

# UC San Diego

## UC San Diego Previously Published Works

### Title

Enabling or Limiting Cognitive Flexibility? Evidence of Demand for Moral Commitment

### Permalink

<https://escholarship.org/uc/item/0874x1wz>

### Journal

American Economic Review, 113(2)

### ISSN

0002-8282

### Authors

Saccardo, Silvia  
Serra-Garcia, Marta

### Publication Date

2023-02-01

### DOI

10.1257/aer.20201333

### Copyright Information

This work is made available under the terms of a Creative Commons Attribution-NonCommercial-NoDerivatives License, available at <https://creativecommons.org/licenses/by-nc-nd/4.0/>

Peer reviewed

# Enabling or Limiting Cognitive Flexibility? Evidence of Demand for Moral Commitment\*

Silvia Saccardo<sup>†</sup> and Marta Serra-Garcia<sup>‡</sup>

August 2022

## Abstract

Moral behavior is more prevalent when individuals cannot easily distort their beliefs self-servingly. Do individuals seek to limit or enable their ability to distort beliefs? How do these choices affect behavior? Experiments with over 8,900 participants show preferences are heterogeneous – 30% of participants prefer to limit belief distortion, while over 40% prefer to enable it, even if costly. A random assignment mechanism reveals that being assigned to the preferred environment is necessary for curbing or enabling self-serving behavior. Third parties can anticipate these effects, suggesting some sophistication about the cognitive constraints to belief distortion.

**JEL Classification:** D83, D91, C91.

**Keywords:** Belief distortion, morality, sophistication, commitment, experiments.

---

\*We would like to thank the editors and reviewers for their excellent comments on this manuscript. We also thank Johannes Abeler, Saurabh Bhargava, Christine Exley, Laura Gee, Russell Golman, David Huffman, Alex Imas, Michel Marechal, George Loewenstein, Kirby Nielsen, Theo Offerman, Ricardo Perez-Truglia, Peter Schwardmann, Eric van Damme, Jeroen van de Ven, Joel van der Weele, Lise Vesterlund, Roberto Weber, Florian Zimmerman, and participants at several conferences and workshops for helpful comments and suggestions. We would also like to thank Ehsan Amozegar, Daniel Henderson, Mandy Lanyon, Sarita Raghunath, Ben Schenk, Phillip Tan, and Stars Xu for excellent research assistance. This research benefited from funding from the National Science Foundation under award number 1926043 (Saccardo) and UCSD Research Award 84132 (Serra-Garcia).

<sup>†</sup>Department of Social and Decision Sciences, Carnegie Mellon University and CESifo. Email: ssaccard@andrew.cmu.edu

<sup>‡</sup>Rady School of Management, UC San Diego and CESifo. Email: mserragarcia@ucsd.edu.

# 1 Introduction

The fundamental desire to preserve a positive identity often leads individuals to engage in motivated reasoning, distorting their beliefs to enable desired behaviors (e.g., Kunda, 1990; Bénabou and Tirole, 2011, 2016). The resulting belief distortion can explain phenomena such as managerial overconfidence (e.g., Malmendier and Tate, 2005), partisan polarization (e.g., Kahan, 2013), or collective denial of wrongdoing in organizations (e.g., Bénabou, 2013). Individuals can protect cherished beliefs by avoiding inconvenient information (e.g., Dana, Weber and Kuang, 2007; Golman et al., 2017). And, when information cannot be avoided, they can distort their beliefs ex-post through cognitive processes like attention and memory (e.g., Eil and Rao, 2011; Zimmerman, 2018; Huffman, Raymond, and Shvets, 2020; Amasino, Pace, and van der Weele, 2021; Möbius et al., 2022). Yet, there are contextual limits to the ability to distort beliefs (Sloman, Fernbach, and Haggmayer, 2010; Epley and Gilovich, 2016). An important open question about motivated cognition is: Do individuals, in anticipation of limits to belief distortion, attempt to limit belief distortion to commit to more accurate beliefs or would they rather seek out the cognitive flexibility needed to distort beliefs? And how do their choices affect their subsequent behavior?

We investigate these questions in the domain of moral behavior (e.g., Abeler, Nosenzo and Raymond, 2019; Cohn et al., 2019), where there is evidence that individuals distort their beliefs to act self-servingly.<sup>1</sup> If informative signals cannot be avoided, belief distortion is enabled when individuals have “cognitive flexibility”: the cognitive ability to pay less attention to and underweight potentially undesired signals. While previous findings suggest that some individuals may desire cognitive flexibility, little attention has been given to the possibility that some people may prefer to constrain belief distortion as a way to commit to moral behavior. In this paper, we investigate individuals’ willingness to constrain or seek out belief distortion, and study how being assigned to *experience* commitment or flexibility affects

---

<sup>1</sup>A large literature suggests that self-serving behavior is more likely when decisions can be rationalized by exploiting ambiguity or subjectivity in the decision environment (e.g., Konow, 2000; Haisley and Weber, 2010; Shalvi et al., 2011; Exley, 2015; Gneezy, Saccardo, and van Veldhuizen, 2018; Gneezy et al., 2020; Falk, Neuber, and Szech, 2020), by avoiding information about how their choices affect others (e.g., Dana, Weber and Kuang, 2007; Grossman, 2014; Grossman and van der Weele, 2017; Serra-Garcia and Szech, 2021), or by conveniently forgetting unpleasant news (Saucet and Villeval, 2019; Carlson et al., 2020). These belief processes can lead to self-deception, enabling self-serving behavior (see, for example, Bodner and Prelec, 2003; Mijovic-Prelec and Prelec, 2010; Bénabou and Tirole, 2016; Bénabou, Falk and Tirole, 2018).

self-serving behavior.

We conduct a series of experiments where participants in the role of advisor ( $N = 8,923$ ) face a potential moral dilemma and can choose the order with which they receive a sequence of signals. In many moral dilemmas, individuals receive information about what is in their best interest as well as information about what is best for another party. The order of information can constrain cognitive flexibility: Assessing what is best for another party without knowing one’s own incentives might raise attention to information about the other party’s outcome, committing individuals to a first unbiased judgment (e.g., Goldin and Rouse, 2000) and restricting the temptation to act self-servingly once new information is received (e.g., Babcock et al., 1995; Gneezy et al., 2020; Schwardmann, Tripodi and van der Weele, 2021). Consider experts – financial advisors, attorneys, accountants, expert witnesses, or reviewers – who have the ethical responsibility to make unbiased recommendations but may succumb to the temptation of favoring their private interests. When evaluating new information (e.g. new investment funds, insurance policies, new cases or materials), experts who anticipate being tempted to violate their duty may actively commit to accurate beliefs by first assessing the information while being blind to their incentives. Or, they may seek out the cognitive flexibility needed to distort their beliefs by first examining potentially biasing information.

In our experiments, an advisor recommends one of two products to an uninformed client and faces a potential conflict of interest. The payoff distribution of one of the products, which we refer to as “quality,” is uncertain. The advisor receives two pieces of information: a signal about the quality of the uncertain product and information about her private incentive (i.e., which product the advisor is incentivized to recommend). If no quality signal is provided, the advisor can recommend the incentivized product without facing a moral dilemma as both products have the same expected payoff. However, all advisors receive both pieces of information before making their recommendation. We study their choice of order with which to receive information. Seeing the quality signal first may increase the attention paid to this piece of information, thereby reducing the scope for bias in the processing of the signal and self-serving recommendations.<sup>2</sup> To explain the effects of information order on behavior, we present a stylized theoretical framework that builds on Bénabou

---

<sup>2</sup>The important role of attention and salience in economic choices has been shown in Gabaix et al. (2006), Bordalo, Gennaioli, and Shleifer (2012), Schwartzstein (2014), among others. There is also work on motivated attention (e.g., Sicherman et al., 2016; Golman et al., 2021). Some of this research has shown that new information not only informs decision-making but it can also focus attention on certain beliefs.

and Tirole (2002), in which quality signals receive more attention when they are seen first, in line with the literature on first impressions (Asch 1946, Anderson, 1965; Yates and Curley, 1986; Tetlock, 1983), work on anchoring and insufficient adjustment (e.g., Tversky and Kahneman 1974), and evidence on the effect of information order on self-serving behavior (e.g., Babcock et al., 1995). Advisors who first see quality information pay more attention to quality signals and have therefore less scope to self-servingly suppress signals that are in conflict with their incentive, which leads to less self-serving behavior. If ethical advisors anticipate that they may be tempted to provide a selfish recommendation, they may prefer to see the signal of quality first. By contrast, if they anticipate that they would like to enable self-serving information processing, they may prefer to see the incentive first and exploit the cognitive flexibility provided by this information order.

We begin by empirically establishing that, in our context, exogenously assigning advisors to an information sequence affects their likelihood of engaging in self-serving behavior. In line with prior work and the theoretical framework, when there is a conflict of interest, advisors are more likely to make recommendations that are in the client’s best interest when they assess the signal about quality first, compared to when they receive information about their incentives first. There is no effect of information order when advisors’ interests are aligned with those of the client.

Our main experiment investigates preferences, recommendations, and beliefs when advisors have the option to *choose* the sequence of information. First, we investigate preferences for information order. We use data from (a) a sample of professionals who self-report being employed in the finance (including insurance) and legal services industries, and (b) from a general (convenience) sample of online participants.<sup>3</sup> Across both samples, we find substantial heterogeneity in preferences. If the choice is costless, 45% of advisors in the convenience sample and 55% of advisors in the sample of professionals commit to more accurate beliefs by choosing to see quality first (with the remaining 55% and 45%, respectively, seeking out cognitive flexibility). Since advisors’ preferences are close to 50%, a concern is that their preferences indicate indifference. However, indifference is not a prominent self-reported explanation of advisors’ choices of information order. Moreover, when we introduce costs, advisors reveal a strict preference: 30% of advisors are willing to incur a financial cost to receive quality information first, committing to more accurate beliefs, and 41% of advisors are willing to incur a financial cost to see the incentive first, pursuing

---

<sup>3</sup>While we cannot verify their work status and experience, Huber and Huber (2020) compare one of our samples to a verified proprietary sample and find similar dishonesty in behavior.

cognitive flexibility.

Advisors' preferences to see quality information first are strongly correlated with advisors' morals, as measured in a separate task in which advisors always face a conflict of interest. They are also correlated with advisors' willingness to take up a stronger form of moral commitment: Advisors who prefer to assess quality first are more likely to blind themselves from learning about their incentive altogether. This evidence is in line with our theoretical framework and suggests that individuals anticipate that seeing quality first favors moral behavior.

Next, we investigate advisors' behavior: How does seeking out commitment or flexibility affect the rate of self-serving recommendations? To answer this question, in the experiment we implemented advisors' preferred information sequence with 75% chance. When advisors are assigned their preferred information sequence and are faced with a conflict of interest, there is a 19-20 percentage point gap in recommendations of the incentivized product between advisors who seek out flexibility and those who seek out commitment. Yet, there is no gap when advisors are not assigned to see information in their desired order. Conditional on preferences, being assigned to *experience* flexibility (vs. commitment) is crucial to advisors' ability to behave self-servingly, suggesting that behavior observed among those who are assigned their preferred information order does not just reflect sorting. For advisors who seek out commitment, being assigned to see quality first significantly reduces self-serving recommendations. This result confirms that altering the order of information to assess quality first can be an effective moral commitment strategy.

The behavior of advisors who seek out flexibility speaks to an important open question about the dynamics of self-deception: whether individuals can *intend* to self-deceive without rendering such intentions ineffective (Mele, 1987 and 2001; Bermúdez, 2000; see also Mijovic-Prelec and Prelec, 2010). Although in economics some theoretical models assume this type of self-deception is possible (e.g., Bénabou and Tirole, 2002), empirical evidence is lacking. A prominent hypothesis in the philosophical literature is that actively seeking flexibility might prevent individuals from subsequently being able to self-deceive and engage in self-serving behavior. In contrast with this hypothesis, our results suggest that actively seeking flexibility by choosing to see incentive information first does not impede advisors' ability to engage in self-serving behavior: Advisors who prefer and are assigned to see the incentive first are significantly more likely to make the self-serving recommendation than those who seek out flexibility but are not assigned to experience it.

Advisors' beliefs about product quality are in line with their recommendations. In line with motivated attention, when advisors receive a signal that conflicts with their financial interests, their beliefs are closer to the prior (as if they had not received a signal), compared to signals that are aligned with their interests. Further, advisors who pursue and get cognitive flexibility exhibit beliefs closer to the prior both when signals are in conflict and when they are aligned with their interests, consistent with signals of quality that are seen later receiving less attention. The theoretical framework highlights that when advisors pursue and get cognitive flexibility they can exploit the lower attention to signals to engage in (even) more motivated attention. In the data we find directional, though weak, evidence that they do. The findings are broadly consistent with cognitive flexibility enabling advisors to pay less attention to informative signals and thereby engage in more self-serving behavior.

Advisors' preferences and recommendations are consistent with a proportion of them being sophisticated about the effect of information order on behavior. In two additional experiments, we provide evidence in support of this interpretation. First, we test whether preferences for cognitive flexibility or commitment respond to changes in advisors' incentives (see also, Coutts, 2019). When we reduce the potential gains from distorting beliefs, reducing advisors' incentives to demand cognitive flexibility, very few advisors (13%) demand to see their incentives first. Yet, when the gains from belief distortion further increase we do not see a similar increase in the demand for cognitive flexibility. This concavity is consistent with advisors experiencing less moral conflict as their incentives increase and, hence, the increase in demand for cognitive flexibility responding less to the incentive increase. Second, we test whether third party participants (the Information Architects, IAs) anticipate the effect of information order on advisors' behavior. IAs do not receive information but choose the order in which advisors learn about their incentives and the quality signal. We vary IAs' incentives to be aligned with the advisors' or the clients' payoffs, and ask them to choose the order of information for advisors. Our findings reveal that IAs are more likely to have advisors first assess quality without seeing the incentive when their own incentives are aligned with those of the client.

Our research contributes to a growing literature on the malleability of moral behavior. While prior work has documented individuals' tendency to behave self-servingly despite an overall desire to feel moral (e.g., Gino, Norton and Weber, 2016), an open question is whether, in anticipation of the conditions that facilitate belief distortion, individuals desire to constrain belief distortion to uphold their

morals. Our findings suggest that some advisors anticipate that changes to the way information is presented can constrain belief distortion, and that moral individuals are significantly more likely to take up opportunities for moral commitment, choosing to blind themselves from incentive information when making their initial judgments.

Understanding how to mitigate the negative consequences of information asymmetries in presence of conflicts of interest (see, e.g., Darby and Karni, 1973; Crawford and Sobel, 1982), has implications for the design of expert systems. Experts across a variety of professions – such as financial or legal professionals, expert witnesses, reviewers evaluating scientific research, and admission officers assessing candidates’ qualifications – are often called to make judgments that may be biased by private interests (e.g., Robertson and Kesselheim, 2016). Our findings suggest that some individuals prefer to learn about potentially biasing information first, which provides them with more scope for self-serving behavior, but others are willing to *temporarily* blind themselves from this information as a way to commit to moral behavior. Even if, over time, those experts may learn their incentives, first impressions can affect experts’ quality assessments and have a long-lasting effect on expert behavior (e.g., Chen and Gesche, 2017).

Our findings have implications for the self-selection of experts into organizations as well as for organizational design. They suggest that experts could self-select into types of organizations according to their practices or policies to prevent bias, consistent with evidence that social and moral preferences correlate with selection into different industries (e.g., Hanna and Wang, 2017; Barfort et al. 2019). In addition to the importance of self-selection, our findings suggest that experiencing flexibility (or commitment) is key. Even within the same industry, those who make decisions in organizations often have some discretion in designing the informational structures and institutional arrangements that govern their behavior, from deciding whether potentially biasing information about candidates is available to hiring managers, to deciding what information different experts have available when making their assessments. Our findings suggest that these individuals may make such design decisions with commitment or flexibility goals in mind.

## 2 Experimental Design

Our aim is to investigate individuals’ willingness to constrain or seek out belief distortion and examine how these choices affect self-serving behavior when unwelcome



information cannot be avoided. Studying these questions requires an environment in which (i) individuals are tempted to put their own interests above those of another party, and (ii) that provides them with the cognitive flexibility needed to pursue private gains. Further, it requires an environment where (iii) individuals can actively pursue cognitive flexibility (or, conversely, mitigate it), when given the choice, and (iv) that allows studying the effect of this active choice on subsequent behavior and beliefs. Our experiment is designed to accommodate these four features.

## 2.1 The Advice Game

The advisor recommends one of two products, A and B, to an uninformed client. Each product is presented as an urn containing five balls, as displayed in Figure 1. Product A has three \$2 balls and two \$0 balls. That is, product A pays \$2 with prob 0.6, and \$0 otherwise (an expected return of \$1.20). Product B's payoff depends on the state, which we refer to as product B's quality and that can be high ( $H$ ) or low ( $L$ ). We denote quality by  $s \in \{H, L\}$ , and the probability that  $s = H$  is 0.5. If  $s = H$ , then B has four \$2 balls and one \$0 ball. It thus yields a higher probability of receiving \$2 than product A, as it pays \$2 with prob 0.8, and \$0 otherwise, for an expected return of \$1.60. If  $s = L$ , then B has two \$2 balls and three \$0 balls. It thus yields a lower probability of receiving \$2 than product A, as it pays \$2 with prob 0.4, and \$0 otherwise, for an expected return of \$0.80. The quality of product B ( $s$ ) is unknown to the advisor.

Before making the recommendation, the advisor receives a signal about quality: a ball that is randomly drawn from product B, which allows the advisor to update her beliefs about whether  $s = H$  or  $s = L$ . Upon learning the signal, the advisor chooses which product (A or B) to recommend to the client. After receiving the recommendation, the client chooses whether to follow the advice and is paid according to one of the balls randomly selected from the product she selects.

The advisor receives an incentive ( $\iota = \$0.15$ ), for recommending either product A or product B. Depending on what product is incentivized and on which signal is drawn from product B, the advisor may face a conflict of interest. If the commission is for product B and the signal is a \$0 ball, the advisor faces a conflict between pursuing the commission (i.e., recommending product B) and making the recommendation that is in the clients' best interest (i.e., recommending product A). Similarly, if the commission is for product A and the signal is a \$2 ball, the advisor has to choose between maximizing her payoff (i.e., recommending product A) or

making the recommendation that is best for the client (i.e., recommending product B). In the remaining cases, the advisor does not face a conflict of interest.

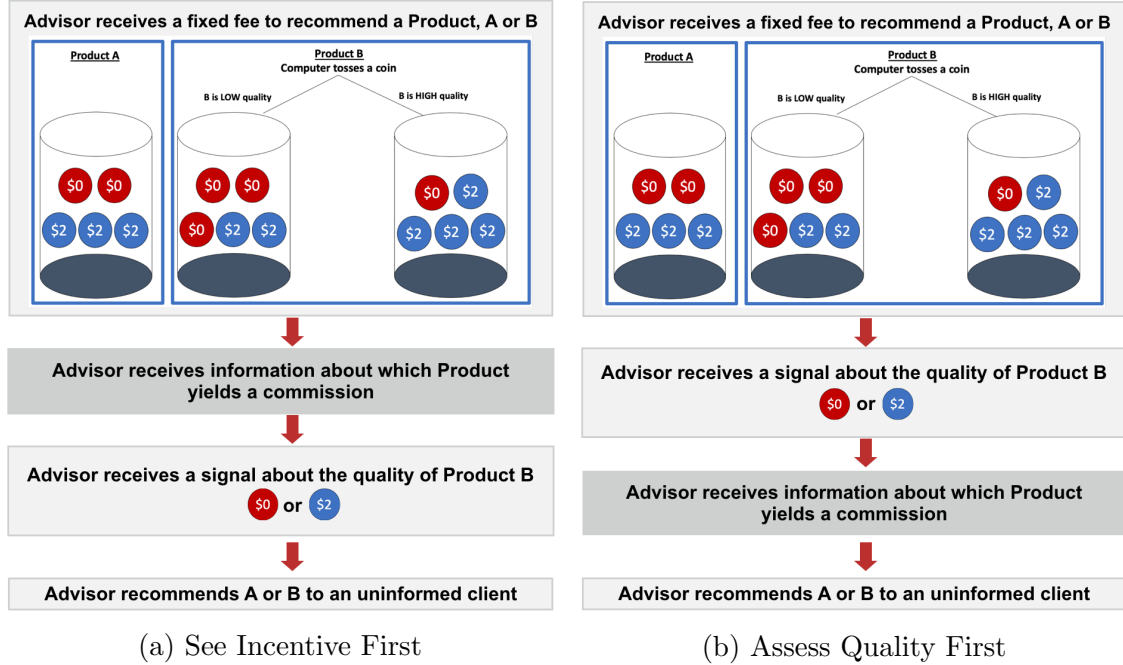


Figure 1: The Advice Game

## 2.2 Main Experiments

We conduct four online experiments, as summarized in Table 1. We first present the two main experiments, NoChoice and Choice. In Section 3, we present a stylized theoretical model that provides a lens through which to view the effect of information order in those experiments, guiding our main hypotheses. In Section 4, we describe the two additional experiments and the experimental procedures.

**A. The NoChoice Experiment.** The goal of the first experiment is to establish that cognitive flexibility varies with the order of information. This experiment has two treatments. In the See Incentive First treatment, the advisor first receives information about which product recommendation is incentivized (Figure 1a) and then, on a later screen, sees the quality signal about product B. In the Assess Quality First treatment, the advisor first sees the quality signal about product B and only later, on the recommendation screen, learns about her incentive (Figure 1b). In both treatments, the evaluation of the signals only occurs in the advisors' mind.

The incentive is always shown on the recommendation screen, with what varies being whether the incentive information also appears before the quality signal.<sup>4</sup>

Table 1: Experimental Design Outline

Experiment	Treatment	What do advisors see first?	N
<b>Documenting Cognitive Flexibility: Information Order Affects Recommendations</b>			
NoChoice	See Incentives First	Incentive	152
	Assess Quality First	Quality Signal	147
<b>Preferences for Information Order: Cognitive Flexibility or Moral Commitment?</b>			
<i>Choice</i> <i>Main Treatments</i>	Choice Free - Professionals	Advisor’s Choice	712
	Choice Free	Advisor’s Choice	2377
	Incentive First Costly	Advisor’s Choice	1358
	Quality First Costly	Advisors’ Choice	1067
<i>Robustness (in Appendix)</i>	Choice Free - High Stakes (10-fold)	Advisor’s Choice	275
	Choice Free - High Stakes (100-fold)	Advisor’s Choice	110
	Choice Free - Replication	Advisor’s Choice	385
	Choice Free - Deterministic	Advisor’s Choice	369
<b>Additional Evidence</b>			
Choice Stakes	Low Incentive	Advisor’s Choice	484
	Intermediate Incentive	Advisor’s Choice	511
	High Incentive	Advisor’s Choice	478
Information Architect	IA-Advisor	Third Party Choice	245
	IA-Client	Third Party Choice	253

*Notes.* This table summarizes the main experiments in the paper. Table B.1 in Online Appendix B provides the detailed description of the sample, waves, pre-registrations, and treatments in each of the experiments. The N indicated in this table refers to sample sizes after pre-registered exclusions.

**B. The Choice Experiment.** In this experiment we elicit advisors’ preferences for information order in the advice game, and examine how being assigned to experience a given order affects recommendation decisions. To estimate the effect of information order on recommendations, conditional on advisors’ preferences, advisors’ choices are implemented probabilistically. With 50% chance, the advisors choice is implemented, while with the remaining 50% chance, the advisor receives a 50/50 randomization. Advisors are informed that their preference is implemented with 75% probability. In this experiment, there are three conditions. In the *Choice Free* treatment, advisors make a simple choice between seeing the incentive first or assessing quality first. We conducted this experimental treatment with a sample of individuals who self-report to work in industries in which advice is frequent—finance

<sup>4</sup>In Online Appendix D, we report the design and result of an additional wave of data collection that tests the effect of receiving the quality signal and incentive information simultaneously.

(including insurance), and legal services (*Choice Free - Professionals*) as well as with individuals from a convenience sample (recruited from Amazon Mechanical Turk or AMT via Cloudresearch). Varying the sample allows us to compare the preferences and recommendations of individuals who are likely to deal with conflicts of interest in their professional lives to those of participants who may have such experiences less often.

To examine whether advisors have strict preferences to see the incentive first or to assess quality first, we introduce a cost of seeing the incentive first (*Incentive First Costly* treatment) and a cost for assessing quality first (*Quality First Costly* treatment), within the AMT sample. In each treatment, advisors forgo an additional payment, equivalent to a third of their commission (\$0.05), if they choose to see their incentive or the signal of quality first, respectively.

As part of this experiment, we conduct two robustness tests. First, we examine whether the probabilistic implementation of advisors' preferences affects their recommendations. We find that when implementing their preferences with certainty (in the *Choice Free-Deterministic* treatment), the effect of information on recommendations is not significantly different from that observed for advisors who were assigned their preference (in the *Choice Free - Replication* treatment, see Online Appendix E). Second, a concern in the Choice experiment is that the incentives in the experiment are relatively small. Previous work has shown that even small incentives can influence expert decisions (DeJong et al. 2016; Malmendier and Schmidt, 2017; Marechal and Thöni, 2019) and that cognitive biases tend to persist across a variety of incentive sizes (e.g., Enke et al., 2021). Since incentives for experts may vary in size and often be larger, we implemented two variations of the *Choice Free* experiment that increased the stakes in the experiment by a factor of 10 (*High Stakes - 10 fold*) or 100 (*High Stakes - 100 fold*). We find no significant change in the effect of information order on recommendations, suggesting that the results are robust to larger incentives (see Online Appendix C.4).

### 3 Theoretical Framework

To explain how an advisor can leverage the order of information to restrict or enable self-serving behavior, we present a stylized theoretical framework. We adopt the framework of self-deception by Bénabou and Tirole (2002), based on attention

management and an inner conflict in the advisor’s morality.<sup>5</sup> To reduce notation, we modify the advice game to focus on the distinction between the presence or absence of conflict between the advisor’s incentive and the quality signal. In this simplified game, the signal the advisor can receive either indicates a conflict with the incentive ( $\sigma = c$ ) or no conflict with the incentive ( $\sigma = nc$ ). The prior likelihood that the signal is  $\sigma = c$  is  $\phi$ . We assume clients follow the advisor’s recommendation.

### 3.1 Limited and motivated attention

Attention is often limited (e.g., Kahneman, 1973) and motivated (e.g., Lang, Bradley, Cuthbert, 1997; Karlsson, Loewenstein and Seppi, 2009; Amasino, Pace and van der Weele, 2021). The literature on first impressions indicates that it may be automatic to pay more attention to the first piece of information individuals receive (e.g., Asch 1946; Anderson, 1965; Yates and Curley, 1986; Tetlock, 1983).<sup>6</sup> We hence propose that cognitive flexibility varies with the order with which information is presented.<sup>7</sup>

Seeing the signal of quality  $\sigma$  first ( $f = q$ ) increases the likelihood that the advisor encodes (or remembers, pays attention to) this signal relative to seeing the incentive first ( $f = i$ ). The reason is that, when the signal of quality is seen first, the incentive is not known, and the advisor is more likely to encode the quality signal. By contrast, seeing information about the incentive first leads the advisor to focus her attention on the incentive and pay less attention to the signal of quality. Formally, the probability that the quality signal is encoded  $\lambda^f < 1$ . Encoding of the quality signal is more likely when the quality signal is seen first  $\lambda^q > \lambda^i$ . If the signal of quality is encoded, it can be in conflict ( $\sigma = c$ ) or not in conflict ( $\sigma = nc$ ) with the incentive. If the signal is not encoded, the advisor does not know the signal, leading to  $\sigma = \emptyset$ . Incentive information is assumed to always be encoded, since all advisors are shown the incentive information on the recommendation screen.

---

<sup>5</sup>We thank the editor and review team for encouraging us develop a theoretical framework that formalizes our predictions and guides our analyses.

<sup>6</sup>Note that there is also a literature finding evidence of recency effects (Benjamin, 2019). Hogarth and Einhorn (1989) propose a belief-adjustment model for updating beliefs. Existing evidence in Gneezy et al. (2020) and in our first (NoChoice) experiment suggests that primacy effects dominate in the advice game we study.

<sup>7</sup>This assumption is in line with our empirical data, where advisors’ belief updating patterns are in line with the work on first impressions. Consistent with attention playing an important role, some advisors self-reported (in an open ended question) that seeing the incentive first “*gives it more salience*” or “*might make me pay less attention to what I was learning*” and that seeing quality first would make them “*pay closer attention,*” allowing them to “*have better knowledge about the products*” and preventing the incentives from “*clouding their judgment.*”

### 3.2 Unstable morality

The advice game aims to capture the moral dilemma that arises when the product that the advisor is incentivized to recommend yields a lower expected payoff to the client. Advisors who recommend the incentivized product may feel immoral, and experience a moral cost (or disutility in monetary units)  $m$ . This moral cost can be viewed as akin to lying costs in sender-receiver games (e.g., Gneezy, 2005; Abeler, Nosenzo and Raymond, 2019), because a large majority of the clients follow advisors' recommendations.

Many individuals care about behaving morally, but moral behavior is often unstable (for a review, see, Gino, Norton and Weber, 2016). Recent work highlights that acting self-servingly may be tempting for some individuals (e.g., Bénabou, Falk and Tirole, 2018), while others may fear being too generous. In the context of the advice game, individuals who feel conflicted about the right behavior may initially want to act selfishly or morally, but anticipate that once they learn about their incentive and the quality signal their recommendation may change (tempting them to act more morally or selfishly).<sup>8</sup>

To illustrate the advisor's inner conflict, we adopt a dual-self framework (Bénabou and Tirole, 2002; Bodner and Prelec, 2003), by which the advisor's Self 0 and Self 1 may differ in their moral costs. Specifically, moral costs are randomly drawn for Self 0 and Self 1, who are both risk neutral. Let  $m_t$  be the moral cost of Self  $t \in \{0, 1\}$ . We assume that  $m_t$  is distributed uniformly on  $[0, M]$ , and independently drawn, with  $M > \iota$ , where  $\iota$  is the advisor's incentive payment. This stylized formulation of the inner conflict does not include an explicit concern for self-image (see, e.g., Bénabou, Falk and Tirole, 2018, which includes both self-image and temptation; and models of self-image by Bénabou and Tirole, 2011; and Grossman and van der Weele, 2017), to simplify exposition, while allowing Self 0 to worry that after the information is presented her moral preferences may change.<sup>9</sup>

---

<sup>8</sup>In line with this intuition, our data show that when advisors are not assigned to see quality first, though they prefer it, they behave more self-servingly. Similarly, when advisors are not assigned to see the incentive first, though they prefer it, they behave more morally. Echoing this behavior several advisors report that the commission would tempt them to be less moral, e.g., *"I felt it was better to learn (my incentive) after so that I wasn't tempted to make a decision out of greed."*, while some advisors mentioned wanting to know the commission first to avoid feeling tempted to go with what was best for the client: *"I wanted to know which one had a commission upfront so I could be less tempted by the randomized drawing of product B."*

<sup>9</sup>Qualitatively similar predictions would result if Self 1's moral costs would be modeled with a  $\beta$  (temptation) parameter relative to Self 0's moral costs, e.g.  $m_1 = \beta m_0$ , where  $\beta$  could be larger or smaller than 1.

Self 0 manages attention to the signal of quality, knowing  $m_0$  but not  $m_1$ , while Self 1 makes the recommendation decision based on the signal received from Self 0. At the beginning of the advice game, Self 0 encodes the signal of quality  $\sigma$  with probability  $\lambda^f$  and sends  $\hat{\sigma}$  to Self 1. Based on  $\hat{\sigma}$ , Self 1 forms a belief about the likelihood that the signal is in conflict with the incentive ( $r(\hat{\sigma})$ ). Self 1 chooses whether to recommend the incentivized product ( $x = 1$ ) and receive the incentive  $\iota$ , or not ( $x = 0$ ). Her utility is:

$$U_1(x|\hat{\sigma}, m_1) = [\iota - m_1 r(\hat{\sigma})]x.$$

From the perspective of Self 0, her utility at the recommendation stage may differ from that of Self 1 due to a difference in moral costs. Self 0 knows the signal that was encoded initially ( $\sigma$ ), leading to:

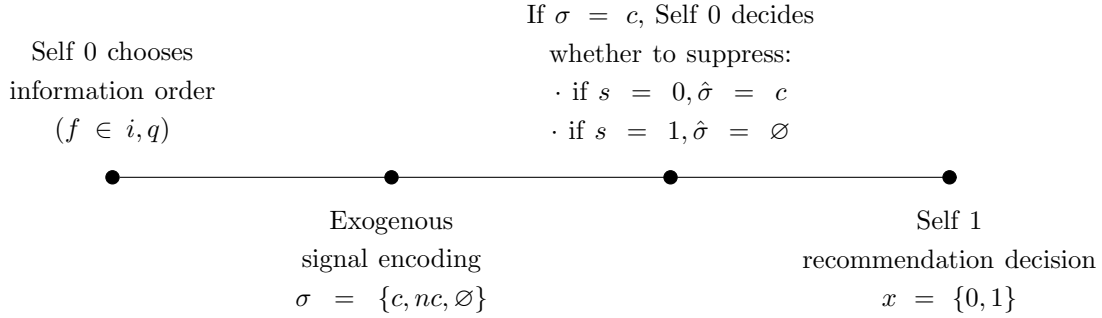
$$U_0(x|\sigma, m_0) = [\iota - m_0 r(\sigma)]x.$$

The potential conflict in moral costs between Self 0 and Self 1 may lead Self 0 to prefer to “manage” Self 1’s attention. If Self 0 starts the advice game with a high moral concern – that is, she initially draws high moral costs  $m_0$  –, she may anticipate that her later Self 1 may have a lower moral concern (low  $m_1$ ) and prefer “moral commitment,” by increasing attention to signals of quality. However, if Self 0 starts the advice game with a low moral concern (low  $m_0$ ), this would motivate her to seek to pay less attention to informative signals about quality.

Since a signal of quality may not be encoded exogenously, Self 0 can exploit this limited attention to engage in “motivated attention,” as in Bénabou and Tirole (2002). When the signal  $\sigma$  is actually encoded and it is in conflict with the incentive, Self 0 chooses whether to “suppress” to the signal ( $s = 1$ ) or not ( $s = 0$ ). By suppressing, Self 0 attempts to act as if it was never encoded to begin with – a form of reality denial (see, e.g., Bénabou and Tirole, 2016). If Self 0’s signal is  $\sigma = c$ , Self 0 can choose  $\hat{\sigma}$  to be  $c$  or  $\emptyset$ . Otherwise,  $\hat{\sigma} = \sigma$ . Suppressing an encoded signal is costless, and Self 0 suppresses with probability  $p_s \in [0, 1]$ . When Self 1 does not receive a signal (it is  $\hat{\sigma} = \emptyset$ ), she uses Bayes’ Rule to form a belief about the likelihood that it is in conflict with the incentive,  $r(\emptyset)$ , as follows:

$$r(\emptyset) = \Pr(\sigma = c|\hat{\sigma} = \emptyset, f, p_s) = \frac{\lambda^f p_s \phi + (1 - \lambda^f) \phi}{\lambda^f p_s \phi + (1 - \lambda^f)},$$

where  $\phi$  is the prior likelihood that the signal is in conflict with the incentive. If the signal received by Self 1 is in conflict with the incentive, then  $r(c) = 1$ , and if it is not in conflict with the incentive,  $r(nc) = 0$ . Figure 2 presents a timeline of the model when Self 0 chooses the information order.



*Notes:* This figure shows, from left to right, the steps in the model when Self 0 decides the information order. Signal encoding occurs exogenously, depending on the information order  $f$ . If  $\sigma = c$ , Self 0 decides whether to suppress the encoded signal. Lastly, Self 1 makes her recommendation decision. If Self 0 does not decide (NoChoice experiment), the first step would be removed.

Figure 2: Timeline of the model, when advisors choose the information order

If Self 1 receives a signal that is in conflict with the incentive  $\hat{\sigma} = c$ , Self 1 chooses  $x = 1$  only if  $m_1 \leq \iota$ . If the signal received is not in conflict with the incentive, there are no moral costs and Self 1 chooses  $x = 1$ . If Self 1 does not receive a signal, her inference about  $r(\emptyset)$ , the risk of recommending a product that is in conflict with the incentive, determines her decision to recommend the incentivized product. She recommends the incentivized product if:

$$m_1 \leq \frac{\iota}{r(\emptyset)}.$$

As in the experiment, we assume that, if Self 1's belief about the likelihood that the signal is in conflict with the incentive is the same as the prior, Self 1 recommends the incentivized product, i.e.,  $\iota - \phi M > 0$ .

### 3.3. No Choice of Information Order

We start by considering first the case where Self 0 cannot choose the information order (as in the NoChoice experiment). When Self 0 has a high moral cost,  $m_0 > \iota$ , she always conveys the signals that are encoded and never suppresses. This minimizes the likelihood that Self 1 recommends the incentivized product when the signal is in conflict with the incentive, providing a form of “moral commitment.”



When Self 0 is selfish and has low moral costs,  $m_0 < \iota$ , Self 0 has an incentive to suppress signals that are in conflict with the incentive. Denote the probability that Self 1 recommends the incentivized product when she does not receive a signal of quality ( $\hat{\sigma} = \emptyset$ ) by  $q = \frac{\iota}{r(\emptyset)M}$ . The expected utility of Self 0 from choosing to suppress with  $p_s$  is

$$E(U_0) = \lambda^f \left( (1 - \phi)\iota + \phi((1 - p_s)((\iota - m_0)\frac{\iota}{M}) + p_s(\iota - m_0)q) \right) + (1 - \lambda^f)(\iota - \phi m_0)q.$$

Self 0 suppresses the signal of quality as often as possible, as long as Self 1 still recommends the incentivized product, and hence chooses,

$$p_s^* = \min\left\{\frac{(1 - \lambda^f)(\iota - \phi M)}{\lambda^f \phi (M - \iota)}, 1\right\}.$$

Because the signal of quality is encoded less often when the incentive is seen first ( $f = i$ ), a selfish Self 0 can exploit the lower attention to engage in more motivated attention (suppression), while still persuading Self 1 to recommend the incentivized product in the absence of a signal. The lower attention and increased ability to suppress thus imply that seeing the incentive first provides (more) cognitive flexibility. This result is summarized in Proposition 1 (further details in Online Appendix A).

**Proposition 1.** *When the signal of quality is shown first, the advisor is less likely to suppress it and less likely to recommend the incentivized product when it conflicts with the incentive, than when the information about the incentive is shown first.*

Hence, when there is a conflict of interest, the likelihood of recommending the incentivized product increases when advisors see their incentive first. When there is no conflict of interest, the advisor recommends the incentivized product under both information orders.<sup>10</sup> This yields our first hypothesis.

**Hypothesis 1 (NoChoice experiment).** If advisors are assigned to See the Incentive First, the likelihood with which advisors recommend the incentivized product

---

<sup>10</sup>This result highlights that the difference between recommendations is expected to be present when the signal is in conflict with the incentive, due to our focus on the role of attention management to signals as the mechanism through which information order affects recommendations. A difference in recommendations when there is no conflict of interest may also arise if advisors who see the quality signal first are less likely to pay attention to incentive information. We find little evidence for this in our data. If there is no conflict of interest, the difference between information orders is either absent or small, between 20 to 30 percent of that observed when there is a conflict of interest.

when the signal is in conflict with the incentive is higher than when they are assigned to See Quality First.

### 3.3 Advisor’s Choice of Information Order

Given the effects of information order on the attention process and recommendation decisions, what order of information does the advisor prefer?<sup>11</sup> We first consider the case of a sophisticated advisor who correctly anticipates the decrease in attention when the incentive is seen first. As shown in Proposition 2, if Self 0 has low moral costs, she prefers to see the incentive first, since it affords more “cognitive flexibility”. In contrast, if Self 0 has high moral costs, she prefers to see the quality signal first to have more “moral commitment.”

**Proposition 2 (Sophisticated Advisors).**

- *If Self 0 is selfish ( $m_0 \leq \iota$ ), she chooses to see the incentive first ( $f^* = i$ ). This order increases the likelihood that Self 1 recommends the incentivized product when the signal is in conflict with the incentive.*
- *If Self 0 is moral ( $m_0 > \iota$ ), she chooses to see quality first ( $f^* = q$ ), which decreases the likelihood that Self 1 recommends the incentivized product when the signal is in conflict with the incentive.*

How would this prediction change if advisors are not sophisticated about the malleability of attention? We define a naïve advisor as one who believes that the order of information does not affect attention. Formally, the advisor believes that her attention is as limited when seeing the incentive first as when seeing quality first,  $\hat{\lambda}^q = \hat{\lambda}^i = \lambda^i < 1$ . If the advisor were naïve, then she would not anticipate any effect of information order, leading to Proposition 3.

**Proposition 3 (Naïve Advisors).** *If the advisor does not anticipate the effect of information order on attention, she is indifferent between seeing the incentive first or seeing quality first.*

These results yield Hypothesis 2, for the Choice experiment.

**Hypothesis 2 (Choice experiment).**

---

<sup>11</sup>We assume that the advisor’s choice is implemented with certainty to simplify exposition. We discuss the case in which the advisor’s preference is not implemented in Online Appendix A.

- (a) Preferences: If advisors are sophisticated, those who are more selfish (lower moral costs) are willing to pay to see the incentive first, while advisors who are more moral (higher moral costs) are willing to pay to see quality first. If advisors are naïve, they are not willing to pay for any information order.
- (b) Recommendations: Advisors who actively choose (and pay) to see the incentive first are more likely to recommend the incentivized product if the signal is in conflict with the incentive. Conversely, advisors who choose (and pay) to assess quality first are less likely to recommend the incentivized product if the signal is in conflict with the incentive.

The theoretical framework we proposed relies on two simplifying assumptions the validity of which we explore in the data analyses. The framework assumes that advisors' active choice of information order does not restrict their ability to suppress signals that are in conflict with their incentives. Philosophers, however, have argued that intentionality can decrease the scope for self-deception (e.g., Mele, 1987 and 2001).

Our experiments allow us to better understand the role of intentionality in self-serving recommendations, which we test in two ways. First, we test whether advisors who choose to see the incentive first are equally able to distort recommendations when they *experience* more cognitive flexibility, relative to those assigned to see quality first, although they potentially intended to self-deceive. Second, focusing on advisors who are assigned their preferred order, we test whether the gap in recommendations between advisors who prefer to see the incentive first and those who prefer to see quality first is larger in the Choice experiment than in the NoChoice experiment. This comparison allows to measure whether information order affects advisors similarly when such information order is directly chosen by advisors.

The theoretical framework also assumes belief distortion occurs through advisors' limited and motivated attention to the signal of quality of product B. This approach complements existing research on self-serving biases showing that individuals may distort their beliefs about what is fair in a self-serving manner (e.g., Babcock et al., 1995; Gneezy et al., 2020) allowing them to maintain a self-image as moral (e.g., Bénabou and Tirole, 2011; Grossman and van der Weele, 2017; Bénabou, Falk and Tirole, 2018). If belief distortion takes place through attention to quality signals, we would first expect that advisors who choose and are assigned to see the incentive first hold beliefs regarding the quality of the product that are closer to the prior

(as they pay less attention) than those of advisors who prefer to assess quality first. If advisors exploit lower attention to engage in more suppression, when signals are in conflict with the incentive, advisors' beliefs should be (even) closer to the prior when they choose and are assigned to see the incentive first.

## 4 Additional Experiments and Procedures

### 4.1 Additional Evidence of Anticipation

**A. The Choice Stakes Experiment.** We test whether advisors' preference to see the incentive first responds to their incentive to recommend the incentivized product. Based on the theoretical model, if the gains from flexibility decrease (due to a decrease in the advisor's incentive), and advisors are sophisticated, we would expect their preference to see the incentive first to drop. In contrast, if the advisor's incentive increases, the effect on her preference is ex-ante unclear. Their preference to see the incentive could increase, since there is a larger gain from flexibility. Or, their preference could be weakened, if the incentive is large enough and the advice decision no longer presents a moral dilemma (see Online Appendix A for details about the predictions of the theoretical framework). In the experiment, we keep the payoffs for the client the same, and vary the incentive for the advisor to be either low, \$0.01 in the Low Incentive treatment, the same as in the Choice experiment, \$0.15 in the Intermediate Incentive treatment, or double it to \$0.30 in the High Incentive treatment. Throughout, choosing to see the incentive first is costly as in the Incentive First Costly treatment.

**B. The Information Architect Experiment.** We introduce third-party participants in the role of Information Architects (IAs), who are matched with an advisor and choose the order in which advisors receive information in the advice game (see Instructions in Online Appendix G). To investigate whether IAs anticipate the effect of information order on behavior, we either align the IAs' incentive with that of the advisor or that of the client. In the *IA-Advisor* treatment, IAs receive a \$0.15 payment if the advisor recommends the incentivized product. In the *IA-Client* treatment, IAs receive a \$0.15 payment if the advisor recommends the product with the highest expected payoff for the client. If IAs anticipate the effect of information order, and they are only motivated by the incentives in each treatment, we would expect them to prefer to see the incentive first more often in the *IA-Advisor* treat-

ment than the *IA-Client* treatment, since their incentive in *IA-Advisor* is to increase the chance that the advisor recommends the incentivized product.

In this experiment, the IA chooses the order of information for the advisor without ever learning the realized incentivize and quality signal. In doing so, we can remove curiosity from driving preferences for information order. To further examine whether individuals anticipate the effect of information order on recommendations, in Online Appendix F we report an additional experiment where we ask third party individuals to predict recommendation rates under the different information orders (see, e.g., DellaVigna and Pope, 2018).

## 4.2 Experimental Procedures

We conducted all experiments except the Choice Free - Professionals treatment, on CloudResearch (Litman, Robinson, and Abberbock, 2016), a platform that allowed us to recruit a sample of high quality participants from Amazon Mechanical Turk (AMT).<sup>12</sup> All experiments on AMT were pre-registered on aspredicted.org. The Choice experiment was conducted in three waves, with each wave pre-registered separately. Online Appendix B provides pre-registration numbers, detailed design information, recruitment procedures, and exclusion criteria for the experiments. The sample of professionals was drawn from individuals who self-report to work in two industries in which advice is very frequent: finance and insurance, and legal services. We used Prolific Academic (Palan and Schitter, 2018) and CloudResearch to target the experiment to professionals in these industries.<sup>13</sup>

Participants received a base payment of either \$0.50 or \$1 for participating in a 5-7 minute study. As detailed above, in most of the advice game experiments, all advisors received a \$0.15 commission depending on their recommendation, and one out of 10 advisors was matched with a client. In the Choice Free - Professionals treatment and in a subsample in the Choice Free AMT treatment, we implemented

---

<sup>12</sup>Existing research shows that classic behavioral experiments have been successfully replicated on this platform (Paolacci, Chandler, and Ipeirotis, 2010), which is more and more commonly used by economists (e.g., DellaVigna and Pope, 2018) and allows us to recruit a large sample of participants.

<sup>13</sup>Prolific has their own sample of participants, and we recruited as many professionals as possible. CloudResearch draws professionals from AMT, and again we recruited as many professionals as possible. We pool these two samples since choices regarding the preferred sequence of information did not vary significantly across them ( $p=0.308$ ), and recommendations did not differ either ( $p=0.820$ ). Concurrent work focusing on truth-telling among financial professionals on Prolific and a proprietary pool consisting of financial professionals (portfolio managers, financial advisors, etc.) found similar behavior across pools (Huber and Huber, 2020).

a probabilistic payment structure, keeping the expected payoff unchanged. We paid 1 out of 100 advisors a \$15 commission, and matched all of the randomly selected advisors with a client. In these treatments, the payoffs of product A or product B were scaled up to \$0 or \$20.<sup>14</sup>

Since our main interest is in the cases where advisors faced a conflict of interest, we predetermined which product yielded a commission in a way that maximized the number of cases in which advisors faced a conflict of interest. All advisors randomly assigned to having a low-quality product B received a commission for recommending product B; all advisors randomly assigned to having a high-quality product B (i.e., four blue (\$2) balls and one red (\$0) ball) received a commission for recommending product A. By this design, 70% of advisors faced a conflict between maximizing their gains and providing advice that was in the best interest of the client.

At the end of the experiments, we randomly selected advisors according to the procedures of each experiment and sent each advisor’s recommendation to a client. We recruited clients ( $N = 924$ ) later and informed them that advisors had received information about the two products and had made a recommendation.<sup>15</sup> Clients saw their advisor’s recommendation and then made a choice between the two products; they received no other information about the products. Overall, 84% of clients followed the advisor’s recommendation.

#### 4.2.1 Additional measures

After the recommendation stage, we collected additional measures.

**Beliefs.** We elicited advisors’ beliefs about the likelihood that the quality of B was low by asking advisors (i) to choose one of ten 10 percentage-point intervals, and (ii) to indicate the exact likelihood by entering a number from 0 to 100. The first

---

<sup>14</sup>To test whether advisors display different responses to probabilistic incentives, in one of the waves of the Choice experiment, we recruited 1,053 participants and randomized whether incentives were probabilistic as in the professional sample and whether the incentivized product was presented on the left side or the right side of the screen. We found no effect of incentive size, order or their interaction on the preference to see the incentive first ( $t - \text{stat} = -1.46$ ,  $p = 0.144$  for incentive size,  $t - \text{stat} = 1.41$ ,  $p = 0.159$  for order, and  $t - \text{stat} = -0.03$ ,  $p = 0.980$  for the interaction of the two). We also found no effect of incentive size, order or their interaction on recommendations ( $t - \text{stat} = 0.34$ ,  $p = 0.733$  for incentive size,  $t - \text{stat} = 0.45$ ,  $p = 0.652$  for order, and  $t - \text{stat} = 0.85$ ,  $p = 0.396$  for their interaction). Hence, we pool the data and control for these design variations in all regression analyses.

<sup>15</sup>Following the instructions, we recruited 1 out of 10 clients for all treatments other than the Choice Free - Professionals treatment and a subsample of the Choice Free treatment in the second wave of the experiment, where we recruited 1 out of 100 clients, and the High Stakes - 100 fold treatment where we matched each advisor with one client.

measure was incentivized: Advisors received \$0.15 for a guess in the correct range.<sup>16</sup>

**Moral costs.** We measured advisors’ moral concern for providing a recommendation that helps the client, when there is a conflict of interest, using a multiple price list, in all experiments except for *Choice Free - Professionals*. Advisors made five recommendation decisions to a newly matched participant, the “advisee.” There were two products, X and Y. Product Y had the same payoffs as product B in the experiment. Advisors were incentivized to recommend Y, with a \$0.15 commission, and received a signal of quality of product Y that indicated that a \$0 had been drawn from Y. Product X varied across 5 different decisions. It paid \$2 with probabilities 1, 0.8, 0.6, 0.4, and 0 respectively, and \$0 otherwise. Given the payoffs of X, recommending Y harmed the client if X paid \$2 with a probability of 0.6 or higher. In those decisions, if the advisor chose to recommend Y, she could suffer moral costs. We consider this measure to capture the moral costs of Self 0, within the theoretical framework, because the signal of quality was presented at the same time as products X and Y. For simplicity, we refer to this (standardized) measure as the advisor’s overall selfishness.<sup>17</sup> In all of the main analyses, following our pre-registrations, we focus on attentive advisors who gave consistent responses in this task, excluding those who switched multiple times. The results remain qualitatively similar if we include them, as shown in Online Appendix C.

**Blinding.** In the third wave of data collection of the Choice Experiment, we measure take up of a stronger form of moral commitment in a separate task. The task was conducted after participants took part in the main experiment and completed the elicitation of moral costs. Participants were assigned the role of advisor, and gave advice to a new participant in the role of advisee about a different set of products.<sup>18</sup> Advisors knew that the incentive and the signal of quality would be drawn again. Advisors then either chose to blind themselves, and receive information about their incentive only *after* providing her recommendation, or not to blind themselves, which implied that they received information about the signal of quality and the incentive

---

<sup>16</sup>The payment was \$15 in the Choice Free treatments in which 1 out of 100 advisors was selected for payment

<sup>17</sup>At the end of the experiment, we randomly selected one out of 10 advisors, randomly picked one of the 5 recommendations, and showed them to a client. For this purpose, we recruited a total of 866 clients across all the experiments reported in Table 1. Of these, 80% of clients followed the advisor’s recommendation.

<sup>18</sup>At the end of the experiment, we randomly selected one out of 10 advisors and send their recommendation to an advisee. For this purpose, we recruited 188 advisees. Of these, 84% followed the advisors’ recommendation.

at the same time, before providing their recommendation. We consider preferences for blinding in this task as a stronger form of moral commitment than choosing to see quality first because advisors choose not to know their incentive at all prior to their recommendation decision.

**Explanations of Choice.** In the second wave of data collection of the Choice Experiment (Choice Free treatment) and among *Choice Free - Professionals*, we added an open ended question asking participants to explain how they made their decision about order of information. Two independent raters, blind to advisors' choices, coded the responses of 1,747 advisors (including  $N = 712$  professionals) and classified their responses into four categories, which apply to 91% of the responses. The remaining 9% consists of empty or unrelated comments according to both raters. The two independent raters (see Online Appendix B for the coding categories and procedures) agreed in over 82% of their classifications, leading to an interrater agreement  $\kappa$  of 0.76. We average their ratings to examine how advisors' explanations vary with their preference of information order.

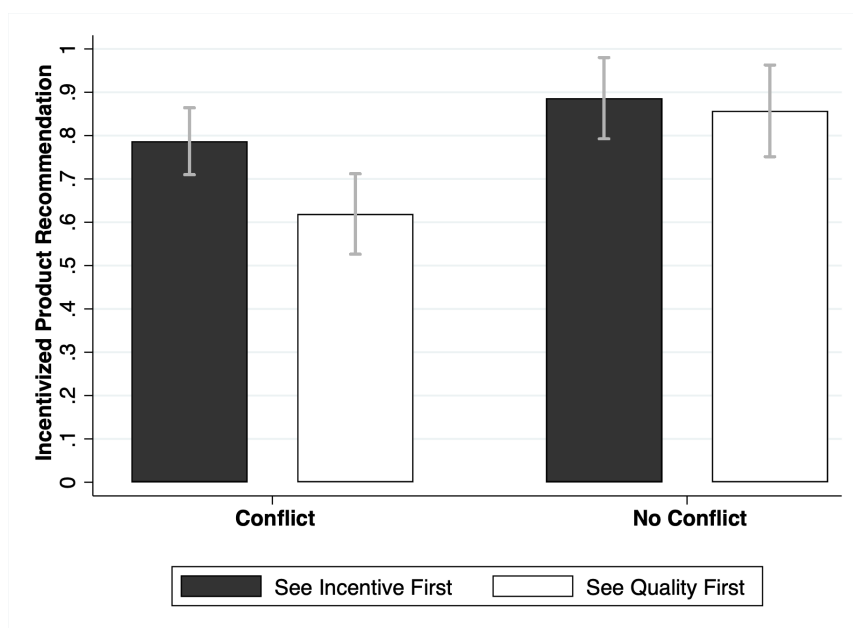
## 5 Does the Order of Information Affect Advice?

We first test whether exogenously assigning a given order of information affects advice in the NoChoice experiment.

When advisors face a conflict of interest (i.e., the quality signal is in conflict with their own incentive), the rate of self-serving recommendation depends on the order in which information is presented to them. Figure 3 shows that in the See Incentive First treatment, 79% of advisors recommend the incentivized product. In the Assess Quality First treatment, 62% of advisors recommend the incentivized product. This 17 percentage point difference is significant ( $Z$ -stat = 2.69,  $p = 0.007$ ,  $N = 213$ ). When advisors do not face a conflict of interest, the order of information does not affect recommendations. Advisors in the See Incentive First treatment recommend the incentivized product 89% of the time, while those in the Assess Quality First treatment recommend the incentivized product 86% of the time ( $Z$ -stat =  $-0.41$ ,  $p = 0.685$ ,  $N = 86$ ). These results are robust to controlling for demographics and advisor's selfishness (see Online Appendix C.1).

Throughout, advisors exhibit a preference for product A recommending it 16% of the time, even when the quality signal is a \$2 ball and the advisor is incentivized to recommend B. Nevertheless, the effect of information order on behavior is similar





Notes: This figure shows the fraction of recommendations of the incentivized product, when there is a conflict of interest between the advisor and the client, by treatment. In the See Incentives First treatment the advisor is presented first with information about her incentive. In Assess Quality First treatment she receives the signal about the quality of product B first.  $\pm$  Error bars show the 95% CI of the mean,  $N = 213$  for cases of conflict and  $N = 86$  for cases of no conflict.

Figure 3: Recommendation of Incentivized Product, by Treatment

regardless of what product is incentivized.<sup>19</sup> Consistent with Hypothesis 1, these results suggest that, when there is a conflict of interest, seeing the incentive first provides more cognitive flexibility, enabling advisors to recommend the incentivized product more often than when the signal of quality is assessed first.

**Result 1.** When there is a conflict of interest, advisors who are assigned to See Incentive First are significantly more likely to recommend the incentivized product than advisors who are assigned to See Quality First.

This experiment and its results set the stage for our main research questions: Which sequence of information do advisors prefer, and how does this choice affect their subsequent recommendations?

<sup>19</sup>In Online Appendix D, we report data from the additional wave of the study that tests effect of presenting both information about incentives and the quality signal simultaneously. The results show that, when both pieces of information are presented on the same screen, advisors behave similarly to the *See Incentive First* treatment, suggesting that in order for advice to be less influenced by incentives, advisors need to first process the quality signal without knowing their incentives.

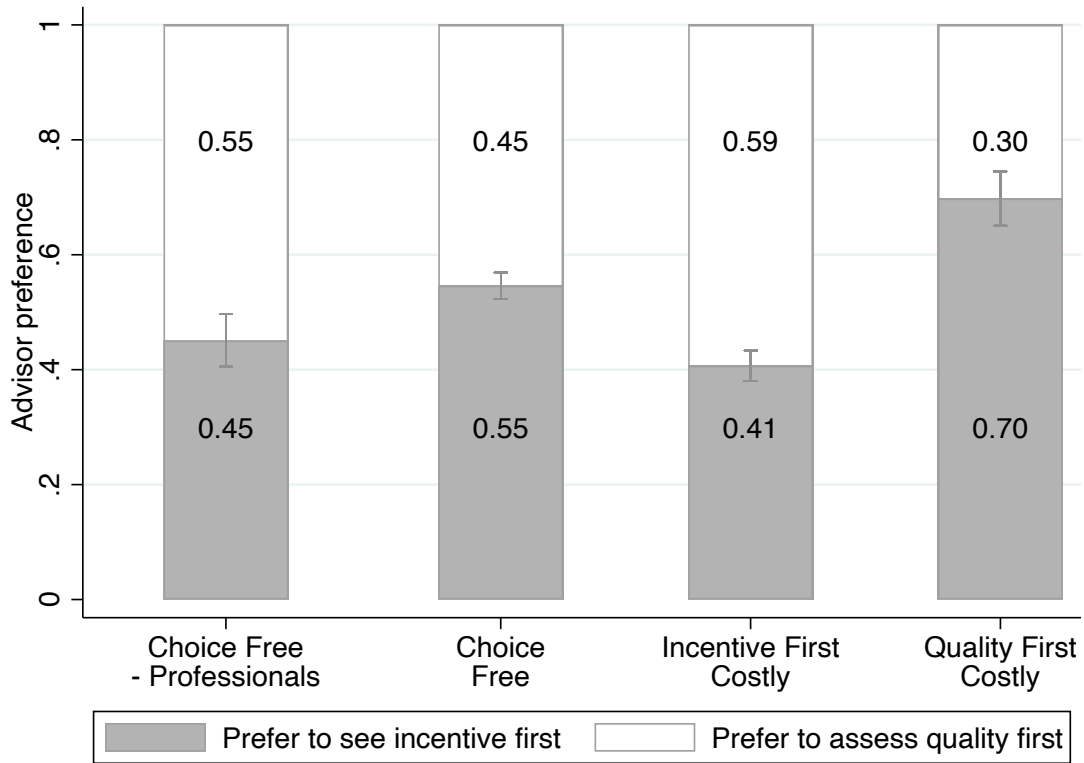
## 6 Preferences for Information Order: Cognitive Flexibility or Moral Commitment?

When choosing the information order is free, advisor preferences for information order are split between seeing the incentive first and seeing quality first, as shown in Figure 4. Since we conducted the experiment in several waves, the figure shows covariate-adjusted demand, controlling for wave, advisor gender and age (disaggregated results are shown in Online Appendix C.2).

Among professionals, 45% of advisors prefer to see the incentive information first, and among AMT participants, 55% of advisors exhibit the same preference. Conversely, between 55% and 45% of advisors choose to see the quality signal first, indicating that a substantial fraction of advisors would rather delay information about their own incentive.

When seeing the incentive first is costly, 41% of advisors are still willing to pay the cost (a third of their commission) to see the incentive first and have cognitive flexibility when assessing the signal. This suggests that the preference to see the incentive first, when it is free, is not driven only by indifference, as a substantial fraction of advisors shows a strict preference. Similarly, when see the quality signal is costly, 30% of advisors are willing to pay a cost to see the quality signal first, limiting cognitive flexibility. We interpret this choice as a form of moral commitment to accurate beliefs. Compared to when choice is free, when seeing the incentive first is costly, there is a 14-percentage-point drop in demand to see the incentive first ( $t - \text{stat} = -7.84$ ,  $p < 0.001$ ), as shown in Table 2. When seeing quality first is costly, there is a 15-percentage-point increase in demand to see incentive first ( $t - \text{stat} = 5.17$ ,  $p < 0.001$ ).

Table 2 shows the determinants of the preference to see the incentive first, and columns (2)-(3) investigate its relationship with advisor selfishness. In line with Hypothesis 2(a), advisors who make more selfish choices in the task designed to measure advisors' moral costs prefer to see the incentive first significantly more often.



*Notes:* This figure presents covariate-adjusted demand of advisors to see the incentive first or assess quality first estimated using OLS-regression. The covariates include the wave of data collection in the Choice experiment, whether the incentives were probabilistic or not, whether product *A* was presented on the left-hand side of the screen, the age and the gender of the advisor. Preferences by experimental wave are shown in Online Appendix C.2. Error bars indicate 95% confidence intervals.

Figure 4: Advisor Preference

## 6.1 Preferences for Information Order and Preferences for Blinding

To examine whether advisors' preference to assess quality first is indicative of a desire for moral commitment, we test whether the preference to see quality first predicts take up of a stronger form of moral commitment: choosing to blind oneself from incentives altogether. For this purpose, we focus on the subset of participants who took part in the blinding task. Advisors who prefer to assess quality first are significantly more likely to also prefer to blind themselves in the blinding task. As shown in Figure 5, 54.5% of advisors who choose to assess quality first also prefer to blind themselves. For advisors who prefer to see the incentive first, only 31.0%, choose to blind themselves, a significant difference compared to advisors who prefer

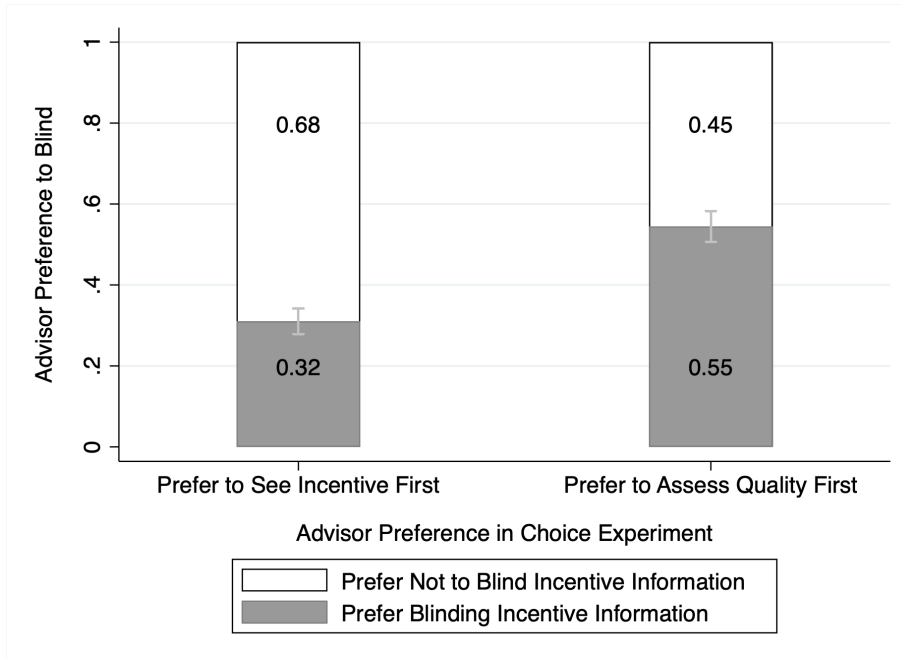
Table 2: Preference for Information Order

	(1)	(2)	(3)
	<b>Prefer to See Incentive First</b>		
See Incentive First Costly	-0.139*** (0.018)	-0.140*** (0.018)	-0.140*** (0.018)
Assess Quality First Costly	0.152*** (0.029)	0.152*** (0.029)	0.152*** (0.029)
Choice Free – Professionals	-0.095*** (0.026)		
Selfishness		0.028*** (0.007)	0.039*** (0.009)
See Incentive First Costly X Selfishness			-0.022 (0.016)
See Quality First Costly X Selfishness			-0.021 (0.018)
Female	-0.029** (0.013)	-0.024* (0.014)	-0.023* (0.014)
Age	-0.003*** (0.001)	-0.002*** (0.001)	-0.002*** (0.001)
Constant	0.674*** (0.024)	0.662*** (0.025)	0.661*** (0.025)
Observations	5908	5196	5196
$R^2$	0.034	0.040	0.040

*Notes:* This table displays the estimated coefficients from linear probability models on the preference to see the incentive first. See Incentive First Costly and Assess Quality First Costly are indicator variables that take value 1 in the respective treatment, 0 otherwise. Selfishness was elicited at the end of the experiment, using a multiple price list (MPL) with 5 decisions. The variable is a standardized measure of the number of times the advisor chose to recommend the incentivized product in the MPL task. The regression models in columns (2) and (3) include individual controls for the advisor's gender and age, each wave of the experiment, whether incentives were probabilistic, the position of the products on the screen and the interaction between these two variables. Robust standard errors (HC3) in parentheses. \*  $p < .10$ ; \*\*  $p < .05$ ; \*\*\*  $p < .01$

to assess quality first ( $Z$ -stat = 9.11,  $p < 0.001$ ,  $N = 1484$ ).

The difference in preference to blind between advisors who prefer to see incentive first and those who prefer to see quality first remains large (22 percentage points) and significant in regression analyses that control for treatment, gender, age, advisors facing a conflict of interest in the main experiment, and for being assigned to their preferred order in the main experiment ( $t$ -stat =  $-7.18$ ,  $p < 0.001$ ; see Online Appendix C.2). Altogether, these findings provide support for the interpretation that preferring to see the signal of quality first is a form of moral commitment, which correlates with the take up of a stronger form of commitment: blinding oneself from incentives altogether.



*Notes:* This figure presents the fraction of advisors who chose to blind themselves from the incentive information, in the blinding task, and those who chose not to blinding themselves, conditional on their preference for information order in Wave 3 of the Choice experiment ( $N = 1484$ ). Error bars indicate 95% confidence intervals.

Figure 5: Take up of Blinding by Preferences for Information Order

## 6.2 Explanations for Choice of Information Order

To gather further evidence on whether individuals choose to see quality first to commit to moral judgment, we make use of advisors' self-reported reasons for their choices between information orders collected for a sub-sample of the Choice Free experiment. The average classification of two independent raters reveals that advisors in the experiment rarely report that they are indifferent between seeing the incentive first or assessing quality first (on average, 10% of the comments), which suggests that indifference is not a main driver of choices. Further, advisors who choose to see the quality signal first are more likely to report doing so to limit bias in their evaluation, as compared to those preferring to see the incentive first (an average of 41% of AMT participants and 53% of professionals versus 5% of AMT participants and 7% of professionals, respectively,  $\chi^2$ -stat = 403.6,  $p < 0.001$ ). These findings are consistent with the interpretation that many advisors anticipate the effect of seeing quality first, and prefer to commit to accurate and therefore moral judgement. Conversely, advisors who choose to see the incentive first report to be interested in the commission (an average of 36% of the cases for both AMT and for professionals) or

to be motivated by other reasons (more details in Online Appendix C.2).

**Result 2(a).** 41% of advisors are willing to pay to see the incentive first, while 30% of advisors are willing to pay to see quality first. Their choices correlate with overall morality, with more selfish advisors being more likely to prefer to see the incentive first, and with preferences for blinding.

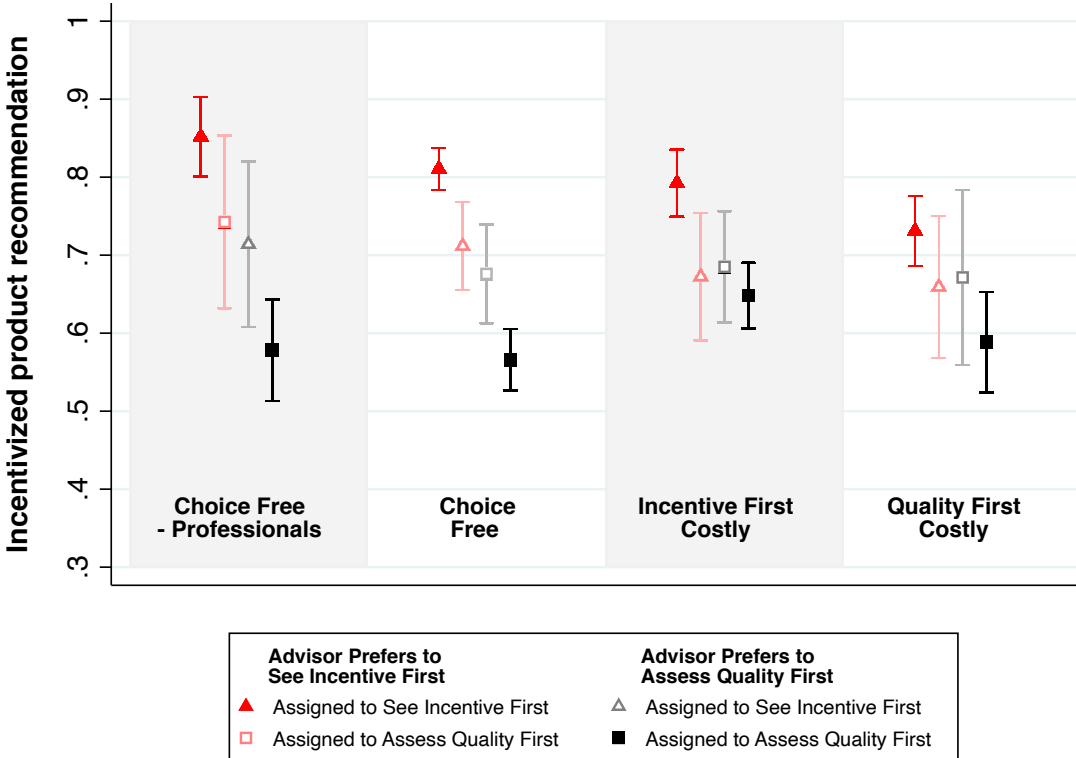
## 7 Does Experiencing Flexibility or Commitment Affect Advice?

Given the heterogeneity in preferences for information order, a central question is how choosing a particular information order affects recommendations. What is the effect of *experiencing* commitment or flexibility?

Figure 6 displays advisors' recommendation decisions conditional on their preference for and assignment to an information order, focusing on cases in which there is a conflict between the signal of quality about product B and the advisor's incentive (in Online Appendix C.2 we also provide the figure for cases in which there was no conflict). For advisors who are assigned their preference, recommendation decisions are significantly different depending on the information order. Across all treatments, advisors who prefer and are assigned to see the incentive first (left-most triangle in each cluster in Figure 6) recommend the incentivized product at highest rate. By contrast, those who prefer and are assigned to see quality first (right-most square in each cluster in Figure 6), recommend the incentivized product significantly less often in all cases ( $t$ -test, all  $p < 0.001$ ). These results are confirmed by the regression analysis reported in Table 3, where we report coefficient estimates of a linear probability model of the advisor's decision to recommend the incentivized product for advisors who are assigned their preferred order (column (1)) and those who are not (column (2)), and all together (column (3)). If advisors are assigned their preference, those who prefer to see the incentive first are 19.5 percentage points more likely to recommend the incentivized product than those who prefer to see quality first ( $t$ -stat = 12.17,  $p < 0.001$ ). There is no difference for advisors who do not receive their preferred order. These results reveal that differences in recommendation are not only due to sorting and that *experiencing* information in the desired order is central to the ability to provide self-serving recommendations or constrain them.

In the absence of conflict, advisors are significantly more likely to recommend

the incentivized product, and the difference between advisors who prefer to see the incentive first and those who prefer to see quality first is significantly smaller. Overall, advisors exhibit a preference for recommending product A, despite the absence of a conflict of interest. Despite this preference, the difference in recommendations between advisors who prefer to see the incentive first and those who prefer to see quality first remains qualitatively similar focusing on cases in which the incentive is to recommend product A or product B, as shown in Online Appendix C.2.



*Notes:* This figure presents the covariate-adjusted recommendations of the incentivized product when there is a conflict between the signal of quality and the advisor’s incentive, with the same covariates as in Figure 4. Error bars indicate 95% confidence intervals.

Figure 6: Advisor Recommendations

To examine whether actively choosing an information order that provides more cognitive flexibility could reduce the scope for rationalizing self-serving behavior, we conduct two sets of analyses. First, we investigate whether advisors who prefer to see the incentive first are more likely to recommend the incentivized product when they are assigned to see information in their desired order. On average, advisors who choose to see the incentive first are 9.8 percentage points more likely to recommend

Table 3: Advisor Recommendations

	(1)	(2)	(3)
	<b>Recommend incentivized product</b>		
<i>Assignment:</i>	Assigned Pref.	Not Assigned Pref.	Both
Prefer to See Incentive First	0.195*** (0.016)	0.003 (0.029)	0.181*** (0.015)
Not Assigned Preference			0.060*** (0.021)
Prefer to See Incentive First X Not Assigned Preference			-0.140*** (0.026)
No Conflict	0.256*** (0.020)	0.202*** (0.033)	0.236*** (0.018)
No Conflict X Prefer to See Incentive First	-0.137*** (0.025)	0.012 (0.045)	-0.098*** (0.022)
No Conflict X Not Assigned Preference			0.019 (0.025)
Choice Free–Professionals	-0.026 (0.025)	0.051 (0.044)	-0.006 (0.022)
See Incentive First Costly	0.035** (0.017)	0.020 (0.031)	0.031** (0.015)
Assess Quality First Costly	0.004 (0.030)	0.093* (0.052)	0.027 (0.026)
Incentive for B	-0.171*** (0.013)	-0.187*** (0.023)	-0.175*** (0.011)
Female	0.005 (0.013)	-0.015 (0.023)	-0.001 (0.011)
Age	-0.002*** (0.001)	-0.003*** (0.001)	-0.002*** (0.000)
Constant	0.737*** (0.027)	0.864*** (0.048)	0.755*** (0.025)
Observations	4448	1460	5908
$R^2$	0.106	0.083	0.097

Note: This table displays the estimated coefficients from linear probability models on the advisor’s decision to recommend the incentivized option. Column (1) focuses on individuals who are assigned their preference, while column (2) focuses on individuals who are not assigned their preference. Both groups are merged in column (3). Prefer to See Incentive First is an indicator of the advisor’s preference, and Not Assigned Preference is an indicator for not receiving the preferred order. No Conflict is an indicator for the cases in which the signal of quality is not in conflict with the advisor’s commission. See Incentive First Costly and Assess Quality First Costly are indicator variables that take value 1 in the respective treatment, 0 otherwise. All regression models include individual controls for the advisor’s gender and age, each wave of the experiment, whether incentives were probabilistic, the position of the products on the screen and the interaction between these two variables. The same analysis including a measure of advisor’s selfishness are shown in Online Appendix C. Robust standard errors (HC3) in parentheses. \*  $p < .10$ ; \*\*  $p < .05$ ; \*\*\*  $p < .01$

the incentivized product if they are assigned their preferred order ( $t - \text{stat} = 3.66$ ,  $p < 0.001$ ). This evidence indicates that, even if individuals’ actively choose to have more cognitive flexibility, they still benefit from experiencing it.

Second, we compare the size of the gap in recommendations between advisors who choose flexibility or commitment and are assigned their preference to the gap in recommendations observed in the NoChoice experiment, where individuals are randomized to a given information order. To compare the two experiments, we focus



on the Choice Free treatment conducted on AMT, since it has the same incentives and sample of the NoChoice experiment. In the Choice experiment, we estimate a 23.5 percentage point gap, which is not significantly different from the gap estimated in the NoChoice experiment ( $t - \text{stat} = 1.26$ ,  $p = 0.207$ ), but directionally larger by about 7 percentage points.<sup>20</sup>

These two sets of analyses show that actively pursuing cognitive flexibility, by choosing to see the incentive first, does not fully remove the advisors' ability to leverage that information order to their advantage to make self-serving recommendations, though it may directionally limit it.

We also examine whether pursuing commitment, by choosing to see quality first, is an effective strategy for preventing self-serving behavior. Our results reveal that it is: Conditional on preferring to see quality first, those actually assigned to assess quality first are less likely to make the incentivized recommendation. Relative to advisors who are assigned to receive cognitive flexibility, those who are assigned moral commitment are 9 percentage points less likely to recommend the incentivized product ( $t - \text{stat} = 3.05$ ,  $p = 0.002$ ). This result suggests that limiting self-serving behavior requires temporarily blinding these individuals from receiving information on their incentive.

**Result 2(b).** Advisors who choose and are assigned to see the incentive first are significantly more likely to recommend the incentivized product than advisors who choose and are assigned to see quality first. When advisors are not assigned their preferred information order there is no significant difference in recommendations.

## 7.1 Evidence of Belief Distortion

To examine whether advisors exhibit biases in belief updating after pursuing and getting flexibility or commitment, we study how individuals update their beliefs from the prior of 0.50 after seeing the signal of quality. For this analysis we merge the beliefs of all advisors in the Choice experiment and follow the approach of Möbius et al. (2022) to examine belief updating relative to Bayes' Rule. For this purpose, we use the continuous belief measure (0-100) that we elicit after our incentivized

---

<sup>20</sup>We thank a reviewer for suggesting this comparison. We note that the NoChoice experiment has a smaller sample than the Choice experiment and, as a result, has wider confidence intervals (6-28 percentage points), which overlap with the more precise estimate obtained in the Choice experiment (19-28 percentage points). We provide a detailed comparison of recommendation behavior across these two experiments in Online Appendix section C.3

belief measure (which is in bins). In Online Appendix C.2 we report the analyses that leverages the incentivized belief measure showing qualitatively similar results.

We test whether belief updating about the signal of quality among advisors who prefer and are assigned to see the incentive first differs from that of those who prefer and are assigned to see the quality signal first.<sup>21</sup> In the experiment, the advisor could get a signal that was in conflict with her incentive ( $\sigma = c$ ) or one that was aligned with her incentive ( $\sigma = nc$ ). We denote the advisors' posterior belief about the likelihood of product B being low with  $\hat{\mu}$ . Möbius et al. (2022) show that the relationship between the advisor's logit belief about quality and the Bayesian benchmark can be estimated using a linear model that includes the loglikelihood ratio of each possible signal. We denote  $\gamma_C$  as the log likelihood ratio of a signal in conflict with the incentive and  $\gamma_{NC}$  the log likelihood ratio of a signal not in conflict with the incentive.<sup>22</sup> Conditional on the advisor's preference and assignment, we estimate the following model of belief updating:

$$\text{logit}(\hat{\mu}) = \beta_C \cdot I\{\sigma = c\} \cdot \gamma_C + \beta_{NC} \cdot I\{\sigma = nc\} \cdot \gamma_{NC} + \epsilon_i$$

where the parameters  $\beta_C$  and  $\beta_{NC}$  indicate the responsiveness of the advisor's beliefs to a signal in conflict with the incentive or not in conflict with the incentive, respectively, relative to the Bayesian benchmark. If individuals are Bayesian,  $\beta_C = \beta_{NC} = 1$ .

In Panel A of Table 4, we report estimates of the aforementioned parameters. Column (1) focuses on advisors who are assigned their preference, while column (2) focuses on those who are not assigned their preference. Columns (3) and (4) conduct the same analysis restricting the sample to exclude advisors who update in the wrong direction, from the prior of 0.5, given the signal.

Panel A of Table 4 shows that, similar to Möbius et al. (2022), beliefs exhibit conservatism, as all coefficients are significantly smaller than 1. In our aggregate sample we find evidence for a directional bias in updating: column (1) shows that advisors are more responsive to signals that are not in conflict with the incentive ( $\beta_{NC} = 0.380$ ) than to signals in conflict with the incentive ( $\beta_C = 0.307$ ,  $F$  – stat = 5.57,  $p = 0.018$ ). The estimated parameters are similar for the case in

---

<sup>21</sup>Beliefs about quality are one of the potential beliefs that individuals distort; others include beliefs about ethicality, which we did not measure in the experiment.

<sup>22</sup>In our experiment, when the signal was a \$2 ball, we have  $\gamma = -\log(2)$ , when the signal is \$0, we have  $\gamma = \log(3)$ . Whether these likelihood ratios are considered conflict or no conflict depends on whether the commission was for product A or B.

Table 4: Belief Updating

	(1)	(2)	(3)	(4)
	<b>Log-odds Belief</b>			
<i>Assignment:</i>	Assigned Pref.	Not Assigned Pref.	Assigned Pref.	Not Assigned Pref.
<i>Data:</i>	All	All	Excl. update in wrong direction	Excl. update in wrong direction
<b>Panel A: Pooled</b>				
$\beta_C$	0.305*** (0.016)	0.312*** (0.028)	0.549*** (0.014)	0.575*** (0.024)
$\beta_{NC}$	0.380*** (0.027)	0.378*** (0.046)	0.644*** (0.023)	0.646*** (0.038)
<b>Panel B: By Choice of Information Order</b>				
$\beta_C^{f=i}$	0.267*** (0.022)	0.299*** (0.038)	0.525*** (0.019)	0.567*** (0.033)
$\beta_C^{f=q}$	0.346*** (0.023)	0.327*** (0.040)	0.574*** (0.020)	0.583*** (0.035)
$\beta_{NC}^{f=i}$	0.324*** (0.038)	0.405*** (0.067)	0.626*** (0.033)	0.677*** (0.055)
$\beta_{NC}^{f=q}$	0.444*** (0.039)	0.347*** (0.063)	0.664*** (0.033)	0.609*** (0.054)
Observations	4385	1447	3674	1193
$\beta_C^{f=q} = \beta_C^{f=i}$	0.014	0.613	0.078	0.743
$\beta_{NC}^{f=q} = \beta_{NC}^{f=i}$	0.029	0.533	0.417	0.374

*Notes:* The outcome in all regressions is the log belief ratio.  $\beta_C^f$  and  $\beta_{NC}^f$  are the estimated effects of the log likelihood ratio for conflict and no conflict signals, respectively, for advisors who prefer order  $f$ . Order  $f = i$  indicates a preference to see the incentive first, and  $f = q$  indicates a preference to see quality first. Columns(1) and (2) include all advisors. Columns (3) and (4) exclude advisors who updated in the wrong direction. Columns (1) and (3) include only advisors who were assigned their preference, while columns (2) and (4) include only advisors who were not assigned their preference. Robust standard errors (HC3) in parentheses. \*  $p < .10$ ; \*\*  $p < .05$ ; \*\*\*  $p < .01$ .

which advisors are not assigned to their preferences, though the estimates are less precise and therefore the difference is not statistically significant. Although there is higher responsiveness to signals when advisors who update in the wrong direction are excluded (columns (3) and (4)), the gap between signals in conflict and not in conflict with the incentives persists ( $F - \text{stat} = 12.06$ ,  $p < 0.001$ ). This finding is consistent with advisors engaging in suppression, which is part of our theoretical framework and consistent with prior work on motivated attention (e.g., Eil and Rao, 2011; Möbius et al., 2022).

To study whether individuals who pursue and get to receive information about their incentive first exhibit more distorted beliefs, both in the form of conservatism and directional bias, we estimate the model separately for advisors who prefer order  $f \in \{i, q\}$ . The results are reported in Panel B of Table 4. We find evidence that the order of information affects belief distortion. Column (1) of Panel B shows that, for advisors who receive information in their desired order, seeing the incentive first

(as opposed to quality first) leads to a lower responsiveness to signals in conflict with the incentive ( $\beta_C^{f=i} = 0.267$  versus  $\beta_C^{f=q} = 0.346$ ,  $t - \text{stat} = 2.45$ ,  $p = 0.014$ ), as well as signals that are not in conflict with the incentive ( $\beta_{NC}^{f=i} = 0.324$  while  $\beta_{NC}^{f=q} = 0.444$ ,  $t - \text{stat} = 2.19$ ,  $p = 0.029$ ). When we exclude advisors who update in the wrong direction, we find that seeing the incentive first leads to a directionally larger and marginally significant decrease in attention to signals in conflict with the incentive ( $t - \text{stat} = 1.76$ ,  $p = 0.078$ ), and a smaller directional decrease in response to signals that are not in conflict with the incentive ( $t - \text{stat} = 0.81$ ,  $p = 0.417$ ). These results provide suggestive, though weak, evidence that advisors exploit the lower attention to quality signals when they see the incentive first to engage in more suppression. Notably, no differences in updating appear when advisors are not assigned to receive information in their desired order, as displayed in columns (2) and (4). Overall, the findings are broadly in line with the theoretical framework, as they show that advisors pay less attention to signals when the incentive is seen first, particularly when these conflict with the incentive.<sup>23</sup>

## 8 Additional Tests of Sophistication

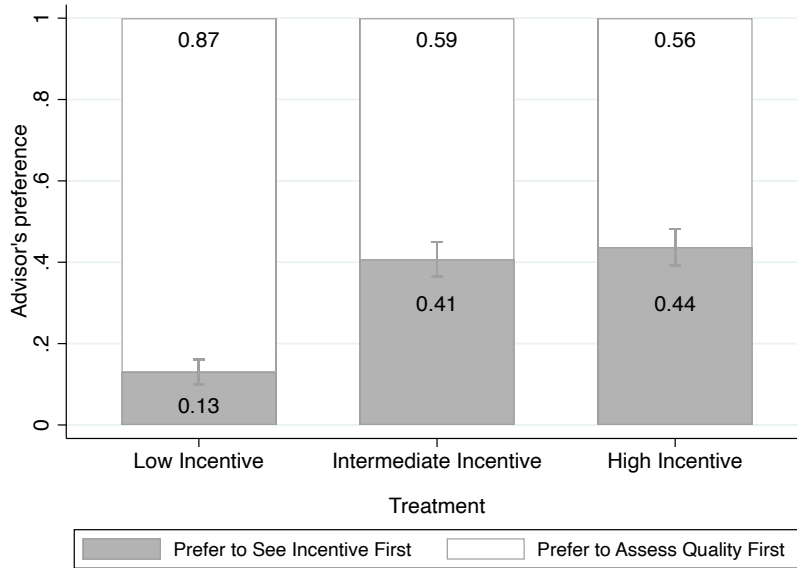
### 8.1 Advisors' Preferences and Incentives

In the Choice Stakes Experiment, we test whether advisors' demand to see the incentive first responds to the financial gain from recommending the incentivized product. If the gains from recommending the incentivized product decrease, advisors have a smaller incentive to distort their beliefs, making the demand for cognitive flexibility (seeing the incentive first) less desirable. Figure 7 shows the advisors' preference to see the incentive first. In the Intermediate incentive treatment, 41%

---

<sup>23</sup>In Online Appendix C we separate the analysis by signal and provide more detailed results on directional bias in updating. The evidence suggests there is more suppression when the signal is \$0 and the advisor sees the incentive first, but not as much when it is \$2. For signals of \$0, when there is a conflict of interest, advisors who see the incentive first hold beliefs closer to the prior,  $\beta_C^{f=i} = 0.419$  compared to  $\beta_C^{f=q} = 0.479$  ( $p = 0.079$ ). When there is no conflict of interest, information order does not affect updating,  $\beta_{NC}^{f=i} = 0.552$  compared to  $\beta_{NC}^{f=q} = 0.555$  ( $p = 0.964$ ). For signals of \$2, beliefs are not significantly closer to the prior, both if there is a conflict and no conflict of interest ( $p > 0.1$ ). The difference in updating patterns between the two signals could arise from the differences between product *A* and *B*. Recommending *B* requires advisors to dismiss "bad news" about the quality of product *B*. Recommending *A* following a \$2 signal, by contrast, can be done using other justifications, such as the fact that the quality of *A* was certain. This potential explanation is in line with our findings of substantially stronger preferences for *A* in our experiment even when advisors did not face a conflict of interest.

of advisors prefer to see the incentive first, replicating our finding in the Incentives First Costly treatment of Choice Experiment. This fraction decreases significantly in the Low Incentive treatment, to 13% ( $Z$  – stat = 9.79,  $p < 0.001$ ). In the High Incentive treatment, the advisors’ preference to see the incentive first increases by only 3 percentage points, to 44% ( $Z$  – stat = 0.96,  $p = 0.337$ ), despite the fact that the commission is doubled. These results are confirmed in regression analyses in Online Appendix C.6.



Notes: This figure shows the fraction of advisors who prefer to see their incentive first. In the Low Incentive treatment the commission for learning before is \$0.01, in the Intermediate Incentive treatment it is \$0.15, and in the High Incentive treatment it is \$0.30. Seeing the incentive first costs \$0.05 in all treatments, as in the Incentive First Costly treatment of the Choice experiment. Error bars show the 95% confidence interval.

Figure 7: Advisor’s Preference to See Incentive First, by Treatment

We conduct exploratory analyses on advisors’ recommendations of the incentivized product in each treatment, shown in Online Appendix C.6. We pool all treatments together, and test whether, when assigned to their preferred information order, advisors who prefer to see the incentive first are more likely to recommend the incentivized product. When advisors are assigned their preferred order, they are 14 percentage points more likely to recommend the incentivized product. However, when not assigned to their preferred order, advisors who expressed a preference to see the incentive first are no more likely to recommend the incentivized product than advisors who indicated the opposite preference.

Most models of motivated cognition assume that belief distortion is driven by incentives (e.g., Brunnermeier and Parker, 2005; Bénabou and Tirole, 2011), yet some evidence suggests that sometimes belief distortion is insensitive to stakes (e.g., Coutts, 2019; Engelmann et al., 2019). This experiment shows that advisors’ preferences to see the incentive first respond to incentives to recommend the incentivized product, in line with our theoretical framework and other models of motivated beliefs (e.g., Brunnermeier and Parker, 2005; Bénabou and Tirole, 2011). When doubling the commission of the advisor, however, the preference to see the incentive first increases by only 3 percentage points, less than 10 percent. Our experiment thus suggests that demand for cognitive flexibility increases concavely with the incentive to recommend the incentivized product. This evidence can be useful for further theoretical and empirical work on self-deception to better understand the role of incentives in belief distortion.

## 8.2 Do Third Parties Anticipate the Effect of Information Sequence on Advisor’s Behavior?

To better understand the motives driving advisors’ preferences for information order, we investigate whether third parties anticipate the effect of information order. In the Information Architect Experiment we focus on choices of information order by information architects (IAs) who have incentives that are either aligned with those of the advisors (*IA-Advisor*) or with those of the client (*IA-Client*). Our findings show that, in the *IA-Advisor* treatment, the fraction of IAs who choose for the advisor to see their incentive first is significantly larger than in the *IA-Client* treatment, where advisors’ incentives are aligned with the client (58% vs 44%,  $N = 498$ ,  $Z - \text{stat} = 3.23$ ,  $p = 0.001$ ), and this difference is robust to controlling for demographics (see Online Appendix C.7). These findings are suggestive that third parties anticipate the effect of information order on behavior. We further find that the fraction of IAs who chose for the advisor to see the incentive first in the *IA-Advisor* treatment is similar to the average fraction of advisors who prefer to see the incentive first in the *Choice Free* treatment of the Choice experiment (56%) ( $Z - \text{stat} = 0.497$ ,  $p = 0.62$ ). Since IAs did not receive any information about the realized incentive and quality signal, this result suggests that choices to see the incentive first in the Choice Free treatment are not entirely explained by individuals choosing to see the incentive first to satisfy curiosity.

## 9 Conclusion

A large body of research has shown that self-serving behavior becomes more likely when individuals can distort their beliefs but that there are cognitive constraints to such ability to distort beliefs. In this paper we ask whether individuals actively take action to constrain their ability to distort beliefs, a form of commitment to moral behavior, or rather seek out the cognitive flexibility needed to distort beliefs, and investigate how, conditional on preferences, being assigned to *experiencing* commitment or flexibility affect self-serving behavior.

We find that a sizable fraction of advisors (30-45%) is willing to take up an opportunity to constrain belief distortion by seeing quality information first, even when this choice is costly. These preferences are correlated with the take-up of stronger forms of moral commitment and with advisors' morals, measured by their choices when a conflict of interest is always present. An interesting avenue for future research would be to investigate whether the take up of moral commitment is correlated with the take up of commitment outside the moral context, in domains such as saving (e.g., Ashraf, Karlan, and Yin, 2006), health (e.g., Giné, Karlan, and Zinman, 2010) or food choice (e.g., Sadoff, Samek and Sprenger, 2020).

Alongside the preference for moral commitment expressed by some advisors, we find that a considerable share of advisors (40-55%) actively seek out cognitive flexibility by asking to see their incentive before making quality assessments, even when doing so is costly. Actively seeking such cognitive flexibility does not entirely preclude individuals from being able to distort their beliefs, indicating that individuals can intend to distort beliefs for self-serving reasons and still be successful at doing so. Altogether, our findings suggest that at least a portion of individuals can anticipate some cognitive constraints to belief distortion, suggesting some level of sophistication about their ability to distort their beliefs when potentially inconvenient information cannot be avoided.

Experts across professions are often called to make partially subjective judgments and variety of incentives could influence their judgment. Such incentives can vary in size (see, e.g, Campbell et al., 2007) and can also assume less tangible forms (e.g, hiring a candidate for reasons other than their qualifications, using information other than merit, such as the authors' names, to evaluate the quality of a research proposal). In our experiments, we mimic such conflict of interests using small monetary incentives. We find that such small incentives can bias judgement and recommendations, leading some advisors to seek out commitment. Whether the

effects documented in this paper apply to settings where experts face substantially higher or less tangible incentives than the ones we used in our experiments is an empirical question that could be investigated in future work.

Altogether, our research suggests that how information provision is structured plays an important role in determining the extent of bias in evaluations, and that a proportion of individuals is willing to temporarily blind themselves from potentially biasing information to ensure the fair and moral behavior. Existing work that focuses on hiring managers and academic reviewers provides suggestive evidence in line with our findings. For instance, a vast majority of reviewers support double-blind peer review (Yankauer, 1991; Regehr and Bordage, 2006), but demand for double-blind review is quite limited among authors, especially those who work at more prestigious institutions (McGillivray and De Rainieri, 2018). In the domain of hiring, although some studies report very high take up of blinding in mock up hiring tasks (e.g., 91.3% in Fath, Larrick, and Soll, 2022), such policies are rare in organizational settings (Bortz, 2018). This evidence could reflect the heterogeneity of preferences we document in our experiment.

The information structure an organization ultimately implements is important, as experiencing commitment or flexibility can alter the extent of self-serving behavior in organizations. As our Information Architect experiment shows, third parties can anticipate the effects of different information orders. Therefore, the findings in this paper can have important implications for the design of expert systems, suggesting that both organizational design and the selection of experts into organization may occur with commitment or flexibility goals in mind.

## Reference List

- Abeler, J., Nosenzo, D., and Raymond, C. (2019). Preferences for truth-telling. *Econometrica*, 87(4), 1115–1153.
- Amasino, D., Pace, D., and van der Weele, J. (2021). Fair shares and selective attention. Tinbergen Institute Discussion Paper 2021-066.
- Anderson, N. H. (1965). Primacy effects in personality impression formation using a generalized order effect paradigm. *Journal of personality and social psychology*, 2(1), 1.
- Asch, S. E. (1946). Forming impressions of personality. *Journal of Abnormal and Social Psychology*, 41, 258–290.



- Ashraf, N., Karlan, D. and Yin, W., (2006). Tying Odysseus to the mast: Evidence from a commitment savings product in the Philippines. *The Quarterly Journal of Economics*, 121(2), 635–672.
- Babcock, L., Loewenstein, G., Issacharoff, S., and Camerer, C. (1995). Biased judgments of fairness in bargaining. *The American Economic Review*, 85(5), 1337–1343.
- Barfort, S., Harmon, N.A., Hjorth, F., and Olsen, A.L. (2019). Sustaining Honesty in Public Service: The Role of Selection. *American Economic Journal: Economic Policy*, 11(4), 96–123.
- Bénabou, R. (2013). Groupthink: Collective delusions in organizations and markets. *Review of Economic Studies*, 80(2), 429–462.
- Bénabou, R., and Tirole, J. (2002). Self-confidence and personal motivation. *Quarterly Journal of Economics*, 117(3), 871–915.
- Bénabou, R., and Tirole, J. (2011). Identity, morals and taboos: Beliefs as assets. *Quarterly Journal of Economics*, 126(2), 805–855.
- Bénabou, R., and Tirole, J. (2016). Mindful Economics: The production, consumption and value of beliefs. *Journal of Economic Perspectives*, 30(3), 141–164.
- Bénabou, R., Falk, A., and Tirole, J. (2018). Narratives, Imperatives and Moral Reasoning. NBER Working Paper #24798.
- Benjamin, D.J. (2019). Errors in probabilistic reasoning and judgment biases (Chapter 2). *Handbook of Behavioral Economics: Applications and Foundations* Vol. 2, 69–186.
- Bermúdez, J.L. (2000). Self-deception, intentions, and contradictory beliefs. *Analysis*, 60(4), 309–319.
- Bodner, R., and Prelec, D. (2003). Self-signaling and diagnostic utility in everyday decision making. *The Psychology of Economic Decisions*, 1(105), 26.
- Bordalo, P., Gennaioli, N., and Shleifer, A. (2012). Salience theory of choice under risk. *Quarterly Journal of Economics*, 127(3), 1243–1285.
- Bortz, D. (2018). Can Blind Hiring Improve Workplace Diversity? *HR Magazine*, March 20, 2018. Society for Human Resource Management.
- Brunnermeier, M.K., and Parker, J.A. (2005). Optimal expectations. *American Economic Review*, 95(4), 1092–1118.

- Campbell, E. G., Gruen, R. L., Mountford, J., Miller, L. G., Cleary, P. D., and Blumenthal, D. (2007). A national survey of physician – industry relationships. *New England Journal of Medicine*, 356(17), 1742–1750.
- Carlson, R.W., Marechal, M., Oud, B., Fehr, E., and Crockett, M. (2020). Motivated misremembering of selfish decisions. *Nature Communications*, 11(1), 1–11.
- Cohn, A., Marechal, M.A., Tannenbaum, D., and Zund, C.L. (2019). Civic honesty around the globe. *Science*, 365(6448), 70–73.
- Coutts, A. (2019). Testing models of belief bias: An experiment. *Games and Economic Behavior*, 113, 549–565.
- Chen, Z. and Gesche, T. (2017). Persistent bias in advice-giving. Mimeo.
- Crawford, V., and Sobel, J. (1982). Strategic information transmission. *Econometrica*, 50(6), 1431–1451.
- Dana, J., Weber, R. A., and Kuang, J.X. (2007). Exploiting moral wriggle room: Experiments demonstrating an illusory preference for fairness. *Economic Theory*, 33(1), 67–80.
- Darby, M.R., and Karni, E. (1973). Free competition and the optimal amount of fraud. *The Journal of Law and Economics*, 16(1), 67–88.
- DellaVigna, S., and Pope, D. (2018). Predicting experimental results: Who knows what? *Journal of Political Economy*, 126(6), 2410–2456.
- DeJong, C., Aguilar, T., Tseng, C. W., Lin, G. A., Boscardin, W. J., and Dudley, R. A. (2016). Pharmaceutical industry-sponsored meals and physician prescribing patterns for Medicare beneficiaries. *JAMA internal medicine*, 176(8), 1114–1122.
- Eil, D. and Rao, J.M. (2011). The good news-bad news effect: Asymmetric processing of objective information about yourself. *American Economic Journal: Microeconomics*, 3(2), 114–138.
- Engelmann, J., Lebreton, M., Schwardmann, P., van der Weele, J. J., and Chang, L. A. (2019). Anticipatory anxiety and wishful thinking. Mimeo.
- Enke, B., Gneezy, U., Hall, B., Martin, D., Nelidov, V., Offerman, T., and van de Ven, J. (2021). Cognitive Biases: Mistakes or Missing Stakes? *The Review of Economics and Statistics*, 1–45.
- Epley, N., and Gilovich, T. (2016). The mechanics of motivated reasoning. *Journal of Economic Perspectives*, 30(3), 133–40.

- Epley, N., and Tannenbaum, D. (2017). Treating ethics as a design problem. *Behavioral Science and Policy*, 3(2), 72–84.
- Exley, C.L. (2015). Excusing selfishness in charitable giving: The role of risk. *The Review of Economic Studies*, 83(2), 587–628.
- Falk, A., Neuber, T., and Szech, N. (2020). Diffusion of being pivotal and immoral outcomes. *The Review of Economic Studies*, 87 (5), 2205–29.
- Fath, S., Larrick, R.P., and Soll, J.B. (2022). Blinding curiosity: Exploring preferences for “blinding” one’s own judgment. *Organizational Behavior and Human Decision Processes* 170, 104135.
- Gabaix, X., Laibson, D., Moloche, G., and Weinberg S. (2006). Costly information acquisition: Experimental analysis of a boundedly rational model. *American Economic Review*, 96(4), 1043–1068.
- Giné, X., Karlan, D. and Zinman, J., (2010). Put your money where your butt is: a commitment contract for smoking cessation. *American Economic Journal: Applied Economics*, 2(4), 213–35.
- Gino, F., Norton M., and 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., and van Veldhuizen R. (2018). Bribery: Behavioral drivers of distorted decisions. *Journal of the European Economic Association*, 17(3), 917–946
- Gneezy, U., Saccardo S., Serra-Garcia, M., and van Veldhuizen R. (2020). Bribing the self. *Games and Economic Behavior*, 120, 917–946.
- Goldin, C. and Rouse, C. (2000). Orchestrating impartiality: The impact of “blind” auditions on female musicians. *American Economic Review*, 90(4), 715–741.
- Golman, R., Hagmann, D., and Loewenstein, G. (2017). Information avoidance. *Journal of Economic Literature*, 55(1), 96–135.
- Golman, R., Molnar, A., Loewenstein, G., and Saccardo, S. (2021). The demand of, and avoidance of information. *Management Science*. Published Online: 15 Dec 2021.
- Grossman, Z. (2014). Strategic ignorance and the robustness of social preferences. *Management Science*, 60(11), 2659–2665.

- Grossman, Z., and van der Weele, J.J. (2017). Self-image and willful ignorance in social decisions. *Journal of the European Economic Association*, 15(1), 173–217.
- Haisley, E.C., and Weber, R.A. (2010). Self-serving interpretations of ambiguity in other-regarding behavior. *Games and Economic Behavior*, 68(2), 614–625.
- Hanna, R., and Wang, S. (2017). Dishonesty and Selection into Public Service: Evidence from India. *American Economic Journal: Economic Policy* 9 (3), 262–90.
- Hogarth, R.M. and Einhorn, H.J., (1992). Order effects in belief updating: The belief-adjustment model. *Cognitive psychology*, 24(1), pp.1-55.
- Huber, C., Huber, J. (2020). Bad bankers no more? Truth-telling and (dis) honesty in the finance industry. *Journal of Economic Behavior Organization*, 180, 472-493.
- Huffman, D., Raymond, C., and Shvets, J. (2020). Persistent overconfidence and biased memory: Evidence from managers. Working paper.
- Kahan, D. (2013). Ideology, motivated reasoning, and cognitive reflection: An experimental study. *Judgment and Decision Making*, 8, 407–424.
- Kahneman, D. (1973). *Attention and Effort*. Prentice-Hall, USA.
- Karlsson, N., Loewenstein, G., and Seppi, D. (2009). The ostrich effect: Selective attention to information. *Journal of Risk and Uncertainty* 38, 95–115.
- 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.
- Lang, P. J., Bradley, M. M., and Cuthbert, B. N. (1997). Motivated attention: Affect, activation, and action. *Attention and orienting: Sensory and motivational processes*, 97, 135.
- Litman, L., Robinson, J., and Abberbock, T. (2016). TurkPrime.com: A versatile crowdsourcing data acquisition platform for the behavioral sciences. *Behavior Research Methods*, 1–10.
- Malmendier, U., and Tate, G. (2005). CEO overconfidence and corporate investment. *Journal of Finance*, 60(6), 2661–2700.
- Malmendier, U., and Schmidt, K. (2017). You owe me. *American Economic Review*, 107(2), 493–526.

- Marechal, M.A., and Thoni, C. (2019). Hidden Persuaders: Do Small Gifts Lubricate Business Negotiations? *Management Science* 65 (8), 3470–3469.
- McGillivray, B., and De Ranieri, E. (2018). Uptake and outcome of manuscripts in Nature journals by review model and author characteristics. *Research Integrity and Peer Review* 3 (5).
- Mele, A. (1987). *Irrationality: An Essay on Akrasia, Self-Deception, Self-Control*. Oxford: Oxford University Press.
- Mele, A. (2001). *Self-Deception Unmasked*. Princeton: Princeton University Press.
- Mijovic-Prelec, D., and Prelec, D. (2010). Self-deception as self-signaling: A Model and experimental evidence. *Philosophical Transactions of the Royal Society B*, 365,227–240.
- Möbius, M., Niederle, M., Niehaus, P. and Rosenblat, T. (2022). Managing self-confidence: Theory and experimental evidence. *Management Science*, Published Online: 24 Mar 2022.
- Palan, S., and Schitter, C. (2018). Prolific.ac – A subject pool for online experiments. *Journal of Behavioral and Experimental Finance*, 17, 22–27.
- Paolacci, G., Chandler, J., and Ipeirotis, P.G. (2010). Running experiments on Amazon Mechanical Turk. *Judgment and Decision Making* 5(5), 411–419.
- Regehr, G., and Bordage, G. (2006). To blind or not to blind? What authors and reviewers prefer. *Medical education* 40 (9), 832–839.
- Robertson, C.T., and Kesselheim, A.S. (2016). *Blinding as a Solution to Bias: Strengthening Biomedical Science, Forensic Science and Law*. Elsevier, USA.
- Sadoff, S., Samek, A. and Sprenger, C. (2020). Dynamic inconsistency in food choice: Experimental evidence from two food deserts. *The Review of Economic Studies*, 87(4), 1954–1988.
- Saucet, C., and Villeval, M.C. (2019). Motivated memory in dictator games. *Games and Economic Behavior*, 117, 250–275.
- Schwardmann, P., Tripodi, E. and van der Weele, J.J. (2021). Self-Persuasion: Evidence from Field Experiments at Two International Debating Competitions. *American Economic Review*, 112(4), 1118–46.
- Schwartzstein, J. (2014). Selective attention and learning. *Journal of the European Economic Association* 12(6), 1423–1452.

- Serra-Garcia, M., and Szech, N., (2021). The (in) elasticity of moral ignorance. *Management Science*. Published Online: 22 Nov 2021.
- Shalvi, S., Dana, J., Handgraaf, M. J., and 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.
- Sharot, T., Korn, C.W., and Dolan, R.J., (2011). How unrealistic optimism is maintained in the face of reality. *Nature Neuroscience*, 14(11), 1475–1479
- Sicherman, N., Loewenstein, G., Seppi, D.J., and Utkus, S.P. (2016). Financial Attention. *Review of Financial Studies*, 29(4), 863–897.
- Sloman, S.A., Fernbach, P.M., and Haggmayer, Y. (2010). Self-deception requires vagueness, *Cognition*, 115(2), 268–281.
- Tetlock., P.E. (1983). Accountability and the perserverence of first impressions. *Social Psychology Quarterly*, 285-292.
- Trivers, R. (2011). *The Folly of Fools: The Logic of Deceit and Self-Deception in Human Life*. Basic Books.
- Tversky, A., Kahneman, D. (1974). Judgment under Uncertainty: Heuristics and Biases: Biases in judgments reveal some heuristics of thinking under uncertainty. *Science*, 185(4157), 1124-1131.
- Yankauer, A. (1991). How blind is blind review? *American Journal of Public Health* 81 (7), 843–845.
- Yates, J. F., Curley, S. P. (1986). Contingency judgment: Primacy effects and attention decrement. *Acta psychologica*, 62(3), 293-302.
- Zimmerman, F. (2018). The dynamics of motivated beliefs. *American Economic Review*, 110(2), 337–61.