A neuroeconomic theory of (dis)honesty *

Isabelle Brocas
University of Southern California and CEPR

Juan D. Carrillo
University of Southern California and CEPR

August 2018

Abstract

We develop a theory of dishonesty based on neurophysiological evidence that supports the idea of a two-step process in the decision to cheat. Formally, decisions can be processed vía a costless "honest" channel that generates truthful behavior or vía a costly "dishonest" channel that requires attentional resources to trade-off costs and benefits of cheating. In the first step, a decision between these two channels is made based on ex-ante information regarding the expected benefits of cheating. In the second step, decisions are based on the channel that has been selected and, when applicable, the realized benefit of cheating. The model makes novel predictions relative to existing behavioral theories. First, adding external complexity to the decision-making problem (e.g., in the form of multi-tasking) deprives the individual from attentional resources and consequently decreases the propensity to engage in dishonest behavior. Second, higher expectations about the benefits of cheating results in a higher frequency of trial-by-trial cheating for any realized benefit level. Third, multiplicity of equilibria (characterized by different levels of cheating) emerges naturally in the context of illegal markets, in which expected benefits are endogenous.

Keywords: neuroeconomic theory, control network, dishonesty, bribery.

JEL Classification: D73, D87.

PsycINFO Classification: 2340.

*Address for correspondence: Department of Economics, University of Southern California, 3620 S. Vermont Ave., Los Angeles, CA 90089, USA, emails: <brocas@usc.edu> and <juandc@usc.edu>. We are grateful to members of the Los Angeles Behavioral Economics Laboratory (LABEL) and the audience at various seminars for comments and suggestions.

1 Introduction

The economics of dishonesty, corruption and collusion has been the subject of significant research in economics, from the classical analyses of dishonesty in politics, bureaucracies and other institutions (Ackerman, 1978; Klitgaard, 1988) to the formal models of corruption in firms (Lui, 1986; Tirole, 1996). Dishonesty has been studied from experimental (Sánchez-Pagés and Vorsatz, 2007; Coricelli et al., 2010) empirical (Fisman and Gatti, 2002; Fisman and Miguel, 2007), and theoretical (Tirole, 1992; Kofman and Lawarree, 1993; Dal Bó and Terviö, 2013) perspectives. The literature is so extensive that our short review is inevitably partial and incomplete.

This research has received a renewed interest in recent years, with the attempt to disentangle between the material and psychological payoffs of cheating. Researchers have developed a new experimental paradigm, where subjects self-report the success or failure of an event whose outcome is privately observed –such as the roll of a dice (Fischbacher and Föllmi-Heusi, 2013) or the answer to some question (Mazar et al., 2008)– and are rewarded on the basis of their unverifiable report. The experiment has been replicated and performed under numerous variants. While experimental procedures affect significantly the results, one robust finding is the coexistence of subjects who never cheat, always cheat and cheat "a bit" (see Rosenbaum et al. (2014) and Abeler et al. (2016) for detailed surveys of this rapidly growing literature).

This experimental line of investigation has also attracted the attention of neuroscientists, who have utilized the lying paradigm to study the neural basis of dishonest choices (Baumgartner et al., 2009; Greene and Paxton, 2009). Brain data has revealed different patterns of activity in the control network (an interconnected system of brain areas involved in cognition and attention) between subjects who report accurate predictions too often (supposedly dishonest) and those who do not. The specific patterns of activity are consistent with a two-step decision-making process involving an ex-ante decision to become dishonest and, conditional on being dishonest, a trial-by-trial decision to cheat or not given trial specific information.

In this paper, we build a theory of dishonesty based on the aforementioned evidence. Our goal is not to depart from the traditional neoclassical approach where subjects evaluate the costs and benefits of illegal actions and behave on the basis of it. Instead, we extend that setting by incorporating neurophysiological considerations, and obtain a novel set of

behavioral predictions. This is related but different from the recent attempts by researchers in behavioral economics, who have successfully explained many of the observed patterns in the data thanks to the inclusion of non-traditional elements in the utility function, such as image concerns (see the discussion below).

Formally, we propose the following two-step decision process that matches the evidence reported above. In the first step, a decision to use the "dishonesty" pathway is evaluated. In the second step, accuracy of predictions are reported. If the dishonesty pathway is not selected, truthful responses are automatically provided in the second step. If the dishonesty pathway is selected instead, (costly) attentional and cognitive resources are recruited to trade-off the monetary and non-monetary costs and benefits of cheating. The model is similar to other dual-process theories in which decisions are routed either through a costless and automatic process or a costly cognitive pathway that compares options to formulate a decision (Brocas and Carrillo, 2014). In this setting, given that different individuals have different propensity to cheat (or different psychological costs associated with cheating), some will choose the dishonest route and end up cheating frequently (though not necessarily always) while others will remain honest. Such heterogeneity in behavior is consistent with the behavioral evidence obtained in the neuroscience study of interest.

Using this simple framework, our model delivers a set of novel predictions. First, increasing the attentional load of subjects –for example, by requiring them to engage in multi-tasking– makes dishonesty relatively more costly and therefore less prevalent (Proposition 1). In the presence of cognitive overload, it is efficient to avoid the effortful activity of trading-off costs and benefits of cheating and, instead, to take the costless honesty route. At the same time, only subjects who have an intrinsically higher propensity to cheat follow the dishonesty route, thereby increasing the average level of cheating among dishonest subjects. Second, the distribution of rewards associated with dishonest behavior affects the ex-ante decision to choose the dishonesty route. More precisely, if one expects higher rewards, it is relatively more interesting to engage the costly control network to evaluate cheating. Once this cost is sunk, dishonest behavior is more likely to follow for any given reward level (Proposition 2). This result is in contrast with classical theories where the realized payoffs (current cost and current benefit) matter but not the set of possible options from which they are drawn. It is also the opposite of an externality-driven argument in a standard dynamic economic model. Indeed, suppose that a subject who is caught cheating is fired and therefore foregoes future income. Fixing the current

payoff of cheating, higher future gains for cheating increases the opportunity cost of being fired and therefore decreases the propensity to behave dishonestly in the first place. Third, we embed the previous findings in a multi-person dishonesty game in which bribes are set endogenously by potential bribers. This extended model exhibits multiple self-fulfilling equilibria (with high bribes and high dishonesty or low bribes and low dishonesty) based exclusively on differences in expected rewards (Proposition 3). If high bribes are expected, it is efficient to choose the dishonesty route and to keep the option to cheat. This in turn makes it worthwhile for bribers to offer high compensations. Conversely, if low bribes are expected, it is efficient to choose the honesty route, which makes it less valuable for bribers to offer high compensations. This exercise rationalizes an observation that has proved difficult to predict within classical theories in the absence of contrived modeling pieces. It shows that brain-based models of behavior may be useful not only to understand and predict behavior in the laboratory but also to explain behavior in real life environments.

Our general approach departs from the classical model developed to explain dishonest behavior, where the individual derives utility from monetary benefits but incurs a psychological cost of cheating (Lui, 1986; Andvig and Moene, 1990). The model applies to the coin flip setting and it will be used as the benchmark of comparison for our brain-based model. Our approach also departs from the more recent belief-based models interested in fitting the data from the above mentioned die-roll experiments. A main feature in this paradigm is the tendency of some subjects to lie partially and to choose between lies of different magnitudes. The problem is conceptualized vía belief-dependent utility functions that capture the disutility associated with being *perceived* as a liar. The situation becomes a psychological game in which the subject cares intrinsically about image (Dufwenberg and Dufwenberg, 2018). In related versions (Gneezy et al., 2018; Khalmetski and Sliwka, 2017), the individual is also subject to lying costs that depend on the size of the lie. These models derive optimal behavioral strategies that parsimoniously capture the stylized facts from experiments.[1] The major difference between behavioral models and brain-based models is a methodological one. The former presuppose the existence of preferences that contain the essence of the trade-off required to fit the evidence (benefits and costs). These preferences are represented with extended version of the neoclassical utility functions. The latter do not impose preferences and, instead, model the brain as a processor of information that

[1]Given we do not have evidence about brain activity in settings in which the trial-by-trial choice is richer than deciding whether to cheat or not, we cannot directly compare the predictions of those models.

represents features. The context of the experiment determines endogenously which information is represented and shapes behavior. In the lying paradigm, a brain-based model can predict that the same individual sometimes act as if he cares about self-image and sometimes as if he does not, again depending on the context. We view both approaches as complementary: they enrich the traditional setting in different dimensions and provide a spectrum of implications and testable predictions.

The article is organized as follows. In section 2, we develop the standard model of dishonesty and our extended version based on neurophysiological evidence. In section 3, we present the comparative statics results of our new model. In section 4, we discuss the implications of the model in a bribery game. In section 5, we offer some concluding comments.

2 A model of costly dishonesty

2.1 The canonical model

In the canonical economic model of corruption (Lui, 1986; Andvig and Moene, 1990), individuals trade off the costs and benefits of engaging in an illegal, immoral, or otherwise reprehensible activity (from now on, we will generically refer to as "cheating"). Formally, the payoff of such activity is captured by the following utility function:

$$u = b - \theta \tag{1}$$

In this equation, $b \in [\underline{b}, \overline{b}] \subset \mathbb{R}^+$ represents the net *monetary* benefit associated with cheating. It may refer for instance to a bribe received net of the expected punishment, or a payment for an illegal service. The (possibly unobserved) parameter $\theta \in [\underline{\theta}, \overline{\theta}] \subset \mathbb{R}^+$ captures the intrinsic *honesty* level of the individual, that is, the non-monetary disutility of cheating. It may represent a psychological cost (guilt, shame, aversion to lying, loss of self-respect, etc.) or an economic cost (reputation loss vis-à-vis others, trustworthiness that affects future trade possibilities, etc.).

A key characteristic of the canonical model is that individuals differ in their level of honesty θ (Tirole, 1996; Carrillo, 2000a,b). In particular, an individual whose parameter θ is sufficiently high never cheats whereas an individual whose parameter θ is sufficiently low always cheats.[2] We denote by $G(\theta)$ the cumulative distribution of honesty levels

[2]This is under the implicit assumption that $\overline{b} \leq \overline{\theta}$ and $\underline{b} \geq \underline{\theta}$ but it can be easily generalized.

in the population. In this framework and for a given monetary benefit b, a fraction of individuals $G(\check{\theta})$ chooses to cheat and a fraction $1 - G(\check{\theta})$ chooses not to. Given the preferences described in (1), we have $\check{\theta} = b$. Not surprisingly, the amount of cheating in the population is increasing in the size of the monetary gain ($dG(\check{\theta})/db > 0$).

2.2 A neurophysiologically based model of dishonesty

Although this basic model is an excellent first approximation and has delivered numerous insights, it incorporates dishonesty concerns in a reduced form. The objective of this section is to open the black-box of "costly dishonesty" by relying on recent neuroscience evidence and developing a more comprehensive decision-making model. More precisely, we use evidence to build a stylized model capable of representing how the brain evaluates the decision to cheat on a given occasion. Our theory relies on the experimental findings obtained by Baumgartner et al. (2009) (hereafter, [B.al]) and Greene and Paxton (2009) (hereafter, [GP]) that identify *neural correlates of dishonesty*.

[GP] perform a coin flip version of the lying experiment proposed by Fischbacher and Föllmi-Heusi (2013). Subjects are asked to predict the outcome of computerized coin flips. In the baseline condition, they record their predictions in advance and are compensated based on accuracy. In the treatment condition, predictions are not recorded and rewards are based on self-reported accuracy. Subjects are behaviorally classified as "honest", "dishonest" or "ambiguous" depending on whether the self-reported success rate in the 70 trials of the treatment condition is average ($\leq .59$), improbably high ($\geq .69$) or in-between. As in most of the experimental literature, the majority of dishonest individuals show statistical evidence of cheating but not in every trial.

A comparison of neural activity between the baseline and treatment conditions reveals two important findings. First, there is no significant difference in activity between the baseline and treatment conditions for subjects categorized as honest. Second, there is increased activity in the so-called "control network" (anterior cingulate, dorsolateral prefrontal cortex and ventrolateral prefrontal cortex) in the treatment condition for subjects categorized as dishonest. This extra activity occurs both for self-reported successes (which pools correct predictions and truthful reports with incorrect predictions and false reports) and self-reported failures (cases unambiguously characterized by incorrect predictions and truthful reports).[3] The experiment in [B.al] is methodologically different but yields similar

[3]Another recent study where participants can earn money by cheating on a die-rolling task during

qualitative results and reveals similar patterns of activation.[4]

Overall, a model that aims at representing neurobiological processes of dishonesty must capture two key features. First, a dedicated system –the control network– is recruited during the entire experiment in subjects who cheat significantly and not at all in subjects who do not cheat. Second, behavioral responses are evaluated in each trial without any differential activation of the control network. This evidence is represented by a two-step choice process in which a decision to be open to dishonesty is selected beforehand (step 1) and a behavioral response is formulated in each trial given the earlier decision (step 2). We formalize these ideas with the following model.

Before the experiment starts, the subject knows the distribution of benefits $F(b)$ but not the benefit in each trial (in [GP], for example, the benefit of a correct guess is \$4, \$5, \$6 or \$7 with equal probability).[5] He also knows the ex ante probability α that his true prediction is incorrect.[6] In step 1, a decision to be dishonest or not is made based on this ex-ante information, resulting in the recruitment of the control network or not. If the control network is recruited, the subject is set up to be dishonest in the rest of the experiment, that is, to *selectively* choose when to reveal truthfully or lie. More precisely, in each trial of step 2, the decision to cheat and to not cheat are evaluated and compared based on the specific stakes b in that trial (again, in [GP] the choice of a subject is made after learning whether the benefit in that block is \$4, \$5, \$6 or \$7). If the control

a transcranial direct current stimulation over the right dorsolateral prefrontal cortex also implicates the control network (Maréchal et al., 2017). The design, however, does not classify "honest" and "dishonest" subjects and therefore it does not allow to compare the effect of the stimulation across types. The absence of a dedicated fMRI study of the die-rolling paradigm (where the size of the lie may have noticeable neural correlates) makes it also difficult to interpret the results. Indeed, the stimulation may have affected the *ex ante* representation of gains and losses, which in turn increased the proportion of people who ended up not cheating. However, it is also possible that the stimulation affected the *ex-post* representation of gains and losses of subjects who self-selected into dishonesty, hence modulating the size of the lie. Overall, while the study establishes a causal relationship between one element of the control network and cheating, it does not allow to pinpoint its nature.

[4]More precisely, [B.al] consider a modified trust game. They also studies neural correlates before the honesty choice stage (when subjects decide whether to make a promise that they will later break or not, and when subjects observe the reaction of others to the promise they plan to break or not). Dishonest subjects break their promise of sending back half the money around 70% of the time. They also show activity when breaking the promise in remarkably similar areas (anterior cingulate, dorsolateral prefrontral cortex and amygdala).

[5]For the formal analysis, the results would be qualitatively identical if we assumed that b is known while θ is drawn from a distribution in each trial.

[6]In [GP], subjects predict a coin toss, so $\alpha = 1/2$. The experimenter can easily manipulate these objective probabilities (e.g., by using a dice toss as it is typically done in the experimental literature) or make them contingent on the subject's expertise.

network is not recruited, the subject is set up to be honest in the rest of the experiment, that is, to always reveal truthfully his prediction in step 2, independently of the realized stakes. Overall, and consistent with the evidence in [GP], it is the conclusive decision to entertain the possibility of cheating that engages the control network and not the cheating itself. Last, recruiting the control network involves a cost c that reflects the attentional or cognitive resources necessary to evaluate the options. The process is summarized in Figure 1.

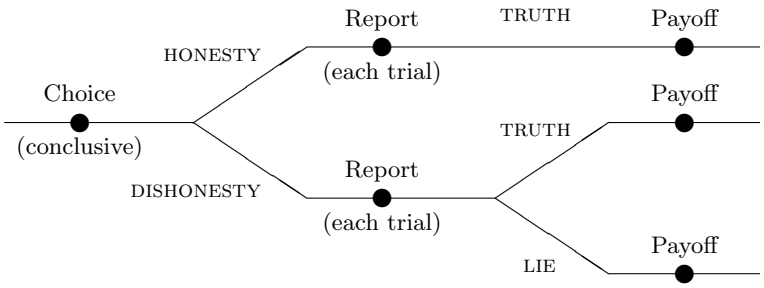

Figure 1. Timing

In this framework, the trial-by-trial decision is context dependent. If the control network has not been recruited, the subject always reports truthfully his prediction. He obtains a payoff b when it coincides with the true outcome, and no payoff otherwise. The expected utility of honesty in step 1 is independent of θ and does not reflect any cost:

$$u_H = (1-\alpha)\, E[b] \tag{2}$$

By contrast, if the control network has been recruited, two cases are possible. When the prediction is correct, there is no incentives to misreport it. When the prediction is incorrect, cheating is optimal if and only if $b > \theta$.[7] The expected utility of dishonesty in step 1 reflects the trial-by-trial trade-offs and the cost of engaging the control network to evaluate options.

$$u_D(\theta) = (1-\alpha)\, E[b] + \alpha \int_{b=\theta}^{\bar{b}} (b-\theta) f(b) db - c \tag{3}$$

[7]This deterministic cheating rule differs from what we observe in practice both in [GP] and in [B.al], where trials with identical payoffs elicit different behaviors. To account for such differences, one could assume that θ can change from trial to trial (so that identical situations may result in different decisions) or that the utility has some random or unobserved component (as in the standard econometric Random Utility Model). The key property, which holds generally in this model and is also supported by the behavioral results, is that lying increases (deterministically or stochastically) with the stakes of the trial.

It is important to note the difference between our brain-based model and the standard model of dishonesty. Depending on the decision to tax the control network, the subject acts as if his psychological cost is "infinite" (which rationalizes never cheating) or "intermediate" (which rationalizes cheating for high enough benefits).

3 Analysis

Given the physiological model outlined in section 2, the decision to engage the control network is driven by the comparison of (2) and (3). The control network is recruited if and only if:

$$u_D(\theta) - u_H \geq 0 \quad \Leftrightarrow \quad V(\theta) \equiv \alpha \int_{b=\theta}^{\bar{b}} (b-\theta) f(b) db - c \geq 0 \tag{4}$$

To be in the interesting scenario, let us assume that the expected benefit is sufficiently high: $E[b] > \underline{b} + c/\alpha$. This means that $V(\underline{b}) > 0$. Since $V(\bar{b}) = -c < 0$ and $V'(\theta) = -\alpha[1 - F(\theta)] < 0$, it is immediate that there exists an interior cutoff $\theta^* \in (\underline{b}, \bar{b})$ given by:

$$V(\theta^*) = 0 \quad \Leftrightarrow \quad \int_{b=\theta^*}^{\bar{b}} (b-\theta^*) f(b) db = c/\alpha \tag{5}$$

such that:

$$\begin{cases} u_H > u_D(\theta) & \text{if} \quad \theta > \theta^* \\ u_H < u_D(\theta) & \text{if} \quad \theta < \theta^* \end{cases} \tag{6}$$

In words, the control network is not recruited (and the subject is hence honest) if the intrinsic cost of cheating is high enough $\theta > \theta^*$. The control network is recruited (and the subject is dishonest) in all other cases $\theta \leq \theta^*$. Then, conditional on being dishonest and making an incorrect prediction, the subject will choose to cheat if the benefit in a particular trial is $b > \theta$ and not to cheat otherwise. Our first result is the following.

Proposition 1 *If the cost c of engaging the control network or the probability of succeeding by chance $1-\alpha$ increase then (i) fewer subjects are dishonest and (ii) the total amount of cheating decreases, but (iii) subjects who choose to be dishonest cheat, on average, more often.*

Proof: Differentiating $V(\theta^*) = 0$ we get:

$$\frac{d\theta^*}{dc} = -\frac{\partial V/\partial c}{\partial V/\partial \theta^*} = -\frac{1}{\alpha[1 - F(\theta^*)]} < 0.$$

By definition, the proportion of dishonest subjects is $G(\theta^*)$ so:

$$\frac{dG(\theta^*)}{dc} = g(\theta^*)\frac{\partial \theta^*}{\partial c} < 0.$$

The total amount of cheating $J(\theta^*)$ is given by:

$$J(\theta^*) = \int_{\theta=\underline{\theta}}^{\theta^*} \int_{b=\theta}^{\bar{b}} f(b)g(\theta)db\, d\theta \;\Leftrightarrow\; J(\theta^*) = \int_{\theta=\underline{\theta}}^{\theta^*} g(\theta)[1-F(\theta)]d\theta$$

Differentiating with respect to c, we get:

$$\frac{dJ(\theta^*)}{dc} = g(\theta^*)[1-F(\theta^*)]\frac{d\theta^*}{dc} < 0.$$

Finally, the amount of cheating among the dishonest subjects $D(\theta^*)$ is:

$$D(\theta^*) = \int_{\theta=\underline{\theta}}^{\theta^*} \int_{b=\theta}^{\bar{b}} f(b)\frac{g(\theta)}{G(\theta^*)}db\, d\theta \;\Leftrightarrow\; D(\theta^*) = \frac{1}{G(\theta^*)}\int_{\theta=\underline{\theta}}^{\theta^*} g(\theta)[1-F(\theta)]d\theta$$

Again differentiating with respect to c, we get:

$$\frac{dD(\theta^*)}{dc} = -\left[\frac{g(\theta^*)}{G(\theta^*)^2}\int_{\theta=\underline{\theta}}^{\theta^*} G(\theta)f(\theta)d\theta\right]\frac{d\theta^*}{dc} > 0.$$

The differentiation with respect to α is analogous. □

The result is intuitive. If engaging the control network becomes more costly, then dishonesty becomes a less attractive option. As a result, there is a larger set of types who prefer to avoid trading off cost and benefits in each trial and, instead, choose to be always honest. This decreases the total amount of cheating. Interestingly, cheating increases among the subjects who decide to be dishonest. The reason is simply that an increase in c makes the marginal types switch from dishonesty to honesty. But these subjects were the least likely to cheat before the increase of c. In other words, only dishonest subjects who are very likely to cheat remain dishonest. Hence, the average amount of cheating by the dishonest subjects is increased. The argument regarding α is analogous. If success is more likely to occur by pure luck, subjects have less incentives to incur the cost of becoming dishonest but, again, those who remain dishonest cheat more frequently.

Proposition 1 provides a testable implication of the theory because c can be manipulated in an experiment. A simple way would be to tax systems involved in the the control network, such as the dorsolateral prefrontal cortex itself involved in working memory.

Manipulations based on dual task experiments are known to affect behavior compared to single-task behavior. We conjecture that asking subjects to perform a cognitive task in conjunction to the computerized coin flip task would increase the cost of trading-off truthful revelation vs. lying. We should then observe a decrease in total cheating and, at the same time, an increase in average cheating by subjects who cheat some of the time. Instead, a classical or a belief-based model of dishonesty would predict no relationship between dishonesty and other orthogonal activities.[8] An individual participating sequentially in an experiment featuring a single coin flip task then an experiment featuring an additional cognitive task would be fitted with different parameters. Differences would be interpreted in terms of different degrees of dishonesty or aversion to perceived cheating across conditions. However, these differences would result from variations in experimental design and they would be orthogonal to image or psychological factors.

This first conclusion is interesting and novel. However, the prediction does not rely on the process being a two-step process. Indeed, the comparative statics would be identical if we assumed instead a one-step process with two costs, θ and c. As such, our first result cannot be used as a test of our two-step model. To provide further testable implications, suppose now that the monetary benefit of cheating can be drawn from one of two different distributions, $F_1(b)$ and $F_2(b)$, such that the latter first-order stochastically dominates the former: $F_2(b) < F_1(b)$ for all $b \in (\underline{b}, \overline{b})$. In words, the individual ex ante knows whether prizes are, in a stochastic sense, high (F_2) or low (F_1). Denote by θ_i^* the dishonesty cutoff when benefits are drawn from distribution $F_i(b)$. Using (5), these cutoffs are such that:

$$\begin{aligned}
& V^*_{F_1}(\theta_1^*) = 0 \qquad \text{and} \qquad V^*_{F_2}(\theta_2^*) = 0 \\
\Leftrightarrow \quad & \int_{b=\theta_1^*}^{\overline{b}} (b - \theta_1^*) f_1(b) db = \int_{b=\theta_2^*}^{\overline{b}} (b - \theta_2^*) f_2(b) db \\
\Leftrightarrow \quad & \int_{b=\theta_1^*}^{\overline{b}} \Big[1 - F_1(b)\Big] db = \int_{b=\theta_2^*}^{\overline{b}} \Big[1 - F_2(b)\Big] db \\
\Leftrightarrow \quad & \theta_2^* > \theta_1^*
\end{aligned}$$

The result is closely related to Proposition 1. A (stochastic) increase in benefit has the same effect as a (deterministic) decrease in cost: it makes dishonesty a relatively more

[8] Notice also that these theories would predict (just like ours) less overall cheating as the probability of succeeding by chance increases. However, unlike ours, they would not predict a higher average level of cheating among those who choose to be dishonest.

interesting option. More interestingly, if we consider one particular trial, we obtain the following result.

Proposition 2 *For a given benefit b, the likelihood of cheating is (weakly) higher if that benefit is drawn from $F_2(b)$ than if it is drawn from $F_1(b)$.*

Proof: Fix b. If it is drawn from F_i, the subject cheats if and only if $\theta < \min\{\theta_i^*, b\}$. Since $\theta_2^* > \theta_1^*$, then for all $b > \theta_1^*$ more cheating occurs under F_2 than under F_1. □

Given our two-step process, the trade-off between honesty and dishonesty is a function of the distribution of benefits whereas the final choice between cheating and not cheating depends on the realization of the benefit. If benefits are likely to be high, recruiting the control network and becoming dishonest is relatively more advantageous. After that choice is made, subjects are locked into a higher likelihood of cheating. Conversely, if benefits are likely to be low, recruiting the control network and becoming dishonest is relatively less advantageous. The individual is therefore more willing to be honest and avoid cheating entirely. In other words, potential rewards frame the mind of subjects on the issue of honesty, and this affects their choice once rewards are announced.

It is easy to see that Proposition 2 holds only under a two-step process. It therefore offers a further test. Indeed, the model predicts that, *controlling for the size of the reward offered*, we should unambiguously observe (behaviorally) more cheating and (neurally) more activity in the control network when the distribution of rewards is tilted towards high values than when it is tilted towards low values. Furthermore, for a given reward b, the same individual could be honest and not display any activity in the control network in an experiment where the ex ante announced distribution of rewards is F_1 and, at the same time, cheat and display activity in the control network in an experiment where the ex ante announced distribution of rewards is F_2. This prediction could be tested by varying the distribution. It should also be noted that classical models cannot predict these variations in behavior. Given trade-offs are always made after the realization of the trial specific information, an individual is predicted to cheat equally for specific realizations, independently of the experimental condition. Therefore, an individual participating sequentially in an experiment featuring distribution F_1 then in an experiment featuring distribution F_2 and who behaves differently in both, would be fitted with different parameters. Behavioral differences would be interpreted in terms of different costs and benefits of cheating while

they in fact result from variations in experimental conditions. Last, given belief-based models are meant to capture the relationship between the size of a lie and the aversion to perceived cheating, they are not directly comparable to our setting in which lying is a binary decision. At the same time, we do not have the neuroscience evidence needed to model decision-making in die-roll paradigms. It is therefore difficult to make adequate predictions and compare them. Based on the general features of these models, we conjecture the following. On the one hand, changing experimental conditions in the context of belief-based models should affect the incentives of a subject to lie, hence the inferences an observer may draw and the self-image concerns. On the other hand, such concerns are always traded-off in these models while they can be avoided in ours through the first step honesty decision. Therefore, variations in experimental conditions are likely to affect the overall amount of cheating in belief-based models but less likely to lead a subject to be fully honest under some conditions and sometimes dishonest under others.

4 Implications for an economic model of dishonesty

In previous sections, the size of the reward associated with cheating has been exogenously set. In an economic game, however, such reward –hereafter referred to as a bribe– is likely to be provided in exchange of a service. In that case, the amount offered is determined endogenously. The objective of this section is to extend the previous individual decision-making problem of costly dishonesty to the case of a bribery game between two actors.

We consider a simplified version of the model developed in sections 2 and 3, where types are discrete. The honesty parameter θ of the potential "bribee" is private information and it can takes three values $\theta \in \{0, \theta_l, \theta_h\}$, with probabilities p, q and $1-p-q$ respectively. The value g of the service for the potential "briber" is also private information and can take two values $g \in \{g_l, g_h\}$, with probability $1-\mu$ and μ respectively. The value of not obtaining the service illegally is normalized to 0. The new parameter g captures for instance the differences in valuations or opportunity cost of individuals.

For simplicity, instead of considering a continuous function $F(b)$, we assume that bribers can only offer two levels of bribes $b \in \{\underline{b}, \overline{b}\}$. Also, in order to be in the interesting situation, we impose the following assumption:

$$0 < \underline{b} < \theta_l < \theta_h < \overline{b} < g_l < g_h \tag{A}$$

Assumption (A) guarantees that a fully dishonest individual ($\theta = 0$) is willing to accept any bribe, whereas the other two types of potential bribees (θ_l and θ_h) find it profitable to accept a high bribe ($\overline{b}$) but not a low bribe ($\underline{b}$). At the same time, while both types of bribers get net benefits even if they pay a high bribe, they obviously prefer to pay a low bribe.

We consider the following stylized two-stage game. In stage 1, one briber and one bribee randomly meet. They each know their own type but they only know the distribution from which the type of the other person is drawn. The bribee decides as a function of his honesty parameter θ whether to pay the cost c of dishonesty or not. As in previous sections, not paying the cost implies remaining honest for sure (in this case, not accepting any bribe and therefore getting a payoff of 0 with certainty). Paying the cost translates into keeping the option of accepting the bribe in exchange of the service later in the game. At the same time, the briber makes a take-it-or-leave-it offer b for the service that depends on his privately known gain. In stage 2, a bribee who paid the cost c decides whether to accept the bribe and provide the service or not. A bribee who did not pay the cost can only behave honestly. Either way, the game ends. We solve the game by backward induction using the standard Perfect Bayesian Equilibrium concept. We obtain the following result.

Proposition 3 *There exists a set of conditions such that the bribery game has multiple self-fulfilling equilibria, with either small bribes and low levels of cheating or large bribes and high levels of cheating.*

Proof. First, notice that by (A), both players anticipate that an individual who has paid the cost c of being dishonest in stage 1 will then in stage 2 accept $\overline{b}$ and refuse $\underline{b}$ if $\theta = \theta_h$ or $\theta = \theta_l$, and he will accept both $\overline{b}$ and $\underline{b}$ if $\theta = 0$.

Let's construct first an equilibrium where $b(g_h) = \overline{b}$ and $b(g_l) = \underline{b}$. Given such offer and the anticipated behavior in stage 2, an equilibrium where the bribee in stage 1 decides to be honest (not pay c) if $\theta = \theta_h$ and to be dishonest (pay c) if $\theta = \theta_l$ or $\theta = 0$ exists if:

$$\mu(\overline{b} - \theta_h) < c \qquad \text{and} \qquad \mu(\overline{b} - \theta_l) > c \tag{C1}$$

Given this behavior by bribees and assumption (A), $b(g_h) = \overline{b}$ (thereby attracting bribees of type 0 and θ_l in stage 2) and $b(g_l) = \underline{b}$ (thereby attracting bribees of type 0 only in stage 2) are incentive compatible in stage 1 if:

$$(g_h - \overline{b})(p + q) > (g_h - \underline{b})p \qquad \text{and} \qquad (g_l - \overline{b})(p + q) < (g_l - \underline{b})p \tag{C2}$$

Overall under (C1) and (C2) there exists an equilibrium where in stage 1 (i) bribees with type θ_h are honest and bribees with type 0 or θ_l are dishonest and (ii) bribers with type g_l offer $\underline{b}$ and bribers with type g_h offer $\overline{b}$.

Let's construct now an equilibrium where in stage 1 $b(g_h) = b(g_l) = \overline{b}$. Given such offer and the anticipated behavior in stage 2, an equilibrium where all types of bribees decide to be dishonest in stage 1 (pay c) exists if:[9]

$$\overline{b} - \theta_h > c \tag{C3}$$

Given this behavior by bribees, it is indeed in the interest of both types of bribers to offer $\overline{b}$ in stage 1 and subsequently attract all bribees in stage 2 rather than offer $\underline{b}$ and subsequently attract only those with honesty parameter 0 if:[10]

$$g_l - \overline{b} > (g_l - \underline{b})p \tag{C4}$$

Thus, under (C3) and (C4) there exists an equilibrium where (i) all bribees are dishonest and (ii) all bribers offer high bribes.

Finally, since conditions (C1)-(C2)-(C3)-(C4) are compatible, when all four are satisfied we have the coexistence of both equilibria. □

Multiplicity of equilibria is a consequence of our two-step sequential process and would not occur if $c = 0$. Indeed, in our model, the potential bribee chooses the dishonest path (and recruits the network system at cost c) *before* learning the size of the bribe offered by the briber. Once this cost is sunk, the trade-off is simply between bribe b and disutility θ. Consider an individual with a moderately high disutility of engaging in bribery (θ_h in our model). If he knows that all potential partners will offer high bribes, he will be willing to incur the dishonesty cost in anticipation of large benefits from the exchange. In that case, it is in his interest to provide the service. Conversely, if he realizes that he will be offered a high bribe only with some probability, it is optimal to remain honest. At the same time, a briber who enjoys a moderately low benefit from illegal services (g_l in our model) is willing to offer a high bribe only if it ensures obtaining the service for sure. If there is a risk involved anyways, he is better-off saving some money by proposing a smaller bribe. These incentives create a coordination problem, which naturally gives rise to multiple equilibria.

[9] Notice that $\overline{b} - \theta_h > c$ automatically implies $\overline{b} - \theta_l > c$ and $\overline{b} > c$.

[10] Notice again that $g_l - \overline{b} > (g_l - \underline{b})p$ automatically implies $g_h - \overline{b} > (g_h - \underline{b})p$.

The result is interesting because an important feature of illegal markets emphasized in the empirical literature (Klitgaard, 1988) is the existence of multiple equilibria. Given a set of initial conditions, markets may end up exhibiting a high or a low level of illegal activities for no obvious reasons. Temporary measures to combat illegal activities may also result in the market switching from a short run high to a long run low level of illegal activities. These observations suggest that both high and low levels of illegal activities may be equilibria in some markets. The existing economics literature has investigated different extensions of the basic model to account for multiplicity. All involve extra modeling pieces such as externalities between the number of corrupt individuals and the probability of catching each of them (Lui, 1986; Andvig and Moene, 1990) or dynamic considerations (Sah, 1991; Tirole, 1996; Carrillo, 2000b; Dal Bó and Terviö, 2013).[11] Our model proposes a new rationale for multiplicity of equilibria in a bribery game. In our model, the anticipation of the behavior of potential bribers has a self-fulfilling effect on the decision to consider the option of cheating. This extension also shows that economic behavior in the context of games and markets can be predicted by modeling individual decisions along the evidence provided by neuroeconomic research.

5 Discussion

Interdisciplinary research between neuroscience and economics has received tremendous attention in recent years, leading to the development of a new field of study: neuroeconomics. However, some economists claim that neuroscience has little to add to our knowledge of economic decision-making (Gul and Pesendorfer, 2008). While some authors argue that understanding neural mechanisms can help make new behavioral predictions (Camerer, 2007) others insist that neuroeconomics will be useful only when it provides out-of-sample predictions in contexts of importance for economists (Bernheim, 2009).

The debate, however, has almost exclusively centered on the potential value of experimental neuroeconomics. This paper falls in a parallel agenda in neuroeconomic theory (Brocas and Carrillo, 2008; Alonso et al., 2013). It contributes to the discussion by demonstrating the methodological advantages for the analysis of individual decision-making of combining the existing empirical fMRI evidence from neuroscience with the theoretical

[11] More precisely, those articles focus on the effect of the anticipation of future equilibrium bribery on current incentives to accept bribes.

optimization tools of economics. Indeed, modeling the neural correlates of choices allows to better understand the mechanisms underlying decisions. This is particularly useful to correctly explain why certain decisions are made in specific contexts, to predict behavior out of sample and to rationalize within subject variations. Usually, brain-based models predict patterns of behavior that cannot be easily reconciled with standard utility-based models. The reason is simply that utility models are designed to represent behavior. Behavior that is not consistent with standard axioms of rationality is represented by augmented utility functions set to reflect (unobservable) psychological factors. The minute a subject is endowed with such utility, he is set to incur those psychological factors. Brain-based models do not presuppose the existence of utility functions and do not intend to fit behavior. This flexibility allows relevant features to be processed only when efficient and predicts within subject behavior heterogeneity. The latter is expected to result from the manipulation of experimental conditions that have no impact on the predictions of standard models. These differences across theories offer valuable testable predictions.

For the specific case of dishonesty, our model delivers three testable behavioral predictions that depart from traditional analyses. All implications are driven by one single mechanism that allocates resources optimally to make decisions. First, adding external complexity to the decision-making problem overloads the control network and consequently decreases the propensity to engage the already costly (in terms of attention) dishonest behavior. Second, higher expectations about future rewards results in an increase in the likelihood of cheating. Better prospects make trial-by-trial trade-offs more valuable and it becomes optimal to recruit the control network more often. In a given experiment, the proportion of dishonest agents increases further resulting in a higher frequency of trial-by-trial cheating for any bribe level. Finally, our brain-based model can be effectively used to model economic phenomena that arise outside the laboratory. In particular, by modeling behavior in a way consistent with how decisions are actually formed (rather than capturing it in a reduced form utility representation), it is possible to predict well documented multiple equilibria that arise in the context of illegal markets.

Last, the model outlined here is consistent with the growing neuroscience evidence of how options are evaluated, selected and implemented in decision-making paradigms. There is converging evidence that different information channels are available and the selection between those is based on task demands. Generally, decision-making tasks can be ordered in terms of their complexity and channels can be organized in terms of the

cognitive resources they allocate to a task. Simple decision-making tasks are processed through channels that incorporate the simple features of a task or its low-order attributes. On the other hand, complex decision-making tasks recruit systems capable of allocating cognitive and attentional resources to represent higher-order attributes. This general organization has been observed in the case of consumption decisions (Hare et al., 2009), dual-task performance paradigms (D'esposito et al., 1995; Szameitat et al., 2002) or memory management (Balconi, 2013) among others. In all these examples, simple versions of the problem involve specific regions directly relevant to the task while the complex versions tax regions involved in attention and cognition. Importantly, regions involved in complex processing overlap in all those examples. Furthermore, these regions overlap with the control network discussed in this article. In the case of dishonesty, a simple task consists in reporting the information truthfully and this solution can be implemented automatically. By contrast, a complex task consists in solving a trial-by-trial trade-off and requires attention. Depending on task expectations (how likely it is to predict correctly, how much may be earned) and individual preferences (costs of dishonesty, image concerns) that are represented beforehand, either an effortless automatic "truthful" processing path or a cognitively costly "dishonest" path is chosen. Overall, the theory developed in this study is not an isolated attempt to model specific patterns of brain activation but rather an example of how decision-making should be conceptualized based on widespread properties of brain-processing.

References

Johannes Abeler, Daniele Nosenzo, and Collin Raymond. Preferences for truth-telling. Working Paper, 2016.

Susan Rose Ackerman. Corruption: a study in political economy. *New York: Academic Pres*, 1978.

Ricardo Alonso, Isabelle Brocas, and Juan D Carrillo. Resource allocation in the brain. *Review of Economic Studies*, 81(2):501–534, 2013.

Jens Chr Andvig and Karl Ove Moene. How corruption may corrupt. *Journal of Economic Behavior & Organization*, 13(1):63–76, 1990.

Michela Balconi. Dorsolateral prefrontal cortex, working memory and episodic memory processes: insight through transcranial magnetic stimulation techniques. *Neuroscience bulletin*, 29(3):381–389, 2013.

Thomas Baumgartner, Urs Fischbacher, Anja Feierabend, Kai Lutz, and Ernst Fehr. The neural circuitry of a broken promise. *Neuron*, 64(5):756–770, 2009.

B Douglas Bernheim. On the potential of neuroeconomics: A critical (but hopeful) appraisal. *American Economic Journal: Microeconomics*, 1(2):1–41, 2009.

Isabelle Brocas and Juan D Carrillo. The brain as a hierarchical organization. *American Economic Review*, 98(4):1312–1346, 2008.

Isabelle Brocas and Juan D Carrillo. Dual-process theories of decision-making: A selective survey. *Journal of economic psychology*, 41:45–54, 2014.

Colin F Camerer. Neuroeconomics: using neuroscience to make economic predictions. *Economic Journal*, 117(519), 2007.

Juan D Carrillo. Corruption in hierarchies. *Annales d'Economie et de Statistique*, pages 37–61, 2000a.

Juan D Carrillo. Graft, bribes, and the practice of corruption. *Journal of Economics & Management Strategy*, 9(3):257–286, 2000b.

Giorgio Coricelli, Mateus Joffily, Claude Montmarquette, and Marie Claire Villeval. Cheating, emotions, and rationality: an experiment on tax evasion. *Experimental Economics*, 13(2):226–247, 2010.

Ernesto Dal Bó and Marko Terviö. Self-esteem, moral capital, and wrongdoing. *Journal of the European Economic Association*, 11(3):599–633, 2013.

Mark D'esposito, John A Detre, David C Alsop, Robert K Shin, Scott Atlas, and Murray Grossman. The neural basis of the central executive system of working memory. *Nature*, 378(6554):279, 1995.

Martin Dufwenberg and Martin A Dufwenberg. Lies in disguise–a theoretical analysis of cheating. *Journal of Economic Theory*, 175:248–264, 2018.

Urs Fischbacher and Franziska Föllmi-Heusi. Lies in disguise: an experimental study on cheating. *Journal of the European Economic Association*, 11(3):525–547, 2013.

Raymond Fisman and Roberta Gatti. Decentralization and corruption: evidence across countries. *Journal of Public Economics*, 83(3):325–345, 2002.

Raymond Fisman and Edward Miguel. Corruption, norms, and legal enforcement: Evidence from diplomatic parking tickets. *Journal of Political Economy*, 115(6):1020–1048, 2007.

Uri Gneezy, Agne Kajackaite, and Joel Sobel. Lying aversion and the size of the lie. *American Economic Review*, 108(2):419–53, 2018.

Joshua D Greene and Joseph M Paxton. Patterns of neural activity associated with honest and dishonest moral decisions. *Proceedings of the National Academy of Sciences*, 106 (30):12506–12511, 2009.

Faruk Gul and Wolfgang Pesendorfer. The case for mindless economics. *The foundations of positive and normative economics: A handbook*, 1:3–42, 2008.

Todd A Hare, Colin F Camerer, and Antonio Rangel. Self-control in decision-making involves modulation of the vmpfc valuation system. *Science*, 324(5927):646–648, 2009.

Kiryl Khalmetski and Dirk Sliwka. Disguising lies: Image concerns and partial lying in cheating games. Working paper, 2017.

Robert Klitgaard. *Controlling corruption*. Univ of California Press, 1988.

Fred Kofman and Jacques Lawarree. Collusion in hierarchical agency. *Econometrica*, pages 629–656, 1993.

Francis T Lui. A dynamic model of corruption deterrence. *Journal of public Economics*, 31(2):215–236, 1986.

Michel André Maréchal, Alain Cohn, Giuseppe Ugazio, and Christian C Ruff. Increasing honesty in humans with noninvasive brain stimulation. *Proceedings of the National Academy of Sciences*, 114(17):4360–4364, 2017.

Nina Mazar, On Amir, and Dan Ariely. The dishonesty of honest people: A theory of self-concept maintenance. *Journal of marketing research*, 45(6):633–644, 2008.

Stephen Mark Rosenbaum, Stephan Billinger, and Nils Stieglitz. Let's be honest: A review of experimental evidence of honesty and truth-telling. *Journal of Economic Psychology*, 45:181–196, 2014.

Raaj K Sah. Social osmosis and patterns of crime. *Journal of Political Economy*, 99(6): 1272–1295, 1991.

Santiago Sánchez-Pagés and Marc Vorsatz. An experimental study of truth-telling in a sender–receiver game. *Games and Economic Behavior*, 61(1):86–112, 2007.

André J Szameitat, Torsten Schubert, Karsten Müller, and D Yves Von Cramon. Localization of executive functions in dual-task performance with fmri. *Journal of cognitive neuroscience*, 14(8):1184–1199, 2002.

Jean Tirole. Collusion and the theory of organizations. In J.-J. Laffont, editor, *Advances in Economic Theory*. Cambridge University Press, 1992.

Jean Tirole. A theory of collective reputations (with applications to the persistence of corruption and to firm quality). *The Review of Economic Studies*, 63(1):1–22, 1996.