The Deceiving Game[*]

Shlomo Cohen and Ro'i Zultan[†]

This version was accepted for publication in the *Journal of the American Philosophical Association*, DOI: 10.1017/apa.2020.33.

One of the central questions debated in the ethics of deception involves the moral comparison of different types of deception, based on the form (mode, venue) of communication employed. The three forms of deception that comprise (at least) the vast majority of deceptions are: (a) lying, or asserting falsehoods,[1] (b) falsely implicating, or communicating truths that in a given context will predictably cause false beliefs, and (c) nonverbal deception, or nonverbal action whose predicted interpretation is intended to create false beliefs.[2] The debate, then, is whether it makes a moral difference how one deceives, given the different forms of deception.

The leading position in this debate has been that lying is morally worse than the other forms of deception. This view has a rich cultural history, it was held by such great thinkers as Augustine, Aquinas, and Kant, and it continues to be the prominent view among contemporary philosophers (Chisholm and Feehan, 1977; Bok, 1989; Adler, 1997; Strudler, 2010; Webber, 2013; Shiffrin, 2014; Berstler, 2019). We shall refer to this position as the Classical View (CV). Justifications for CV include the idea that others only have a right to the truth vis-à-vis what one asserts, that lying to a person's face is more disrespectful and shameless, that lying entails a greater loss of credibility as communicator, and more. A second important position is that it makes no

[*] The authors wish to thank: Shaul Shalvi for his important role in initiating this research, Ori Regev, for providing great assistance in conducting Experiment 1, Uri Leibowitz and Max Hayward for very useful comments on an early draft, and John Doris for insightful suggestions in conversation. Various versions of the paper were presented at: DMEP-Ratio workshop, Jerusalem 2016; Buffalo Annual Experimental Philosophy Conference, University at Buffalo, 2016; Rocky Mountain Ethics Congress, University of Colorado – Boulder, 2017; Economic Science Association European meeting, Dijon 2019; Israel Organizational Behavior Conference, Tel Aviv 2020; London Judgment and Decision Making Seminar, 2020; ECONtribute workshop on social image and moral behavior, Cologne 2021. We thank participants in all these venues for helpful comments. We also thank two dedicated anonymous referees for the Journal of the American Philosophical Association.

[†] Ben-Gurion University of the Negev. Cohen: shlomoe@bgu.ac.il; Zultan: zultan@bgu.ac.il.

[1] Mahon (2016) articulates the traditional definition of lying as: "to make a believed-false statement to another person with the intention that the other person believe that statement to be true." There is a debate in the philosophical literature as to whether lying requires an intention to deceive. In this paper, by "lying," we mean "deceptive lying."

[2] "Nonverbal" is often used interchangeably with "nonlinguistic," but it can in effect be linguistic (when there is a robustly established meaning to the nonverbal gesture). This distinction is at times difficult to draw, but for this paper it is *inconsequential*.

difference morally *how* one deceives (Williams, 2002; Saul, 2012a,b). This can be readily understood as a particular application of the widely held moral intuition that what determines moral wrongness is some function of intention and consequences only, so that if the intent to deceive is the same and the result (the false belief created) is the same, moral evaluation must be similar. We will refer to this position as the Equivalence Thesis (ET). Jonathan Adler succinctly sums up the dilemma of the moral comparison between CV and ET: "From one angle, there is no moral difference. If you are going to mislead, just go ahead and lie. From another angle, the [non-lying] deceiver does manage to avoid a far worse wrong, even if his means are tainted" (1997, p. 446). Against the benchmark moral intuition that the precise manner of wronging someone is morally irrelevant as such, the aim of our paper is to check whether lying indeed constitutes "a far worse wrong."[3]

There has to date been no attempt to move forward in this debate using empirical methods. This may be anything but surprising: given that it is a *normative* debate between two moral positions, it would seem wrongheaded to attempt to solve it in the lab. Normative debates cannot be reduced to analysis in descriptive terms only (transgressing against this is known as the "Naturalistic Fallacy"). Nonetheless, this paper argues that empirical evidence can legitimately inform the normative debate in meaningful and even decisive ways. We support this argument by describing an experiment that we devised, and then explaining in detail the significant ways in which it allows to make headway in the normative debate.

We should emphasize at the outset that our project here is fundamentally different from that of experimental work on deception and dishonesty in the social sciences (see Abeler, Nosenzo, and Raymond, 2019; Gerlach, Teodorescu, and Hertwig, 2019, for meta-analyses on those). The objective there is, roughly, to detect existing norms against deception, and to examine the various factors that influence compliance with them. Our objective here, in contrast, is to adjudicate normatively between moral positions, i.e. to exercise moral judgment regarding the relative rightness or wrongness of the given positions. This is a completely different undertaking. Hence, the experimental results that we present are not in themselves our aim; they are input in the service of the normative analysis that follows, which constitutes the heart of our endeavor.[4]

Beyond the employment of empirical methods, our approach to the problem of comparing morally the forms of deception is innovative in a second methodological sense too. The growing field of experimental moral philosophy has not yet tapped into the resources offered by experimental economics,[5] as far as using them as tools for *normative* inferences. Methods

[3] The two central views in the debate, CV and ET, are not the only positions possible. Clea Rees (2014) has argued that falsely implicating (in her words: "merely deliberately misleading") is worse than lying, and other positions are also theoretically possible (e.g. that nonverbal deception is morally the worst). As will become clear, since the substantive argument in what follows will be the *rejection* of CV (not the adoption of any view), the existence of such (minor or theoretical) alternative positions will not affect our argument or conclusions.

[4] There is little literature in psychology that investigates the comparison of different forms of deception specifically (e.g. Rogers et al., 2017) but that literature does not engage in normative analysis as such. There is also little experimental philosophical work on people's usage of the relevant *concepts* (e.g. Weissman and Terkourafi, 2019), but that work too does not refer to our *normative* debate.

[5] The method of experimental economics creates a micro-economic system in laboratory conditions. By designing the way in which participants' decisions translate into actual payoffs, the experimenter is able to not only control the environment and the institutions of the system, but also induce preferences. See, e.g., Smith (1994).

of experimental economics have been used in experimental moral philosophy (e.g. Bicchieri, 2006) to investigate *descriptive/comparative* ethics, i.e. social norms, not, however, to investigate *normative/prescriptive* ethics, i.e. to establish moral rightness and wrongness. Skepticism about the prospects of common questionnaire-type methods of experimental philosophy to allow normative inferences about the moral dilemma regarding modes of deception led us to an alternative paradigm: we devised a strategic game to test this question experimentally. While questionnaire studies survey people's moral views, we constructed a situation where participants are incentivized to use deception, enabling us to learn about people's actual *behaviors*. As our discussion will make clear, this experimental approach allows to study the implicit empirical assumptions (”commitments”) of the different normative views; it consequently allows to proceed much further in gaining normative insights from empirical results.[6] Indeed, a second main objective of this paper—beyond contributing to the debate on the moral comparison of forms of deception—is to demonstrate with respect to a concrete problem, how far empirically-informed moral reasoning can advance, while respecting the logical constraint of the descriptive-normative gap.

1. Experimental Scheme – The *Deceiving Game*

The *Deceiving Game* is a strategic game that takes the form of a financial consulting interaction. One player, the investor, chooses how much out of an endowment of 100 points to invest in a virtual project. The project either pays a return of 250% on the investment, or loses the investment altogether. The other player—the consultant—has access to certain information about the odds of the project's success, and can advise the investor. The consultant receives remuneration equal to the latter's investment, and thus has an incentive to misrepresent information that conditions for investment are unfavorable.

We implemented the Deceiving Game as follows. A computer program selects one of two urns randomly to determine the outcome of the investment. The “blue” urn contains three blue and two orange balls, while the “orange” urn contains three orange and two blue balls. The selected urn represents the state of the world. The project succeeds if and only if the blue urn was chosen by the computer (a straightforward interpretation would be that the blue and orange urns represent bullish and bearish markets, respectively).

The consultant is shown three balls drawn from the chosen urn. These always include one blue and one orange ball, with the third ball randomly picked from the remaining three balls. A consultant who applies Bayes' rule after observing two blue and one orange balls assigns a

[6] To wit, since normative positions typically aim to be prescriptive to creatures like us, they ought to presuppose some understandings of how, in fact, we can and do operate; these presuppositions (or a subgroup of these) are their “empirical commitments.” The following passage, describes this lucidly with regard to the question of moral motivation: “accounts of moral motivation typically presuppose commitments regarding the nature of psychological states such as beliefs, desires, choices, emotions, and so on, together with commitments regarding the functional and causal roles they play.

Observations about the nature and the functional and causal roles of psychological states, it seems to us, are as much empirical as they are philosophical. At least, it is rather obscure how such claims are to be understood, if they are not to be understood as involving substantial empirical elements.” (Schroeder, Roskies, and Nichols, 2010, pp. 78–79) probability of two thirds to the chosen urn being the blue urn, and similarly for the orange urn when observing two orange and one blue ball. Since the consultant's payoff equals the amount invested by the investor, the consultant has an incentive to persuade the investor that she observed two blue and one orange balls, even if she didn't.

Next, the consultant must choose one of two communication options. These vary across three experimental conditions, corresponding to the three modes of deception outlined above:

- In the *Lies* (henceforth *LY*) condition, the consultant chooses whether to send to the investor the message "I saw two blue balls" or "I saw two orange balls."
- In the *Falsely Implicating* (*FI*) condition, the consultant chooses between sending "I saw blue" or "I saw orange." Because (per the game's design) the consultant always observes at least one blue and one orange ball, both messages are always literally true. The former, however, implicates that a majority of blue balls was observed.
- In the *Nonverbal Deception* (*ND*) condition, the consultant chooses whether to place a small bet of 5 points, which pays 10 points if and only if the chosen urn is the blue urn, or not to place a bet. This choice is (known to be) revealed to the investor, who may be expected to draw conclusions accordingly.

We shall refer to the first option in each condition, which communicates having observed a majority of blue balls, as a BLUE message, and to the second option as an ORANGE message.

A general methodological note is in order. In the Deceiving Game, people choose *whether* to deceive but not *how* to deceive, as only one form of deception in available to each participant. Although in real life people are typically free to choose how to deceive, we believe nonetheless that our design provides the cleanest and most direct comparison between the three forms of deception. In the alternative, "choice of form of deception" paradigm, the choice of one form depends not only on the attractiveness of that form, but also on that of the available alternatives. In our design, in contrast, each form is judged independently of the others without cross-contamination of preferences. This independent measurement allows *quantifying* the willingness to deceive in each form, whereas a choice between forms only stands to inform which form is more attractive (even if merely infinitesimally so). Thus in the latter design we risk losing important information. Moreover, the "choice of form of deception" paradigm is susceptible to demand characteristics. That is, if participants were asked to explicitly choose between forms, their answers would have likely been influenced by what they consider the experimenters to expect from them, thus biasing our results. In a choice whether to deceive, in comparison, we avoid this most central form of bias; decisions then better represent intrinsic preferences regarding form of deception.

We now describe two independent experiments we conducted that studied behavior in the Deceiving Game, and their results. First, however, we should articulate our experimental hypotheses. In Section 3, below, we present and discuss various empirically-testable conditions, whose existence would support CV (they are "empirical commitments" of CV). If these conditions in fact hold, then we would expect: (1) less lying, compared to other forms of deception, and (2) more trust placed in assertions (which potentially can be lies) than in other forms of communication (which potentially can be non-lying deceptions). We operationalize the two expectations in the following hypotheses:

Hypothesis 1: The percentage of choices to deceive by consultants is lower in *LY* than in *FI* and *ND*.

Hypothesis 2: The mean difference in investment between receiving a BLUE message and an ORANGE message is higher in *LY* than in *FI* and *ND*.

If these hypotheses are confirmed by the experimental results, they provide grounding for normative arguments in favor of CV, as we will explain. Conversely, if these hypotheses are not confirmed by experimental results, then this means the conditions that constitute CV's empirical commitments do not in fact hold, and then support for CV is undermined.

We should emphasize the following regarding *Hypothesis 2*. *Hypothesis 2* expresses the thought that people assign a higher probability to the message (or the inferred state) being true when the message, if deceptive, is a lie. In line with classical decision theory, best developed and articulated by Leonard Savage (1954), we interpret "assigning a higher probability" as a higher willingness to bet on the outcome with which the probability in question is associated.[7]

2. Experiments and Results

2.1. Experiment 1

The first experiment was conducted as a classroom experiment among 139 economics students (78 females and 61 males, mean age 24) at Ben-Gurion University of the Negev. Subjects were randomized into the three experimental conditions, with each subject participating in one condition. The experimenter entered the classroom towards the end of the class and offered a chance to participate in a short experiment for money. The students decided whether to participate before learning the details of the experiment. The experimenter handed out the written instructions for the consultant role, and explained the structure of the game (without referring to the content of the message, as different participants received instructions for different conditions; the instructions referred to "sender" and "receiver" rather than "consultant" and "investor" they did not mention deception or any other morally loaded terms). See Appendix 1 for the complete translation.

All participants answered comprehension questions, and made two decisions in the role of the consultant, conditional on observing a majority of blue or a majority of orange balls. For the second part of the experiment, the instructions simply indicated that the roles are reversed, and instructed the participants to make two more decisions, now in the role of the investor, conditional on receiving a BLUE or an ORANGE message. After collecting all decision forms, we randomly assigned the participants in pairs of consultant and investor to calculate payoffs. Payoffs were stated in New Israeli Shekels (NIS), and paid out in class a week after the experiment took place; we contacted participants who were not in attendance to arrange payment separately.

[7] Simply put, if people prefer a gamble that will give them a desirable outcome on event A over a gamble that will give them the same outcome on event B, we say that they assign a higher probability to event A than to event B. The application to the Deceiving Game is straightforward, where the events in questions are the conditional events "The blue urn was chosen (by the computer), given the consultant's action X," with the content of X varying across conditions.

This experimental design allowed us to measure two central variables of interest. First, the tendency to deceive in each of the conditions was measured as the proportion of consultants who sent the BLUE message after observing a majority of orange balls. Second, the level of trust of the investors in the messages sent by the consultants was measured as the increase in their investments after receiving the BLUE message compared to their investment after receiving the ORANGE message.

2.1.1. Results of Experiment 1

Were consultants less likely to deceive when deception involved a lie, thus providing support for CV? The left panel of Figure 1 presents the proportion of consultants who chose to send the BLUE message when observing one blue and two orange balls. We see that deception rates are, if at all, higher in the *LY* condition, with 31 of 48 (64.6%) participants choosing to deceive compared to 24 of 47 (51.1%) and 25 of 44 (56.8%) in the *FI* and *ND* conditions, respectively. The differences between the three conditions are not significant ($\chi^2(2) = 1.79, p = 0.408$). The proportion of participants who choose to deceive in the *LY* condition is 10.7 percentage points higher than in *FI* and *ND* combined, with a 95% confidence interval of $[-27.7, 6.2]$ (Koopman, 1984). That is, we can significantly reject the hypothesis that deception rates in the *LY* condition are lower than in the other two conditions by 6.3 percentage points or more.

The right panel of Figure 1 presents the investors' reactions to the consultants' choices. That is, the mean difference in investment between observing a BLUE and an ORANGE message. This was almost identical in *LY* and *ND* (30.5 points and 30.7 points, respectively) and slightly and non-significantly higher in *FI* (39.3 points; $F(2, 136) = 0.76, p = 0.468, \eta^2 = 0.011, \omega^2 = 0.00$ for the one-way ANOVA). A non-negligible share of investments are left (/right) censored following an ORANGE (/BLUE) message (25.9% and 32.37%, respectively). A tobit regression on investment on condition and message and their interaction, censoring at 0 and 100 yields essentially identical results.

2.1.2. Establishing equivalence

The lack of significant evidence in support of CV is not sufficient, in itself, to reject the hypotheses underlying CV, as there can always be a small and undetectable effect in the hypothesized direction. Nonetheless, we can test whether—if such effects exist—they are of negligible magnitude at best. For our main analyses we conduct inferiority tests.[8] That is, we test the null hypothesis that the effect size, as measured by Cohen's d for the difference between LY and the other two conditions in the predicted direction, is larger than a minimal benchmark, which we set based on effect sizes observed in the relevant psychological literature (Cohen, 1988). In Appendix 2, we report the full details of the analysis, including results for a more conservative benchmark.

[8] Inferiority tests are the one-sided version of equivalence tests (Lakens, Scheel, and Isager, 2018), and are appropriate when the question of interest is whether there is an effect in a predicted direction (cf. Rothmann, Wiens, and Chan, 2011). Indeed, in this experiment we cannot reject with confidence the hypothesis that consultants deceive *more* in *LY*, however since we are testing the arguments in favor of CV, this does not reflect on our conclusions.

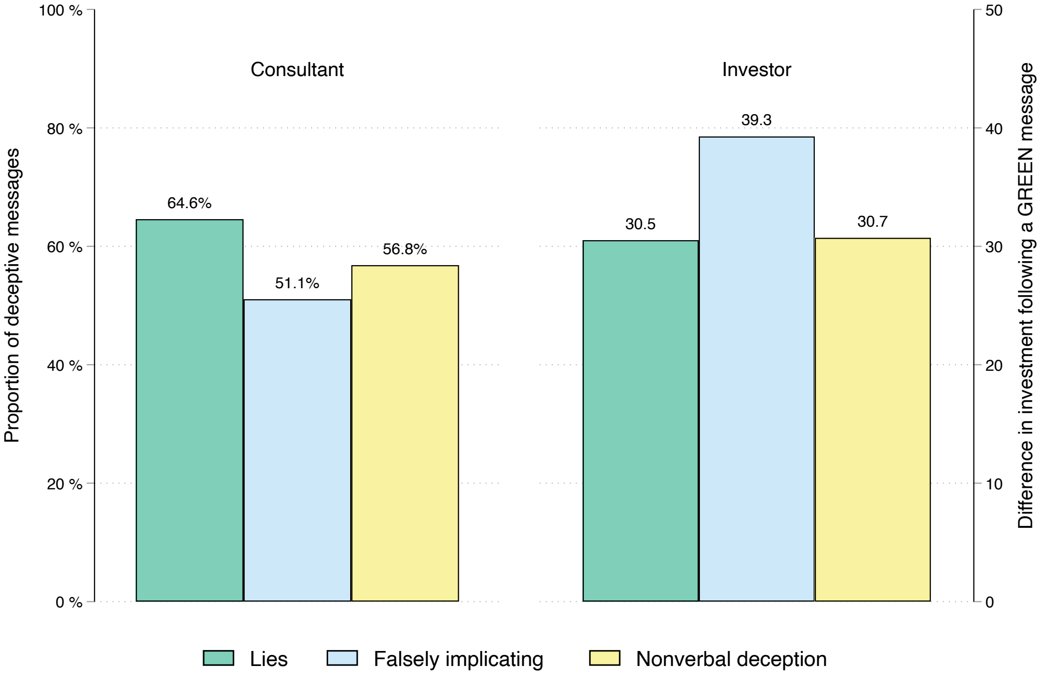

Figure 1: Deception and trust in Experiment 1.

The results of the inferiority tests are significant, at $p < .001$ for consultants; and $p = .001$ for investors. Accordingly, we conclude that the mode of deception had no significant effect either on deception or on trust in the direction supporting CV. The power to detect the benchmark effect sizes (for both consultants and investors) given our benchmark is $1 - \beta = .781$.

2.2. Experiment 2

We ran the second experiment as a laboratory experiment. In addition to allowing us to corroborate the results of the classroom experiment, a laboratory experiment has several advantages. The laboratory setting provided ample time (approximately 75 minutes per session) for guaranteeing participants' understanding of the instructions (thanks to detailed explanation, answering clarification questions, and testing understanding in control questions). Each participant played repeatedly ten times in each role, increasing the statistical power and providing further opportunity for learning to take place and for testing whether experience alters behavior in the game.

The basic game was the Deceiving Game described above. At the beginning of the experiment, participants were randomly allocated to roles of consultant and investor. In each round, the computer randomly (re)matched participants in pairs of consultant and investor within matching groups of eight participants (a standard practice in experimental economics aimed to ensure statistical independence between matching groups(. The consultant saw three balls

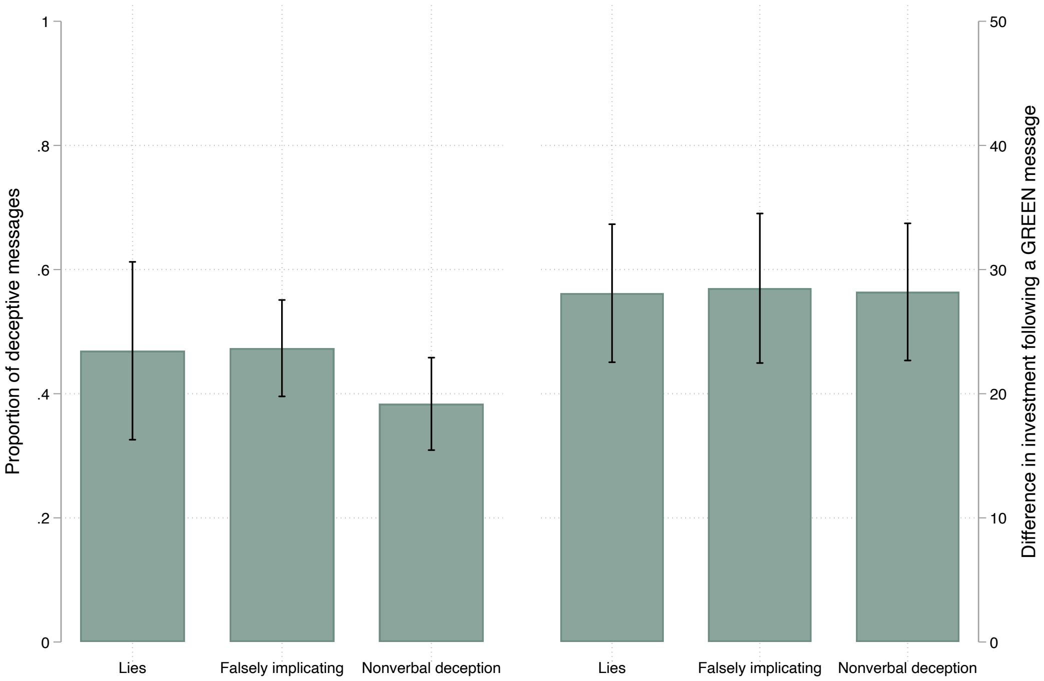

Figure 2: Deception and trust in Experiment 2.

and chose a message by clicking on the message (or betting option) presented on the screen. Next, the matched investor was informed of the consultant's action and chose an investment. At the end of the round, both participants received feedback regarding the chosen urn, the consultant's message, the investment, and their round payoffs. After ten rounds, the roles were reversed for an additional ten rounds. The final payoff in points was the participant's total earnings in five randomly selected rounds out of the 20 rounds. The payoff was converted to NIS at a conversion rate of 100 points = 10 NIS and added to a 15 NIS base fee. The average final payoff was 55.60 NIS (approximately 16 USD). A total of 168 participants were recruited using ORSEE (Greiner, 2015); the experiment was programmed in zTree (Fischbacher, 2007).

2.2.1. Results of Experiment 2

Figure 2 presents the results, with 95% confidence intervals based on a mixed-effects linear regressions with random effects for participants and robust standard errors clustered on matching groups. The left panel presents the estimates for the proportion of deceptive choices by condition. The right panel presents the estimates for the marginal effect of the message on investment by condition (i.e., the difference in mean investments depending on receiving an ORANGE or a BLUE[9] message as predicted by the regression model). *The results corroborate the findings in Experiment 1, with no apparent condition effects.* Willingness to deceive and trust in the message are practically identical in *LY* and *FI*. Willingness to deceive is lower in the *ND*

[9] The actual color in Experiment 2 was green rather than blue. We continue to refer to the "high" signal as BLUE for consistency.

condition, though the difference is not statistically significant. (This difference may be due to the small cost of deception incurred in this condition.)

As in experiment 1, we conducted inferiority tests of the null hypothesis that there is no non-trivial effect in the direction supporting CV. The inferiority test yields highly significant results of $p = .002$ for consultants and $p = .001$ for investors. Power is $1 - \beta = .962$ for consultants and $1 - \beta = .993$ for investors. (Again, see Appendix 2 for details.)

2.3. Conclusions from the two experiments

Two independent experiments, using different subject pools and protocols, comprised of 307 participants who made, in total, $1,819$ decisions in each role yielded similar behavior in the different conditions of the Deceiving Game. In particular, we find (a) that people are not less likely to deceive when the only way to do so involves explicit lies, i.e. we can reject *Hypothesis 1*; and (b) that people are not more trusting in explicit messages that, if deceptive, are outright lies, i.e. we can reject *Hypothesis 2* (In the sense that we can reject the hypotheses of meaningful differences between *LY* and the other two conditions).

3. Normative Insights

We now move to the central task of extracting valid normative conclusions from our experiments. (Since both experiments yielded relevantly similar results, the discussion conveniently applies to both.) Prior to this, we should mention that we can also draw traditional conclusions, in terms of *descriptive* ethics: since we did not find less lying or more trusting behavior in the lying condition than in the other two, we can conclude that CV does not represent folk moral commitments (rather, ET seems to reflect them best). While this is an interesting result, the focus of our analysis here is different—it is to draw *normative* conclusions. The general idea is to identify the "empirical commitments"—i.e. the empirically testable elements—of the moral principles that are assumed in this debate, to show how these are addressed by our experiments, and to analyze the normative import of the results.

In one trivial sense, descriptive observations are always relevant to moral judgment, viz. in determining whether a moral principle at all applies to the given situation (e.g., trivially, the application of "murder is wrong" is only relevant when (roughly) one person intentionally killing another is the issue at hand). In our experiment, however, our task is to adjudicate between two normative positions, i.e. *to judge which view is more correct, morally speaking*; the appeal to empirical facts is hence *prima facie* suspicious, and therefore interesting.

3.1. Conclusions from Equivalence among Investors across Conditions

In this section, we first present the inference from the experimental finding of equivalence among investors across the three conditions to the normative conclusion; then we discuss methodological and theoretical assumptions that support this inference, i.e. a set of insights which, taken together, entail that our results indeed justify deriving moral judgment regarding the relative wrongness among the three modes of deception.

The first argument for CV that we consider is that people trust the veracity of assertions more than of other forms of communication, and that therefore lying amounts to a greater betrayal of trust (and is for that reason morally worse).

The similar increases in investments in *LY*, *FI*, and *ND* following a BLUE (compared to an ORANGE) signal can be naturally taken to show similar levels of trust/mistrust in the consultants across the conditions. The investors understood perfectly well that the consultants have an incentive to deceive and are therefore more or less likely to do so; they also understood by what means deceptions would be carried out. If the investors' baseline level of trust in the propensity to truthfulness of the consultants and in the veracity of the messages sent by them had been lower in one condition compared to another, they would have been correspondingly more averse to take the risk involved in investing (i.e. losing the entire investment), and would have invested on average less. To the extent that the levels of trust of the deceived are similar across the conditions—to which the experimental results indeed testify (see discussion below)—the breach of their trust by the different kinds of deceptions is of similar magnitude. Now, since the wrongness of deception is, *ex hypothesi*, a function of betrayal of trust (betrayal of trust is a wrong-making feature of deception), and since the design of our experiment allows us to compare betrayals of trust across the three modes of deception, then having found similar levels of betrayal of trust, we are allowed to draw a normative conclusion: our results, which seem to align with ET (similar wrongness across forms of deception), undermine support for CV. The pivotal idea here is that since the wrong-making feature is a function of a psychological state (of the investors), it *can* be assessed empirically.

A general remark about the formal nature of our conclusions is in order. Failing to support a reason for X (CV, in our case) is different, logically speaking, from providing a reason against X. That being granted, we should also stress that when X competes against rival positions, undermining X can *de facto* testify against it, by making it inferior to existing alternative explanations. This point is especially relevant in our case, where the default position onto which we fall back when failing to support CV is prima facie in line with ET. In addition, systematically undermining plausible reasons for X *can* amount to a reason against X, in the sense of rendering X exceedingly implausible (as long as some further reasonable hypothesis is not put forth).

We now turn to discuss theoretical underpinnings of our inference from experimental results to normative reason. First, we will clarify a basic methodological point; then, we will explain the theoretical grounds for our analysis in terms of trust.

Notice, importantly, that we are not attempting to derive from the empirical results an answer to the question of whether, as a rule, betraying (justified) trust is morally wrong. We rather accept and presuppose the validity of the—hardly controversial—moral judgment "betraying (justified) trust is morally wrong," but then focus on the empirical dimension that betraying trust has, and attempt to assess and analyze its contribution to solving the normative debate as to which form of deception is morally worse. In other words, while the debate whether lying is or isn't a greater moral wrong is indeed a normative one, our approach is to identify a ground-level moral principle that underlies this debate, identify the empirical (psychological) dimensions involved in observing that principle, show how our experiment can measure those empirical elements, and use this to arrive at a moral verdict. Since the intuitive ground-level moral principle is presupposed to be true, deriving a normative conclusion from the empirical investigation does not ultimately transgress against the Naturalistic Fallacy.

The possibility of drawing normative conclusions that we describe is not unique (which might have made it suspicious). A moral agent has, for instance, moral reason to avoid greater rather than lesser harming of others; yet what in fact constitutes greater or lesser harming is arguably determined (at least partly) by the psychology of people, i.e. by what they experience as a greater or lesser drawback to their interests or welfare. In parallel, there is moral reason to avoid greater rather than lesser betrayal of trust, yet that which in fact constitutes greater or lesser betrayal of trust is determined (at least partly) by the psychology of trusting. Since our experiment operationalizes this psychological attitude, it can legitimately inform the normative debate between CV and ET, without committing a Naturalistic Fallacy. (Empirical psychology can validly inform normative debates in more ways, not restricted to the "greater than" form. For instance, when debating between two actions to perform, there is moral reason to prioritize the action that is one's duty over that which is over-demanding and hence supererogatory; yet what counts as over-demanding is determined, at least partly, by the psychology of moral agents, i.e. by what in fact compromises agents' basic interests or adversely affect their welfare to an unreasonable degree.)

One might worry that, since "betraying trust" can mean different things, our use of it might be conceptually untidy. We therefore turn now to clarify the theoretical grounds, and consequently the validity, of our use of the notions of trust and betrayal of trust.

(a) It is widely accepted that trust is never placed on someone non-specifically but always with respect to some particular kind of performance. As Russell Hardin (2002, p. 9) put it, trust is "a three-part relation: A trusts B to do X." This then is also the case with respect to the investors in our experiment; and in the context of the extremely specific interaction they have with the consultants, it seems prima facie clear that their trust, to the extent it exists, refers to the expectation that the consultants not deceive them.

(b) All deception necessarily involves betrayal of trust (in some sense; correspondingly, the very possibility of deception presupposes a background of trust). This underlies the cogency of the comparison we draw among the three forms (modes) of deception. Let us examine this. It has been argued (e.g. Chisholm and Feehan, 1977) that assertions constitute a unique "invitation to trust"; but even if that is true, it does not follow that betrayal of trust is restricted to lying. As Bernard Williams put it: "Truthfulness is a form of trustworthiness, that which relates in a particular way to speech." He stresses: "Trustworthiness is more than the avoidance of lying." This is so since asserting is but a restricted part of speech. "There may be special circumstances in which it is understood that a hearer is to ignore everything about an assertion except its content, but they are very special. In general, in relying on what someone said, one inevitably relies on more than what he *said*." (Williams, 2002, pp. 94, 97, 100). Trusting ("relying on") others to be truthful forms part of the bedrock of human communication in all its linguistic manifestations; hence, deceiving, by any means, involves betrayal of trust. (A closely related intuition on the non-uniqueness of lying with respect to conversational trust is found in Saul, 2012b, pp. 75–79). These intuitions receive systematic support from Paul Grice's theory of language, according to which linguistic exchange relies on the assumption that interlocutors are (usually, to some degree) engaged in a cooperative enterprise. Hence typical linguistic exchange presupposes the Cooperative Principle, "Make your conversational contribution such as is required, at the stage at which it occurs, by the accepted purpose or direction of the talk exchange in which you are engaged."(Grice, 1989, p. 26) This refers to conversation in all its linguistic aspects—to the *way* things are said, not merely to *what* is said. Since the assumption of cooperation is an assumption of trustworthiness, trust refers to all aspects of linguistic expression. In parallel, deception via all our three forms constitutes a betrayal of trust.

According to Collin O'Neil, although a special invitation to trust is associated with some forms of deliberate communication and not others (and deception via these forms "misuses" and "abuses" trust), deception by any form of communication "consists in *failing to perform as one is trusted to perform*," and the wrong associated with this is "betrayal of trust"; indeed, "trust need not be invited for a betrayal of trust to occur." (O'neil, 2012, pp. 306, 318) Hence, deception via all three forms betrays trust. Lastly, the intuitions above regarding trust are corroborated and enhanced by analysis in terms of "warrant of truth" (Carson, 2010). Deception is not even possible in theatre play performance, in "bull-sessions," etc. since in such interactions there is no presumption that truth is warranted. Conversely, if truth *is* warranted, then deception betrays trust, regardless of the form of deception.

(c) Having established that the forms of deception are in principle comparable in terms of betrayal of trust, we can now address a skeptical challenge to the effect that our experimental method lacks sufficient information to render a moral verdict. The idea is that there may be different senses of (betrayal of) trust in play, and that each may have different moral weight. For instance, assuming that O'Neil's analysis (above) is right, the experimental results may fail to distinguish between the moral effects of "abusing trust" versus of "betraying trust." In response, the beautiful thing about the experimental setup is that it is not vulnerable to this potential difficulty. The various senses of trust (corresponding to the various senses of breaching trust), whatever they may be, de facto converge to an all-things-considered level of trust that is expressed in the bottom-line readiness to count on the consultant's word and stake an investment despite the risk (losing the entire investment). The crucial point is that the wrongness of betrayal of trust correlates directly with this all-things-considered position of trust—a bottom-line position of *making oneself vulnerable to the other*—which expresses the aggregate level of trusting (whatever its internal breakdown), and whose behavioral expression our experiment is constructed to measure! In other words, the wrongness of betrayal of trust is the *wrongness* of exploiting the position of vulnerability that the trusting other agrees to put herself in vis-à-vis the agent—and *that* vulnerability is precisely embodied in agreeing to take the risk of investing money that will be entirely lost, if the consultant decides to deceive.[10]

(There are additional dimensions of trust and therefore betrayal of trust: notably, those relating to the kind of relationship one has with others (e.g. betraying friends rather than strangers normally involves greater betrayal of trust), and to the type of scenario people are in (e.g. deceiving under oath normally involves greater betrayal of trust). Those dimensions are excluded from our experimental scheme, which focuses exclusively on comparing *forms* of deception. But even if they weren't, the randomization of participants would have an equalizing effects on them, and then it would create no problem if they too entered into the all-things-considered aggregate trust that is experimentally measured.)

(d) Based on the considerations above, we conclude, our experimental results fail to support an important *normative* reason for CV. Previous work has rarely used experimental findings to

[10] An even *stronger* position on this issue would be that the bottom-line (betting) behavior (is not merely correlated with the wrongness of betraying trust, but rather) constitutes the very *meaning* of "trust" (in the truthfulness of reporting). This could express venerable philosophical views such as Gilbert Ryle's, or Ludwig Wittgenstein's "meaning in use."

adjudicate normatively between rival moral positions, and none we know of has done so via methods of experimental economics. Having reached a normative conclusion, we achieved all we can hope for from experimental moral philosophy; and yet, we may inquire further about the *relative weight* of that normative conclusion in an overall comparison of the three modes of deception. Comparing the modes of deception as such, we *ex hypothesi* keep intentions and results constant, and focus on candidates for being the intrinsic wrong-making features of deception (i.e. the grounds of the judgment "the deceptive act as such is wrong"). When we do that, we find betrayal of trust as arguably the prominent candidate. We cannot argue for this view in this space, only mention that it seems much in the spirit of views such as Bernard Williams' ("Truthfulness is a form of trustworthiness…"). Other candidates admittedly exist (though not many)—for instance: manipulativeness, or an aesthetic flaw with ethical dimensions (Pepp, 2019). We need not (and cannot) review all theoretical possibilities here;[11] what is important to emphasize, however, is that *to the extent* that philosophical reflection finds betrayal of trust as the only or most important intrinsic wrong-making feature of deception, then our experimental results produced not merely *a* moral reason, but *the* moral verdict on the dilemma we are investigating.

3.2. Conclusions from Equivalence among Consultants across Conditions

The experimental results show equivalence in the consultants' rate of deceiving across *LY*, *FI*, and *ND*. While this means that actual behavior does not follow CV, this piece of (merely) descriptive ethics is *not* what we are here after. Our interest is very different: it is in whether the equivalence in the consultants' rate of deceiving across the conditions can undermine *normative* (moral) reasons for CV (and thus help adjudicate morally in the debate between CV and ET). The answer depends on the specific grounds for adopting CV. Below we discuss three salient possible grounds, and explain with respect to each how it can support drawing normative conclusions from our results.

(A) The first ground for CV is that lying is worse per convention. That norms relating to truthfulness change geographically and historically seems well-established (e.g. Gächter and Schulz, 2016; Hugh-Jones, 2016); the question of whether lying is worse or not may well be one aspect of that phenomenon. If it is, then all there is to say on the matter should be fully discoverable by experimental observation. That, in itself, would not further our purposes, however, since if experimental results of levels of deception are attributable to acting according to convention (descriptive ethics), then this *cannot* by itself satisfy the challenge of judging the soundness of moral reasons. Mere convention cannot command true normative (prescriptive) authority. Yet, while a "mere" convention indeed does not furnish normativity, the following scheme can do so: (a) a plurality of relevant considerations fails to converge on one rational bottom line moral conclusion; (b) the conventional principle based on the summation of the relevant considerations, which is as justifiable as other alternatives, is reflectively endorsed as providing the obligating norm. Once a norm has undergone such a process, there can be strong moral reasons of fairness to act accordingly, i.e. for one to reciprocate by doing one's fair share

[11] Our discussion in the next section will cover additional possibilities.

for the success of the social enterprise.[12] Now it is plausible that such an account indeed holds true for the question of assessing CV versus ET. If it does, then our experimental results *can* yield a normative reason.

The specific story in our case could for instance be described roughly along the following lines.[13] Lying demands less preparation, effort, and imaginativeness than other, craftier deceptions; it can therefore be more easily and readily produced, and for that reason poses a greater threat to social cooperation. On the other hand, a lie is a less deniable form of deceiving, and is as such more vulnerable to exposition and hence *less* of a social threat. From a different perspective, lying is worse, since the success of other forms of deception depends on the inferences others make, which shifts part of the responsibility away from the non-lying deceiver. On the other hand, the lie is not worse, as it is at least a more "authentic" way of concealing the truth, without resorting to treacherous techniques that implicate others in their own deception. From yet another perspective, lying expresses a worse attitude toward truthfulness, as the evasive quality of all other forms of deception is the result of maneuvers aimed at not lying—thus, ironically, those other forms of deception confirm the value of truthfulness. On the other hand, since the lie needs less preparation (as mentioned above), it can be more mindlessly executed, and so offers less of a testimony to lack of respect for truthfulness. And so on. Now it is entirely sensible to argue that there can be no way to sum up these various opposing considerations reliably into one rational objective conclusion, and that we therefore normatively endorse the prevailing social norm that expresses a holistic sensibility about this issue, whatever it happens to be. Jonathan Adler, in a similar vein, speaks in this context of a progression from social norm to ethical norm. After claiming that "a norm corresponding to the lessened demands of truthfulness for implicatures would be desirable for all," Adler hastens to add: "Such a norm of conversation acquires moral force" (Adler, 1997, p. 451). If such (or sufficiently similar) is the ground offered for CV, then the equivalence in the consultants' rate of deceiving across our experimental conditions undermines a normative reason for CV.

(B) A second ground for CV is that lying is worse because it reveals a deeper antisocial attitude and as such is more sinister an expression of moral character and motivation. According to this view, it is psychologically more difficult for a decent person to lie than to deceive in other ways.[14] Normal upbringing includes a long process of conditioning to not say what is false; this results in greater psychological difficulty to utter falsehoods compared with uttering truths, and this holds even when those truths are in the service of deception. Now normally and other things being equal, for a person to overcome greater inhibitions in order to commit a wrong suggests greater *malice* and to that extent exhibits greater deficiency of moral goodness and worth. In the terminology of economics, we would say that the psychological cost of lying is greater compared to non-lying deception, and hence that *ceteris paribus* lying testifies to a stronger motivation to

[12] The non-reciprocator may not only be guilty of free-riding on others but may risk harming them too—think e.g. of disrespecting the admittedly arbitrary norm of driving on the right-hand side of the road.

[13] The following arguments, extracted from the literature on deception, are cited here to demonstrate the *de facto* plurality of non-converging views; it is not our intention or indeed business in this context to argue for any of them over any other.

[14] The conventional and psychological grounds are not mutually exclusive, but they are different. The convention may be a direct function of social value or utility that is irreducible to individual psychology.

deceive. This insight yields (one interpretation of) the CV.[15] The source of deeper inhibitions ultimately lies in the social psychology of communication. While non-lying deceptions are rather evasive ways of misleading others, which otherwise decent people may adopt in delicate social predicaments or under duress, outright lying is a more daring interpersonal position that requires a more shameless disposition. "The liar is more brazen," observes Jonathan Adler (Adler, 1997, p. 442). This explains why many who find themselves unable to utter falsehoods in someone's face intentionally, resort to saying misleading truths or to performing various nonlinguistic actions intended to mislead. Basic competence in human communicative norms makes people acutely sensitive to the fact that uttering falsehoods in front of others is a more flagrant and jarring disruption of human communicative expectations, lies inspire a special "sense of violation or outrage" (Frankfurt, 2009, p. 50); they consequently encounter deeper inhibitions to lie. Again, the upshot is that lying testifies to a looser moral stance regarding deception. In other words: although lying is not inherently more evil, being the kind of creatures that we are, we experience it as more offensive; therefore, going ahead and lying involves *ipso facto* greater meanness, and by virtue of *this* is morally worse. This view supports CV.

If the ground of CV is a function of the greater psychological difficulty in lying, we would expect less deception in *LY* compared to the other conditions. Rates of deception, however, were not significantly different. We can therefore conclude that to the extent that the moral explanation of CV is the greater malice in lying (as explained), our experimental results of equivalence in rates of deception in the three groups, again, undermine CV.

(C) Finally, a third ground for CV involves the *intrinsic nature* of lying. The idea is that lying is inherently more deceptive, in the sense that the distance between falsehood and the truth is greater than the distance between a misleading truth and the truth (or between nonverbal gestures, whose truth value is vaguer, and the truth). Lying is thus a greater evasion of the truth and simultaneously an intrinsically greater deviation from truthfulness. It is thus more seriously immoral. Now if this is the conceptual ground for CV, then *empirical* findings about people's actual attitudes toward *LY*, *FI*, and *ND* can neither support nor oppose CV, normatively speaking. Accepting this ground seems therefore to finally draw a limit to the moral relevance of the experimental approach.

We argue that this impression is *false*. We ought to ask how we are to understand the idea that lying is "inherently more deceptive." As explained, this presumably invokes the idea that what grounds the moral status of the different forms of deception is a conceptual truth regarding the epistemic properties of lying versus the other deceptions. But even if such a conceptual account about the greater distance between "lying" and "truth" is sound, it is not directly determinative of "level of deceptiveness." The latter seems rather to be the empirical fact regarding the degree to which people *are* actually deceived by each of the different forms of deception. And it should be this latter fact concerning potency that is directly significant for the possible *moral* import (wrongness) of "inherent deceptiveness." The logical or epistemic status of the different forms of deception vis-à-vis the notion of the truth is a different issue from the question of which form of deception as a matter of fact conceals the truth more effectively; only the latter correlates directly with moral wrongness. Now the question of which form of deception is more potent, or more effective in deceiving people, is not a question for armchair theorizing but a testable aspect of human communicative interaction. If it turns out that people generally

[15] Motivation in turn may influence the rightness of actions, as cogently argued by Sverdlik (1996).

deceive more successfully by performing non-lying deceptions, then for all practical purposes it is not true that lying is "more deceptive."

Suppose, for example, that I say "I climbed the rope" (which I did), while simultaneously motioning climbing a ladder (which I didn't). If, empirically, in such a situation, people tend to believe the motioning more, implicitly assuming that it is more reliable (i.e. assuming that I must have mistakenly said "rope" when in fact intended "ladder"), then if I know this fact about human information processing and use it to deceive more efficiently via nonverbal deception (as contrasted to lying by saying "ladder" while correctly motioning climbing a rope, which would result in less misleading on average), then we must conclude that nonverbal deception is worse, as far as the parameter of being singularly more deceptive is concerned. Another, more fantastic, example: if we all lived in Pinocchio's world, where lies (only) would immediately and universally manifest themselves on our noses, then lying would be consistently *less* deceptive than the other forms of deception. The point here is that the questions raised by such examples are empirical, despite the conceptual guise of "(level of) inherent deceptiveness."

We argue that the combination of our results regarding investors and consultants suggests an interesting conclusion in this respect. Recall that the equivalence among *investors* meant that people trust potential deceivers (to send truthful messages) to a similar degree in all three conditions. Since the extent of deceiving by consultants in the three conditions was also similar, the combination of the results regarding investors and consultants suggests that the level of deceptiveness of the three forms of deception is, as a matter of fact, similar (i.e. if similar amount of deceiving among conditions has similar deceptive impact, then the *singular potency* of deceptiveness is equivalent among conditions)! Hence, what seemed prima facie to be a conceptual point that poses a rigid limit to the empirical investigation of the ethics of deception, turns out upon further reflection to be a function of our very experiment. Our experimental scheme can deliver a normative verdict even with respect to a parameter that seemed beyond empirical reach. Our results seem to align with ET—or again, strictly speaking, undermine another possible reason for CV.

4. Concluding remarks

While the normative question of the moral gradation of forms of deception has been debated for long, it has thus far not been recognized that this question can to a significant extent, if not predominantly, be addressed empirically. Demonstrating this was the objective of this research. Moreover, the potential fruitfulness in using methods of experimental economics in experimental moral philosophy, with the aim of drawing *normative* conclusions, has thus far remained untapped. In this paper we have operationalized the normative dilemma regarding forms of deception in terms of a strategic game and were able to draw normative conclusions from our results.

Typical vignette-cum-questionnaire-based experimental studies poll people's moral judgments, yet normative conclusions obviously cannot be inferred validly from such psychological facts. At best, the common folk view could figure as one component in a holistic "inference to the best explanation" (which is the limited kind of normative yield one may attempt to find in experiments in moral philosophy, usually). In contrast, our method has been to identify implicit behavioral assumptions ("empirical commitments") of moral views and to test those experimentally. Thus, given that breaching trust is morally wrong and that breaching greater trust is *ceteris paribus* morally worse than breaching lesser trust, we can then test which kind of deception breaches greater trust, by operationalizing level of trust as amounts of money invested in response to a potentially informative yet potentially deceptive message. This can yield *moral* arguments about degrees of wrongness.[16]

A potential concern could be that participants perceive the strategic interaction as a "mere game" where deception need not be considered morally problematic (as in, say, a game of poker). We believe this is not a significant problem for several reasons. First, the norm against deception is both strong and deeply entrenched. While it can be waived in particular situations, no indication for such waiving was hinted at in explaining the experiment to the subjects. (We should add that the title "the Deceiving Game" chosen for this paper was not used in the experiment.) Moreover, deception in the experiment causes monetary loss to the deceived, which invokes the even stronger norm against fraudulent behavior. Up-to-date meta-analyses show substantial evidence that behavior of participants in laboratory experiments conforms to moral norms against deception (Abeler, Nosenzo, and Raymond, 2019; Gerlach, Teodorescu, and Hertwig, 2019). In particular, participants' behavior in our experiment attests that deception was indeed perceived as misconduct. Sans moral considerations, the game-theoretical solution of our game requires that the messages be ignored as non-informative. The intuition behind this is clear: if the investors respond to the messages, consultants can only gain by choosing the message that maximizes investments, regardless of what they observe (cf. Crawford and Sobel, 1982). In contrast, *de facto*, messages are strongly contingent on the observed balls, and investors increase their average investment by approximately one third of their endowment if the consultant chooses the BLUE message. Both results cannot be explained if participants perceive the situation as a game free of ethical constraints.

CV has been the prominent view about forms of deception for millennia, and still is today. Our experimental results together with our normative analyses pose a new (*kind of*) challenge to CV.

To be sure, the conclusions reached here are not final pronouncements. Our results should be extended and tested in some notable directions. These include the following. (1) We identified testable empirical commitments at the basis of the normative arguments. To the extent that behavior varies across cultures, the normative conclusions might vary correspondingly. We tested these commitments with subjects hailing from a WEIRD (Western, Educated, Industrialized, Rich, and Democratic) society (Henrich, Heine, and Norenzayan, 2010); additional tests with more diverse populations are required to ascertain the generality of the conclusions. (2) There is value in extending the tests to different contexts or scenarios, beyond the rather abstract and impersonal presentation of the original Deceiving Game. (3) Allowing repeated interactions and cross-examinations between subjects would simulate an important dimension of real-life communication. And so on. The contribution of this paper is rather in presenting the basic experimental idea, showing how it can be put into actual practice, reporting seminal results, and providing a detailed analysis of how normative conclusions can be derived from them. This should provide a firm basis for future treatments.

The holy grail of experimental moral philosophy is in offering support for normative judgments. Yet experimental results cannot establish what the relative weight of such moral reasons

[16] As mentioned, our experiment does *also* yield results in traditional terms of descriptive ethics—these suggest that CV does not reflect folk normative commitments.

is or, a fortiori, that a normative reason is trumping all others and is therefore the all-things-considered moral recommendation. Only philosophical analysis relying on philosophical theory can yield such conclusions. The more certain we are that the experimental design covers all plausible moral hypotheses, the weightier will be the normative conclusions derived from the experiment. And then if all results point to the same conclusion, we can hope to approach normative knowledge asymptotically. This paper did not, as it cannot, refute CV; it did, however, illustrate how experimental results can inform and influence the normative debate between two moral positions.

References

Abeler, Johannes, Daniele Nosenzo, and Collin Raymond (2019). Preferences for truth-telling. *Econometrica* 87(4), pp. 1115–1153.

Adler, Jonathan (1997). Lying, deceiving, or falsely implicating. *Journal of Philosophy* 94(9), pp. 435–452.

Berstler, Sam (2019). What's the good of language? on the moral distinction between lying and misleading. *Ethics* 130(1), pp. 5–31.

Bicchieri, Cristina (2006). *The grammar of society: the nature and dynamics of social norms.* New York: Cambridge University Press.

Bok, Sissela (1989). *Lying: moral choice in public and private life.* Vintage Books.

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

Chisholm, Roderick and Thomas Feehan (1977). The intent to deceive. *Journal of Philosophy* 74(3), pp. 143–159.

Cohen, Jacob (1988). *Statistical power analysis for the behavioral sciences.* Hillsdale, NJ: Erbaum.

Crawford, Vincent P. and Joel Sobel (1982). Strategic information transmission. *Econometrica: Journal of the Econometric Society* 50(6), pp. 1431–1451.

Fischbacher, Urs (2007). z-Tree: Zurich toolbox for ready-made economic experiments. *Experimental Economics* 10(2), pp. 171–178.

Frankfurt, Harry G. (2009). *On bullshit.* Princeton: Princeton University Press.

Gächter, Simon and Jonathan F Schulz (2016). Intrinsic honesty and the prevalence of rule violations across societies. *Nature* 531(7595), p. 496.

Gerlach, Philipp, Kinneret Teodorescu, and Ralph Hertwig (2019). The truth about lies: a meta-analysis on dishonest behavior. *Psychological bulletin* 145(1), pp. 1–44.

Greiner, Ben (2015). Subject pool recruitment procedures: organizing experiments with orsee. *Journal of the Economic Science Association* 1(1), pp. 114–125.

Grice, H Paul (1989). *Studies in the way of words.* Cambridge, MA: Harvard University Press.

Hardin, Russell (2002). *Trust and trustworthiness.* New York: Russell Sage Foundation.

Henrich, Joseph, Steven J. Heine, and Ara Norenzayan (2010). The weirdest people in the world? *Behavioral and brain sciences* 33(2-3), pp. 61–83.

Hugh-Jones, David (2016). Honesty, beliefs about honesty, and economic growth in 15 countries. *Journal of Economic Behavior & Organization* 127, pp. 99–114.

Koopman, P. A. R. (1984). Confidence intervals for the ratio of two binomial proportions. *Biometrics*, pp. 513–517.

Lakens, Daniël, Anne M Scheel, and Peder M Isager (2018). Equivalence testing for psychological research: a tutorial. *Advances in Methods and Practices in Psychological Science* 1(2), pp. 259–269.

Mahon, James Edwin (2016). "The definition of lying and deception". *The stanford encyclopedia of philosophy*. Ed. by Edward N. Zalta. Winter 2016.

O'neil, Collin (2012). Lying, trust, and gratitude. *Philosophy & public affairs* 40(4), pp. 301–333.

Pepp, Jessica (2019). The aesthetic significance of the lying-misleading distinction. *The British Journal of Aesthetics* 59(3), pp. 289–304.

Rees, Clea F. (2014). Better lie! *Analysis* 74(1), pp. 59–64.

Rogers, Todd, Richard Zeckhauser, Francesca Gino, Michael I Norton, and Maurice E Schweitzer (2017). Artful paltering: the risks and rewards of using truthful statements to mislead others. *Journal of personality and social psychology* 112(3), p. 456.

Rothmann, Mark D, Brian L Wiens, and Ivan SF Chan (2011). *Design and analysis of non-inferiority trials*. CRC Press.

Saul, Jennifer (2012a). Just go ahead and lie. *Analysis* 72(1), pp. 3–9.

— (2012b). *Lying, misleading, and what is said: an exploration in philosophy of language and in ethics*. Oxford: Oxford University Press.

Savage, Leonard J. (1954). *The foundations of statistics*. New York: John Wiley & Sons, pp. 188–190.

Schroeder, Timothy, Adina L Roskies, and Shaun Nichols (2010). "Moral motivation". *The moral psychology handbook*. Ed. by John Doris. Oxford University Press.

Shiffrin, Seana Valentine (2014). *Speech matters: on lying, morality, and the law*. Princeton University Press.

Smith, Vernon L (1994). Economics in the laboratory. *Journal of Economic Perspectives* 8(1), pp. 113–131.

Strudler, Alan (2010). The distinctive wrong in lying. *Ethical Theory and Moral Practice* 13, pp. 171–179.

Sverdlik, Steven (1996). Motive and rightness. *Ethics* 106(2), pp. 327–349.

Webber, Jonathan (2013). Liar! *Analysis* 73, pp. 651–659.

Weissman, Benjamin and Marina Terkourafi (2019). Are false implicatures lies? an empirical investigation. *Mind & Language* 34(2), pp. 221–246.

Williams, Bernard (2002). *Truth and truthfulness.* Princeton: Princeton University Press.

Appendix 1: Experimental instructions

<INSTRUCTIONS FOR EXPERIMENT 1>

Instructions for participants

Payment: In this experiment you will receive payment based either on your decisions in the first part or on your decisions in the second part (as will be determined randomly), and the decisions of another participant that will be randomly matched with you. Each point in the experiment is worth 20 Agorot, and in addition each participant will receive 5 NIS.

First part

The participants are randomly matched in pairs. In each pair, one participant is in the role of *Sender* and the other participant in the role of *Receiver*. The Sender has exclusive access to information (as explained shortly), and is free to choose which of two messages to pass to the *Receiver* regarding that information, as explained below.

Each participant knows what instructions the other participant received.

You are in the role of ***Sender***.

A computer randomly chooses, with equal probabilities, one of the following two urns: a *blue* urn, which contains two orange balls and three blue balls, or an *orange* urn, which contains three orange balls and two blue balls. [You will be informed of which urn was chosen next week, when you get paid].

Although no participant will know for certain which urn was chosen until next week, the *Sender* receives the following relevant information: s/he observes three balls drawn from the urn, as follows:

First, one blue ball and one orange balls are drawn. Additionally, one ball is randomly drawn of the three remaining balls. That is, if the computer chose the blue urn, there is 66.6% chance that the third ball drawn will be blue, and 33.3% that it will be orange, and vice versa for the orange urn.

<LY CONDITION>

The task of the participants is:

The *Sender* observes the three drawn balls, and then chooses to send the *Receiver* one of these two messages:

1. Send to the *Receiver* the message: “I saw two blue balls and one orange ball”.
2. Send to the *Receiver* the message: “I saw two orange balls and one blue ball”.

The *Receiver* does not know which balls were drawn; the only information available to him/her is the message sent by the *Sender*. After receiving this message, the *Receiver* decides how many points to invest.

<FI CONDITION>

The task of the participants is:

The *Sender* observes the three drawn balls, and then chooses to send the *Receiver* one of these two messages:

1. Send to the *Receiver* the message: "I saw two blue".
2. Send to the *Receiver* the message: "I saw two orange".

The *Receiver* does not know which balls were drawn; the only information available to him/her is the message sent by the *Sender.* After receiving this message, the *Receiver* decides how many points to invest.

<ND CONDITION>

The task of the participants is:

The *Sender* observes the three drawn balls, and then chooses on which of the two urns to bet. if s/he bets on the correct color (the color of the chosen urn) s/he will earn a bonus of 5 points. In addition, this color will be revealed to the *Receiver.* That is, the *Sender* chooses between the following two options:

1. Bet that the chosen urn is blue, a bet that will earn the *Sender* a 5 points bonus if the chosen urn is indeed blue. In addition, the *Receiver* will learn that the *Sender* bet on blue.
2. Bet that the chosen urn is orange, a bet that will earn the *Sender* a 5 points bonus if the chosen urn is indeed orange. In addition, the *Receiver* will learn that the *Sender* bet on orange.

The *Receiver* does not know which balls were drawn; the only information available to him/her is the betplaced by the *Sender.* After learning which color the *Sender* bet on, the *Receiver* decides how many points to invest.

The monetary payoff of the ***Sender*** is determined thus: The Sender will receive one point for each point that the Receiver will invest, regardless of the colour of the drawn ball.

The ***Receiver*** will win or lose in this investment according to the following:

The *Receiver* receives at the beginning of the experiment 100 points. After receiving the message that the *Sender* chose to send, s/he decides how many points out of the 100 to invest in a gamble, which will win or lose according to these rules:

- If the blue urn was chosen, (to be clear, the *Receiver* cannot gamble on the orange urn), the *Receiver* wins 2.5 points for each point invested, so that his/her payment is the sum of points s/he decided to invest times 2.5, in addition to the amount that s/he chose not to invest (100 minus the invested amount).
- If the orange urn was chosen, the *Receiver* wins 0 points for each point invested, so that his/her payment is 100 minus the amount that s/he chose to invest in the blue urn.

Questions to ascertain understanding of the instructions (please mark the correct answer):

1. The *Receiver* invested 100 points, and the chosen urn was blue. How many points would each participant receive?

   (a) The *Receiver* receives 200 points and the *Sender* receives 100 points.

   (b) The *Receiver* receives 200 points and the *Sender* receives 250 points.

   (c) The *Receiver* receives 250 points and the *Sender* receives 100 points.

2. What is the correct order?

   (a) The *Sender* observes three balls (one blue, one orange, and one randomly drawn), the *Sender* chooses a message, the *Receiver* bets on the number of points that would be multiplied by a factor of 2.5 if the chosen urn turns out to be blue.

   (b) The *Sender* observes three balls (one blue, one orange, and one randomly drawn), the *Sender* chooses a message, the *Receiver* bets on the number of points that would be multiplied by a factor of 2.5 if the chosen urn turns out to be orange.

   (c) The *Sender* chooses a message, the *Sender* observes three balls (one blue, one orange, and one randomly drawn), the *Receiver* bets on the number of points that would be multiplied by a factor of 2.5 if the chosen urn turns out to be blue.

THE DECISION:

<LY CONDITION>

Indicate your decision if two blue balls and one orange ball are drawn:

○ Send the following message to the *Receiver*: “I saw two blue balls and one orange ball”.

○ Send the following message to the *Receiver*: “I saw two orange balls and one blue ball”.

Indicate your decision if one blue ball and two orange balls are drawn:

○ Send the following message to the *Receiver*: “I saw two blue balls and one orange ball”.

○ Send the following message to the *Receiver*: “I saw two orange balls and one blue ball”.

<FI CONDITION>

Indicate your decision if two blue balls and one orange ball are drawn:

○ Send the following message to the *Receiver*: “I saw blue”.

○ Send the following message to the *Receiver*: “I saw orange”.

Indicate your decision if one blue ball and two orange balls are drawn:

○ Send the following message to the *Receiver*: “I saw blue”.

○ Send the following message to the *Receiver*: “I saw orange”.

<ND CONDITION>

Indicate your decision if two blue balls and one orange ball are drawn:

○ Bet on the blue urn, a bet that will earn you a 5 points bonus if it turns out that the chosen urn is blue, and will be revealed to the *Receiver*.

○ Bet on the orange urn, a bet that will earn you a 5 points bonus if it turns out that the chosen urn is orange, and will be revealed to the *Receiver*.

Indicate your decision if one blue ball and two orange balls are drawn:

○ Bet on the blue urn, a bet that will earn you a 5 points bonus if it turns out that the chosen urn is blue, and will be revealed to the *Receiver*.

○ Bet on the orange urn, a bet that will earn you a 5 points bonus if it turns out that the chosen urn is orange, and will be revealed to the *Receiver*.

Second part

<handed out after the decision forms for the first part were collected>

The instructions for this part are identical to the instructions in the previous part, with the exception that you are in the role of ***Receiver***.

You will choose how many points out of 100 you want to invest (recall that this amount will be invested in blue, and will be multiplied by a factor of 2.5 if the blue urn is chosen). You can condition your decision on the decision of the ***Sender***:

<LY CONDITION>

1. How many points do you want to invest if the *Sender* chose for you to see the message "I saw two blue balls and one orange ball"?
   ___________ points.
2. How many points do you want to invest if the *Sender* chose for you to see the message "I saw two orange balls and one blue ball"?
   ___________ points.

<FI CONDITION>

1. How many points do you want to invest if the *Sender* chose for you to see the message "I saw blue"?
   ___________ points.
2. How many points do you want to invest if the *Sender* chose for you to see the message "I saw orange"?
   ___________ points.

<ND CONDITION>

1. How many points do you want to invest if the *Sender* bet that the blue urn was chosen?
   ___________ points.
2. How many points do you want to invest if the *Sender* bet that the orange urn was chosen?
   ___________ points.

Personal details: Sex: Male Female

Age: _____

Last five digits of your ID number: _______________

<INSTRUCTIONS FOR EXPERIMENT 2>

Welcome to the experiment!

Please read the instructions carefully

The experiment includes 20 rounds. In each round you can earn money according to the following rules. Payment will be in points, which will be converted to money at the end of the experiment at a rate of 10 points = 1 NIS. At he end of the experiment, the computer will randomly draw 5 rounds. Your payoff in this part will be the sum of your earnings in the five drawn rounds in addition to a base payment of 15 NIS.

At the beginning of the experiment you will be allocated into roles. Half of the participants will be in the role of Sender and the other half will be in the role of Responder. These roles will remain fixed during the first ten rounds, and will switch in the last ten rounds. In each round you will be matched in pairs of Sender and Responder. You will be rematched a the beginning of each round.

Instructions for the round

Responders

In each round, the Responder receives 100 points and decides how many of them to invest in a lottery. If the lottery is successful, the Responder wins 2.5 points for each point invested, so that his or her payoff is the sum of points invested times 2.5 plus the sum of points not invested (100 minus the amount invested). If the lottery fails, the Responder will lose the points he or she invested, so that his or her payoffs is 100 minus the amount invested.

Regardless of the outcome of the lottery, the Sender will receive the amount of points that the Responder chose to invest.

Determining the outcome of the lottery

In each rouns, the computer chooses with equal probabilities one of the following two urns: a green urn, which contains three green balls and two orange balls, and an orange urn, which contains three orange balls and two green balls.

The lottery is successful if the green urn is chosen, and the lottery fails if the orange urn is chosen.

Senders

The Sender observes three balls drawn from the chosen urn as follows: First, the computer draws one green ball and one orange ball. In addition, the computer draws one ball randomly chosen out of the three remaining balls in the urn. That is, if the green urn was chosen, then there is a chance of two thirds that the third ball that the Sender observes is green and a one third chance that the third ball is orange, and vice versa for the orange urn.

Note that this means that if the Sender observes two green balls, it is possible to deduce that there is a chance of two thirds that the chosen urn is the green urn. If the Sender observes two orange balls, it is possible to deduce that there is a chance of two thirds that the chosen urn is the orange urn.

The computer will present the Sender with the three balls drawn from the urn. Then, the Sender will choose one of two options:

<LY CONDITION ONLY>

- Send to the Responder the message: "I saw two green balls and one orange ball".
- Send to the Responder the message: "I saw two orange balls and one green ball".

<FI CONDITION ONLY>

- Send to the Responder the message: "I saw green".
- Send to the Responder the message: "I saw orange".

<ND CONDITION ONLY>

- bet that the green urn was chosen. This bet will earn the Sender 5 bonus points if the green urn was chosen.
- bet that the orange urn was chosen. This bet will earn the Sender 5 bonus points if the orange urn was chosen.

<LY AND FI CONDITIONS ONLY>

The Responder does not observe the balls drawn from the urn, but does see the message of the Sender. After receiving the message, the Responder decides how many points to invest in the lottery that wins if the green urn was chosen.

<ND CONDITION ONLY>

The Responder does not observe the balls drawn from the urn, but does see the bet of the Sender. That is, the Responder knows which of the two urns the Sender bet on, and how this affects the bonus points of the Sender. After receiving the message, the Responder decides how many points to invest in the lottery that wins if the green urn was chosen.

End of the round

At the end of the round, the computer will present to you the chosen urn, the Sender's choice, the Responder's investment, and the profits of the two participants in the round.

The end of the experiment

After the end of the experiment you will be asked to fill in a short general questionnaire. This questionnaire, as well as your decisions during the experiment, is anonymous. Please wait in your seat until we call you to get paid.

We will read out the instructions in a minute. If later you have remaining questions, please raise your hand and the experimenter will approach you and answer you in private.

Appendix 2: Statistical appendix

To confirm that the results do not support the hypotheses underlying CV, we conducted inferiority tests to test the null hypothesis that the effect size, as measured by Cohen's d for the difference between LY and the other two conditions, is larger than a minimal benchmark (Cohen, 1988).

To determine the minimal benchmark (which we also use to conduct our power analyses), we use the data collected by Schäfer and Schwarz (2019), who estimated the distributions of effect sizes in published psychology papers by subdisciplines.[1] Following Schäfer and Schwarz (2019), we set our benchmark to be the lower median (i.e., the median of the low third of observations, or the 16.65% quantile) of effect sizes found in experimental or quasi-experimental between-subjects studies in the relevant sub-disciplines. Because we are interested not only in an expected effect size, as is typically the case with equivalence tests and power calculations, we also report here results based on a more conservative benchmark calculated as half of the main benchmark. Our benchmarks are, accordingly, 0.434 and 0.217. We additionally report the statistical power of our design to detect the benchmark effect sizes in the hypothesized direction.

In experiment 1. the inferiority test yields a highly significant result of$p < .001$ for the main benchmark and$p = .008$ for the conservative benchmark for consultants. Furthermore, we can similarly reject at a confidence level of $.90$ any effect size of 0.013 or higher. For investors, the inferiority test yields a significant result of$p = .001$ for the main benchmark and$p = .031$ for the conservative benchmark. The test is significant for any effect size of 0.112 or above.

In experiment 2, the inferiority test yields a highly significant result of$p < .002$ for the main benchmark and$p = .036$ for the conservative benchmark for consultants. Furthermore, we can similarly reject at a confidence level of $.90$ any effect size of 0.123 or higher. For investors, the inferiority test yields a significant result of$p = .001$ for the main benchmark and$p = .017$ for the conservative benchmark. The test is significant for any effect size of 0.063 or above.

To calculate power in Experiment 2, we cluster standard errors on subjects and estimate the interclass correlations from the data. For consultants, the power to detect the benchmark effect sizes is $1 - \beta = .962$ for the main benchmark and $1 - \beta = .527$ for the conservative benchmark. For investors, the power to detect the benchmark effect sizes is $1 - \beta = .993$ for the main benchmark and $1 - \beta = .656$ for the conservative benchmark.

Finally, we calculate the joint power for the two experiments taken together, i.e., the probability of obtaining a significant result in at least one experiment. The joint power using the main benchmark is $1 - \beta = .992$ for consultants and $1 - \beta = .998$ for investors. With the conservative benchmark, the power is $1 - \beta = .684$ for consultants and $1 - \beta = .770$ for investors.

[1] The sub-disciplines are defined according to the Social Sciences Citation Index (SSCI). The most relevant sub-discipline is "Psychology: Multidisciplinary". Because the number of relevant studies in the data is small, we extend the sample to the "Psychology: Experimental" and "Psychology: Social psychology" sub-disciplines, which yield a very similar lower median as taking just the Multidisciplinary sub-discipline.

References

Cohen, Jacob (1988). *Statistical power analysis for the behavioral sciences.* Hillsdale, NJ: Erbaum.

Schäfer, Thomas and Marcus Schwarz (2019). The meaningfulness of effect sizes in psychological research: differences between sub-disciplines and the impact of potential biases. *Frontiers in Psychology* 10, p. 813.