave of the experiment in which the matrix puzzle game was modified in an attempt to lower expected cognitive dissonance after dishonest behaviour, as compared with the wave 1 version. As stated in Hypothesis 3, since we attempted to induce lower cognitive dissonance in the wave 2 version of the game, we expect to observe higher levels of dishonesty in the wave 2 version of the matrix puzzle game than in the wave 1 version. The last two columns of table 1 show that the mean number of matrices reported to be solved in the modified matrix puzzle game is 7.881 (7.059) in the control group whereas it is 5.438 (5.401) in the treatment group for the low (high) incentive task. This result shows a significant decrease in dishonesty after the self-awareness induction and an increase in overall dishonesty when it is compared with the original version of the matrix puzzle game.

To delve deeper into this effect, we present a regression analysis in table 2 which supports the results obtained from the mean value comparison tests. The first two columns of the table represent linear regression results on the number of matrices reportedly solved in wave 1 and wave 2. The last two columns present the marginal effects from the probit regression on a dummy variable which takes the value of 1 if a subject sent a dishonest message in the sender-receiver game in wave 1. Model 1 includes only the main variables for both of the regressions whereas model 2 adds demographic control variables that were collected at the beginning of the experiment. In both models we merged the data from waves 1 and 2 and since only wave 1 features a sender-receiver game, the last two regressions do not include wave 1 variables or interactions. "Wave 1" is a dummy variable which takes the value of 1 if the observation is drawn from wave 1 of the experiment and 0 if it is drawn from wave 2. The "treatment" variable takes the value of 1 if the observation belongs to the treatment groups and 0 if it belongs to the control group. We treat the low incentive and high incentive tasks as one and include a dummy variable "incentive" which takes the value of 1 if the observation is from the high incentive version of the game and 0 if it is from the low incentive version. The regressions also include an interaction of the treatment variable with wave 1 to observe whether the effect of the treatment differs among the two different versions of the matrix puzzle game across waves and an interaction of the incentive variable with wave 1 to observe whether the effect of material incentives differs among waves.[6]

Column 1 of table 2 suggests that subjects in the treatment group report 2.051 fewer total matrices than subjects in the control group ($p < 0.01$) which is in line with Result 2 above. We also note an increase in the number of matrices reported in wave 2 as compared to wave 1. Subjects in wave 2 report 1.84 additional matrices on average than subjects in wave 1 ($p < 0.01$). This result supports Hypotheses 3 which states that since the induced level of cognitive dissonance (the psychological cost of lying) is lower in the wave 2 version of the matrix game, the level of dishonesty should be higher. The interaction of the treatment and wave variables is also significant at the 10% level. Combining the three significant variables in our analysis suggests that the level of dishonesty is significantly higher in the control group than treatment group by 0.731 units in wave 1 and by 2.051 units in wave 2. Moreover, dishonesty is significantly higher in wave 2 than in wave 1 by 0.621 units in the

[6]In an unreported regression, we added the interaction between the treatment variable with the incentive variable to the models in Table 2. However since the variable was not significant ($p > 0.10$) in any of the regressions, we did not include it in table 2.

treatment groups and by 1.84 units in the control group. Column 2 adds control variables to model 1: subjects in the treatment group report 1.932 fewer matrices than subjects in the control group ($p < 0.01$) and subjects in wave 2 report 1.558 more matrices solved than subjects in wave 1 ($p < 0.05$). The interaction of the treatment with the wave variable remains significant at the 1% level which suggest that the level of dishonesty is significantly higher in the control group than treatment group by 0.653 units in wave 1 and by 1.932 units in wave 2. Dishonesty is significantly higher in wave 2 than in wave 1 by 0.279 unit in the treatment groups and by 1.558 in the control group. Taken together this generates our next result which indicates a higher level of dishonesty in the matrix puzzle game in Wave 2 than Wave 1:

Result 3: The level of dishonesty is higher in Wave 2 than Wave 1 for both control and treatment groups.

Our design allows us to detect lying at the individual level if subjects report an infeasible number of solved matrices. In particular, some subjects reportedly solved more than 10 matrices in both version of the game which is not possible since there were only 10 matrices that included two numbers which summed to 10. While this is perhaps not as robust a distinction as in our more general results above (because individuals can lie while remaining within the bounds of feasibility) we can see some further support for our findings coming from this quarter, for example, our data reveals that 13.6% of the wave 1 control group (of 284 subjects) reportedly solved more than 10 matrices, with this number rising to 20.8% for those in wave 2 which are different at the 5% level if significance ($p < 0.028$). We make further use of this feature of the experiment later when we try to identify key characteristics of liars which requires lying to be detectable at the individual level.

We next move on to examining the role of additional material incentives on decision-making. The incentive variable and the interaction of incentive with the wave variable are not significant for the matrix puzzle game in any of the models. This suggests that increasing material benefits does not affect the level of dishonesty which is in line with the results in Mazar et al. (2008). Column 3 presents the marginal effects from a probit regression on the decision to send a dishonest or honest message in the sender-receiver game in wave 1. In both model 1 and model 2, the treatment effect is significant and positive suggesting that subjects in the treatment group are more likely to send a dishonest message than subjects in the control group ($p < 0.001$). As stated in Hypothesis 2, we observe an increase in dishonesty when subjects are induced with self-awareness. This is in line with our earlier argument that the sender-receiver game incorporates a relatively low psychological cost of lying attributable to cognitive dissonance. Moreover, as opposed to the results from the matrix puzzle game, material benefit *is* a significant determinant of whether to send a dishonest or honest message in this game. The results suggest that there is an increase in the probability of sending a dishonest message by 7.58% (7.73%) in the sample when the material benefit of lying is increased from \$0.2 to \$2.0. This result is consistent with the findings in Gneezy (2005). This leads us to our next result:

Result 4: Material incentives do not play a significant role in behaviour in the matrix puzzle game but do play a significant role for the sender-receiver game.

B Individual-level Results

Next we compare the detectable incidence of lying across tasks. We label someone as a "detectable liar" if they reported more than 10 solved matrices in the matrix puzzle game: something that cannot possibly be true. While most of the results above are based on group comparisons since we cannot generally know how many matrices an individual truly solved, this classification gives us access to individual level data and allows us to consider the characteristics of those who report more than 10 solved matrices. Figure 1 classifies the subjects as "never a detectable liar" if they did not report more than 10 solved matrices in any of the low or high incentive matrix puzzles in wave 1, "once a detectable liar" if they reported more than 10 matrices in only one of the matrix puzzles and "always a detectable liar" if they reported more than 10 matrices in both low and high incentive matrix tasks. Note that we cannot rule out that individuals lied even if they reported less than 10 solved matrices and that is the basis of the group level comparisons above which instead considers the average number solved.

FIGURE 1: Comparison of Liars Across Dishonesty Tasks

(a) Low Incentive **(b)** High Incentive

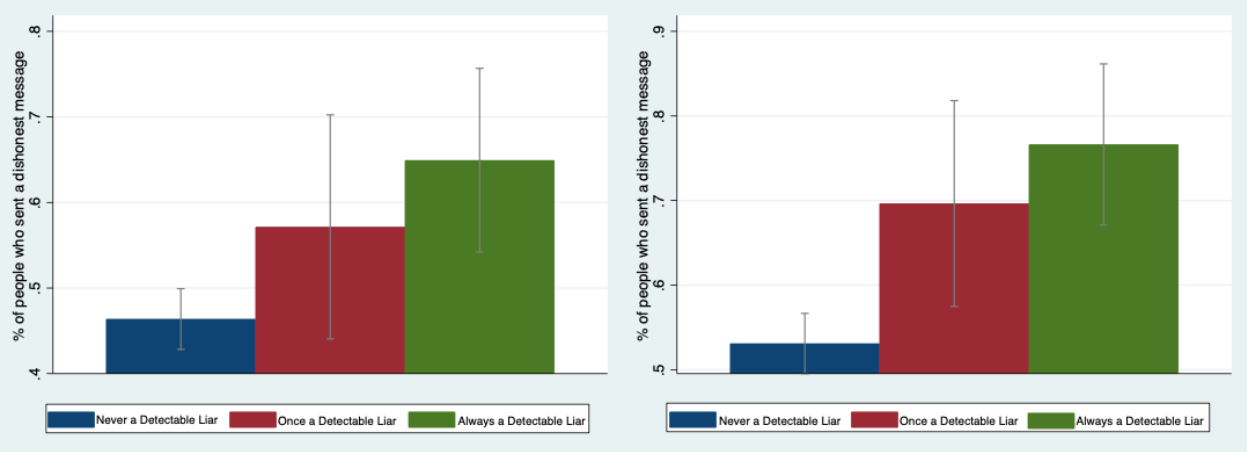

Figure 1 shows that the proportion of people who sent a dishonest message in the sender-receiver game increases with the frequency of reporting more than 10 matrices in the matrix puzzle game. 46.38% (53.10%) of subjects who are never detectable liars in the matrix puzzle games sent a dishonest message in the low (high) incentive sender-receiver game whereas 57.14% (69.64%) of subjects who were once detectable liars sent a dishonest message in the low (high) incentive sender-receiver game and that percentage shifts to 64.94% (76.62%) for those who are always a detectable liar. For the low incentive task, the confidence intervals and a two-sided t-test indicate a significant increase in the proportion of people who sent a dishonest message in the sender-receiver game from those who are never a detectable liar to those who are always a detectable liar ($p < 0.01$). For the high incentive task, this relationship is significantly different for those who are never a detectable liar as compared to the other two groups ($p < 0.01$) which suggests consistency across tasks.

TABLE 3: Who are the Liars?

	Sender-Receiver Game		
	(1)	(2)	(3)
	Never Lied	Lied Once	Always Lied
Dishonesty Variables			
No of matrix (Low Inc.)	4.583	5.183	5.9***
Detectable Liars % (Low Inc.)	7.82	9.36	16.86***
No of matrix (High Inc.)	4.547	5.306*	5.931***
Detectable Liars % (High Inc.)	5.54	11.49**	17.43***
Personality Variables			
BFI-Extraversion	0.261	0.296**	0.281
BFI-Conscientiousness	0.488	0.488	0.512**
BFI-Neuroticism	0.284	0.276	0.258*
BFI-Agreeableness	0.419	0.426	0.405
BFI-Openness	0.460	0.435*	0.448
Other Survey Variables			
Ethic Score	0.742	0.689**	0.716
Integrity Score	0.446	0.432	0.434
People take advantage	0.555	0.508**	0.481***
Amount to put in risky option	0.398	0.471**	0.447*
Extreme Sports	0.074	0.082	0.080
Amount to keep in UG	0.509	0.534	0.583***
Altruism Variables			
Donation to charity	0.189	0.178	0.100***
Donation to researcher	0.148	0.132	0.075***
No of observation	307	235	350

Mean values are represented in the table. A two-sided t-test is used where Column 2 and Column 3 are compared with Column 1, separately. Stars indicate statistical significance as follows: * p<0.10, ** p<0.05 and *** p<0.01.

Table 3 classifies subjects based on their dishonest behaviour in the two different dishonesty tasks and compares their personal and other characteristics. Subjects are classified as "never a liar" if they sent an honest message in both the low and high incentive versions of the sender-receiver game, "once a liar" if they sent only one dishonest message in either the low or high incentive task and "always a liar" if they sent a dishonest message in both. The table compares the "never a liar" group of subjects with the "once a liar" and "always a liar" groups separately and reports significance levels from a two-sided t-test. We note that people who sent more dishonest messages in the sender-receiver game report more matrices being solved in the matrix puzzle game. Subjects who always sent an honest message in the sender-receiver game report 4.583 (4.547) solved matrices on average in the low (high) incentive matrix task whereas subjects who sent a dishonest message once report 5.183 (5.306) solved matrices, and those who send a dishonest message twice report 5.9 (5.931) solved matrices. We also note that the proportion of subjects who report more than 10 matrices solved in the low (high) matrix puzzle game (the detectable liars) is higher for the groups who sent at least one dishonest message in the sender-receiver game than for the group of individuals who always sent an honest message. Putting this all together we have our next

result which confirms Hypothesis 5:

Result 5: A subject who lies in the sender-receiver game is more likely to be a detectable liar in the matrix puzzle game and vice-versa.

In order to check whether Hypothesis 6 (the moral balancing argument) holds, we compare the level of lying to our two donation variables. After the dishonesty tasks in both waves of the experiment, we ask subjects how much they would like to donate from their bonus to *MacMillan Cancer Support* and also how much they might wish to donate to the team of researchers to conduct more sessions in the experiment. We opted for two different measures in an attempt to disentangle reciprocity which seems more likely to apply to donations to the researcher, from general altruism. In the first instance we opted to be as open as possible to the possibility that moral balancing applies, biasing our results as heavily as we could in that direction: then if Hypothesis 6 fails we would have the strongest possible case against moral balancing. On that basis we consider subjects who lied the most in the sender-receiver game (in which lying is most easily detectable) and who donate more money *either* to the charity or to the researchers (despite the fact that this could entail reciprocity rather than moral balancing) to be attempting to balance out their dishonesty with a moral action. According to Table 3, while the mean level of donations made to charity is 18.9% for the group of people who did not send a dishonest message in the sender-receiver game, it is 17.8% for those who lied once and 10.0% for those who always lied. This amounts to a statistically significant decrease in the donation made to charity and to the researchers for the people who always sent a dishonest message in the sender-receiver game (in both cases $p < 0.01$). Our results suggest that the most extreme liars, rather than attempting to morally balance out their actions, instead donate *less* to the charity *and* to the researchers than people who did not lie in the sender-receiver game. Therefore, Hypothesis 6 (the moral balancing argument) is not supported by our data:[7]

Result 6: The moral balancing argument does not hold in our sample.

We also have a number of other tests and scales that we can use in an attempt to tease out any interesting further results though these do not link to any particular hypotheses. First we consider the results from a short version of the Big Five Inventory (Rammstedt & John 2007). Comparing the subjects who never sent a dishonest message in the sender-receiver game with the groups who sent a dishonest message once or twice, we do not observe any prevalent pattern in terms of personality. However, there are some important differences among groups. Subject who lied once in the sender-receiver game appear to be more extravert ($p < 0.05$) and open ($p < 0.10$) than subjects who never lied. Also, subjects who always sent a dishonest message in the sender-receiver game are more conscientious ($p < 0.05$) and less neurotic ($p < 0.10$) than subjects who never sent a dishonest message. Second, we asked subjects questions about the justifiability of unethical actions and created an ethics score which is higher if they answer that more of these actions are unjustifiable. We observe that people who lied once in the sender-receiver game have a lower ethics score than people

[7]Given the lower level of donations to the researchers, nor does there appear to be any reciprocity stemming from those who lied the most.

who never lied ($p < 0.05$). However, this relationship does not hold true for subjects who always lied. We also asked subjects whether they had engaged in some unethical actions such as avoiding public transport fares within the last 12 months and created an integrity score which is higher if they had engaged in fewer of these activities. The results do not show any significant difference between groups. In our test of risk aversion we observe that subjects who never lied in the sender-receiver game invest less money in the risky option, displaying more risk-averse behaviour, than the other two groups of subjects who lied once ($p < 0.05$) and twice ($p < 0.10$). They also believe more in the idea that people are likely to be fair rather than trying to take advantage than the other two groups. Finally, subjects who always lied in the sender-receiver game retained more money in the ultimatum game than subjects who opted never to lie in the sender-receiver game ($p < 0.01$).

Finally, we can also check our findings relating to the demand effect. The demand effect is supposedly driven by a feeling of reciprocity towards the researcher: a desire to help them in their research by behaving as expected (Zizzo 2010, Quidt et al. 2018). To help us understand if this might play a role in explaining our findings, after the dishonesty tasks we asked subjects to respond to two questions.[8] The first question asks: "If someone realizes they have done something dishonest, how likely is it that they will behave more or less honestly in the future?". We refer to the answer to this question as "Subject's Expectation" in table 4. The second question asks: "What is the percentage chance that the experimenter expected you to behave honestly in the various tasks you had to undertake?". We refer to the answer to this question as "Researcher's Expectation" in table 4. Subjects selected from a scale with any number below 50% reported by the subjects indicating more dishonest behaviour and above 50% indicating more honest behaviour.

TABLE 4: Mean Value Comparisons of Demand Effect Variables

	Wave 1		**Wave 2**	
	Subject's Expectation	*Researcher's Expectation*	*Subject's Expectation*	*Researcher's Expectation*
Mean Values				
Control Group	59.736	65.866	73.703	82.307
Honesty Treatment	53.18	67.273	54.842	61.961
Low Dishonesty Tr.	56.317	67.736	60.260	67.740
High Dishonesty Tr.	54.856	60.344	56.851	67.207
Treatment Group	54.791	65.209	57.607	65.921
T-test[1]				
Control Group	0.001***		0.000***	
Treatment Group	0.000***		0.000***	

[1] p-values from a two-tailed t-test are reported where the null hypothesis is Researcher's Expectation=Subject's Expectation.
* p<0.10, ** p<0.05 , *** p<0.01

As reported in table 4, subjects in all treatment and control groups and in both waves on average expected that if someone had done something dishonest in the past, they would behave more honestly in the future. Subjects also reported on average that researchers expected them to behave more honestly in the future after being dishonest in the past. At the same time, subjects think that researchers expect significantly more honest behaviour than subjects think they will behave ($p < 0.001$). However, our results indicate that subjects

[8] We are grateful to Chris Roth for early discussions about how to best deal with the potential demand effect at an early stage in the design of the experiment.

do not follow either their own average expectation about behaviour, or their beliefs about the researcher's more optimistic expectations about behaviour. In particular, a demand effect should push up the incidence of honesty in all of our tasks for those in treatment groups, however this was not the case. Moreover, our design also includes an even more direct test of the demand effect. Our design admits the possibility of direct reciprocity towards the researcher through a donation. As described above in result 5, subjects who behaved more dishonestly in the sender-receiver game donated significantly less to the researcher than the control group. In other words, even where subjects had a direct opportunity to donate in situations when they behaved more dishonestly they did not do so. Recall that this is exactly the situation in which subjects declared that they believed that honesty would rise: and so not only did honesty fall, but so too did direct reciprocity towards the researcher.

VII Conclusion

What happens when individuals become more aware of their own dishonesty? We asked this question in the introduction, and our results seem to give a clear but subtle answer: it depends upon the context. In a competitive in which subjects could earn more by lying, inducing them to think more about honesty pushes them towards become more dishonest. This might be viewed as a form of acceptance, spurred on by the psychological costs of cognitive dissonance, or attempting to hold two very different views at the same time. On the other hand, in a single-player task self-awareness reduces dishonesty. This is true when we compare across both games in wave 1 of our experiment, and true when we purposefully make one of our tasks more competitive in wave 2.

Attempting to derive policy implications from our findings we would urge caution when attempting to use self-awareness as a weapon against dishonesty. This can easily backfire especially in competitive contexts, forcing dishonest individuals to accept their true nature in order to avoid cognitive dissonance and making it even easier for them to behave dishonestly in the future. Competitive contexts do of course exist in a variety of different important real-world situations: politicians attempting to defeat or outmaneuver their political opponents by wining elections, celebrities trying to garner more popularity and fame than their rivals, or newspapers attempting to drive competitors out of business. In each of these cases, "naming and shaming" may still be useful in highlighting the dishonesty of others but we cannot assume that it will also bring with it a period of moral balancing by those caught out in a lie.

References

Abeler, J., Becker, A. & Falk, A. (2014), 'Truth-telling: A representative assessment', *Journal of Public Economics* **113**, 96–104.

Ariely, D. (2009), *Predictably Irrational: The hidden forces that shape our decisions*, Harper, New York.

Ariely, D. (2012), *The (Honest) Truth about Dishonesty: How we lie to everyone - especially ourselves*, Harper, New York.

Benabou, R. & Tirole, J. (2016), 'Mindful economics: The production, consumption, and value of beliefs', *Journal of Economic Perspectives* **30 (3)**, 141–164.

Bucciol, A. & Piovesan, M. (2011), 'Luck or cheating? a field experiment on honesty with children', *Journal of Economic Psychology* **32**, 73–78.

Cohn, A., Marechal, M. A. & Noll, T. (2015), 'Bad boys: How criminal identity salience affects rule violation', *The Review of Economic Studies* **82(4)**, 1289–1308.

Cohn, A., Marechal, M. A., Tannenbaum, D. & Zünd, C. L. (2019), 'Civic honesty around the globe', *Science* **365**, 70–73.

Diener, E. & Wallbom, M. (1976), 'Effects of self-awareness on antinormative behavior', *Journal of Research in Personality* **10**, 107–111.

Erat, S. & Gneezy, U. (2012), 'White lies', *Management Science* **58(4)**, 723–733.

Fenigstein, A. & Levine, M. P. (1984), 'Self-attention, concept activation and the causal self', *Journal of Experimental Social Psychology* **20(3)**, 231–245.

Fischbacher, U. & Föllmi-Heusi, F. (2013), 'Lies in disguise: An experimental study on cheating', *Journal of European Economic Association* **11**, 525–547.

Franzen, A. & Pointner, S. (2013), 'The external validity of giving in the dictator game: A field experiment using the misdirected letter technique', *Experimental Economics* **16**, 155–169.

Gino, F., Norton, M. I. & Weber, R. A. (2016), 'Motivated bayesians: Feeling moral while acting egoistically', *Journal of Economic Perspectives* **30 (3)**, 189–212.

Gneezy, U. (2005), 'Deception: The role of consequences', *American Economic Review* **95(1)**, 384–394.

Gneezy, U. & Potters, J. (1997), 'An experiment on risk taking and evaluation periods', *Quarterly Journal of Economics* **112(2)**, 631–645.

Houser, D., Vetter, S. & Winter, J. K. (2012), 'Fairness and cheating', *European Economic Review* **56**, 1645–1655.

Levit, S. D. (2006), 'White-collar crime writ small: A case study of bagels, donuts and the honor system', *Academic Economic Review* **96**, 290–294.

Mazar, N., Amir, O. & Ariely, D. (2008), 'The dishonesty of honest people: A theory of self-concept maintenance', *Journal of Marketing Research* **45 (6)**, 633–644.

Mazar, N. & Zhong, C. (2010), 'Do green products make us better people?', *Psychological Science* **21**, 494–498.

Mead, N. L., Baumeister, R. F., Gino, F., Schweitzer, M. E. & Ariely, D. (2009), 'Too tired to tell the truth: Self-control resource depletion and dishonesty', *Journal of Experimental Social Psychology* **45(3)**, 594–597.

Ploner, M. & Regner, T. (2013), 'Self-image and moral balancing: An experimental analysis', *Journal of Economic Behavior and Organization* **93**, 374–383.

Pruckner, G. J. & Sausgruber, R. (2013), 'Honesty on the streets. a field study on newspaper purchasing', *Journal of the European Economic Association* **45**, 661–679.

Quidt, J., Haushofer, J. & Roth, C. (2018), 'Measuring and bounding experimenter demand', *American Economic Review* **108**, 3266–3302.

Rabin, M. (1994), 'Cognitive dissonance and social change', *Journal of Economic Behaviour and Organization* **23**, 177–194.

Rammstedt, B. & John, O. P. (2007), 'Measuring personality in one minute or less: A 10-item short version of the big five inventory in english and german', *Journal of Research in Personality* **41**, 203–212.

Rosenbaum, S. M., Billinger, S. & Stieglitz, N. (2014), 'Let's be honest: A review of experimental evidence of honesty and truth-telling', *Journal of Economic Psychology* **45**, 181–196.

Shu, L., Mazar, N., Gino, F. & Bazerman, M. (2011), 'When to sign on the dotted line? signing first makes ethics salient and decreases dishonest self-reports', *Harvard Business School: Negotiation, Organizations Markets Unit Working Paper Series* **45 (6)**, 633–644.

Stoop, J. (2014), 'From the lab to the field: Envelopes, dictators and manners', *Experimental Economics* **17**, 304–313.

Sutter, M. (2009), 'Deception through telling the truth?! experimental evidence from individuals and teams', *The Economic Journal* **119**, 47–60.

West, M. D. (2005), *Law in everyday Japan: Sex, sumo, suicide and statues.*

Whiteley, P. (2012), 'Are britons getting more dishonest'.
URL: *http://www.west-info.eu/files/Research11.pdf*

Yezer, A. M., Goldfarb, R. S. & Poppen, P. J. (1996), 'Does studying economics discourage cooperation? what what we do, not what we say or how we play.', *Journal of Economic Perspectives* **10**, 177–186.

Zizzo, D. (2010), 'Experimenter demand effects in economic experiments', *Experimental Economics* **13**, 75–98.

Appendix A

TABLE 5: Timeline of the Events

	Control Group	**Honesty Treatment**
First Stage	Demographic Questionnaire, Ethic and Integrity Questionnaire[1], Risk preference[2], Ultimatum Game, the Big Five	
Second Stage	-	Writing about an event in their own life which involves complete honesty
Third Stage	Dishonesty Tasks[3], Dictator Game, Demand Effect questions	Dishonesty Tasks[1], Dictator Game, Demand Effect questions
	Low Dishonesty Tr.	**High Dishonesty Tr.**
First Stage	Demographic Questionnaire, Ethic and Integrity Questionnaire[1], Risk preference[2], Ultimatum Game, the Big Five	
Second Stage	Writing about an event in their own life in which they decided not to be completely honest to benefit themselves but it did not ended up harming someone else	Writing about an event in their own life in which they decided not to be completely honest to benefit themselves but it ended up harming someone else
Third Stage	Dishonesty Tasks[3], Dictator Game, Demand Effect questions	Dishonesty Tasks[1], Dictator Game, Demand Effect questions

[1]The integrity questionnaire is taken from Whiteley (2012). [2]To elicit preferences towards risk we follow the method outlined in Gneezy & Potters (1997). [3]Honesty tasks include the Matrix Puzzle and the Cheap Talk Sender-Receiver Game in wave 1 and the modified Matrix Puzzle in wave 2. Subjects complete these tasks in a randomized order.

TABLE 6: Descriptive Statistics - Wave 1 and Wave 2

	Wave 1				Wave 2			
	Control		*Treatment*		*Control*		*Treatment*	
	Mean	Std.Dev.	Mean	Std.Dev.	Mean	Std.Dev.	Mean	Std.Dev.
Demographic Variables								
Age	38.15	11.66	38.89	12.05	36.78	10.05	36.72	10.60
Female	0.44	0.50	0.45	0.50	0.34	0.47	0.48	0.50
American	0.98	0.14	0.97	0.17	0.91	0.29	0.99	0.09
College degree or more	0.63	0.48	0.59	0.49	0.80	0.40	0.71	0.45
Married	0.60	0.49	0.49	0.50	0.76	0.43	0.50	0.50
Other Variables								
Amount to put in safe option	0.57	0.39	0.56	0.36	0.66	0.56	0.58	0.35
Engaging in extreme sports	0.10	0.15	0.07	0.13	0.23	0.20	0.09	0.13
Amount to keep for yourself (UG)	0.54	0.19	0.55	0.21	0.55	0.22	0.54	0.18
BFI-Extraversion	0.28	0.17	0.28	0.18	0.30	0.14	0.29	0.18
BFI-Conscientiousness	0.48	0.15	0.50	0.15	0.41	0.15	0.49	0.15
BFI-Openness	0.43	0.16	0.46	0.15	0.37	0.13	0.45	0.17
BFI-Agreeableness	0.40	0.16	0.42	0.16	0.39	0.14	0.43	0.16
BFI-Neuroticism	0.29	0.18	0.26	0.19	0.30	0.14	0.27	0.19
People take advantage of others?[1]	0.53	0.24	0.51	0.22	0.63	0.24	0.52	0.22
Integrity score[2]	0.42	0.11	0.44	0.09	0.35	0.13	0.42	0.10
Ethics score[3]	0.67	0.31	0.74	0.28	0.44	0.34	0.71	0.31
Donation to charity (%)	19.32	24.86	13.11	23.02	33.30	21.62	15.36	23.11
Donation to researcher (%)	16.20	23.55	9.30	19.58	31.05	21.16	12.46	20.30
Observations	284		608		101		267	

[1] A higher number means fairer and less advantageous. [2] The integrity score is created using participants' ratings on 15 dishonest actions. [3] The ethics score is created using participants' self-reports on whether they had engaged in various unethical actions. More explanation about how these variables are defined can be found in the main text.

Appendix B - Experimental Instructions

Participation Agreement

You have been invited to take part in a research study run by researchers at the University of Warwick. Please read the following statements carefully and answer the question below.

Our commitments and privacy policy:

We never deceive participants. For example, if we inform you that another participant is making a choice on which you can then react, this is indeed the case. We keep our promises made to participants. For example, if we promise a certain payment, participants will indeed receive it. In the event that we are responsible for a mistake that is to the disadvantage of participants, we will inform and compensate the respective participants. We design, conduct and report our research in accordance with recognized scientific standards and ethical principles.

We adhere to the terms of our privacy policy as stated below:

The data in the participants' database will only be used for the purpose of the study. There is no link between the personal data in the participants' database and the data collected during a study. The generated anonymous data will be used for analysis. The end product will be publicly available. Your participation in this study is purely voluntary, and you may withdraw your participation or your data at any time without any penalty to you. Please note that the software (Qualtrics) automatically notes the time you spent on each question and this data will be made available to researchers for analysis.

If you would like to make a complaint about the way you have been dealt with during the study or any possible harm you might have suffered please address your complaint to the person below, who is a senior University of Warwick official entirely independent of this study:

Head of Research Governance,
Research Impact Services,
University House, University of Warwick,
Coventry CV4 8UW
Tel: 024 76 522746
Email: researchgovernance@warwick.ac.uk

If you are happy to proceed please tick the "I agree" button below to continue.

First Stage

Demographic Questionnaire

Please answer the following questions.
Age:
Gender:
Marital status:
Highest educational attainment:
Nationality:
English as a native language:

Do you think most people would try to take advantage of you if they got a chance, or would they try to be fair? 1 means that "people would try to take advantage of you," and 10 means that "people would try to be fair" :
Please write "purple" if your favourite colour is asked later on this study.

Ethic Questionnaire

Which of these things, if any, have you done in the past 12 months?
i) Avoided a fare on public transport
ii) Made something up on a job application
iii) Downloaded music or videos without paying for them
iv) Called in sick to work/ to school when not actually unwell

How often do you participate in extreme sports? (Extreme sports include bungee-jumping, para-gliding, parachute jumping, gliding, rafting, diving and other dangerous sports.) :
What is your favourite colour according to the statement written before in this study?

Integrity Questionnaire

	Always justified	Sometimes justified	Rarely justified	Never justified
Claiming government benefits to which you are not entitled	○	○	○	○
Buying something which you know it is stolen	○	○	○	○
Taking cannabis	○	○	○	○
Keeping money that you found in the street	○	○	○	○
Lying in your own interest	○	○	○	○
Having an affair when you are married	○	○	○	○
Having sex under the legal age of consent	○	○	○	○
Failing to report accidental damage you have done to a parked vehicle	○	○	○	○
Throwing away litter in a public place	○	○	○	○
In order for us to check you are reading instructions, please select "Always justified" for this statement.	○	○	○	○
Driving under the influence of alcohol	○	○	○	○
Avoiding a fare on public transport	○	○	○	○
Cheating on taxes if you have a chance	○	○	○	○
Someone accepting a bribe in the course of their duties	○	○	○	○
Driving faster than the speed limit	○	○	○	○
Making up things on a job application	○	○	○	○

Bonus Payments: The next few tasks involve the chance for you to win a bonus payment. One of them will be selected at random and depending upon your choices and which question is selected you stand the chance to win a bonus. The precise nature of the bonus will be made explicit during each task but please remember that only one of these tasks will end up

paying out a bonus.

Risk Preference

For your next task please consider the following scenario. You have been endowed with $1. You are asked to allocate this amount among 2 options: Option A and Option B. The amount you put in Option A will stay as it is (the value of the money you placed in this option will not change). The amount you put in Option B will be determined by the following rule:

A random whole number between 1 and 6 will be generated. If the number is less than or equal to 3, the amount you put in Option B will be multiplied by 2. If the number is greater than 3, it will be 0 (zero).

Your potential bonus from this task will be the sum of the amount you put in Option A and the final amount resulted in Option B. Please indicate how you would like to allocate the $1 among Options A and B.

Ultimatum Game

For your next task, imagine that you are randomly matched with another participant in this experiment. You are Player 1. You need to decide how to allocate $2 between yourself and Player 2. You need to offer an allocation. If Player 2 rejects your offer, both players will receive a potential bonus of $0. If Player 2 accepts the offer, both players will receive a potential bonus according to the allocation you offered.
Please select how to allocate $2 between yourself and the other player.

Now, you are assigned to the role of Player 2 and randomly matched with another participant Below, you will see different allocations offered by Player 1. If you reject the offer, both players will receive a potential bonus of $0. If you accept the offer, both players will receive a potential bonus according to the allocation Player 1 offered.

Big Five

Please indicate how well do the following statements describe your personality.

I see myself as someone who...

	Disagree strongly	Disagree a little	Neither agree nor disagree	Agree a little	Agree strongly
is reserved	○	○	○	○	○
is generally trusting	○	○	○	○	○
tends to be lazy	○	○	○	○	○
is relaxed, handles stress well	○	○	○	○	○
has few artistic interests	○	○	○	○	○
is outgoing, sociable	○	○	○	○	○
tends to find fault with others	○	○	○	○	○
does a thorough job	○	○	○	○	○
gets nervous easily	○	○	○	○	○
has an active imagination	○	○	○	○	○
In order for us to check you are reading instructions, please select "Agree a little" for this statement.	○	○	○	○	○

Second Stage

Honesty Treatment

For your first task we would like to ask you to write about a real life event. Please think about an event in your own life (preferably in the last 12 months) in which you decided to be completely honest. We would like to ask you to write about this event below.

Low Dishonesty Treatment

For your first task we would like to ask you to write about a real life event. Please think about an event in your own life (preferably in the last 12 months) in which you decided not to be completely honest in order to benefit yourself, but where you felt that this dishonesty did not harm anyone else. We would like to ask you to write about this event below.

High Dishonesty Treatment

For your first task we would like to ask you to write about a real life event. Please think about an event in your own life (preferably in the last 12 months) in which you decided not to be completely honest in order to benefit yourself, and where this dishonesty ended up harming someone else (a little or lot). We would like to ask you to write about this event below.

Third Stage

Matrix Task

In the next page, you will be a given an image that consists of 20 matrices . Your task is to find two numbers that add up to 10 in these matrices. The potential bonus you can make in this task is \$0.10 (\$0.30) per correct answer. You will have 5 minutes to complete this task. Once you press the button to continue, your time will start. You will see a timer on top of your screen. Once your time is over, you will be directed to the next page where you need to report how many pairs of numbers you had found. Once you are ready please press the button to proceed to the matrix task.

Please write how many pairs of numbers you had found that sum up to 10.

Cheap Talk Sender Receiver Game

In this task, you are asked to imagine yourself in the following situation:
You are matched with another MTurk worker. Neither of you will ever know the identity of the other. Two possible monetary payments are available to you and your counterpart and are described as "Option A" or "Option B". The choice rests with the other MTurk worker who will have to choose either Option A or Option B and this will determine how much money you make and how much they will make. However, the only information your counterpart will have about the money that can be made is information sent by you in a message. That is, he or she will not know the monetary payments associated with each choice. You are the only one who knows about the monetary payments associated with each option. The potential bonus payments for this task depend upon the actual amounts specified in the two options: option A and option B and what is selected by the other MTurk worker. You need to decide which message to sent to your counterpart.

Which message would you send if the two payment options were:
Option A: \$1 to you and \$1.2 to the other player.
Option B: \$1.2 to you and \$1 to the other player.

i) Message 1 : "Option A will earn you more money than option B."
ii) Message 2: "Option B will earn you more money than option A."

Which message would you send if the two payment options were:
Option A: \$1 to you and \$3 to the other player.
Option B: \$3 to you and \$1 to the other player.

i) Message 1 : "Option A will earn you more money than option B."
ii) Message 2: "Option B will earn you more money than option A."

Modified Matrix Puzzle

In the next page, you will be a given an image that consists of 20 matrices. Your task is to find two numbers that add up to 10 in these matrices. The potential bonus you can make in this task will be decided by your ranking among all participants.If you are in the top 50% of the distribution, then you will receive \$0.72 (\$2.13). You will have 5 minutes to complete this task. Once you press the button to continue, your time will start. You will see a timer on top of your screen. Once your time is over, you will be directed to the next page where you need to report how many pairs of numbers you had found. Once you are ready please press the button to proceed to the matrix task. Please remember the number of pairs you had found at the end of the task.

Please write how many pairs of numbers you had found that sum up to 10.

Dictator Games

We would like you to indicate what percentage of your actual bonus from this experiment you would like to donate to "MacMillan Cancer Support". If you decide to donate any percentage of your bonus, the donation will be done by us anonymously to the charitable organisation.
The percentage that I want to keep it for myself :
The percentage that I want to donate to the charity:
What percentage of your actual bonus from this study you would like to give up for researchers to use to conduct more sessions of this experiment.
The percentage that I want to keep it for myself :
The percentage that I want to leave it for researchers :

Demand Effect Questions

If someone realizes they have done something dishonest, how likely is it that they will behave more or less honestly in the future? The slider below indicates the percentage chance of being more honest, moving from 0% (certainly more dishonest) on the left to 100% (certainly more honest) on the right. Please move the slider to the percentage chance that you think is correct.
What is the percentage chance that the experimenter expected you to behave honestly in the various tasks you had to undertake? The slider below indicates the percentage chance of being more honest, moving from 0% (certainly dishonest) on the left to 100% (certainly honest) on the right. Please move the slider to the percentage chance that you think is correct.