BRIBING THE SELF*

Uri Gneezy
Silvia Saccardo
Marta Serra-Garcia
Roel van Veldhuizen

January 14, 2020

Abstract

Expert advice is often biased in ways that benefit the advisor. We demonstrate how self-deception helps advisors be biased while preserving their self-image as ethical and identify limits to advisors' ability to self-deceive. In experiments where advisors recommend one of two investments to a client and receive a commission that depends on their recommendation, we vary the timing at which advisors learn about their own incentives. When advisors learn about their incentives before evaluating the available investments, they are more likely to be biased than when they learn about their incentives only after privately evaluating the investments. Consistent with self-deception, learning about the incentive before evaluating the options affects advisors' beliefs and preferences over the investments. Biased advice persists with minimal justifications but is eliminated when all justifications are removed. These findings show how self-deception can be constrained to improve advice provision.

JEL classification: D03, D83, C91.

Keywords: Advice, Self-Deception, Self-Image, Motivated Beliefs, Laboratory Experiment.

*Gneezy: University of California San Diego, 9500 Gilman Drive, La Jolla, CA 92093 and University of Amsterdam (email: ugneezy@ucsd.edu). Saccardo: Carnegie Mellon University, Department of Social and Decision Sciences, 5000 Forbes Avenue, BP 208, Pittsburgh, PA 15213 (email: ssaccard@andrew.cmu.edu) and CESifo. Serra-Garcia: University of California San Diego, 9500 Gilman Drive, La Jolla, CA 92093 (email: mserragarcia@ucsd.edu) and CESifo. Van Veldhuizen: Lund University, Scheelevägen 15B, 223 63 Lund, Sweden (email:roel.van_veldhuizen@nek.lu.se) and WZB Berlin Social Science Center. This research was conducted under IRB#140762 and benefited from financial support of internal funds at Rady School of Management UCSD and the Handelsbanken foundation. We thank Linda Babcock, Aislinn Bohren, Russell Golman, Alex Imas, George Loewenstein, Justin Sydnor, Joël van der Weele, Peter Schwardmann, Florian Zimmermann, and participants at various seminars, the SITE Experimental Economics, the World Congress of the Econometric Society, the Early-Career Behavioral Economics Conference, the Carnegie-Mellon Belief-Based Utility Conference and the Penn State Moral Psychology Research Group conference, among others, for their comments and suggestions.

1 Introduction

We rely on experts' recommendations for decision-making in a variety of domains, including medical, legal and financial decisions. A common feature of these domains is the informational asymmetry between the expert and the consumer: experts are more knowledgeable about the quality of the goods or services than consumers, and consumers cannot fully assess whether the advice they received was in their best interest (see the literature regarding credence goods; e.g., Darby and Karni, 1973). This informational asymmetry may generate incentives for experts to provide advice that is in their own best interest, and not the clients'.[1] However, being dishonest towards clients may lead to a conflict between advisors' material goals and their desire to maintain a self-image as honest (e.g., Akerlof and Kranton, 2000; Bénabou and Tirole, 2016; Abeler, Nosenzo and Raymond, 2019; Bénabou, Falk, and Tirole, 2018). To attenuate this tension, advisors may self-deceive by distorting their beliefs, convincing themselves that their advice is ethical. In this paper, we study the constraints to advisors' ability to self-deceive, and test the effect of progressively reducing the scope for justifying biased advice as ethical.

Consider, for example, financial advice. Advisors often recommend products for which they are directly compensated (e.g., Anagol, Cole and Sarkar, 2017), and financial advice could even hurt the clients in some cases (Bergstresser, Chalmers and Tufano, 2009; Hackethal, Haliassos and Jappelli, 2011; Chalmers and Reuter, 2012). Some advisors may ignore their own incentives and give an unbiased recommendation to their clients. Others may knowingly bias their recommendations in order to increase their own profits and conceal this behavior to preserve a positive reputation or social-image. Yet, given that financial advice is partially subjective, some advisors may be able to convince themselves that investment recommendations that benefit them financially are actually in the client's best interest, thereby preserving their positive self-image.

In medical decisions, overtreatment is estimated to cost $210 billion in wasteful annual spending in the US (Institute of Medicine, 2013). A partial reason for overtreatment is that doctors recommend unnecessary procedures for which they are directly compensated

[1] See also, Cain, Loewenstein and Moore (2011); Dulleck, Kerschbamer and Sutter (2011); Balafoutas, Beck, Kerschbamer and Sutter (2013); and Dulleck and Kerschbamer (2006) for a review of the literature.

(Emanuel and Fuchs, 2008; Clemens and Gottlieb, 2014). Specific examples include the growing number of surgeries in response to back pain, many of which have been shown to be unnecessary and even harmful (Mafi et al., 2013), or doctors who recommend unneeded C-sections for birth delivery when such procedures are financially compensated (Gruber et al., 1999; Johnson and Rehavi, 2015). DeJong et al. (2016) show that doctors who receive payments from the medical industry tend to prescribe drugs differently than their colleagues who do not. When a reporter asked doctors to comment on the finding,[2] "several doctors who received large payments from industry and had above-average prescribing rates of brand-name drugs said they are acting in patients' best interest" (see also, Sharek, Shoen and Loewenstein, 2012).

In this paper, we study advisors' ability to engage in self-deception to preserve their self-image as ethical, convincing themselves that advice that maximizes their material gains is also the advice the client would prefer. A growing body of work has shown that individuals are likely to engage in self-interested behavior in presence of ambiguity or subjectivity, which allows them to preserve a positive social or self-image (e.g., Kunda, 1990; Konow, 2000; Dana et al., 2007; Haisley and Weber, 2010; Di Tella et al., 2014; Exley, 2015; Grossman and van der Weele, 2017; Gneezy, Saccardo and van Veldhuizen, 2018; Zou, 2018; Bicchieri, Dimant and Sonderegger, 2019). An important open question in the literature regards the factors constraining people's ability to preserve desirable beliefs about their actions while behaving unethically. Individuals cannot simply choose to believe what they want to in all circumstances (Epley and Gilovich, 2017). To better understand the factors constraining self-deception, we systematically reduce advisors' ability to self-deceive and measure the size of the bias in advice when advisors can find multiple, minimal or no justifications for their recommendations.

We first report the results of three laboratory experiments. In each experiment, advisors are tasked with recommending one of two investment options, A and B, and are allocated to one of three treatments. In the first treatment, the advisor is incentivized to recommend A and is told about the incentive before she is presented with the options she

[2] https://www.propublica.org/article/doctors-who-take-company-cash-tend-to-prescribe-more-brand-name-drugs.

needs to consider (the Before treatment). In the second, the advisor is first presented with the two options and asked to consider which she would recommend, and only after privately considering the two options she is told about her incentive to recommend A (the After treatment). The third treatment is a Control treatment in which there is no incentive to recommend A.

By varying the timeline of information about incentives we identify self-deception and show that learning about the incentives after evaluating the investment options constrains advisors' ability to self-deceive. If the advisor is informed about the incentives before having a chance of evaluating the investments, she might inadvertently distort her beliefs by convincing herself that the recommendation she is incentivized to make is also the one the client would prefer. If she instead privately evaluates the investments before learning about her incentives, engaging in self-deception becomes harder. The advisor may no longer be able to recommend the incentivized investment while preserving a positive self-image. The before versus after manipulation builds on Babcock et al. (1995), who use a timing manipulation to study self-serving biases in bargaining. An important difference is that in Babcock et al. (1995) participants' initial evaluations were always observed by the experimenter. Our design in which evaluations of the investment options only take place in the advisor's mind allows us to directly manipulate self-image concerns while keeping social-image and social desirability concerns constant across treatments. By doing so, we contribute to previous work on self-serving behavior under ambiguity, which did not disentangle whether individuals' tendency to engage in such behavior in ambiguous or subjective situations is driven by an increased ability to justify their choices to themselves (self-image concerns) or to others (social-image concerns).

If advisors' ability to self-deceive is crucial to advice provision, and this ability is constrained, then recommendations may no longer be biased when the advisor learns about her incentives after evaluating the investment options. To investigate how much bias remains in the After treatment, we include the Control treatment, in which the advisor no longer has an incentive to recommend A. If recommendations when the advisor has no incentive to self-deceive (Control treatment) are similar to those when the advisor's ability to self-deceive is constrained (After treatment), this would be evidence that a timing manipulation (delaying the information about incentives) can remove the bias in recommendations.

Across the three laboratory experiments, we test a constraint to self-deception by investigating how the order of information about incentives affects advice when the relative value of the investment options to the client changes. In the RiskReturn experiment, there is a risk-return tradeoff between A and B: A has lower variance but also lower expected return than B. In addition, advisors may convince themselves that the sure commission they obtain from recommending A justifies the small loss to the client (in expectation). In this experiment, advisors therefore have two different ways to convince themselves that A is the better option. In the Dominance experiment, we change investment B to eliminate the risk-return tradeoff between the two investments while keeping the difference in expected return between A and B unchanged. Advisors in this experiment can no longer use the risk-return tradeoff to justify their recommendations. However, the trade-off between their commission and the client's expected loss from choosing A over B is identical to RiskReturn. Finally, in the ObviousDominance experiment, we change investment B such that the client's expected loss from choosing A over B is significantly higher, making it much harder to justify such a recommendation. In this setting, advisors should have a significantly reduced scope for self-deception even when evaluating the investment options before learning about the commission, since they have a smaller scope or no scope for rationalizing a recommendation of the strictly inferior investment.

The data from the laboratory experiments show that, in the RiskReturn experiment, advisors recommend the incentivized investment in over 60% of the cases when they receive information about incentives before seeing the options (Before treatment). Consistent with the importance of self-image, the incentivized investment is only recommended 33% of the time in the After treatment, where self-deception is harder. In the Dominance experiment we find that advisors still recommend A at a higher rate when learning about the commission before evaluating the investment options (53% versus 25% when advisors learn about the commission after), suggesting that self-deception can arise even with a minimal justification. However, in the ObviousDominance experiment, the gap between the Before and After treatments closes: the rate of investment A recommendations is about 30% in both treatments. This result implies that all justifications have to be removed to eliminate self-deception in this setting.

In three additional online experiments, we provide some evidence of the mechanisms driving self-deception. We replicate the effect of the Before and After treatments on advisors' recommendations in the RiskReturn, Dominance and ObviousDominance experiments. We also measure advisors' beliefs when they first consider the investments and study their choices when asked to select one of the two investments as a reward for themselves (Chen and Gesche, 2018).[3] Consistent with self-deception, in the Before treatment advisors' beliefs about the recommendation the client would prefer are distorted toward the incentivized investment, when there is scope for self-deception. This distortion becomes smaller when the scope for self-deception is reduced (the After treatments). This bias also spills over into advisors' personal investment decisions: advisors in both the RiskReturn and Dominance experiments are more likely to select investment A as an investment for themselves in the Before treatment than in the After treatment.

Taken together, our findings suggest that advisors distort their beliefs to enable themselves to recommend the investment associated with a commission. Our results highlight some of the constraints in people's ability to preserve desirable beliefs about their actions, and show that, in order to decrease the rate of distorted recommendations, any scope for self-deception must be removed. Understanding the constraints of self-deception is important because advisors who self-deceive may actually be more persuasive, as argued by Trivers (2011) and shown by Schwardmann and van der Weele (2019). Therefore, our findings can have important implications for the design of regulations, and for mitigating the effect of commissions on advice. We show that interventions aimed at reducing the scope for self-deception by reducing subjectivity in judgment or by strengthening the self-image costs that would arise from biased recommendations need to be carefully designed to lessen the extent of dishonest advice.

[3] We thank the referees, Associate Editor and Editor for encouraging us to run these additional experiments.

2 Experimental Design

2.1 The Advice Game

We study a sender-receiver game in which the sender ("advisor") is informed about the details of two investment opportunities, A and B, and is asked to send a recommendation to an uninformed receiver (the "client") regarding which of the two to choose. This game differs from standard sender-receiver games (e.g., Crawford and Sobel, 1982) in that the sender is asked to make a judgment instead of reporting an objective piece of information, such as the state of nature.

The experiments have three treatments, which modify the basic game as displayed in Figure 1. In the Control treatment, advisors receive no additional payment for recommending A or B. In the Before and After treatments, the advisor receives an additional commission of \$1 if she recommends A. The key difference between the Before and After treatments is *when* the advisor is first informed about the additional payment. In the Before treatment, the advisor learns this information before learning the details of the investments. By contrast, in the After treatment, the advisor learns about the commission only after reading about the investments and having already thought about her recommendation, but before making the recommendation.

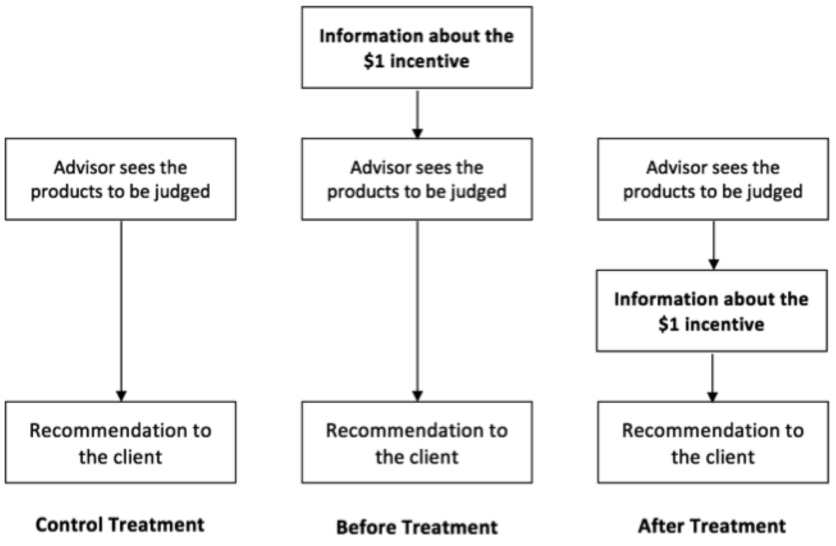

Information about the $1 incentive
Advisor sees the products to be judged
Advisor sees the products to be judged
Advisor sees the products to be judged
Information about the $1 incentive
Recommendation to the client
Recommendation to the client
Recommendation to the client
Control Treatment
Before Treatment
After Treatment

Figure 1: Timeline of the Experimental Treatments

We test the constraints to advisors' ability to self-deceive by examining how these treatments affect recommendations in three laboratory experiments. Across experiments, we fix investment A but vary investment B in order to progressively reduce advisors' ability to justify (to themselves) recommending A in the Before treatment. The payoffs of the lotteries were chosen such that, given the stakes, a majority of advisors would expect the client to prefer B, and thus recommend B in the absence of incentives to recommend A (Control treatment) in all experiments.

In the RiskReturn Experiment, advisors evaluate two investment opportunities, A and B. Investment A is a 50-50 lottery between \$2 and \$4. Investment B is a 50-50 lottery between \$1 and \$7. The expected payoff of B (\$4) is higher than that of A (\$3). However, B has a higher variance.

In the Dominance Experiment, we keep investment A unchanged, but make investment B more attractive for the client. In particular, we change investment B from a 50-50 lottery between \$1 and \$7 to a 50-50 lottery between \$2 and \$6. The expected payoff of investment B remains thus unchanged, but B now yields the same payoff as investment A in the low state, and a higher payoff in the high state.

In the ObviousDominance Experiment, we again keep investment A unchanged, and modify investment B to become even more attractive for the client by changing it to a 50-50 lottery between \$5 and \$7. This way, investment B yields a strictly higher payoff than A in both states of nature.

In the three laboratory experiments, that we present next, one out of every ten recommendations was randomly selected to be delivered to a client.

2.2 Hypotheses

The experimental design allows us to test whether advisors' ability to self-deceive increases their propensity to recommend the incentivized investment, and to distinguish the self-deception mechanism from other potential drivers of advice. First, if advisors are purely guided by their own self-interest, they should recommend A whenever they have an incentive to do so and the timing with which advisors learn about their own incentive should not affect recommendations. In all experiments, the fraction of A recommendations should be the same in the Before and After treatments, and higher than in Control.

Advisors may, however, care about their social or self image (e.g., Bénabou and Tirole, 2006) as ethical, i.e., being individuals who recommend the investment the client would prefer. They may derive utility from making a recommendation that others perceive as ethical (social image) or that they themselves perceive as ethical (self image). If advisors only care about social image, the timing of information about their own incentive should not matter. Since in all treatments the experimenter only observes the advisor's recommendation and not her private judgment, the rate of A recommendations should not differ between the Before and After treatment.

If advisors care about their self-image, the effect of the timing of information on recommendations will depend on whether advisors engage in self-deception. If advisors do not engage in self-deception, and believe that the client would prefer A, they will be unable to recommend A while preserving a positive self-image. Hence, the rate of A recommendations in the presence of incentives should be similar to the rate of recommendations in the Control treatment regardless of the timing at which advisors learn about their own incentives.

By contrast, if advisors who care about their self-image engage in self-deception, the timing of information may have an important effect on recommendations. When advisors learn their incentive *before* learning about the details of investments A and B, they might self-deceive by evaluating investment A as the preferred one for the client, whenever there is enough scope to do so. Instead, when advisors learn about their incentive *after* evaluating A and B, it is substantially harder for them to deceive themselves into believing that recommending A is ethical. In Appendix C we provide a stylized theoretical framework of self-deception, simplifying Bénabou (2015), that provides further detail regarding how self-deception may arise in the Before treatment. In Section 4.3 we discuss the results of our experiments in light of the theoretical framework.

Across the three experiments, we vary the payoffs of B and hence vary the scope for considering A the investment the client would prefer. In the RiskReturn experiment, there are two reasons that could make A the preferred option for the client. First, the advisor could believe the client is (sufficiently) risk averse and thus favors A. Second, A may be seen as the fair recommendation by both parties, since the advisor receives a \$1 commission for recommending investment A. This is the same as the client's cost (in expectation) of following an A recommendation.[4]

Hence, in the RiskReturn experiment we hypothesize that self-deception by advisors who care about their self-image leads to a higher rate of A recommendations in the Before treatment, relative to the After treatment.

Hypothesis 1: *In the RiskReturn experiment, advisors recommend A more frequently in the Before than in the After treatment.*

In the Dominance experiment, investment B yields weakly higher payoffs than A in both states, hence the risk in B is no longer a reason to favor A. However, in this experiment the client's costs (in expectation) of following an A recommendation are the same as in the RiskReturn experiment. This might provide advisors with an excuse for recommending the incentivized investment.

Hypothesis 2A: *If advisors self-deceive about the client's risk preference, in the Dominance experiment advisors recommend A equally frequently in the Before as in the After treatment.*

Hypothesis 2B: *If advisors self-deceive about the client's fairness concerns, in the Dominance experiment advisors recommend A more frequently in the Before than in the After treatment.*

In the ObviousDominance experiment, the cost of recommending A to the client increases substantially. This greatly diminishes the scope for arguing that the client would prefer A. We hence hypothesize that there is no room for self-deception in this experiment. This should be reflected in the gap between the rate of A recommendations between the Before and After treatment, which should decrease.

Hypothesis 3: *In the ObviousDominance experiment, advisors recommend A equally frequently in the Before as in the After treatment.*

[4] In our laboratory experiment the advisor's recommendation is delivered with a 10% chance, which could give additional moral wiggle room to advisors, as the cost is then lower in expectation. However, Charness, Gneezy and Halladay (2016) find that paying one out of several participants does not affect behavior. In our online experiments every advisor is matched to a client, such that the cost of choosing A for the client is equal to the benefit for the advisor, \$1. In both settings, we find a significant Before-After treatment effect.

Finally, across all three experiments, if there is no scope for self-deception in the After treatment, we would expect no difference in recommendations between After and Control. By contrast, if advisors still recommend the incentivized option without self-deception, e.g., because they mainly care about their own monetary incentives, we would observe a difference in recommendations between the After and Control treatments.

2.3 Procedures

We conducted our laboratory experiments at the University of California San Diego in the Spring and Fall of 2015. For each experiment, advisors took part in an hour-long experimental session involving other studies.[5] The experiments were run during a two-week period. The three experiments were conducted sequentially and, within each experiment, randomized assignment to each treatment occurred at the advisor level. Clients were recruited later, and participated in separate sessions.

The procedures were identical for all experiments. The instructions were presented to the advisor on four separate pages on their computer screen (all instructions are provided in Appendix B). First, we introduced the advisor to a study on economic decision-making. Then, the advisor was informed about her role in the experiment and she was told that she would be given a fixed payment of $1 for her participation. She was told that her task in the experiment was to recommend one of two investments (A and B) to another participant in a different session, but she was not told what A and B exactly were. She also learned that the other participant received no information about A or B except her recommendation. In the Before treatment, she was also informed about the $1 commission on this screen. On the third screen of the instructions, advisors were presented with the details of A and B.

In addition to receiving information about the lotteries, the advisor was asked (at the bottom of the third screen) to think about her recommendation and continue to the next screen once she was ready to provide it. Once the advisor moved to the fourth screen, the instructions asked her to raise her hand so that the research assistant could bring her

[5] All other studies in a session were not incentivized and unrelated to our study. They were surveys in the fields of marketing and management. Each of our experiments took place during the course of two weeks, and within each week the other studies remained always the same. The experiment was always either the first or the last within a session, and this was randomized at the session level.

a piece of paper on which she could write her recommendation. Once she received the paper, she was asked to move onto the fifth and final screen. The research assistant was instructed to ensure that participants would move to the next screen before writing anything on paper. In the After treatment, the advisor learned about the $1 commission, and was then asked to provide her recommendation both on paper and on screen. To introduce only one change across treatments, the information on the commission was also presented on the fifth screen in the Before treatment. This procedure had the advisor send a message in her own handwriting, making the recommendation more tangible, and provides us with a direct electronic record of recommendations.[6]

In each experiment, we aimed at collecting 100 observations per treatment. The decision regarding the number of participants was based on findings from a pilot study and the number of subjects that could be recruited for the study from our subject pool. We stopped collecting data at the end of the day in which we achieved 300 observations, giving us a total of 947 participants across the three experiments. Out of these participants, 38 had participated in a previous session and are thus excluded from the analysis.[7] In the RiskReturn experiment, there were 294 participants in the role of advisor, 98 participants in each of the three treatments. In the Dominance Experiment, we had 295 participants in the role of advisors (98 in the Control treatment, 100 in the Before treatment and 97 in the After treatment). In the ObviousDominance experiment we had 320 participants (106 in Control, 105 in Before, and 109 in After). 48% of advisors was female across all three experiments.

One out of every ten recommendations was randomly selected and given to a client in a different session. Advisors knew this incentive structure. In all experiments, clients largely followed the advisor's recommendation (76.7% in the RiskReturn Experiment, 80.7% in the Dominance Experiment and 73.5% in the ObviousDominance Experiment).[8] We found no difference in following depending on the recommendation, A or B across all experiments and treatments (Fisher's exact test, p>0.1).

[6] The piece of paper for the recommendation only included the message "I recommend you to choose Product ____." Advisors were asked to write down A or B.

[7] Our intention was to exclude anyone who had participated in the pilot experiment but a failure in the filter meant that some subjects still participated. The results remain unchanged if these subjects are included.

[8] These fractions are analogous to those of previous work using the sender-receiver game to study deception (e.g., Gneezy, 2005; Sutter, 2009; Dreber and Johannesson, 2008).

3 Results

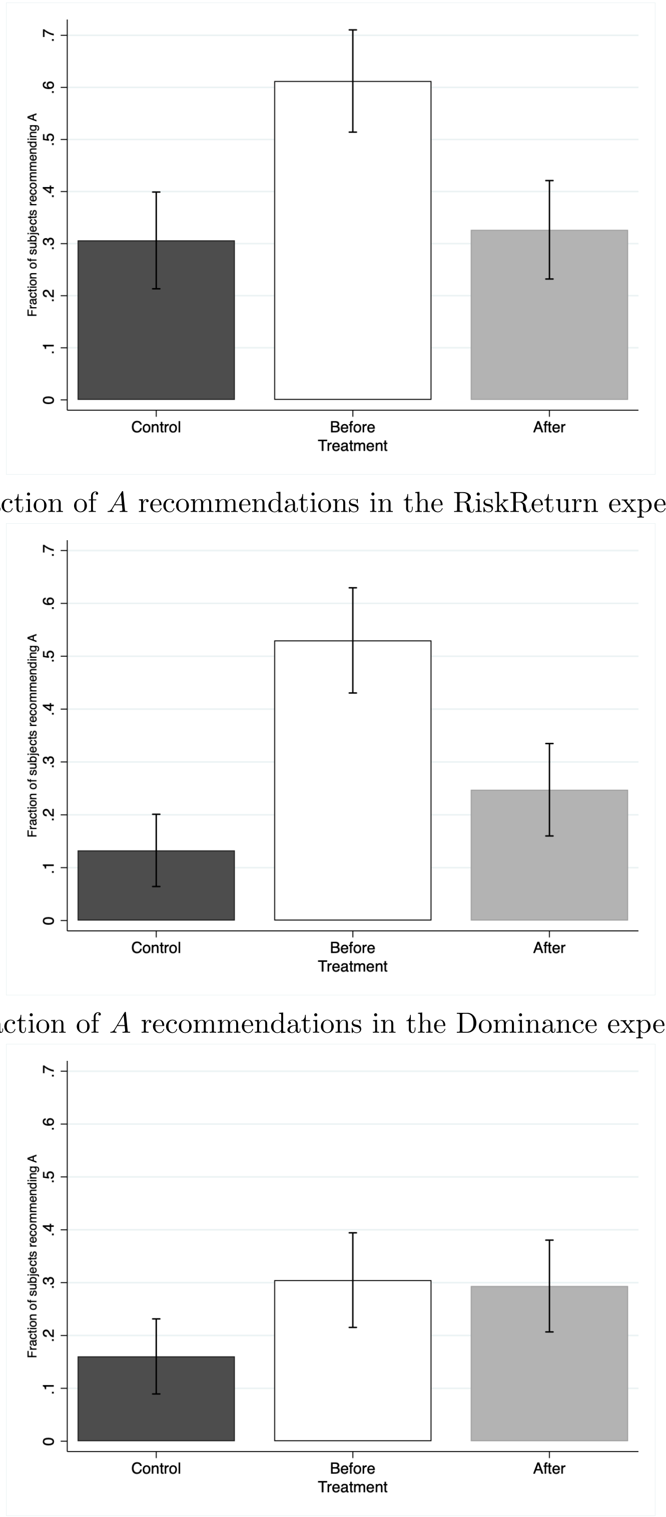
3.1 The RiskReturn Experiment

Figure **??** displays the fraction of advisors recommending investment A in the three treatments of the RiskReturn Experiment. The percentage of advisors recommending A in the Control treatment is 30.6%. When information about the incentive tied to A is provided before reading about A, advisors are significantly more likely to recommend it (test of proportions, $Z = 4.30$, $p < 0.001$). They recommend A in 61.2% of the cases in the Before treatment. In contrast, the rate of A recommendations drops to 32.7% of the cases in the After treatment ($Z = 4.01$, $p < 0.001$). This rate does not differ significantly from that in the Control treatment, 30.6% ($Z = 0.31$, $p = 0.759$).

Table 1 below confirms the results using regression analysis by comparing the likelihood of recommending A in the Before and Control treatment to that of the After treatment in each experiment (columns (1)-(3)). In the RiskReturn experiment, the likelihood of recommending A increases by 28.6 percentage points in the Before treatment, relative to the After treatment, which is the baseline. The difference between the Before and Control treatments is also significant, as shown in the last row of Table 1. A detailed comparison of the Before and After treatments relative to the Control treatment is shown in Appendix A.

Overall, our results are in line with the self-deception hypothesis, Hypothesis 1. When self-deception is easy (the Before treatment), the $1 commission distorts advice, increasing the fraction of A choices by 30.6 percentage points relative to the Control treatment. By contrast, when self-deception is harder (the After treatment), the effect falls to 2.1 percentage points, and is no longer significant. Trivers (2011) suggests that men may be more likely to self-deceive (and act overconfidently) than women. We do not find differences between male and female participants. The results of the analysis of gender differences are presented in Appendix A.

The results of the RiskReturn experiment could also be consistent with two alternative explanations. One alternative explanation is that participants in the Before treatment avoid evaluation altogether and simply recommend the incentivized investment, either because of the incentives per se, or because they perceive the incentives as a signal that the incentivized



(a) Fraction of A recommendations in the RiskReturn experiment

(b) Fraction of A recommendations in the Dominance experiment

(c) Fraction of A recommendations in the ObviousDominance experiment

Figure 2: A recommendations by experiment
Note: the error bars represent 95% confidence intervals.

investment is in fact the better product. Second, the smaller bias could also result from preferences for consistency (Cialdini, 1984; Falk and Zimmermann, 2015). That is, advisors in the After treatment might have a preference to stick to the first judgment they formulate in their minds prior to learning about the incentive. In what follows, we report the results of the experiments that we conducted to study the constraints to self-deception. These experiments also help in ruling out these alternative explanations.

3.2 The Dominance Experiment

Figure **??** displays the fraction of advisors recommending A in the three treatments in the Dominance experiment. In the Control treatment, participants recommended A in 13.3% of the cases. In the Before treatment, advisors recommended A at a higher rate, in 53% of the cases. In contrast, in the After treatment advisors recommended A in 24.7% of the cases, which is significantly smaller than in treatment Before ($Z = 4.06$, $p < 0.001$). The results also reveal that advisors are less likely to recommend A in the Control treatment (13.3%) than in the After treatment (24.7%) in the Dominance experiment ($Z = 2.04$, $p = 0.041$).

The distortion in recommendations in the Before treatment, relative to the After treatment, is similar to that in the RiskReturn experiment (28.3 percentage points), as seen in column (2) of Table 1. Difference-in-difference estimates, which allow us to compare the effect of the Before and Control treatments on recommendations in the Dominance experiment to those in the RiskReturn experiment, are shown in column (4) of Table 1. If there is less scope for self-deception in the Dominance experiment, the coefficient for the Before treatment should be significantly smaller in this experiment. We find no evidence that this is the case. The point estimate for the coefficient *Before treatment X Dominance* is close to zero and not significant. These results therefore suggest that advisors are able to engage in self-deception even when they have only "minimal" reasons to recommend A, in line with Hypothesis 2B.

Table 1: Treatment Effects on the Likelihood that A is Recommended across the Three Experiments

	(1)	(2)	(3)	(4)
		Pr(A is recommended)		
	Risk		Obvious	
Experiment:	Return	Dominance	Dominance	All
Before	0.286***	0.283***	0.011	0.286***
	(0.069)	(0.067)	(0.063)	(0.069)
Control	-0.020	-0.115**	-0.133**	-0.020
	(0.067)	(0.056)	(0.057)	(0.067)
Dominance experiment				-0.079
				(0.065)
Before X Dominance experiment				-0.003
				(0.096)
Control X Dominance experiment				-0.094
				(0.087)
ObviousDominance experiment				-0.033
				(0.065)
Before X ObviousDominance experiment				-0.275***
				(0.093)
Control X ObviousDominance experiment				-0.113
				(0.088)
Constant	0.327***	0.247***	0.294***	0.327***
	(0.048)	(0.044)	(0.044)	(0.048)
Observations	294	295	320	909
R-squared	0.080	0.133	0.023	0.098
Before vs. Control (p-value)	0.000	0.000	0.013	-

Notes: Columns (1) to (4) report the estimates of a linear probability model. Using probit regressions we obtain leads to very similar results, as shown in Appendix A. We use linear probability models estimates due to the difficulty in interpreting interaction effects in probit models (Ai and Norton, 2003). The dependent variable is a dummy that is equal to 1 when A is recommended. The variables 'Before treatment' and 'Control treatment' are dummy variables taking value 1 if the treatment is Before or Control, respectively. The omitted category is the After treatment. The Dominance and ObviousDominance variables take value 1 for the Dominance and ObviousDominance experiments respectively. The final row reports the results of a Wald Test for the equality of the Control and Before coefficients. Robust standard errors are reported in parentheses.
*** Significant at the 1% level; ** Significant at the 5% level; * Significant at the 10% level.

3.3 The ObviousDominance Experiment

Figure **??** displays the fraction of advisors recommending investment A in the ObviousDominance experiment. In this experiment, investment A is recommended 16% of the time in the Control treatment. The fraction of A recommendations observed in the Before treatment (30.5%) is no longer significantly different from the fraction observed in the After treatment (29.4%, $Z = 0.18$, $p = 0.858$). In other words, the timing of information about the incentive no longer affects recommendations in this experiment. Compared to the Control treatment, A is recommended more frequently in the Before treatment ($Z = 2.48$, $p = 0.013$), as well as the After treatment ($Z = 2.33$, $p = 0.020$).

In column (4) of Table 1 we show that the Before-After effect is significantly weaker in the ObviousDominance experiment than in the RiskReturn experiment and the Dominance experiment. The increase in A recommendations observed in the Before treatment of the RiskReturn experiment relative to the After treatment (of 28.6 percentage points) is almost completely eliminated in the ObviousDominance experiment, where it is 27.5 percentage points weaker. A similar result is obtained comparing the Dominance treatment, where the Before treatment increases A recommendations by 28.3 percentage points, to the ObviousDominance experiment, in which the effect of the Before treatment is 27.1 percentage points weaker.

The results of the ObviousDominance experiment rule out the alternative mechanisms we discussed above. Specifically, both consistency and evaluation avoidance would predict a treatment difference between the Before and After treatment, independent of the characteristics of the lotteries. However, we find that the difference between the Before and After treatment is significantly smaller in the ObviousDominance experiment, suggesting that these alternative explanations cannot explain the larger difference observed in the RiskReturn and Dominance experiment.

Taken together, the three laboratory experiments provide novel evidence on some of the factors that constrain people's ability to preserve desirable beliefs about the ethicality of their recommendation. In the RiskReturn and Dominance experiments, we show that learning about the commission after forming an initial judgment limits advisors' ability to self-deceive (i.e., convince themselves that their recommendation matches the clients' preferences), even in presence of ambiguity that could be used to rationalize decisions ex-post. We interpret this as evidence that changing advice after having had an opportunity to form an unbiased evaluation of the two investments makes it substantially harder for advisors to preserve the (biased) belief that their advice is what the client would prefer.

We further document another constraint in advisors' ability to self-deceive. Advisors' ability to self-deceive persists even when there is *minimal* scope for convincing oneself that recommending A is ethical, as in the Dominance experiment. However, advisors are no longer able to justify recommending the investment that allows them to earn more money when one investment option is unambiguously better than the other (as in the ObviousDominance experiment).

4 Mechanisms of Self-Deception: Belief Distortion and Choice for Self

The bias in recommendations in our three laboratory experiments is consistent with self-deception. Yet these experiments provide indirect evidence of self-deception, since we did not measure beliefs. To provide direct evidence of self-deception, we ran additional online experiments. In these experiments we follow the design of the laboratory experiments and collect two additional measures: advisors' beliefs about the investment recommendation the client would prefer and advisors' choices for themselves between the two investments. Since we conducted these experiments online, we also test whether our results from the laboratory replicate in a different sample.

4.1 Procedures

We conducted three experiments (RiskReturn, Dominance and ObviousDominance) on Amazon Mechanical Turk in the Spring of 2019. We focus on the Before and After treatments only (excluding Control), because of our interest in measuring self-deception. We recruited participants to a 3-minute study for a fixed payment of \$0.25. We adapted the instructions to this online setting, dividing all payoffs in the experiment by four. In all experiments, investment A was a 50-50 lottery between \$0.50 and \$1. Investment B was a 50-50 lottery between \$0.25 and \$1.75 in the RiskReturn experiment, between \$0.50 and \$1.50 in the Dominance experiment and between \$1.25 and \$1.75 in the ObviousDominance Experiment. The advisor's commission for recommending A was always \$0.25. Whereas in the lab experiment the advisor only had a 1 in 10 chance to be matched with a client, in the online experiments we implemented a 1 to 1 matching between the advisor and the client.

At the beginning of the experiment, the advisor was informed about her role in the experiment, and was informed about her task of recommending one of two investment options (A and B) to a client. Then, the advisor was told she would receive \$0.25 for her participation. In the Before treatments, the advisors also received information about her own commission. Thereafter, the advisor was presented with the details about A and B. In this online setting, we did not ask the advisor to make her recommendation in writing. Instead, we prompted the advisor to carefully consider her recommendation after seeing the investments, in part by forcing advisors to spend at least 45 seconds on the screen that described the investments before proceeding.

Before asking the advisor to make a recommendation, we elicited the advisor's beliefs. Advisors indicated which recommendation, A or B, they thought the client would prefer.[9] Measuring beliefs before asking subjects to make their recommendation could perhaps distort their recommendation. We chose this procedure, which follows Babcock et al. (1995), because it allows us to capture whether learning about the commission before learning about the investments biases advisors' beliefs at the evaluation stage.[10] We did not incentivize beliefs, using measures such as coordination games (Krupka and Weber, 2013), because we were interested in measuring what the advisor believes herself is preferred by the client, and not what others believe is ethical. Using incentives could also reduce the scope for self-deception. Existing studies in other contexts suggest that simple non-incentivized measures of beliefs perform similarly to incentivized ones (e.g., Friedman and Massaro, 1998; Sonnemans and Offerman, 2001; Trautmann and van de Kuilen, 2014; Hollard, Massoni and Vargnaud, 2016). After the belief elicitation task, the advisor moved to the final screen, which asked her to provide her recommendation to the client. In the After treatment, advisors learned about their commission prior to being asked to provide their recommendation.

At the end of the experiment, we asked advisors to choose between the two investments, A and B, as a reward for themselves. One randomly chosen advisor within each experiment actually received the payoff of A or B at the end of the study. Measuring ad-

[9] The exact question was "Which Product do you think the client would prefer?". The answer was "I believe the client prefers Product A/B."

[10] This measure may introduce social image concerns with respect to the experimenter if advisors wish to make recommendations in line with their beliefs.

visors' choices for themselves follows Chen and Gesche (2018), who show that advisors are more likely to choose an investment product for themselves and for a client after having been previously incentivized to recommended it to another client. The screens for the experiment are presented in Appendix B.

In each experiment, we aimed at collecting at least 100 observations of attentive subjects per treatment. We measured attention after advisors had provided their recommendations by asking them about the payoffs of investment A, using a multiple-choice question. Considering potential exclusions due to inattention, we decided to recruit 150 advisors per treatment in each of the three experiments. We restricted participation to individuals located in the US, with an approval rating higher than 80% on their previous HITs. A total of 900 advisors took part across the three experiments, out of which 899 provided a recommendation. Excluding advisors who answered our attention question incorrectly, there are 269 advisors in the RiskReturn experiment, 283 in the Dominance experiment and 279 in the ObviousDominance experiment. Including inattentive advisors in our analysis does not change the results.

We then recruited an additional 899 participants in the role of clients. A total of 88.7% of clients followed the advisors recommendation (90.4% for product A and 85.4% for product B).

4.2 Results

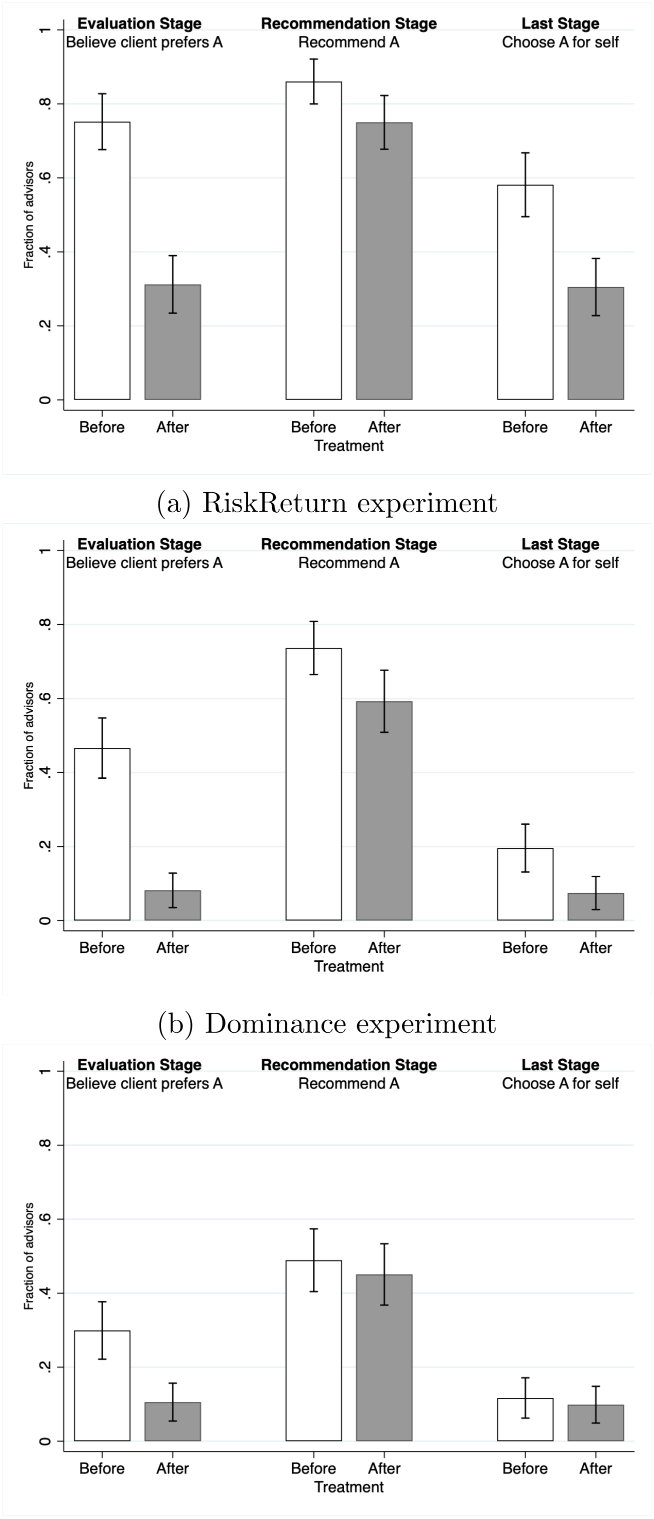
Beliefs. Are advisors engaging in self-deception, convincing themselves that the incentivized recommendation is the recommendation the client would prefer? Figure 3 displays the fraction of advisors who believe A is the recommendation preferred by the client, at the evaluation stage of the Before and After treatment in the RiskReturn, Dominance and ObviousDominance experiments. In the RiskReturn experiment, 75.2% of advisors consider A the client's preferred recommendation in the Before treatment. By contrast, this fraction is only 31.4% after advisors evaluate the lotteries in the After treatment (Z= 7.18, p<0.001). In the Dominance experiment, 46.6% of participants consider A the recommendation the client would prefer whereas this fraction is only 8.1% in the After treatment (Z= 7.18, p<0.001). In the ObviousDominance experiment, the difference in beliefs becomes smaller: 29.9% of participants consider A the investment recommendation the client would prefer, and this fraction is 10.6% in the After treatment (Z= 4.04, p<0.001). Panel B of Table 2 confirms these results and shows that the gap between the Before and After treatment becomes significantly smaller in the ObviousDominance experiment.

Hence, learning about one's incentive prior to evaluating the investments distorts advisors' beliefs toward the investment that allows them to earn a commission. In line with the idea that reducing ambiguity over the better recommendation limits self-deception, advisors are progressively less likely to distort their beliefs when we move from the RiskReturn to the Dominance, and to the ObviousDominance Experiment.

Recommendations. Figure 3 displays the fraction of advisors recommending investment A in the Before and After treatment in the three experiments. The pattern of recommendations in the two treatments replicates the results from the lab experiments, though the gap between the Before and After treatment in this online sample is smaller. In the RiskReturn experiment, 86.0% of advisors recommend A when they learn about their own incentives before evaluating the two investment options, whereas 75.0% of advisors recommend A when they learn about their incentives afterwards (Z= 2.28, p=0.02). In the Dominance experiment, 73.6% of advisors recommend A in the Before treatment, whereas this fraction is only 59.3% in the After treatment (Z= 2.57, p=0.01). Finally, there is no significant treatment difference in A recommendations in the ObviousDominance experiment, with 48.9% of the advisors recommending A in Before and 45.1% of advisors recommending A in After (Z= 0.64, p=0.52).

Panel A of Table 2 confirms these results. In the RiskReturn experiment, the likelihood of recommending A increases by 11 percentage points in the Before treatment. Similarly, in the Dominance experiment it increases by 14 percentage points. By contrast, in the ObviousDominance experiment the difference in A recommendations is small (4 percentage points) and no longer significant. However, given the smaller Before-After gap in the RiskReturn experiment, we are unable to detect a significant decrease in the size of the gap between the Before and After treatments.

Taken together, the results of these experiments qualitatively replicate the pattern



(a) RiskReturn experiment

(b) Dominance experiment

(c) ObviousDominance experiment

Figure 3: Beliefs, recommendations, and choices for self by experiment
Note: the error bars represent 95% confidence intervals.

documented in the lab experiments. Learning about the incentive prior to evaluating the two investments only led to a significantly higher fraction of A recommendations in presence of some scope for self-deception.

Table 2: Recommendations, beliefs and choices for self in the online experiments

	(1)	(2)	(3)
Experiment:	RiskReturn	Dominance	ObviousDominance
Panel A: Pr(A is recommended)			
Before treatment	0.110**	0.144**	0.038
	(0.048)	(0.056)	(0.060)
Constant	0.750***	0.593***	0.451***
	(0.037)	(0.042)	(0.042)
Observations	269	283	279
R-squared	0.019	0.023	0.001
Difference-in-difference			
Before treatment relative to RiskReturn (p-value)	-	0.650	0.347
Panel B: Pr(Advisor believes client prefers A)			
Before treatment	0.438***	0.385***	0.194***
	(0.055)	(0.047)	(0.047)
Constant	0.314***	0.081**	0.106***
	(0.039)	(0.024)	(0.026)
Observations	269	283	279
R-squared	0.192	0.182	0.058
Difference-in-difference			
Before treatment relative to RiskReturn (p-value)	-	0.446	0.001
Panel C: Pr(Advisor chooses A for herself)			
Before treatment	0.274***	0.122***	0.018
	(0.059)	(0.040)	(0.037)
Constant	0.307***	0.074***	0.099***
	(0.039)	(0.023)	(0.025)
Observations	269	283	279
R-squared	0.076	0.031	0.001
Difference-in-difference			
Before treatment relative to RiskReturn (p-value)	-	0.029	0.000

Notes: Columns (1) to (3) report the estimates of linear probability model on recommendations (Panel A), advisor beliefs (Panel B), and advisor choices for herself (Panel C). Each column reports results for one experiment. The dependent variable in Panel A is a dummy that is equal to 1 when A is recommended. The dependent variable in Panel B is a dummy variable that is equal to 1 if the advisor believes A is the recommendation preferred by the client. The dependent variable in Panel C is a dummy that is equal to 1 when the advisor chooses A for herself. The variable 'Before treatment' is a dummy variable taking a value of 1 if the treatment is Before. The omitted category is the After treatment. The difference-in-difference test in the bottom row is based on a regression that interacts the 'Before treatment' with dummy variables for the respective experiments. Robust standard errors are reported in parentheses.
*** Significant at the 1% level; ** Significant at the 5% level; * Significant at the 10% level.

Choice for Self. As an additional test of self-deception, we investigate advisors' choices between the two investments, when asked to pick an investment for themselves. In the RiskReturn experiment, we find that 58.1% of the advisors choose investment A for themselves, whereas this fraction is only 30.7% in the After treatment (Z= 4.53, p<0.001). In the Dominance experiment, the fraction of advisors choosing A in the Before treatment is smaller, 19.6%, but a gap between the Before and After treatment persists: only 7.4% of advisors chooses A in the After treatment (Z= 2.97, p=0.003). This result suggests that a substantial fraction of advisors convince themselves that investment A is the better recommendation, to the point of choosing it for themselves. Consistent with this interpretation, we find that beliefs are significantly correlated with choices for the self, and the difference between the Before and After treatments is no longer significant when controlling for beliefs. These results are shown in Appendix A.

In line with smaller scope for self-deception, in the ObviousDominance experiment we see no gap in A choices between treatments: 11.7% of advisors choose it in Before and 9.9% of advisors choose it in After (Z= 0.49, p=0.62). In all the experiments, the fraction of participants who chooses A for themselves in the After treatment is similar to the fraction of advisors who believe A is preferred by the client in that treatment. Panel C of Table 2 confirms these results using regressions.

Overall, the results from the online experiment confirm the results of the lab and provide additional evidence for the self-deception mechanism. As in the laboratory, we find that a minimal justification is enough to enable advisors to recommend A.

4.3 Discussion

How can the findings in our experiments in the laboratory and online be reconciled within one framework? In Appendix C we develop a stylized framework, that simplifies Bénabou (2015), in which advisors derive self-image utility from providing advice that they believe is in line with what the client would prefer. In the Before treatment, advisors have information about their incentives from the start, and can form motivated beliefs about whether the client would prefer an A recommendation. By contrast, in the After treatment, advisors are initially unaware of the commission, and most advisors evaluate B as the recommendation preferred by the client. When learning about the incentives later, recommending A may come at the cost of losing their initial self-image, which decreases the share of advisors recommending A. Comparing the results across the three experiments provides information about the nature of self-deception costs that can explain the results in this setting.

Specifically, in our framework an internal observer evaluates the actions of the decision maker. The internal observer does not know whether the decision maker truly believes A to be the recommendation preferred by the client, but evaluates the decision maker's stated belief as well as her recommendation to form a belief about her type. Self-deception takes the form of biased updating by the observer, as in Bénabou (2015).

An important question is what determines the cost of self-deception? In our framework we assume that it is constant (and specifically zero). What our empirical evidence reveals is that this is a plausible assumption. Empirically, the extent of self-deception is the same in the RiskReturn and Dominance experiments, although fewer individuals actually believe A to be the client's preferred recommendation. This indicates that a positive share of decision makers who truly believe A is preferred by the client is sufficient for self-deception to arise, consistent with a constant cost. To conclude, our model and the experimental findings suggest that advisors primarily care about the expected payoffs of the client and that the costs of self-deception can be modeled as a constant. These assumptions could be used in future research aimed at finding ways to reduce the scope for self-deception.

5 Conclusion

Understanding the conditions under which experts give biased advice can help structuring policies to reduce such behavior. For example, physicians may believe incentives (e.g., receiving fees for each procedure they perform or gifts from pharmaceutical companies) do not influence their judgment and some financial advisors may believe not to be influenced by commissions. These beliefs allow them to receive incentives while maintaining their self-image as unbiased professionals. The evidence suggests these experts are wrong, as incentives do distort their judgment in many cases (Steinman et al., 2001; Moore et al., 2010; Cain, Loewenstein and Moore, 2011; see also Malmendier and Schmidt, 2017). This biased judgment comes at a cost to the patients, who may not receive the best available treatment or may pay more for it, or to clients, who may end up making decisions that hurt them financially.

Our findings advance the broader literature on belief-based utility (e.g., Loewenstein and Molnar, 2018; Bénabou and Tirole, 2016; Golman et al., 2016; Mobius et al. 2014) by outlining some of the constrains to people's ability to preserve and defend desirable beliefs about themselves. They also have implications for the literature on preferences for truth-telling. People are averse to lying (e.g., Gneezy, 2005; Dreber and Johannesson, 2008; Sutter, 2009) and cheating, even when cheating does not rely on deceiving and changing the beliefs of the receiver (Charness and Dufwenberg, 2009; Fischbacher and Föllmi-Heusi, 2013; Shalvi et al. 2011, 2012). The most striking result in this literature is that people do not lie much, even in presence of strong incentives (Abeler et al., 2019). In these experiments, individuals choose whether to misreport an objective state of nature. Similarly to our ObviousDominance experiment, this environment does not allow room for self-deception. Our data suggests that the relatively low lying rates observed in these experiments may be due to subjects' inability to distort their beliefs about the ethicality of their behavior.

Examples in which expert advice is biased by incentives are plentiful and have a huge impact on efficiency and fairness. One solution for this problem is to limit such incentives when possible. For example, moving physicians from fee-for-service to salary-based compensation schemes may limit the extent of overtreatment. However, such changes may be hard to implement due to pushback from lobbyists, or due to the high costs of monitoring and enforcement. We propose additional approaches that may reduce the effectiveness of incentives in distorting judgment. The first relates to the timing of decisions, by having experts first evaluate the options and only then receive information about the incentives. If the first evaluation of a financial product or service is unbiased, such evaluation could potentially persist over time, provided the environment is sufficiently stable and the scope for a self-serving re-evaluation of the product or service is limited. A second approach involves providing experts with as much information about the client's preferences as possible. In the case of physicians, this could involve eliciting the patient's willingness to try a riskier or longer treatment, which may be cheaper. In the case of financial advisors, this would imply making investor risk profiles as concise as possible.

Taken together, our findings illustrate how people have psychological costs associated with distorting advice. Creating procedures that reinforce the role of self-image costs can reduce unethical behavior by ethical-but-biased individuals (Chugh, Bazerman and Banaji, 2005), and thereby increase the well-being of many, who depend on others to receive medical, financial, legal and policy advice.

References

Abeler, J., Nosenzo, D., & Raymond, C. (2019). Preferences for truth-telling. *Econometrica*, 87 (4), 1115-53.

Ai, C. & Norton, E.C. (2003). Interaction terms in logit and probit models. *Economics Letters* 80 (1), 123-29.

Akerlof, G. A., & Kranton, R. E. (2000). Economics and identity. *Quarterly Journal of Economics*, 115(3), 715-753.

Anagol, S., Cole S., & Sarkar, S. (2017). Understanding the advice of commissions-motivated agents: Evidence from the Indian life insurance market. *Review of Economics and Statistics*, 99(1), 1-15.

Babcock, L., Loewenstein, G., Issacharoff, S., & Camerer, C. (1995). Biased judgments of fairness in bargaining. *The American Economic Review*, 85(5), 1337-1343.

Babcock, L., & Loewenstein, G. (1997). Explaining bargaining impasse: The role of self-serving biases. *The Journal of Economic Perspectives*, 11(1), 109-126.

Balafoutas, L., Beck, A., Kerschbamer, R., & Sutter, M. (2013). What drives taxi drivers? A field experiment on fraud in a market for credence goods. *The Review of Economic Studies*, 80(3), 876-891.

Bénabou, R. (2015). The Economics of Motivated Beliefs. Jean-Jaques Laffont Lecture, *Revue d'economic politique* 125 (5), 665-85.

Bénabou, R., & Tirole, J. (2002). Self-confidence and Personal Motivation. *Quarterly Journal of Economics* 117 (3), 871-915.

Bénabou, R., & Tirole, J. (2011). Identity, Morals and Taboos: Beliefs as Assets. *Quarterly Journal of Economics* 126 (2), 805-55.

Bénabou, R., & Tirole, J. (2016). Mindful Economics: The Production, Consumption and Value of Beliefs. *Journal of Economic Literature* 30 (3), 141-64.

Bénabou, R., Falk, A. & Tirole J. (2018). Narratives, Imperatives and Moral Reasoning. NBER Working Paper #24798.

Bergstresser, D., Chalmers, J.M. & Tufano, P. (2009). Assessing the costs and benefits of brokers in the mutual fund industry. *The Review of Financial Studies* 22 (10), 4129–4156.

Bicchieri, C., Dimant, E., & Sonderegger, S. (2019). It?s not a lie if you believe it: Lying and belief distortion under norm-uncertainty. Mimeo.

Cain, D., Loewenstein, G., & Moore, D.A. (2011). Understanding the Perverse Effects of Disclosing Conflicts of Interest. *Journal of Consumer Research* 37 (5), 836-57.

Chalmers, J. & Reuter, J. (2012). How do retirees value life annuities? Evidence from public employees. *The Review of Financial Studies*, 25(8), 2601-2634.

Charness, G. & Dufwenberg, M. (2006). Promises and partnership. *Econometrica*, 74(6), 1579-1601.

Charness, G., Gneezy, U., & B. Halladay (2016). Experimental methods: Pay one or pay all. *Journal of Economic Behavior & Organization* 131 (A), 141–150.

Chen, Z. and Gesche, T. (2018). Persistent bias in advice-giving. Mimeo.

Chugh, D., Bazerman, M. H., and Banaji, M.R. 2005. Bounded ethicality as a psychological barrier to recognizing conflicts of interest. *Conflicts of interest: Challenges and solutions in business, law, medicine, and public policy*, 74-95.

Cialdini, R. (1984). *Influence, the psychology of persuasion.* Harper Collins, New York.

Clemens, J., & Gottlieb, J. D. (2014). Do physicians' financial incentives affect medical treatment and patient health? *The American Economic Review*, 104(4), 1320-1349.

Crawford, V., & Sobel, J. (1982). Strategic Information Transmission. *Econometrica* 50 (6), 1431-51.

Dana, J., Weber, R. A., & Kuang, J.X. (2007). Exploiting Moral Wriggle Room: Experiments Demonstrating an Illusory Preference for Fairness. *Economic Theory* 33 (1), 67-80.

Darby, M.R., & Karni, E. (1973). Free Competition and the Optimal Amount of Fraud. *The Journal of Law and Economics* 16 (1), 67-88.

DeJong, C., Aguilar, T., Tseng, C. W., Lin, G. A., Boscardin, W. J., & Dudley, R. A. (2016). Pharmaceutical industry-sponsored meals and physician prescribing patterns for Medicare beneficiaries. *JAMA Internal Medicine*, 176(8), 1114-10.

Di Tella, R., Perez-Truglia, R., Babino, A., & Sigman, M. (2015) Conveniently Upset: Avoiding Altruism by Distorting Beliefs about Others' Altruism. *American Economic Review* 105 (11), 3416-42.

Dulleck, U., & Kerschbamer, R. (2006). On Doctors, Mechanics and Computer Specialists: The Economics of Credence Goods. *Journal of Economic Literature* 44 (1), 5-42.

Dulleck, U., Kerschbamer, R., & Sutter, M. (2011). The economics of credence goods: An experiment on the role of liability, verifiability, reputation, and competition. *The American Economic Review*, 101(2), 526-555.

Dreber, A., & Johannesson, M. (2008). Gender differences in deception. *Economics Letters*, 99(1), 197-199.

Emanuel, E. J., & Fuchs, V. R. (2008). The perfect storm of overutilization. *JAMA*, 299(23), 2789-2791.

Epley, N. & Gilovich, T. (2016). The mechanics of motivated reasoning. *Journal of Economic Perspectives*, 30(3),133-40.

Exley, C.L. (2015). Excusing Selfishness in Charitable Giving: The Role of Risk. *The Review of Economic Studies* 83 (2), 587-628.

Festinger, Leon. (1957). *A Theory of Cognitive Dissonance*. Evanston, Il: Row, Peterson.

Falk, A. & Zimmermann, F. (2015). Information processing and commitment. *The Economic Journal*

Fischbacher, U., & Föllmi-Heusi, F. (2013). Lies in disguise? An experimental study on cheating. *Journal of the European Economic Association*, 11(3), 525-547.

Friedman, D., & Massaro, D. (1998). Understanding variability in binary and continuous choice. *Psychonomic Bulletin & Review*, 5(3), 370-389.

Gino, F., Norton M., & Weber, R. (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), 384-94.

Gneezy, U., Saccardo S., & van Veldhuizen R. (2018). Bribery: Behavioral Drivers of Distorted Decisions. *Journal of the European Economic Association*, forthcoming.

Golman, R., Hagmann, D., & Loewenstein, G. (2016). Information Avoidance. *Journal of Economic Literature* 55 (1), 96-135.

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.

Gruber, J., Kim, J., & Mayzlin, D. (1999). Physician fees and procedure intensity: the case of cesarean delivery. *Journal of Health Economics*, 18(4), 473-490.

Haisley, E. C., & Weber, R. A. (2010). Self-serving interpretations of ambiguity in other-regarding behavior. *Games and Economic Behavior*, 68(2), 614-625.

Hackethal, A., Haliassos, M. & Jappelli, T. (2012). Financial advisors: A case of babysitters?. *Journal of Banking & Finance*, 36(2), 509-524.

Hollard, G., Massoni, S. & Vergnaud, J.-C. (2016). In Search of Good Probability Assessors: An Experimental Comparison of Elicitation Rules for Confidence Judgments. *Theory and Decision*, 80(3), 363-387.

Institute of Medicine (2013). Best care at lower cost: The path to continuously learning health care in America. The National Academies Press, Washington DC.

Johnson, E. M., & Rehavi, M. M. (2016). Physicians treating physicians: Information and incentives in childbirth. *American Economic Journal: Economic Policy*, 8(1), 115-141.

Krupka, E. & R. Weber (2013).

Konow, J. (2000). Fair shares: Accountability and cognitive dissonance in allocation decisions. *The American Economic Review*, 90(4), 1072-1091.

Kunda, Z. (1990). The Case for Motivated Reasoning. *Psychological Bulletin* 108 (3), 480-98.

Loewenstein, G., & Molnar, A. (2018). The renaissance of belief-based utility in economics. *Nature Human Behavior* 5(1).

Mafi, J. N., McCarthy, E. P., Davis, R. B., & Landon, B. E. (2013). Worsening trends in the management and treatment of back pain. *JAMA Internal Medicine*, 173(17), 1573-1581.

Malmendier, U., & Schmidt, K. (2017). You owe me. *American Economic Review* 107 (2), 493-526.

Moore, D. A., Tanlu, L., & Bazerman, M. H. (2010). Conflict of interest and the intrusion of bias. *Judgment and Decision Making*, 5(1), 37.

Möbius, M., Niederle, M., Niehaus, P. & Rosenblat, T. (2014). Managing Self-Confidence. Mimeo.

Schwardmann, P., & van der Weele, J. (2019). Deception and Self-Deception. *Nature Human Behavior*, forthcoming.

Shalvi, S., Dana, J., Handgraaf, M. J., & De Dreu, C. K. (2011). Justified ethicality: Observing desired counterfactuals modifies ethical perceptions and behavior. *Organizational Behavior and Human Decision Processes*, 115(2), 181-190.

Shalvi, S., Eldar, O., & Bereby-Meyer, Y. (2012). Honesty requires time (and lack of justifications). *Psychological Science*, 23(10), 1264-1270.

Sharek, Z., Robert E. S., & Loewenstein G. (2012). Bias in the evaluation of conflict of interest policies. *The Journal of Law, Medicine & Ethics*, 40(2), 368-382.

Sonnemans, J. & Offerman T. (2001) Is the Quadratic Scoring Rule Really Incentive Compatible? *Unpublished Manuscript*.

Sutter, M. (2009). Deception through telling the truth?! Experimental evidence from individuals and teams. *The Economic Journal*, 119(534), 47-60.

Steinman, M. A., Shlipak, M. G., & McPhee, S. J. (2001). Of principles and pens: attitudes and practices of medicine housestaff toward pharmaceutical industry promotions. *The American Journal of Medicine*, 110(7), 551-557.

Trautmann, S. & G. van de Kuilen (2014). Belief elicitation: A horse race among truth serums. *Economic Journal* 125 (89), 2116–2135.

Trivers, R. (2011). *The Folly of Fools: The Logic of Deceit and Self-Deception in Human Life*. Basic Books.

Zimmerman, F. (2018). The Dynamics of Motivated Beliefs. *American Economic Review*, forthcoming.

Zou, W. (2018). Motivated Belief Updating of Norms: Theory and Experimental Evidence. Mimeo.

Appendix

The online version of this article contains additional analyses (Appendix A), the experimental instructions (Appendix B) and the theoretical framework (Appendix C). All of the supplementary material can be found at https://osf.io/sm3u6/?view_only=f8effdaabf94425bb6e06454a191a