Routledge
Taylor & Francis Group

Philosophical Psychology

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/cphp20

Watching the watchmen: Vigilance-based models of honesty fail to explain it

Camilo Ordóñez-Pinilla & William Jiménez-Leal

To cite this article: Camilo Ordóñez-Pinilla & William Jiménez-Leal (01 May 2023): Watching the watchmen: Vigilance-based models of honesty fail to explain it, Philosophical Psychology, DOI: 10.1080/09515089.2023.2206852

To link to this article: https://doi.org/10.1080/09515089.2023.2206852

Published online: 01 May 2023.

PHILOSOPHICAL PSYCHOLOGY
https://doi.org/10.1080/09515089.2023.2206852

Routledge
Taylor & Francis Group

Watching the watchmen: Vigilance-based models of honesty fail to explain it

Camilo Ordóñez-Pinilla [a,b] and William Jiménez-Leal [b]

[a]Philosophy Program, Universidad El Bosque, Bogotá, Colombia; [b]Department of Psychology, Universidad de los Andes, Bogotá, Colombia

ABSTRACT

Promoting honesty is considered a key endeavor in the betterment of our societies. However, our understanding of this phenomenon, and of its evil twin, dishonesty, is still lacking. In this text, we analyze the main tenets assumed by empirical models of vigilance and sanctions. We approach our analysis in three sections. Initially, we investigate the concept of honesty as assumed by commonly used methodologies in studying honesty. This then leads us to identify the previously overlooked but essential element of epistemic privilege in characterizing honesty. In the third part, we delve into how current explanatory models of honesty lack sufficient consideration of epistemic privilege, resulting in incomplete narratives about honesty. Our analysis of the extant literature suggests that both internal (including the self-concept maintenance theory) and external vigilance models fall short of explaining honesty and dishonesty because of both conceptual problems and empirical inadequacy. Identifying these shortcomings allows us to suggest some possible directions of research.

ARTICLE HISTORY
Received 18 October 2021
Accepted 19 April 2023

KEYWORDS
Vigilance; honesty; norms; motivation; self-concept maintenance theory

1. Introduction

Dishonesty, in its different guises, is recognized as a social and public problem around the world (ONU, 2004). From notable corruption cases in government and private companies, to micro-dishonest acts, such as free-riding public transportation, or paying classmates for homework, it seems a general agreement that our societies and personal lives would be better off with less dishonesty among us (Helliwell et al., 2017). In consequence, there are several initiatives to develop programs and interventions that promote honesty and discourage dishonesty. Their success depends, however, on a correct understanding of crucial factors involved in honesty promotion and dishonesty blocking.

CONTACT Camilo Ordóñez-Pinilla ordonezpcamilo@unbosque.edu.co; ca.ordonez@uniandes.edu.co
Philosophy Program, Universidad El Bosque, Ak. 9 #131a-2, Bogotá, Colombia

In recent decades, different methodologies and theories have been proposed to explain honesty. In our view, the study of honesty is fragmented: the underlying concept of honesty behind experimental methods and experimental tasks is not clear, as well as the connection between the proposed theories and the empirical evidence.

In this text, we present two general objectives. First, we propose an analysis of honesty from three sources: the way in which empirical research seems to assume what honesty is, the intuition that people seem to have in mind when they use the concept of honesty to describe behaviors, and a conceptual analysis of the conditions of the behavioral contexts in which it makes sense to use honesty/dishonesty as predicates of behaviors. This analysis will conclude with the proposal of a definition of honesty based on the notion of epistemic privilege. Second, we use this definition of honesty to propose a critical evaluation of the two most widely used explanatory models of honesty: the external sanctions model and the self-image defense model. This evaluation frames the explanatory models of honesty as surveillance/vigilance models: models that propose that honesty/dishonesty is motivated by the concern of avoiding being caught engaging in reprehensible behavior by an external or internal vigilance device.

To achieve these objectives, we organize the paper as follows. First, we review the empirical research on honesty and conclude that the notion of epistemic privilege is key to understanding the concept of honesty that underlies all of this research. Second, privileged knowledge, a decision about how to report that knowledge, and its contextual relevance. Third, we apply our version of honesty as involving epistemic privilege to a critical analysis of current explanations of honesty that rely on external or internal vigilance. Finally, we explore other promising ideas to explain honesty: emotions as commitments theory and intrinsically motivated norms theory.

The way we achieve these goals leads the paper to have a main negative thesis: vigilance-based theories of honesty fall short in explaining the phenomenon. While some external or internal *watchers* (i.e., vigilance devices) may play a role in controlling behaviors that are relevant to the analysis of honesty and dishonesty, the model case of honesty, when epistemic privilege is assured, requires a different kind of explanation.

2. How honesty has been empirically studied

In recent decades, scholars have developed interesting empirical approaches to study honesty from a behavioral approach. These approaches have been useful in providing insights into the factors associated with honest behaviors, and they have uncovered key behavioral patterns that constrain theories and explanations. But, in our opinion, there is not always a clear theoretical background on honesty and dishonesty in behavioral sciences.

The behavioral study of honesty can be considered, in a crucial sense, *fragmented*: empirical studies do not always provide a clear and conceptually sound notion of honesty, which has been described in detail, critically analyzed, and compared with other possible notions and is logically connected to the methodologies used to study the phenomenon. The studies and methodologies seem to respond to an underlying intuition whose conceptual structure is not always explicitly described.

Experimental reports of honesty, dishonesty, and deception research usually begin by either giving a very brief definition of its subject (Azar & Applebaum, 2020; Markowitz & Levine, 2020) or by taking for granted the reader knows exactly what is being studied (Dimant et al., 2020). In other cases, it is only meta-analyses that bring to the foreground questions on the implied concepts and operationalizations of the phenomena at hand (Gerlach et al., 2019). Hence, it is key to analyze how honesty is, overtly or covertly, conceptualized in empirical studies. Mainly, we are interested in analyzing the underlying concept of honesty behind the experimental paradigms used to study it.

We identify three main methodologies to empirically study honesty: random games-die-roll tasks (Fischbacher & Föllmi-Heusi, 2013), matrices tasks (Mazar et al., 2008), and deception games (Gneezy, 2005).

In random games, participants gain knowledge of the outcome of a random process and report it. In *die-roll tasks* (Fischbacher & Föllmi-Heusi, 2013), subjects privately roll a die and make a report of the result, and, depending on it, they win or lose a reward. In *mind games* (Jiang, 2013; Potters & Stoop, 2016)—which may be seen as extensions of die-roll tasks—participants are instructed to think a number between one and six and then roll a die, and, if the numbers match, they claim a reward. Similarly, in *coin-flip tasks* (Bucciol & Piovesan, 2011), subjects privately flip a coin, report the result, and receive, or not, a reward associated with this result.

In skill games, participants resolve a task that requires their abilities and report the result. In *matrices tasks* (Mazar et al., 2008), participants must solve a mathematical problem: finding two decimal numbers that add up to 10.00 in a matrix of 9 numbers. Participants then report how many matrices they solved, claiming a reward that increases with each matrix reported as solved. In some variations, participants return the materials of the experiment without any identification (e.g., no names or codes), while in others, participants may destroy or keep the materials. The first option is used for tracing down actual individual performances using secret codes in the materials, allowing an individual measure of dishonesty, but possibly introducing suspicions from participants about anonymity; that is a matter of concern since honesty requires privileged access to one's performance. The second option works better ensuring anonymity from the perspective of participants, but it allows measuring dishonesty only in aggregates. In

another variation that maximizes the perception of anonymity, participants take the materials home, are instructed to solve as many matrices as possible in a fixed time, and pay themselves from the money included in the materials and return the change to a mailbox in an unidentified envelop (Yaniv & Siniver, 2016).

Finally, in *deception games* (Gneezy, 2005), the sender must choose messages to send to a receiver. There are honest (truthful) messages and dishonest (false) messages. The messages are about which option is associated with a higher or lower payoff. The receiver, without certainty, has to decide whether to trust the received messages. Such a decision determines the payoffs for both players. Honest messages are less lucrative for senders than dishonest messages.

An experimental paradigm to study honesty requires defining a task in which people can be honest or dishonest (i.e., the chance to cheat) and a way to determine the presence or the magnitude of honesty or dishonesty (i.e., the measurement of honesty/dishonesty). In random games, participants have the chance to cheat because they know something that no one else knows: the actual outcome of the random process. And, the measurement of honesty/dishonesty is the difference between the actual outcome and the one reported. In skill games, participants have the chance to cheat, again, because they are the only ones (or so they think) who know what their actual performance in the task was. Again, the measurement of honesty/dishonesty is the difference between their real performance and the reported one. Finally, in deception games, participants have the chance to cheat, unsurprisingly, because they know which messages are associated with specific payoffs. And the measurement of honesty/dishonesty depends on a difference between the actual and the reported payoffs. Table 1 summarizes the tasks used to empirically study honesty and dishonesty.

From the above characterization, it seems that the key element of honesty as a behavioral phenomenon is, as we will call it in the next section, an epistemic privilege: a knowledge that a person has, that no other person has, and that gives him or her the opportunity to take advantage of it. In our analysis, having this epistemic privilege is what creates a context, a behavioral context, in which it makes sense for a person to be in a dilemma between being honest or dishonest. We will develop this idea further in the next section.

Table 1. Summary of experimental tasks to study honesty.

Task	Type of Task	Key references
Die-Roll	Random process	Fischbacher and Föllmi-Heusi (2013)
Mind Game	Random process	Jiang (2013); Potters and Stoop (2016)
Coin-Flip	Random process	Bucciol and Piovesan (2011)
Matrix	Skill performance	Mazar et al. (2008)
Deception Game	Social Interaction	Gneezy (2005)

A second implication of our previous analysis is that, from a purely methodological analysis, random games are best suited to study honesty. We expect that an experimental task allows measuring the dependent variable without *noise*: without measuring other variables at the same time; variables that can explain the measured variance. Hence, we expect that experimental tasks to study honesty allow us to measure differences in response to the task as differences in honest/dishonest behavior. But reports of dishonest behavior in skill and interaction games are likely confounded with signaling concerns. In skill games, dishonest participants might want to display and signal ability[1] while in interaction games, participants might be concerned with appearing virtuous.[2] Hence, "signaling a skill" and "signaling a virtue" possibly introduce noisy motivators in the experimental designs: in a skill game, reporting a certain number of solved matrices may be interpreted as honest or dishonest behavior or as signaling a mathematical skill; in an interaction game, that a participant sends a certain message may be interpreted as an effect of being more or less honest but also as an effect of wanting to signal being a person more or less virtuous.

Additionally, noisy motivators that could taint the effect of independent variables over honest and dishonest behaviors are not the only issue at hand. Characteristics of the task may also threaten the epistemic privilege condition. The different manners in which participants implement their reports are sensible to this condition insofar as they may be seen by them as compromising their anonymity depending on how much they require producing traceable records. The tasks summarized in Table 1 can involve reporting performance to experimenters having the chance to destroy records (e.g., if the participants may recycle or conserve all used material), reporting to experimenters with no records (e.g., in mind games or in random tasks with few trials that do not require taking notes of results for a final report), or reporting without interaction with the experimenters having (e.g., in tasks using envelopes where participants return materials or remaining money) or not having previous records of performance (as in self-payed versions of random and skill games). Anonymity, as a way to secure the presence of the conditions for honesty, especially the epistemic privilege, will be best ensured if there are not any traceable records or at least if participants have full control of disposing the materials used in the experiment.

Gerlach et al. (2019) report meta-analytical comparisons between results from the different tasks to study honesty, finding relevant differences. Interaction games showed the higher rate of dishonesty (M: 51% of false messages), followed by coin-flip tasks (M: 31% of reports higher than expected by probability), die-roll tasks (M: 30% over the expected) and matrices tasks that have the lowest rate of dishonesty (M: 17% of unsolved

matrices reported as solved). In addition, the difference between the coin-flip task and die-roll task was not statistically significant. These considerations, perhaps, suggest that, as we previously argued, the type of task is indeed determining the dishonesty rates, introducing suspicions on whether one can fairly say that they measure the same construct or if they covertly introduce factors that distort it.

In summary, we argue that the different tasks used in honesty research present structural differences affecting the study of the phenomenon. On the one hand, particularities of some tasks imply artificially introducing aspects that possibly affect the interpretation of their results in terms of factors involved in honesty and dishonesty, mixing considerations about signaling skillfulness or virtue. In this sense, random games—called cheating games by Kajackaite and Gneezy (2017)—are the type task that allows a better interpretation of results in terms of factors affecting honesty and dishonesty.[3] On the other hand, anonymity, as a way to ensure epistemic privilege, can be affected by the ways in which the report of the behavior is requested from participants. Experimental setups where participants produce records of their behavior may be perceived as threatening anonymity and, as a consequence, as threats to their epistemic privilege. Then, it is crucial that experiments give participants the total control over the records of their behavior, if its production is necessary.[4]

According to our analysis, epistemic privilege is the key element presented in empirical studies to characterize behavioral contexts where it makes sense to use honesty and dishonesty as descriptions of behavior. Perhaps, this notion can be used to develop a more structured conceptualization of what honesty is or, at least, what it is assumed in behavioral sciences that honesty is. In the next section, we will to develop this idea, proposing a definition of honesty based on the idea of epistemic privilege.

3. What is honesty? – an epistemic definition

Honest and *good* —in a moral sense—are both used as predicates of persons or behaviors, even though *honest* is a narrower predicate than *good*: all honest things are considered good, and all dishonest things are considered bad, but there are good and bad things that are neither honest nor dishonest. *Feeding their cats every day with fresh food* may be considered a good action of a good person but, only with this information, we may not fairly judge whether it is honest or dishonest. If we add to the situation another agent with whom there is an informational transaction, honesty seems more salient as a predicate: feeding their cats every day with fresh food and *telling that* to people at the animal shelter is clearly an honest act. Forgetting to provide cats with fresh food every day while *telling* people at the animal shelter that you indeed fed them is clearly an immoral and also dishonest

act. Likewise, *playing the guitar every day* may not be easily judged as a good or bad moral action, but if we add an informational context that involves another agent (for example, playing the guitar every day but letting parents believe that that time is being spent studying for exams), honesty becomes a predicate that may be true or false of the action. So, while being good may be understood as depending on coherence between actual actions and recommended actions from a normative code, being honest or dishonest seems to involve, in essence, a *reporting dimension*.

For these reasons, we consider that in a behavioral context where an evaluation in terms of honesty and dishonesty is pertinent, the agent has epistemic privilege[5] characterized by three conditions: (a) he/she has privileged knowledge about their own behavior compared with the knowledge about it that other agents may have and (b) he/she makes a decision on how to report it (in terms of quality and quantity of information conveyed).

A third condition emerges when we consider that a measure of *relevance* is also important for our purpose: that A knows something that B does not is not itself an epistemic privilege. It would count as an epistemic privilege only when A's knowledge is relevant to B. So, characterizing behavioral contexts of honesty and dishonesty in an epistemological framework necessarily implies introducing an (epistemic) interaction between agents, one of whom has privileged knowledge about her own behavior, and (c) one (or various) agents for whom such knowledge is relevant.[6] The relevance of the report for the recipient can be based on a number of factors: general trust (e.g., a client reassuring a shopkeeper he has money), filial relationships (e.g., a kid telling his parents he has done his homework) or social roles (e.g., a citizen informing the police his driving license is not expired). These factors also highlight that the relevance of the report might be entangled in a web of expectations, which might include the consequences of these expectations.

Our definition can shed light over possible differences between being honest and being – in general – morally righteous or morally virtuous, in the context of understanding honesty in the epistemic context of the concept of truth. Some other conceptual definitions of honesty also emphasize the connection between truth and honesty, but highlighting other aspects. For example, Roberts and West (2020) propose that even if honesty is a virtue of truth, it involves more than just a disposition to tell the truth but a concern and emotional sensitivity toward the truth. T. Smith (2003) claims that honesty implies refusal to fake reality in a broader sense. In turn, Carson (2010) conceives honesty as defined by (avoiding) lying and deception, in a negative sense, and openness to reveal information, in a positive sense. And Fritz (2020) takes honesty as, in part, an openness to revealing things about oneself that just can be known by such reveals. In this sense, honesty points to a relationship of trust with another person. Other kinds of

proposals such as Miller (2017) and Miller and West (2020) analyze honesty in two factors: a factor about truth (relative to behaviors as avoiding deception) and a factor about justice (relative to behaviors as avoiding stealing).

Hence, our definition, in part inspired by Hume (2014), distinguishes between honesty and justice, understanding justice as a matter of respecting social agreements as the one of respecting private property and understanding honesty as a matter of an privileged epistemic relation with others. Other conceptualizations of honesty have assumed that a relation with truth is essential to understanding honesty. But, in our consideration, honesty has to be more than a mere attitude (for example, favoring truth) but a kind of recognition of a privilege, an epistemic privilege. For us, honesty is not just being *truthful*, as for example we expect for a scientist or a journalist, in general, but restricted to a kind of knowledge to which we have a privilege compared with other persons to whom this knowledge is relevant. It is this kind of privilege, and the kind of commitments derived from it, that makes honesty relevant. Being truthful imposes some kind of commitments, for example, assuming the consequences of truth, as for example for a scientist publishing that some experiment did not result as expected implying that her theory is not validated. Being honest imposes some commitments derived from truthfulness but others that go beyond. When honesty is more salient than mere truthfulness, the commitments are derived from the fact of recognizing a privilege, and then commitments for honesty are derived from normative considerations about how to deal with such privilege. Fritz (2020) has a conceptualization of honesty closer to ours, when she speaks of honesty in terms of openness and trust. But we want to stress that in our notion honesty implies openness but derived from the recognition of a privilege and a kind of epistemic trust that originates in the recognition that others have epistemic privilege.

As a final point, we agree that one of the consequences of our analysis is that honesty is empirically undetermined, and in that sense, we can be blamed for not being psychological realists. Perhaps, epistemic privilege is something that is not presented in our everyday lives. But we want to highlight that our point is at another level of analysis: on one side, our point is intentional not extensional. Our point does not depend on the real possibility of epistemic privilege but on the possibility that people can actually believe that. Even in the experimental setups – in exception of mind games – experimenters have the possibility to get the real answers. But even in these cases, experiments work because participants indeed believe that they have privilege about the concerned knowledge. On the other side, our definition seems to be in use in the experimental setups to study honesty. Therefore, if our definition lacks psychological realism, the accusation reaches the experimental programs to study honesty. Finally, we suppose our definition captures an intuition in the common use of the term.

If the footballer Lionel Messi were to say that he scored twice in the World Cup final match, it would make no sense to evaluate his utterances as honest or dishonest since his performance is public. But if he talked about his feelings the night prior to the match, it would make sense to evaluate his honesty because we suppose he has a privileged epistemic access to his mental life.

In summary, if our analysis is correct, the notion of epistemic privilege is useful to construct a concept of honesty that is conceptually sound and that allows operationalization in experimental setups. In this sense, being honest is being truthful – assuming a commitment with truth – under some specific conditions: it is necessary to report a knowledge that is privileged and relevant for the persons at stake. Hence, under these conditions, what factors are associated to the decision of making truthful or untruthful reports? The next sections will explore further one widely proposed answer to this question: the key factor to explain honesty is *vigilance*.

4. How honesty may be explained

4.1. Norms and vigilance as explanations

We will start by presenting the norm-vigilance-sanction model to explain action and later analyze whether this model can be applied to build an explanation of honesty and dishonesty. Vigilance and sanctions explain why people do the things they do in the context of appealing to norms as guidelines for action and appealing to vigilance and sanctions as tools for making norms operative. Norms[7] are usually invoked in explanations of action since they explain how behavior is regulated—promoted or refrained—(Bicchieri, 2015) in different situations such as sharing a common good (Olson, 1971; Ostrom, 1990), self-regulation to act correctly (Kant, 1996), curbing effects of negative externalities or promoting positive ones (Bicchieri, 2006, 2017), establishing regularities favored by the evolutive development of social animals (Churchland, 2011), achieving peaceful coexistence between individuals (Mockus, 2002), controlling natural impulses (Freud, 1961), culturally shaping moral intuition (Haidt, 2012) or establishing a minimal rational agreement between the individuals of a social group (Rawls, 1996).

Since the mere existence of a norm does not explain why people are motivated to comply with it (Elster, 1989, 2007), sanctions emerge as the theoretical elements that close this gap, explaining how norms have an instrumental motivational power (Fehr & Gächter, 2000, 2002; Foucault, 1995; Kallgren et al., 2000).[8] Rational Choice Theory (Arrow, 2012; Elster, 1986) proposes that sanctions are *costs*[9] that, once integrated in utility functions, make norm compliance the rational

optimal choice (Ullman-Margalit, 2015). Moral theories usually understand sanctions as emotional negative states, as feeling guilty (Freud, 2019; Nietzsche, 2011). In turn, theories of social norms understand sanctions as the loss of reciprocation in social interactions (Cialdini & Goldstein, 2004; Cialdini et al., 1990); the loss of social privileges reserved to members of the social group (Toby, 1957); or the experience of negative social emotions such as shame (Da Cunha Fortes & Thomé Ferreira, 2014).

If norms require sanctions to be instrumentally operative, sanctions require vigilance (watchers) to be implemented (Becker, 1968; Cohen, 1985; L. Smith & Vásquez, 2015; Sarat, 2005). As classical research by Foucault (1995) shows, vigilance is the device that allows behavior monitoring to allocate sanctions, having as a side effect the deterrence of possible transgressors. For example, legal sanctions require an institutionalized vigilance apparatus structured in surveillance cameras (Kuhns et al., 2012) or the presence of police officers in the streets (Klick & Tabarrok, 2005). In the case of social norms, informal vigilance operated by members of the social groups ensures that transgressors are sanctioned. As Elias (2000) argues, social vigilance over table manners, sexuality, and the disposition of the body was a key element in both the creation and sustain of social norms in Western Societies. In turn, instrumental operativity of moral norms inaugurates a notion of internal vigilance and sanctions: individuals monitor their own behavior, judge it, and, when a moral violation is detected, a moral-emotional sanction arises, such as a—unpleasant—feeling of guilt (Freud, 2019; Nietzsche, 2011).

In short, explaining why people do what they do by saying that doing it this way is complying with some norm (or set of norms) requires a subsequent explanation of the normative force of this norm (Haugeland, 1998). Sanctions are useful to provide this explanation, and vigilance is the device that allows sanctions operativity, identifying the concrete instances of actions where sanctions must be implemented. This is what we call the norm-sanction-vigilance model of explanation of actions.

Can this model be used to develop a successful explanation of honesty and dishonesty? Would the decisions on how to report one's behavior, when an epistemic privilege is present, take part in an awareness of vigilance and the *costs* of transgression? Is it even possible to articulate a notion of vigilance that may be operative in behavioral contexts where honesty and dishonesty may be used as predicates? In what follows, we present answers to these questions, emphasizing the problems inherent in articulating a notion of vigilance—either external or internal—which may be operative in honest and dishonest acts. We suggest, hence, that honesty and dishonesty have to be explained in a framework beyond norms and sanctions.

4.2. Honesty and the external watchers

External vigilance may be understood, in the sense that it is relevant for explaining honest behavior, as a process in which behavior is monitored by *another* agent or institution. Then, we may ask if honesty could be explained as motivated by the vigilance of other agents or institutions. According to the definition introduced in section 3, honesty and dishonesty are salient as possible predicates of behavior or character in contexts where there is an epistemic privilege. Does this definition by itself rule out external vigilance as a possible explanation of honesty?

Two questions should be distinguished here: one has to do with the existence of an empirical link between external vigilance and honesty; the other, with the cognitive mechanism that explains this link. So, first, we are going to inquire whether empirical evidence supports the claim that people's choices on reporting are driven by the aversion of sanctions that may follow from external. Second, we are going to introduce an analysis of the cognitive mechanisms associated with external vigilance in the context of honesty.

As a necessary clarification, to our knowledge, empirical studies have focused to a greater extent on the link between external vigilance and prosocial behavior and to a lesser extent on the specific link we are interested in between external vigilance and honesty. We consider here both kinds of studies, especially those on prosocial behavior that we believe can be informative about honesty.

In psychological research, there are different ways to operationalize the concept of external vigilance to study its effect over honesty. Some studies have used a direct gaze from a real person to a subject performing an experimental task (Hietanen et al., 2018). Even printed images of eyes have been used as watchers, giving rise to what has been called *the watching eyes effect* (Manesi et al., 2018). Additionally, direct scrutiny over the results in a task performed in private (Gino et al., 2013) is an alternate way of operationalizing external watchers to implement vigilance. Also, direct vigilance from a person or a robot has an effect over honest behavior (Hoffman et al., 2015).

There appear to be contradictions between some of the key findings of these studies on the connection between external vigilance and honest behavior. While some studies have found an association between external watchers and honesty (Conty et al., 2016; Hietanen et al., 2018; Jansen et al., 2018), other studies have not found such an effect (Cai et al., 2015). Therefore, it is not clear what the effect of external vigilance on honesty is. It may also be that other mediator and moderator factors between external vigilance and behavior are the ones doing the true explanatory work.

It is possible that the difficulty in finding a clear-cut case of external vigilance's influence on honesty is due to *confounders*. Some studies have proposed that personal traits are of greater importance. For example, Pfattheicher et al. (2018) revealthat honesty-humility trait is a negative predictor of cheating in the absence of a watching eyes effect. In a similar line, Jiang (2013) found that something as simple as the order of the task is associated with differences in lying rates.

In addition, other studies propose that it is not really external vigilance but norm adherence, which explains why honesty is fostered in situations where the possibility of social consequences of actions is made salient. Oda et al. (2015) report a study in which participants roll a die and report the result, knowing that a higher report implies getting a higher reward and knowing that this reward will be donated to a charity. This task creates a situation in which being dishonest is seen as a prosocial act: inflating the result implies giving more money to a charity. Participants in the condition of external vigilance were more honest than participants in the condition of no external vigilance. This may count as evidence for the claim that external vigilance effects are mediated by adherence to norms and not to by the motivation for honesty. In turn, finding contradictory results, Hietanen et al. (2018) proposed a task where being dishonest is explicitly allowed and presented as a possible curse of action from the instructions of the experiment. Participants in the external vigilance condition were less dishonest than participants with no vigilance, implying that vigilance has an effect over behavior mediated by the motivation to be honest and not by pure norm adherence.

Since the effect of external watchers on honesty is not uniform, we believe that it is an unlikely candidate for an *overall* explanation. The evidence is contradictory and, some of it, tangential. There is evidence of people being honest or dishonest, both in the presence and absence of vigilance.

In addition, the explanation of honest behavior as exclusively motivated by the avoidance of sanctions linked to external vigilance yields an empirical prediction: *people would be fully dishonest* in situations when external vigilance is not possible. But, contrary to this prediction, the average results of the studies about honesty and dishonesty, where by design—see section 2—external vigilance is excluded, show a different trend of behavior: it is not common to find 100% dishonesty (Bucciol & Piovesan, 2011; Fischbacher & Föllmi-Heusi, 2013; Gneezy, 2005; Jiang, 2013; Mazar et al., 2007, 2008). Then, honesty seems to be present in contexts where the prediction of the model of external vigilance and external sanctions is not fulfilled.

Can social reputation work as a mechanism to explain how external vigilance may affect honest behavior? At first glance, the very idea of vigilance is at odds with the necessary conditions for honest behavior *per our* definition. If *complete* vigilance was achievable in a given

situation, then *honest* would be irrelevant as a predicate. Epistemic privilege *of some sort* is a requirement to make sense of the description of behavior as *honest*.

In our conception, honesty requires a social context since it involves a relation with other agent, which is structurally different from the social context required for external vigilance to be fully operative. As suggested above, for external vigilance to affect prosocial behavior, agents must be aware of the potential implications of their actions on their social reputation. But the social context required for honesty is precisely one in which the actions (understood as reports of privileged information) of the privileged agent are not likely to have an impact on reputation. If the notion of epistemic privilege has a role in defining honesty it is, precisely, to establish honesty as a *social riddle*: honesty requires a social interaction, the nature of which is incompatible with the social dynamics of social vigilance.

Obviously, the differences between social contexts necessary for honesty and operational vigilance might seem too extreme. Presumably, real-world situations with total epistemic privilege are rare, for example, because our actions always leave a trail that can be objectively tracked down. For this reason, the conceptualization of honesty presented in section 3 must not be taken as an extensional description of a fact, but an intentional conceptual reconstruction of a phenomenon: honesty is more related with how subjects represent their social situations – as involving epistemic privilege – that with how these situations indeed are in objective terms. Perhaps, in causal terms, any social situation is *prone* to social implications since it is always possible to acquire knowledge about the actions of every agent. But what is really important for honesty is that an agent believes that he/she is in a situation where his/her epistemic privilege is ensured. Consequently, the more an agent considers that he/she has an epistemic privilege over his/her performance, the less external vigilance will have impact on his/her decisions about how to report such performance.[10]

So far, we can conclude that external vigilance is unlikely to be the central explanatory variable of honest/dishonest behavior. First, because there is no clear empirical evidence showing that this kind of vigilance promotes honest behavior or discourages dishonesty. Second, because a crucial empirical prediction of the theory is not fulfilled. And, third, because there are conceptual problems in the analytic connection between the definition of honesty and the conditions for the operativity of external vigilance on behavior. However, this does not imply that vigilance has to be ruled out since it is also possible to think about honesty as a product of internal vigilance. Can internal vigilance explain honesty?

4.3. Honesty and internal watchers – the self-concept maintenance theory

The failure of external vigilance in explaining honesty leads us to introduce another version of the norm-vigilance-sanction model where the elements are understood in a psychological perspective. In this version, the model assumes that individuals self-monitor their own behavior and self-inflict sanctions, usually understood in psychological terms, such as feelings of guilt, stress or anxiety, cognitive dissonance, or threats to their moral identity. In this section, we present a critical analysis of the most influential view of honesty, which we interpret as a psychological-internal version of the norm-vigilance-sanction model: the self-concept maintenance theory (SCMT).

SCMT (Mazar et al., 2008) claims that, when anonymity is fully ensured, and social vigilance may not be operative, the self remains vigilant, watching his/her own behavior, ready to display internal sanctions or rewards. Mazar et al. (2008) argues that honesty must be explained by the interaction between internal-psychological factors and external factors, which sets up the motivational structure of the behavior. As external factors, Mazar et al. (2008) identify the utility derived from a dishonest act, and, as internal factors, the attention to moral standards and the malleability of categorization. According to these authors, the defense of moral self-concept is the variable that explains the equilibrium between external and internal factors involved in honest behavior. Being dishonest is useful, hence people will behave dishonestly if they have the chance. However, they will not behave as dishonest as possible because they also want to maintain their moral self-concept. They will, therefore, behave dishonestly up to the point in which an updating of their moral self-concept may be avoided. And malleability of categorization works as a process for such avoiding, giving options to treating transgressions as correct actions or, at least, as less serious than they are.

Mazar et al. (2008) propose an experimental program to support SCMT. Participants in the experiments presented by Mazar et al. (2008) solve matrices with an arithmetical riddle and receive monetary rewards depending on how many matrices they solve in the time allotted. They are able to report successful solution of as many matrices as they wish since no one different from them has access to the worksheets. This design ensures that participants solve matrices in conditions of full social anonymity allowing participants to cheat in order to gain more money, without any social sanction. Since they do not cheat as much as possible (i.e., since participants in general do not report the maximum number of matrices that are possible to solve), Mazar et al. (2008) interpreted that participants refrain themselves of behaving dishonestly as much as possible because they want to avoid being obliged to update their moral self-image.

According to SCMT, the cognitive mechanism people use to maximize the allowed amount of useful dishonesty is moral disengagement. In moral disengagement theory (Bandura, 2016), people avoid behaving in ways contrary to their own moral standards because such kind of behavior will produce self-condemnation. But sometimes people are tempted to behave in ways that would result in violating their own moral principles, but, at the same time, represent desired benefits. And, at times, people give in to such temptation. Since self-condemnation would result psychologically costly, people may morally disengage from violations to their own moral code, blocking the rise of self-sanctions and then, obtaining the benefits of moral violation, without paying the psychological prize of immorality. This may be seen in the context of the distinction between omission and commission: dishonest behavior is less frequent when it is framed in a setup in which dishonesty requires commission than when it just requires omission (Mazar & Hawkins, 2015). Commission makes moral disengagement harder than omission and then is related with less dishonesty. In addition, moral disengagement may not only justify but also motivate dishonest behavior (Mazar & Aggarwal, 2011).

Moral self-concept defense is presented as an explanation of honesty and dishonesty at the individual level. When such behaviors are studied at the group level, other variables need to be considered, such as (1) lower external costs plus higher benefits, (2) lack of social norms, which results in a weak internal mechanism, (3) lack of self-awareness, and (4) self-deception (Mazar & Ariely, 2006).

Hence, SCMT may be seen as a contemporary version of classical theories of moral norms that stress the psychological costs of moral transgressions as the mechanism whereby moral norms control behavior (specifically in the context of honesty/dishonesty). For instance, Nietzsche's (2011) conception of guilt as a way in which the individual behaves cruelly against himself for committing a moral transgression; Freud's (2019) conception of psychic tension generated by the impulses of the *Id* and the moral constraints of the Super-Ego; and moral sentiments theories (A. Smith, 1986; Hume, 2014), which argue that moral concepts obtain their meaning from emotional evaluations of experience. In turn, SCMT establishes as moral mechanism of control, not an emotional activation, but a psychological function: the defense of moral image that is linked to self-concept. According to this theory, since people are highly motivated to defend their own self-concept, they are also highly motivated to defend their own self-moral concept.

It seems clear that the evidence presented in Mazar et al. (2008)—see also Mazar et al. (2007)—shows that, in absence of social vigilance, people are not fully dishonest, even if that would be useful for them. However, this does not, by itself, show that the reason people are still honest in the absence of vigilance is a defense of their own moral self-image. Experiments presented

in Mazar et al. (2008) showed that when people are presented with a moral reminder, or when they have a low chance of describing their dishonest behavior as not dishonest, they cheat less than people in the contrary conditions, even if dishonest behavior could not be discovered and even if they report being conscious of their own dishonesty. Strictly, this evidence shows that, in the absence of external vigilance, people cheat at some level but also remain honest at some level. Hence, this research supports the idea that vigilance decreases dishonest behavior, and also that we need an explanation of honesty beyond mere external vigilance. This does not count, however, as directly testing the defense of moral self-image as an explanation of honesty since they do not control whether such defense is operative when people decide what action to perform and they do not manipulate the strength, for example, of such self-image.

These experiments, in the best-case scenario, show that moral considerations are important when people decide whether to behave honestly or dishonestly, but they are mute regarding the mechanism whereby moral norms affect behavior. Additionally, it is not clear how the levels of tolerated immorality are structured: are they defined in terms of kinds of behavior (i.e., such behaviors are permitted, but such others not)? Are they composed of frequency of behaviors (i.e., subjects may do the action X in some range of times) or are they composed of contextual determinants (i.e., subjects may do the action X in these conditions but not in that others). Finally, evidence for SCMT comes from experiments using matrices tasks that, as argued in section 2, seem not to be well suited to study honesty. The task introduces motivations to be seen as skillful or a good-performer that obscure the interpretation of the results in pure terms of motivations for honesty or dishonesty, the goal of the theory.

In addition, it is not yet clear why people would be motivated to defend their moral self-concept. If research demonstrates that people behave honestly in conditions of no vigilance because doing otherwise would imply changing their moral self-concept, it would remain an open question why people place such high value on such moral self-concept that deprives themselves from maximizing utilities for the sake of preserving a good concept of themselves.

Furthermore, from an empirical point of view, Verschuere et al. (2018) report that, contrary to what is reported by Mazar et al. (2008), moral reminders do not work as motivators for honesty: when cheating is possible, although people do not cheat as much as possible, there is no difference between the dishonesty rate of people who are primed with a moral reminder and the dishonesty rate of people who are primed with a neutral reminder. This evidence casts doubt on the idea that moral considerations are operative when people behave in honest or dishonest ways where external vigilance is not possible.

Moreover, we find failures in the scope of SCMT as an explanation of honesty. First, results reported in Cappelen et al. (2013) show that people occasionally refrain to express Pareto White Lies – lies that would benefit all parties involved in a group task. It is difficult for SCMT to explain this result since in this theory honest-dishonest behavior is generated by a utilitarian calculation between rewards from lying and psychological costs of lying. However, the result seems best interpreted as a deontological concern about lying, which the authors called *aversion to lying*, which is not context sensitive, as the defense of self-image would be. Second, a fundamental phenomenon for SCMT is that in absence of external vigilance, when cheating is possible and useful, people cheat but not all the way out, explaining such residual honesty appealing to the defense of the moral self-concept. However, Pascual-Ezama et al. (2015) present a review of previous research showing that, with complete anonymity, 100% cheating is not unusual, problematizing the key finding for which SCMT emerges as an explanation.

Hence, the conceptual, methodological, and empirical problems in SCMT suggest that the widely accepted explanation of honesty in terms of internal vigilance may not be the best account of honesty and dishonesty. In the best-case scenario, SCMT must be seen as an incomplete explanation of honesty and dishonesty.

The analyses presented above seek to show that external and internal vigilance do not seem to be the key variables for understanding honesty and dishonesty. However, given that vigilance and sanctions were introduced as elements to explain the operationalization of norms, another possible explanatory route would be to appeal to the thesis that honesty and dishonesty are indeed behaviors governed by norms, but thinking of types of norms that are operative outside the framework of vigilance and sanctions. The next section will explore this possibility.

5. Honesty, emotions, and intrinsically motivated norms

Norms are cognitive structures widely used in social and cognitive sciences to explain action. Most theories appeal to extrinsic motivators to explain compliance with norms. These extrinsic motivators are usually understood as sanctions, which in turn require vigilance as a tool to identifying behaviors that deserve them.

As we explained before, vigilance and sanctions are useful notions to explain actions, but external vigilance and external sanctions were found insufficient to build an explanation of honesty: there is no clarity on whether there is an empirical link between external sanctions and honesty, given the contradictory evidence and the possibility of confounders. Furthermore, we analyzed that even if such link existed, the cognitive mechanisms that

explain it requires a kind of social context different from the kind of social context that is relevant for honesty (according to our definition that requires epistemic privilege). Internal vigilance might work as an alternative framework that appeals to a self-monitoring of agents' own behavior, which is usually considered as linked to emotional reactions associated to private norm transgression, as feelings of guilt. The contemporary version of internal vigilance theory that has been developed to explain honesty, SMCT, is an *update* of traditional moral theories that explain norm compliance appealing to internal emotional sanctions. But as our analyses showed, SMCT is also problematic as an explanation of honesty. First, we showed that there is no clear empirical evidence supporting it, and some of their main findings could not be replicated.[11] Second, we found that SCMT lacks some conceptual elaborations – specially about the cognitive mechanisms operative in honesty – making the theory, in the best case, an incomplete picture of honest behavior.

But there is another way in which norms can still be part of an explanation of honesty: if norms can be operative by themselves, they can motivate behaviors out of the scope of the dynamics of external or internal sanctions. We want to explore two theoretical approaches that lead the explanations of honesty in this direction: the theory of emotions as commitments and the possibility of intrinsically motivated norms.

5.1. Norms and emotions as commitments

One possibility to support SMCT as an explanation of honesty is to adopt a response inspired by theories of moral sentiment (A. Smith, 1986; Hume, 2014): there is a negative emotional reaction—e.g., guilt—associated with the recognition of having committed a blameworthy act, and *human nature* always seeks to avoid the displeasure generated by these negative emotions.

But appealing to negative emotions in the context of self-interest models introduces a further problem: unless we assume a purely innate connection between a kind of act and the moral emotions that arise as a response to acts of that kind,[12] we must explain how that connection is established in social and personal development. In this sense, we have to explain *how and why did guilty develop as a control mechanism of morality in honest behavior?* This question is especially tricky since, to reiterate, explaining honesty from moral sentiment theories in the framework of SMCT as a self-interest model would imply that feeling a negative emotion, such as guilty, when committing a dishonest act, is somehow self-serving. Hence, refraining dishonesty through associating it with a negative emotion must be self-serving, even when dishonesty has an objective material reward and does not imply a social reputation cost. How is this possible?

A possible answer may be found in the theory of emotions as commitments (Frank, 1988). According to Frank (1988), honesty is motivated by guilt understood as a commitment. In this explanation, honesty and guilt serve the interest of the individual in the context of character trait cultivation: being seen as good is useful to signal being trustworthy. The best way to be seen as good is indeed being good. And the best way to be good is the cultivation of good character traits, behaving in good ways in all possible occasions. Hence, in the absence of external vigilance, subjects self-vigilate motivated by a desire to cultivate good character traits that, when socially signaled, serve self-interest.

Hence, the theory of emotions as commitments explains that people are honest in absence of external vigilance *because* this is a way to cultivate a moral character that, when it is expressed in social situations, signals to others that we are trustworthy. This serves the self-interest of individuals, as being seen as trustworthy makes them prone to profitable social interactions. Feeling guilty is a commitment to behave honestly, culturally developed as a way to insure the cultivation of a positive moral character.

It is clear that behaving X-ly is the best strategy to signal that X is a trait of our character – if others note our behavior. And, perhaps, the best way to ensure that we will behave X-ly when it is required is developing an agencial habit of doing X, namely, developing X as a real trait of our character. But since honesty is an action defined by an epistemological privilege, its self-serving role must be explained without appealing to the possibility that honesty signals something positive to our social group. Hence, if others will not note our behavior, then developing a trait to signal something associated to such behavior is *innocuous*.

Since Frank (1988) among others (Wilson, 2018) has proposed that honesty is a virtue, a possible solution to this problem could be appealing to a conception of moral-agencial character as a strong unity—known as the thesis of the unity of the virtues (TUV). Aristotle (*NE*, 1144b32, 1145a2) proposes that a virtuous person possesses all the single virtues, thanks to the possession of practical wisdom. In accordance with TUV in the context of the theory of emotions as commitments, honesty is a by-product of the cultivation of those virtues, while it is not socially profitable, will be anyway developed as a behavioral commitment as a part of the all-in-one package of virtues.

TUV has been broadly criticized. Barbosa and Jiménez-Leal (2020) have found that the unity of the virtues is not recognized as a central feature of folk concept of character. In turn, Wolf (2007) argues that our daily experiences reveal the unity of virtues to be psychologically unrealistic. Further, the unity of virtues just entails that having one virtue implies having the necessary practical knowledge to be in possession of other virtues (but not an actual possession of them). In addition, Badhwar (1996) proposes that, if

practical life is divided into different domains, virtues are united just within domains, implying that the TUV is just a limited version of character. Also, authors such as Flanagan (1991) and Foot (1983) claim that an analysis of the practical implications of virtues shows that some of them are incompatible, making the unity of virtues incoherent.

Independently of these problems, the version of honesty implied by the TUV seems strange by itself. Thinking of honesty just as a by-product makes it difficult to explain why an agent would be motivated to be and to become honest, investing resources in self-vigilance and self-control. According to our previous analysis, SMCT lacks an explanation of why people are motivated to be honest since it is not really clear how and why dishonesty affects the moral self-concept. The theory of emotions as commitments may solve this problem appealing to a functional role of emotions as commitments to behave in ways that, though irrational, result in the self-interest of individuals, suggesting that honesty is motivated by a desire to avoid guilt. And since—ideally—honesty occurs with epistemic privilege, it is not useful for social signaling, and it seems generated just in a vision of virtues as unified.

When someone is already honest, and feels guilty, there is an established commitment. Such commitment works to explain why people are prone to self-vigilance and self-control. But when honesty is being cultivated, it is not easy to see how the motivation to adopt it as a trait arises. Why will people adopt a pattern of behavior (investing resources in self-vigilance, self-control, and losing the rewards associated with dishonesty) which is not useful for them in isolation, and which their community is not watching or controlling? In the process of developing other virtues, where social vigilance is operative, is that social dynamic both through rewards and sanctions play the role of motivators for the adoption of a new behavior. Once established, the operativity of external vigilance is not necessary to control behavior. For this reason, authors like Nussbaum (1990) and Rawls (1999) have proposed that the flourishing of virtues requires a proper social environment. But, in the explanation of honesty derived from the theory of emotions as commitments, the social world cannot play that role. Hence, we lack a complete explanation of how honesty may flourish.

5.2. Intrinsically motivated norms

Another possible approach to a norm-based explanation of honesty is to argue for the possibility of intrinsically motivated norms: compliance with norms is not exhausted in – the fear of – sanctions. Kelly (2020) has proposed that, in functional terms, intrinsically motivated norms are internalized norms: their effect over behavior is direct, bypassing practical reasoning. In Kelly's view, the cognitive architecture for norms is not restricted to effortful, slow, rational systems, but norms can operate in the

automatic, effortless, unconscious cognitive system (Kahneman, 2011). According to Kelly and Davis (2018), there are intrinsically motivated-internal norms and instrumental-external norms, and a person actually acquires a norm only when the compliance with the norm is intrinsically motivated: only because it is the right behavioral option.

How have these internal-intrinsically motivated norms been generated? According to Kelly and Davis (2018) and Kelly and Setman (2020), the process of internalization of norms is a process of social learning. In this model, humans are constrained by cultural and evolutionary pressures to acquire, comply with, and enforce certain norms. And, the cognitive mechanism to achieving the required social learning is biased/heuristic learning. Conformity and prestige biases have been identified as two of the most important biased/heuristic learning strategies for norms.

Intrinsically motivated norms seem to be suitable candidates for explaining honesty. In an important sense, they are better explanatory options than norms associated with external sanctions because they could operate in contexts of epistemic privilege. In addition, intrinsically motivated norms can be understood as internal norms associated to an operative mechanism clearer than the mechanism that would explain the compliance to norms in the context of SMCT. The operative mechanism is the automatic activation in the mediation between a perception and a behavioral disposition.

But the lack of clarity in the operative mechanism was not the only problem with internalized norms in the context of the SMCT. The main problem is that epistemic privilege does not seem to create the right context for the success of the internalization process. Suppose that X is a norm that promotes honesty (or dishonesty). Therefore, X is a norm that has to operate under conditions of epistemic privilege. Once X has been internalized as an intrinsically motivated norm, its operation mechanisms can function smoothly. But how can X become an internalized norm? In order for people learn X through conformity or prestige bias, it is necessary that people can notice that most of their social group, and specially the most prestigious individuals, comply with X. And, this can occur in every kind of norm, except the ones that operate in contexts of epistemic privilege. At least in the model cases, where epistemic privilege is assured, it is not clear how biased learning mechanisms can be activated because the public behavior of people would be consistent with X and with ¬X.

Bicchieri et al. (2018) propose a different version of internalization through socialization that may be useful. In this approach, norms are *sluggish*: once a norm is established in a group, it will become sedimented, guiding behavior even in novel situations. People persist in complying with established norms. Keeping with the case stated above, even if people are not sure that other people actually comply with X, they may form the belief that X is the norm that the group openly recommends. And, this may be enough

to adopt X as a personal norm, used to guide behavior in different situations, just because this is the norm set by the group. Thus, people may learn that X is the norm that their group chooses as the correct norm in contexts of epistemic privilege. And, people will adopt X as a guide to behavior because that is how norms work in social groups.

As a final remark, the sluggish version of norms can be useful to explain the role of internalized norms in honest behavior. But this can be considered only an *ad hoc* explanation: sluggish internalized norms explain honesty because that is the way norms work. This seems to indicate that whether internalized norms, in this special sense, can be useful to explain honesty is a question that have to be empirically answered.

Notes

1. Mazar et al. (2008) claims that the task of summing up numbers to solve matrices is not seen by participants as a task showing any skill (not even a mathematical one). Even if this assumption could be controlled in an experimental task, we suppose that a fundamental problem remains: if the task involves a set of defined satisfaction conditions (i.e., if the task itself define what it means to perform correctly or incorrectly that task), the task implies that performances may be interpreted as mixing motives for being honest or dishonest with motives for presenting herself (to others or to him/her selves) as a performer who performs a task correctly or incorrectly.
2. At least motivated by social desirability bias that is clearly present because in interaction games the actual individual performances are openly known by experimenters and because sending false messages is, in the context of the experiment, causing a harm over the other participants since it implies they receive less money than the one they would receive if the sent message were true.
3. We suppose that, by definition, random games introduce an ecological problem in the study of honesty: epistemic privilege is about the results of a random process that, with the mediation of the report, are directly correlated with a (monetary) reward. But, in ordinary life, several cases of honesty and dishonesty arise in contexts where epistemic privilege is about the result of an *agencial process*, namely, cases in which outcomes are not random but in a relation of (causal) dependence with intentional actions that makes sense to notions as accountability or responsibility. In contrast, agents playing random games have no normative relation with the outcome, only with the report. Then, we may consider that random games configurate a situation structurally different from several real cases where honesty and dishonesty are present.
4. Related with these concerns about anonymity and its relations with epistemic privilege, we consider that in online experiments to study honesty, it would be necessary that records were not necessarily produced in the terminal.
5. We want to stress that this epistemic privilege is not referred to mental states, namely, it is not a cartesian first person epistemic privilege. It is just relative to knowing how oneself behaved in cases when no other being has the possibility of having such knowledge or when that other agent has such knowledge is highly unlikely. Then, it is an epistemic privilege over some public outcomes -behaviors-. Even if subjects do not have full control and awareness over all the cognitive and pure-physical processes

involved in producing their behavioral outcomes, they can certainly have full and, in some cases, privileged knowledge about these outcomes.

6. In his classic analysis, Hume (2014) establishes that truthfulness involves the report nature that we are recognizing in our analysis of honesty. Baier (2009) claims that Hume's approach of honesty as truthfulness lacks of a proper analysis of this report dimension since it is not clear if *brutal honesty* is a virtue: is telling to everyone all things that we know that are relevant to them a virtuous action, even if we know that some of them are hurtful or offensive? According with Baier, Hume's approach does not offer a definitive answer and it may include some kind of *hypocrisy* as virtuous when is based in social and interpersonal criteria, such as what is considered matter of private life or rude (Baier, 1984). Our conceptualization about honesty is a characterization of a type of action and it does not include a consideration about the conditions under which a particular expression of honesty can be considered virtuous as a feature of character.
7. It is possible speaking of norms that are not, at least directly, involved in action evaluation and regulation. For example, it is possible to consider that rules of formal logic are norms regulating (our use of/the nature of) the concept of truth (Frege, 1879), or that language rules are norms governing not just over linguistic structures but also over meaning and thinking (Dummett, 1981). In this text, we will use the concept of `norm' just referred to norms directly connected to human agency, providing guidelines both for action and for judging about action.
8. Some theories of norms propose that there are contexts where people are intrinsically motivated to follow norms (see for example Ben-Ner & Putterman, 1998; Gagn, 2007; Hofeditz et al., 2017; Crain and Krawiec, 2011).
9. Following Elster (2007), sanctions are cost but they should be seen as *more than cost* since pure cost does not integrate the sense of *wrongness* that is essential to understand what a sanction is. This is not a flaw in Rational Choice Theory since its goal is to explain why people decide some courses of action, understanding decisions just as maximizations of utility values, so, sanctions as cost represent a proper modeling of sanctions. Nevertheless, in theories where decisions are not just the result of maximizations but cognitive process involving meaningful contents it is necessary an account of sanctions where the sense of wrongness is integrated.
10. We also recognize that epistemic privilege involved in honesty is not necessarily absolute but gradual: agents may be more or less confident about how much their epistemic privilege is guaranteed in a given situation. The less the epistemic privilege is ensured, the less honesty is as a special predicate referring to an action: For that reason, we consider that the model case that requires explanation is the one in which epistemic privilege is ensured, from an intentional point of view.
11. To our knowledge, the only reference of the original proposers of SCMT to the failed replication of their findings presented by Verschuere et al. (2018) is introduced in Amir et al. (2018). Unfortunately, in that text they do not directly respond to the conclusions of the failed replication—since they adduce not knowing these results when redacting the text—but they just made some general considerations about the challenges of conducting replications of experiments on the relation between moral reminders and moral transgressions, which cannot count as a defense of SCMT.
12. Some recent evolutionary approach of morality (Henrich, 2015) acknowledges that, even if there are an innate machinery for morality, the particular connections between some acts and some evaluations have to be developed in ways that are sensible to the historical and social context where individuals have their own developmental trajectories.

Acknowledgements

We thank Diana Acosta Navas, Gino Carmona, Natalia Lara, Santiago Amaya, Tomás Barrero, and Nayib Medina for their insightful comments on a draft of the text.

Disclosure statement

There are no relevant financial or non-financial competing interests to report related to this work.

Funding

This work was financed by Colciencias with the doctoral scholarship awarded to Camilo Ordóñez through the program National Doctoral Grants 2017 and Universidad El Bosque under Grant PCI20179414. W.J.-L. was funded by the James S. McDonnell Foundation award 2020-1200.

ORCID

Camilo Ordóñez-Pinilla http://orcid.org/0000-0003-3000-6553
William Jiménez-Leal http://orcid.org/0000-0002-8824-5269

References

Amir, O., Mazar, N., & Ariely, D. (2018). Replicating the effect of the accessibility of moral standards on dishonesty: Authors' response to the replication attempt. *Advances in Methods and Practices in Psychological Science*, *1*(3), 318–320. https://doi.org/10.1177/2515245918769062

Arrow, K. J. (2012). *Social choice and individual values*. Yale University Press.

Azar, O. H., & Applebaum, M. (2020). Do children cheat to be honored? A natural experiment on dishonesty in a math competition. *Journal of Economic Behavior & Organization*, *169*, 143–157. https://doi.org/10.1016/j.jebo.2019.11.007

Badhwar, N. K. (1996). The limited unity of virtue. *Noûs*, *30*(3), 306. https://doi.org/10.2307/2216272

Baier, A. C. (1984). Hume's excellent hypocrites. *Rivista Di Storia Della Filosofia*, *62*(3), 267–286. https://doi.org/10.2307/44024016

Baier, A. C. (Ed.), (2009). Why honesty is a hard virtue. In *Reflections on how we live* (pp. 85–108). Oxford University Press.

Bandura, A. (2016). *Moral Disengagement*. Worth Publishers.

Barbosa, S., & Jiménez-Leal, W. (2020). Virtues disunited and the folk psychology of character. *Philosophical Psychology*, *33*(3), 332–350. https://doi.org/10.1080/09515089.2020.1719396

Becker, G. S. (1968). Crime and punishment: An economic approach. In *The economic dimensions of crime* (pp. 13–68). Palgrave Macmillan UK. https://doi.org/10.1007/978-1-349-62853-7_2

Ben-Ner, A., & Putterman, L. (1998). Values and institutions in economic analysis. In A. Ben-Ner & L. Putterman (Eds.), *Economics, values, and organization* (pp. 3–72). Cambridge University Press.

Bicchieri, C. (2006). The grammar of society: The nature and dynamics of social norms. In *The grammar of society: The nature and dynamics of social norms*. Cambridge University Press. https://doi.org/10.1017/CBO9780511616037

Bicchieri, C. (2015). Why do people do what they do? A social norms manual for Zimbabwe and Swaziland. *The multi country study on the drivers of violence affecting children*. https://repository.upenn.edu/pennsong/1

Bicchieri, C. (2017). *Norms in the wild: How to diagnose, measure, and change social norms*. Oxford University Press.

Bicchieri, C., Muldoon, R., & Sontuoso, A. (2018). Social Norms. In N. Z. Edward (Ed.), *The Stanford Encyclopedia of Philosophy*. https://plato.stanford.edu/archives/win2018/entries/social-norms/>

Bucciol, A., & Piovesan, M. (2011). Luck or cheating? A field experiment on honesty with children. *Journal of Economic Psychology*, *32*(1), 73–78. https://doi.org/10.1016/j.joep.2010.12.001

Cai, W., Huang, X., Wu, S., & Kou, Y. (2015). Dishonest behavior is not affected by an image of watching eyes. *Evolution and Human Behavior*, *36*(2), 110–116. https://doi.org/10.1016/j.evolhumbehav.2014.09.007

Cappelen, A. W., Sørensen, E. Ø., & Tungodden, B. (2013). When do we lie? *Journal of Economic Behavior & Organization*, *93*, 258–265. https://doi.org/10.1016/j.jebo.2013.03.037

Carson, T. L. (2010). *Lying and deception - theory and practice*. Oxford University Press.

Churchland, P. (2011). *Braintrust: What neuroscience tells us about morality*. Princeton University Press.

Cialdini, R. B., & Goldstein, N. J. (2004). Social influence: Compliance and conformity. *Annual Review of Psychology*, *55*(1), 591–621. https://doi.org/10.1146/annurev.psych.55.090902.142015

Cialdini, R. B., Reno, R. R., & Kallgren, C. A. (1990). A focus theory of normative conduct: Recycling the concept of norms to reduce littering in public places. *Journal of Personality and Social Psychology*, *58*(6), 1015–1026. https://doi.org/10.1037/0022-3514.58.6.1015

Cohen, S. (1985). *Visions of social control : Crime, punishment, and classification*. Polity Press.

Conty, L., George, N., & Hietanen, J. K. (2016). Watching eyes effects: When others meet the self. *Consciousness and Cognition*, *45*, 184–197. https://doi.org/10.1016/j.concog.2016.08.016

Crain, M., & Krawiec, K. (2011). Introduction: For Love or Money? Defining Relationships in Law and Life. *Journal of Law and Policy*, *35*(1), 1–9.

Da Cunha Fortes, A. C., & Thomé Ferreira, V. R. (2014). The influence of shame in social behavior. *Revista de Psicologia Da IMED*, *6*(1), 25–27. https://doi.org/10.18256/2175-5027/psico-imed.v6n1p25-27

Dimant, E., van Kleef, G. A., & Shalvi, S. (2020). Requiem for a Nudge: Framing effects in nudging honesty. *Journal of Economic Behavior & Organization*, *172*, 247–266. https://doi.org/10.1016/j.jebo.2020.02.015

Dummett, M. (1981). *Frege. Philosophy of language*. Harvard University Press.

Elias, N. (2000). *The civilizing process*. Blackwell Publishing.

Elster, J. (editor). (1986). *Rational choice*. New York University Press.

Elster, J. (1989). *Nuts and bolts for the social sciences*. Cambridge University Press.

Elster, J. (2007). *Explaining social behavior : More nuts and bolts for the social sciences*. Cambridge University Press.

Fehr, E., & Gächter, S. (2000). Cooperation and punishment in public goods experiments. *The American Economic Review*, *90*(4), 980–994. https://doi.org/10.1257/aer.90.4.980

Fehr, E., & Gächter, S. (2002). Altruistic punishment in humans. *Nature*, *415*(6868), 137–140. https://doi.org/10.1038/415137a

Fischbacher, U., & Föllmi-Heusi, F. (2013). Lies in disguise - an experimental study on cheating. *Journal of the European Economic Association*, *11*(3), 525–547. https://doi.org/10.1111/jeea.12014

Flanagan, O. (1991). *Varieties of moral personality*. Harvard University Press.

Foot, P. (1983). Moral realism and moral dilemma. *The Journal of Philosophy*, *80*(7), 379. https://doi.org/10.2307/2026455

Foucault, M. (1995). *Discipline and punish : The birth of the prison*. Vintage Books.

Frank, R. H. (1988). *Passions within reasons*. W.W. Norton & Company.

Frege, G.(1879). Logic. In H. Hermes, F. Kambartel, & F. Kaulbach (Eds.), *Posthumous writings* (pp. 1–8). Basil Blackwell.

Freud, S. (1961). *Civilization and its discontents* J. Strachey, (Ed.) W.W. Norton & Company.

Freud, S. (2019). *The ego and the id*. Clydesdale Press.

Fritz, J. H. (2020). Honesty as ethical communicative practice: A framework for analysis. In C. B. Miller & R. West (Eds.), *Integrity, honesty, and truth seeking* (pp. 127–152). Oxford University Press.

Gagn, L. (2007). Non-rational compliance with social norms: Sincere and hypocritical. *Social Science Information*, *46*(3), 445–469. https://doi.org/10.1177/0539018407079726

Gerlach, P., Teodorescu, K., & Hertwig, R. (2019). The truth about lies: A meta-analysis on dishonest behavior. *Psychological Bulletin*, *145*(1), 1–44. https://doi.org/10.1037/bul0000174

Gino, F., Krupka, E. L., & Weber, R. A. (2013). License to cheat: Voluntary regulation and ethical behavior. *Management Science*, *59*(10), 2187–2203. https://doi.org/10.1287/mnsc.1120.1699

Gneezy, U. (2005). Deception: The role of consequences. *The American Economic Review*, *95* (1), 384–394. https://doi.org/10.1257/0002828053828662

Haidt, J. (2012). *The righteous mind : Why good people are divided by politics and religion*. Pantheon Books.

Haugeland, J. (Ed.), (1998). Truth and rule-following. In *Having thought - essays in the metaphysics of mind* (pp. 305–361). Harvard University Press.

Helliwell, J., Layard, R., & Sachs, J. (2017). *WORLD HAPPINESS REPORT 2017*. https://s3.amazonaws.com/happiness-report/2017/HR17.pdf

Henrich, J. (2015). *The secret of our success: how culture is driving human evolution, domesticating our species, and making us smarter*. Princeton University Press.

Hietanen, J. O., Syrjämäki, A. H., Zilliacus, P. K., & Hietanen, J. K. (2018). Eye contact reduces lying. *Consciousness and Cognition*, *66*, 65–73. https://doi.org/10.1016/j.concog.2018.10.006

Hofeditz, M., Nienaber, A. M., Dysvik, A., & Schewe, G. (2017). "Want to" versus "Have to": Intrinsic and extrinsic motivators as predictors of compliance behavior intention. *Human Resource Management*, *56*(1), 25–49. https://doi.org/10.1002/hrm.21774

Hoffman, G., Forlizzi, J., Ayal, S., Steinfeld, A., Antanitis, J., Hochman, G., Hochendoner, E., & Finkenaur, J. (2015). Robot presence and human honesty: Experimental evidence. *ACM/IEEE International Conference on Human-Robot Interaction*, *2015*(March), 181–188. https://doi.org/10.1145/2696454.2696487

Hume, D. (2014). *A treatise of human nature* D. F. Norton & M. J. Norton, (Eds.) Clarendom Press.

Jansen, A. M., Giebels, E., van Rompay, T. J. L., & Junger, M. (2018). The influence of the presentation of camera surveillance on cheating and pro-social behavior. *Frontiers in Psychology*, *9*, 1937. https://doi.org/10.3389/fpsyg.2018.01937

Jiang, T. (2013). Cheating in mind games: The subtlety of rules matters. *Journal of Economic Behavior & Organization*, *93*, 328–336. https://doi.org/10.1016/j.jebo.2013.04.003

Kahneman, D. (2011). *Thinking, fast and slow*. Farrar, Straus and Giroux.

Kajackaite, A., & Gneezy, U. (2017). Incentives and cheating. *Games and Economic Behavior*, *102*, 433–444. https://doi.org/10.1016/j.geb.2017.01.015

Kallgren, C. A., Reno, R. R., & Cialdini, R. B. (2000). A focus theory of normative conduct: When norms do and do not affect behavior. *Personality & Social Psychology Bulletin*, *26* (8), 1002–1012. https://doi.org/10.1177/01461672002610009

Kant, I. (1996). *The metaphysics of morals*. Cambridge University Press.

Kelly, D. (2020). Internalized norms and intrinsic motivations: Are normative motivations psychologically primitive? *Emotion Researcher*, *1*, 36–45.

Kelly, D., & Davis, T. (2018). Social norms and human normative psychology. *Social Philosophy & Policy*, *35*(1), 54–76. https://doi.org/10.1017/S0265052518000122

Kelly, D., & Setman, S. (2020). *The psychology of normative cognition*. Stanford Encyclopedia of Philosophy. https://plato.stanford.edu/entries/psychology-normative-cognition/#BehaEcon

Klick, J., & Tabarrok, A. (2005). Using terror alert levels to estimate the effect of police on crime. *The Journal of Law & Economics*, *XLVIII*(Issue 1), 267–279. https://doi.org/10.1086/426877

Kuhns, J., Blevins, K., & Lee, S. Understanding decisions to burglarize from the offender ' s perspective. (2012). *Forensic Science, Medicine, and Pathology*, *8*(3), 243–251. Issue 2012. https://doi.org/10.13140/2.1.2664.4168

Manesi, Z., Van Lange, P. A. M., Van Doesum, N. J., & Pollet, T. V. (2018). What are the most powerful predictors of charitable giving to victims of typhoon Haiyan: Prosocial traits, socio-demographic variables, or eye cues? *Personality and Individual Differences*, *146*, 217–225. https://doi.org/10.1016/j.paid.2018.03.024

Markowitz, D. M., & Levine, T. R. (2020). It's the situation and your disposition: A test of two honesty hypotheses. *Social Psychological and Personality Science*, *194855061989897* (2), 213–224. https://doi.org/10.1177/1948550619898976

Mazar, N., & Aggarwal, P. (2011). Greasing the Palm. *Psychological Science*, *22*(7), 843–848. https://doi.org/10.1177/0956797611412389

Mazar, N., Amir, O., & Ariely, D. (2007). *Mostly Honest: A theory of self-concept maintenance*. (Unpublished Manuscript; Unpublished Manuscript).

Mazar, N., Amir, O., & Ariely, D. (2008). The dishonesty of honest people: A theory of self-concept maintenance. *Journal of Marketing Research*, *45*(6), 633–644. https://doi.org/10.1509/jmkr.45.6.633

Mazar, N., & Ariely, D. (2006). Dishonesty in everyday life and its policy implications. *Journal of Public Policy & Marketing*, *25*(1), 117–126. https://doi.org/10.1509/jppm.25.1.117

Mazar, N., & Hawkins, S. A. (2015). Choice architecture in conflicts of interest: Defaults as physical and psychological barriers to (dis)honesty. *Journal of Experimental Social Psychology*, *59*, 113–117. https://doi.org/10.1016/j.jesp.2015.04.004

Miller, C. B. (2017). Honesty. In W. Sinnott-Armstrong & C. B. Miller (Eds.), *Moral psychology* (Vol. 5, pp. 237–273). *Virtue and Character*. The MIT Press.

Miller, C. B., & West, R. (2020). Introduction. In C. B. Miller & R. West (Eds.), *Integrity, honesty, and truth seeking* (pp. xv–xl). Oxford University Press.

Mockus, A. (2002). Convivencia como armonización de ley, moral y cultura. *Revista Perspectivas*, *XXXII*(1), 19–37.

Nietzsche, F. (2011). *On the genealogy of morals* W. A. Kaufmann, (Ed.) New York Vintage Books.

Nussbaum, M. (1990). Aristotelian Social Democracy. In R. Douglass, G. Mara, & H. Richardson (Eds.), *Liberalism and the Good* (pp. 203–252). Routledge.

Oda, R., Kato, Y., & Hiraishi, K. (2015). The watching-eye effect on prosocial lying. *Evolutionary Psychology*, *13*(3), 147470491559495. https://doi.org/10.1177/1474704915594959

Olson, M. (1971). *The logic of collective action; public goods and the theory of groups*. Harvard University Press.

ONU. (2004). *UNITED NATIONS CONVENTION AGAINST CORRUPTION*. https://www.unodc.org/documents/brussels/UN_Convention_Against_Corruption.pdf

Ostrom, E. (1990). *Governing the commons : The evolution of institutions for collective action*. Cambridge University Press.

Pascual-Ezama, D., Fosgaard, T. R., Cardenas, J. C., Kujal, P., Veszteg, R., Gil-Gómez de Liaño, B., Gunia, B., Weichselbaumer, D., Hilken, K., Antinyan, A., Delnoij, J., Proestakis, A., Tira, M. D., Pratomo, Y., Jaber-López, T., & Brañas-Garza, P. (2015). Context-dependent cheating: Experimental evidence from 16 countries. *Journal of Economic Behavior & Organization*, *116*, 379–386. https://doi.org/10.1016/j.jebo.2015.04.020

Pfattheicher, S., Schindler, S., & Nockur, L. (2018). On the impact of Honesty-Humility and a cue of being watched on cheating behavior. *Journal of Economic Psychology*, *71*, 159–174. https://doi.org/10.1016/j.joep.2018.06.004

Potters, J., & Stoop, J. (2016). Do cheaters in the lab also cheat in the field? *European Economic Review*, *87*, 26–33. https://doi.org/10.1016/j.euroecorev.2016.03.004

Rawls, J. (1996). *Political liberalism*. Columbia University Press.

Rawls, J. (1999). *A theory of justice*. Harvard University Press.

Roberts, R. C., & West, R. (2020). The virtue of honesty: A conceptual exploration. In C. B. Miller & R. West (Eds.), *Integrity, honesty, and truth seeking* (pp. 97–126). Oxford University Press.

Sarat, A. (2005). *Crime and punishment : Perspectives from the humanities*. Elsevier JAI.

Smith, A. (1986). *The theory of moral sentiments* H. C. Recktenwald, (Ed.) Verl. Wirtschaft u. Finanzen.

Smith, T. (2003). The metaphysical case for honesty. *The Journal of Value Inquiry*, *37*(4), 517–531. https://doi.org/10.1023/B:INQU.0000019033.95049.1e

Smith, L., & Vásquez, J. (2015). Crime and vigilance. *Social Science Research Network Electronic Journal*. https://doi.org/10.2139/ssrn.2629321

Toby, J. (1957). Social disorganization and stake in conformity: Complementary factors in the predatory behavior of hoodlums. *The Journal of Criminal Law, Criminology, and Police Science*, *48*(1), 12. https://doi.org/10.2307/1140161

Ullman-Margalit, E. (2015). *The emergence of norms*. Oxford University Press.

Verschuere, B., Meijer, E. H., Jim, A., Hoogesteyn, K., Orthey, R., McCarthy, R. J., Skowronski, J. J., Acar, O. A., Aczel, B., Bakos, B. E., Barbosa, F., Baskin, E., Bègue, L., Ben-Shakhar, G., Birt, A. R., Blatz, L., Charman, S. D., Claesen, A., Clay, S. L., . . . Yıldız, E. (2018). Registered replication report on Mazar, Amir, and Ariely (2008). *Advances in Methods and Practices in Psychological Science*, *1*(3), 299–317. https://doi.org/10.1177/2515245918781032

Wilson, A. T. (2018). Honesty as a Virtue. *Metaphilosophy*, *49*(3), 262–280. https://doi.org/10.1111/meta.12303

Wolf, S. (2007). Moral psychology and the unity of the virtues. *Ratio*, *20*(2), 145–167. https://doi.org/10.1111/j.1467-9329.2007.00354.x

Yaniv, G., & Siniver, E. (2016). The (honest) truth about rational dishonesty. *Journal of Economic Psychology*, *53*, 131–140. https://doi.org/10.1016/j.joep.2016.01.002