Journal of Risk and Uncertainty (2022) 64:235–250
https://doi.org/10.1007/s11166-022-09380-1

Self-serving dishonesty: The role of confidence in driving dishonesty

Stephanie A. Heger[1] · Robert Slonim[2] · Franziska Tausch[2]

Accepted: 27 April 2022 / Published online: 19 May 2022

Abstract
Ambiguity and uncertainty as an explanation for ethical blind spots is well-documented. We contribute to this line of research by showing that these blind spots arise even when there is naturally occurring uncertainty—that is, when individuals are simply uncertain of the truth they "fill-in" this uncertainty in a self-serving way. To examine *self-serving dishonesty*, we asked a sample of U.S. car owners to respond to an auto insurance underwriting questionnaire that affects their price of insurance (i.e., premium), and investigated how financial incentives affect the honesty of their responses. We find, consistent with the current literature, that people have a strong preference for truthfulness, but only when they are confident of the objective truth. However, when people are not completely certain of the objectively correct answer, significant dishonesty occurs in a self-serving manner. We also find that reports of confidence do not depend on incentives and thus self-serving dishonesty is not strategic.

Keywords Dishonesty · Insurance underwriting · Experiment · Beliefs

JEL C91 · D81

1 Introduction

Dishonest, fraudulent and corrupt behaviors pose substantial costs to many private and public organizations and societal wellbeing—international costs of corruption total $3.6 trillion or more than 5% of global GDP (World Economic Forum, 2018) and 1 in 4 adults across the world report having to pay a bribe to access public services in the past year (International Transparency, 2017). Becker (1968) argued that

✉ Robert Slonim
Robert.Slonim@uts.edu.au

[1] Department of Economics, The University of Bologna, Bologna, Italy

[2] School of Economics, University of Technology Sydney, Sydney, Australia

the decision to commit crime (including fraud and corruption) may be based on a rational cost-benefit analysis. However, while the standard economic model predicts individuals to behave as cold cost-benefit calculators, more recent literature suggests that individual's willingness to cheat, lie and commit fraud for financial gain depends on whether they are perceived by others, and by themselves, as dishonest (Abeler et al., 2019; Cressey, 1986; Mazar et al., 2008). This suggests that people weigh the benefits from unethical behaviour against the costs, which also include the desire to maintain a moral self-image (Rabin, 1995; Ayal et al., 2015; Bénabou & Tirole, 2016)—a tendency that emerges in childhood (Maggian & Villeval, 2016).

However, mounting evidence suggests that "people who appear to exhibit a preference for being moral may in fact be placing a value on *feeling* moral (Gino et al., 2016)." To maintain their moral feeling while behaving dishonestly, people frequently use behavioural strategies to justify, excuse or remain ignorant about the consequences of their dishonest behaviour (Sykes & Matza, 1957; Bandura, 1999; Shu et al., 2011; Ayal et al., 2015; Köneke et al., 2015; Shalvi et al., 2015), resulting in a reduction of "ethical dissonance" (Ayal et al., 2015). In particular, there is growing evidence that individuals use uncertainty, ambiguity and subjectivity when behaving dishonestly to preserve their self-image (Kunda, 1990; Konow, 2000; Dana et al., 2007; Mazar et al., 2008; Haisley & Weber, 2010; Shalvi et al., 2015; Exley, 2016; Grossman & Van Der Weele, 2017; Gneezy et al., 2020). Exley and Kessler (2019) organizes this literature into two forms of rationalization in which individuals engage to justify motivated decisions. The first exploits uncertainty in how individual decisions link to outcomes to rationalize misbehavior (Dana et al., 2007; Mazar et al., 2008; Shalvi et al., 2015). A second stream, to which our paper belongs, uses seemingly innocuous preferences and beliefs to rationalize their decision (Haisley & Weber, 2010; Pittarello et al., 2015; Exley, 2016; Schwardmann & Van der Weele, 2019; Gneezy et al., 2020). For example, Gneezy et al. (2020) focus on the role of self-deception, or the strategic manipulation of beliefs, to neutralize the effect of dishonest behaviour on one's self-image.

In this paper, we join a growing literature demonstrating that increased uncertainty or ambiguity about the truth leads to more self-serving behavior. However, we depart from the current literature in which the level of ambiguity is experimentally induced (Pittarello et al., 2015) and instead show that this finding is robust to naturally occurring ambiguity. We find that when individuals are uncertain of the exact truth of a piece of *objective* information they are asked to provide and when tasked with making an approximation of the truth, they "fill-in" this uncertainty in a self-serving way.[1]

[1] Our notion is related to the literature on elastic justification. Schweitzer and Hsee (2002) report results from an experiment in which subjects were asked to play the role of a hypothetical car salesman. In the experiment, the uncertainty about mileage of the hypothetical car was randomly varied and sellers were more likely to report lower mileage in the high uncertainty condition than in the low uncertainty condition. However, the experiment in Schweitzer and Hsee (2002) was unincentivized and thus it is unclear whether subjects had a preference over their responses.

We show this using an online experiment in which we ask car owners in the U.S. to participate in a survey that resembles an auto insurance underwriting application. The survey is an 11-item questionnaire that asks participants about driving habits, relevant demographics (see survey here) and later elicits their confidence in the accuracy of their answers to each of the 11 items in the questionnaire. Our study includes three (sets of) conditions. In our Control Treatment, participants are asked to respond with no financial consequences tied to their answers. In our Incentive Treatments, we repeat the same questions, but include high stakes financial incentives that proxy for the cost structure behind premiums in underwriting policies, using two different stake levels. Our third set of conditions, the Intervention Treatments, also include financial incentives but at the same time apply a series of interventions that are meant to address common explanations found in the literature for dishonesty: moral wriggle room, attenuation, and rational lying detection. We measure dishonesty by comparing responses in the Control treatment, where there are no financial incentives to lie with responses in the Incentive Treatment. This method for identifying dishonesty or truthfulness is similar to John et al. (2012), thus while we cannot know the truthfulness of any individual response, we can identify how responses change, on average, when there are financial incentives versus when there are no financial incentives.

To identify the effects of self-serving dishonesty on reporting, we included questions for which we expected subjects to hold a range of confidence in the answers. For example, we expect subjects to be certain of their age and gender, less certain about the number of years they have been licensed to drive and the number of speeding tickets received and even less certain of the frequency of their on-street parking habits and the value of their car. After subjects completed the survey, we asked them to indicate the confidence in the correctness of each of their responses. We hypothesize that if participants are subject to self-serving dishonesty, then we will observe more dishonesty on the responses that respondents report less certainty when there are financial incentives compared to when there are not financial incentives.

We find that relative to our Control Treatment, subjects in the Incentive Treatments are significantly more dishonest. On average, when participants have an incentive to lie, they add nearly 30 USD to their payoff in our Incentive Treatments versus what subjects in our Control Treatment would have earned had they also been paid bonuses for their responses. The incentives in the Base Incentive Treatment were structured such that subjects could add or subtract $10 for each change in response (see survey here.) and thus a $30 increase is equivalent to three minimal lies.

Second, we find evidence consistent with self-serving dishonesty; that is, the majority of dishonesty detected between the Control Treatment and the Incentive Treatments is driven by the two questions in which subjects report below average confidence in the accuracy of their answers. For these two questions that the subjects were significantly less confident about, subjects, on average, made a minimal lie (i.e., increased their bonus by $10), whereas for the other nine questions in which subjects were confident, they lied on average by approximately 1/10 of a minimal lie. In other words, dishonesty driven by their own uncertainty in the truth is 10 times larger than when they are almost certain of the truth.

Interestingly, we find that the beliefs about accuracy reported in the Incentive Treatments and the Control Treatment are not statistically different. Thus, we find no evidence that self-serving dishonesty is *strategic*; that is, individuals do not manipulate their confidence in the correctness of their response when there is an incentive to lie (Schwardmann & Van der Weele, 2019; Gneezy et al., 2020). Instead, self-serving dishonesty appears to be nonstrategic.

Self-serving dishonesty appears to be driven by an implicit bias rationalized by beliefs (i.e., low levels of confidence in the truth). However, it could also be that subjects use ambiguity in other ways to justify their dishonesty. To test this, we construct a set of interventions aimed at reducing any ambiguity (or ignorance) that honesty is the expected and appropriate behavior.[2] We design a set of treatments that explicitly informs subjects about the types of behaviors that are deemed unacceptable and dishonest and ask subjects to agree to abide by an honor code of honesty (henceforth: Honor Code Interventions). We find that our Honor Code interventions are only minimally effective at mitigating the dishonesty we detect in the Incentive Treatments, reducing the detected dishonesty by only 16%.[3]

2 Experimental design

We recruited 1,069 participants via the online platform Amazon Mechanical Turk (Mturk) to participate in a survey that resembles an insurance underwriting questionnaire. All participants are car owners and located in the US. They all earned 1 US Dollar for finishing the survey, and five participants are randomly drawn to be paid an additional bonus. This bonus starts at $350 and – depending on the Treatment – may be adjusted depending on the responses the participant indicates in the questionnaire. The bonus payments were $354 on average, ranging from $80 to $500. All Treatments were released on Mturk at the same time. The experimental instructions used for the two main treatments, the Control Treatment and the Incentive Treatments can be found here.

The questionnaire includes questions that are typically asked in car insurance underwriting that are used to determine the premium a person has to pay to receive insurance. Table 1 displays a summary of the questions that are asked and the appendix includes the exact wording and response options.

[2] There is evidence that individuals avoid information and remain purposefully ignorant (Oster et al., 2013; Thunström et al., 2016; Golman et al., 2017).

[3] We also examined two other interventions that were unrelated to ambiguity–rational lying (Becker, 1968) and consequence attenuation (Sykes & Matza, 1957; Köneke et al., 2015; Shalvi et al., 2015; Bellé & Cantarelli, 2017; Fukukawa, 2002). We reserve the description of these strategies, our experimental treatments and the results to the appendix because these strategies are not as developed in the literature and the interventions were hypothetical rather than involving actual financial penalties, which is core to the rational lying hypothesis, and thus we interpret the results with caution. See (Köneke et al., 2015) who provide a comprehensive overview of justification strategies applied in the insurance context.

Table 1 11-ITEM SURVEY QUESTIONS

1.	**Speeding** Number of speeding fines you received in the last five years
2.	**Parking** Your parking habits on and off the street
3.	**Other** Frequency of other drivers with less than five years' experience driving your car in the last year
4.	**Accidents** Number of accidents or incidents involving loss or damage in the last ten years
5.	**Alcohol** Average number of alcoholic drinks you consume in a week
6.	**Miles** Amount of miles you have driven any car in the last five years
7.	**Value** The current value of your car
8.	**Gender** Your gender
9.	**Marital** Your marital status
10.	**Age** Your age
11.	**Licensed** Number of years you have been licensed to drive

2.1 Treatments

For each Treatment condition, the 11 survey questions indicated in Table 1 were presented on a single page of the online survey and always in the same order. Respondents could enter them from top to bottom but could have gone back and forth before answering them all. All 11 questions were forced responses in order to simulate an actual underwriting process in which they would need to answer all questions to be offered a policy.

We implemented three sets of conditions: (1) Control Treatment; (2) Incentive Treatments (Base Incentive Treatment and High Incentive Treatment); and (3) Intervention Treatments. Our Control Treatment provides a baseline for responses to the underwriting questions when there is no monetary incentive to be dishonest. The Incentive Treatments introduce monetary incentives for responses. The Intervention Treatments adds interventions to the Base Incentive Treatment aimed at mitigating dishonesty by weakening the popular justification strategies of moral ambiguity and attenuation. Table 2 lists all of the treatments and sample sizes. Note that since our primary research questions involve comparing the Base Incentive Treatment to every other treatment, we included twice as many observations in the Base Incentive Treatment to increase power (see List et al. (2011) for a discussion of power for this situation). We describe each set of Treatments below.

Control Treatment In only the *Control Treatment*, participants could receive a fixed potential bonus of $350 that is unaffected by their responses to individual survey items. The Control Treatment survey questions use the insurance context, similar to the Incentive Treatment described below, but participants have no monetary incentive to provide a dishonest response to any question. Because there is no monetary incentive for individual responses, the Control Treatment does not explain how payoffs would change with different responses since they, in fact, do not change in the Control Treatment.

Table 2 Sample Sizes by Treatment Assignment

Treatment	Sample Size
Control Treatment	151
Incentive Treatments	
Base Incentive Treatment	308
High Incentive	151
Honor Code Interventions	
Signature	154
Signature PS	153
Check Box	152
Total	1,069

Incentive Treatments Participants in the *Base Incentive Treatment* and all subsequent conditions initially receive information that the questions that are going to follow are typically asked in determining the insurance premium that drivers have to pay to receive auto insurance. They are explained how premiums are determined based on whether people are high risk or low risk based on the answers they provide in the questionnaire, and how this in turn affects premiums, and - in a similar manner - the bonus in the experiment. For each question they are provided with information regarding how each response will affect their potential bonus. For 8 of the 11 of our underwriting survey questions, the marginal impact of riskier (from the insurer's perspective) responses decreased the subject's earnings by an additional $10. The remaining 3 of the 11 questions were framed as good driver discounts, where the marginal impact of a less riskier (from the insurer's perspective) responses increased the subject's earnings by an additional $10. Starting with a bonus of $350, the most extreme responses would result in bonus payments that could range from $500 to $0. We also include one variation of the Base Incentive Treatment, called the *High Incentive Treatment*, in which the basic structure is identical, but we triple the amount added to the bonus for the question on people's parking habits. Because the Base Incentive Treatment and the High Incentive Treatment are nearly identical, we pool them together (henceforth: Incentive Treatments) for our analysis. However, we note that our results are qualitatively equivalent if we just use the Base Incentive Treatment.

Intervention Treatments The Intervention Treatments keep the identical structure as the Base Incentive Treatment, but with the additional information described below. The first set of conditions are aimed at dishonesty that stems from moral ambiguity.

Honor Code Treatments In the *Signature at the Top Treatment* (henceforth: Signature Treatment) participants are asked to confirm an honor statement regarding the truthfulness of their responses by typing their first name. The honor statement is intended to provide a moral cue. This counteracts individuals' strategy to justify dishonesty by referring to a lack of awareness about what behavior is expected from

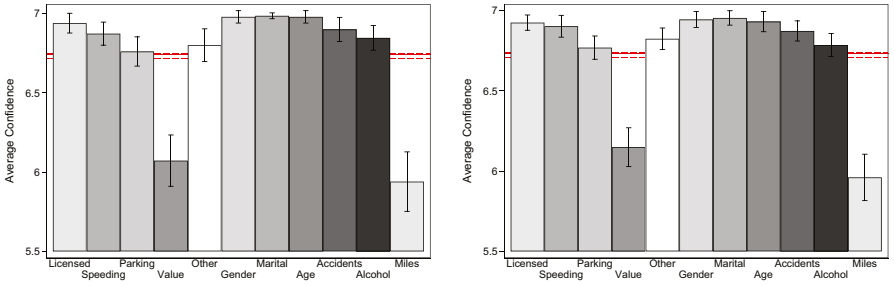
them thereby mitigating moral ambiguity. Signing one's name underneath an honor code has come to be known as an expression of 'social self-presence' - manifesting one's identity on a page as a promise. Shu et al. (2012) find that those who sign such a statement at the beginning of a document, rather than at the end, were more likely to act honestly, while Kristal et al. (2020) find no such effect. E-signatures, however, are found to be less effective. Individuals who gave a handwritten signature indicate they are less likely to breach a contract and are less likely to cheat on simple tasks than those who sign by e-signature (such as PINs, check boxes, or typed names) (Chou, 2015a, b). The evidence on the effectiveness of different types of e-signatures is however limited and mixed, and thus requires further investigation.

In addition to the Signature Treatment, we implemented two additional variations: (1) Signature Previous Screen (henceforth: PS) and (2) Check Box Treatment. In the *Signature PS Treatment*, the honor statement is not on the same page as the questionnaire but on the previous screen. By placing the honor statement on the previous page, participants are required to sign the statement before they can answer the questionnaire. In the *Check Box Treatment*, participants are asked to confirm an honor statement regarding the truthfulness of their responses by clicking a button. This treatment is identical to the Signature Treatment except that subjects only have to check a box rather than type in their name.

2.2 Measuring confidence

After subjects completed the 11-item survey we additionally asked them how confident they are in the accuracy of the responses they provided in the survey for each of the 11 questions (the appendix includes the exact wording). When subjects reached the page with the questions on confidence, they were not permitted to return to the previous page to change their responses to the other questions. They were asked to rate their confidence on a scale from 1 (= Not At All Confident) to 7 (= Completely Confident). We conjectured, based on the notion of self-serving dishonesty, that confidence about the correctness of their response would influence honesty.

The average level of confidence reported across the 11 questions was 6.74 (standard deviation = 0.27) in the Control Treatment and 6.73 (standard deviation = 0.46) in the Incentive Treatments. We find that subjects are significantly less certain of the accuracy of their answers to the "Miles" and "Value" questions relative to the other 9 questions (mean = 6.10 versus mean = 6.90, respectively). Figure 1 shows the average level of confidence reported for each of the 11 questions with 95% confidence intervals. The solid red line represents the average level of confidence across the 11 questions and the dashed red lines represent the 95% confidence interval of the average. As Fig. 1 makes clear, not only do subjects report significantly lower levels of confidence in their reports for the "Values" and "Miles" question, but these two questions are the only questions with significantly lower levels of confidence than the average level of confidence. Unpaired t-tests between the confidence reported in the "Values" and "Miles" questions for the Control Treatment and the Incentive Treatments indicate significantly lower levels of confidence than the average level of



(A) Control Treatment **(B)** Incentive Treatment

Average reported confidence in the accuracy of response with 95% confidence intervals. Scale "1" Not at all confident to "7" Completely confident.

Fig. 1 Confidence Levels

confidence in the two treatments, t = -7.81***, t = -8.16***, t = -8.68***, t = -9.86***, respectively.

3 Main results

3.1 Dishonesty in response to financial incentives

We first ask whether the presence of financial incentives causes the participants to lie in their responses to the insurance claim. To answer this question, we pool together the responses to each of the 11 questions in the survey. We calculate the *change in the bonus* earned in the Incentive Treatments based on what the respondent added to or subtracted from their final payment with their responses. We calculate the bonus earned in the Control Treatment based on what the respondent *would have* added or subtracted from their final bonus, if there had been identical response-based incentives, with their responses. For example, if a participant's responses resulted in an added \$10 on one response, an added \$30 on another response, subtracted \$10 on 3 responses, subtracted \$20 on 2 other responses and had no additional changes on their bonus for the remaining 4 questions, the net effect on their bonus would be subtracting \$30 from their initial starting bonus of \$350. Since the questions varied by how much could be added to the bonus with the responses, we also consider the percentage of the maximum possible payoff rendered by the subject's response. For example, if on one question the participant added \$10 and the most they could have added was \$40, then they added 25% of the maximum possible on this response, whereas if they added \$10 and the maximum they could have added was \$50 then they added only 20%.

In Table 3, we report OLS estimates of a model that regresses *Total \$ of Lying* in column (1) and *Percent of Maximal Bonus* in column (2) on a dummy for the Incentive Treatments, using the Control Treatment as the omitted group. We find that, on average, subjects in the Incentive Treatments distort their answers to increase their bonus payment by \$29.39 above the amount they would have received in the Control

Table 3 Total Dishonesty in Response to Financial Incentives

	Total lie in $	% of Maximal Bonus
Incentive Treatments	29.39***	0.43***
	(3.43)	(0.07)
Constant	-30.07***	7.40***
	(3.05)	(0.06)
Observations	610	610
R^2	0.12	0.06

OLS regression estimates. Robust standard errors in parentheses and *, ** and *** indicate statistical significance at the 10%, 5% and 1% levels, respectively

Treatment (i.e., without any financial incentives) and increase their percentage of maximal bonus by .43 percentage points (6% increase).

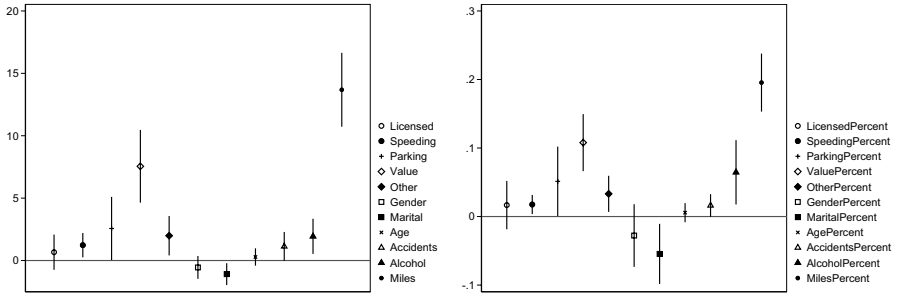
Result 1 Participants' reports are affected by the presence of financial incentives to be dishonest. Participants who have an incentive to lie report answers to the 11-item survey such that they increase their bonus payment, on average, by $29.39 and increase their maximal bonus by 6% relative to those participants who have no incentive to lie.

3.2 Self-serving dishonesty

Table 3 showed that subjects were dishonest in the presence of financial incentives. In this section, we provide evidence that this dishonesty is driven by self-serving dishonesty; that is, subjects distort their responses in a financially self-serving way, but only when they are uncertain of the objective truth. Recall, Fig. 1 shows that subjects exhibit a high level of confidence in the correctness of their responses to each of the 11 survey questions, except for two questions, "Miles" and "Value". In fact, these are the only two questions that subjects give "below" average confidence in the correctness of their answers.

In Fig. 2a, we plot the coefficients from 11 OLS regressions (1 regression for each question) that compare the average responses to each of the 11 survey questions in the Incentive Treatments relative to the Control Treatment. We find that 76% of the increase in the payoff found in Table 3 is driven by the responses given in the "Miles" and "Value" questions. Specifically, by under-reporting "Miles" and "Value", subjects in the Incentive Treatments, relative to subjects in the Control Treatment, earn on average a higher bonus of 14USD and 8USD, respectively.

In Table 4 we estimate a random effects (mixed effects) model in which we pool together the responses for all 11 questions from subjects in the Control Treatment and the Incentive Treatments, resulting in 11 observations per subject for 610 independent observations. We use the same two outcome variables as in Table 3. In columns (1) and (2), we compare the responses in the Incentive Treatments to



(A) Dollar Effect of Dishonesty **(B)** Percentage of Dishonesty

OLS regression coefficients for each of the 11 questions in the survey. Figure 2a shows the dollar amount added to bonus in the Incentive Treatments relative to the non-incentivized Control Treatment. Figure 2b shows the percentage change in the dollar amount added to the bonus in the Incentive Treatments relative to the non-Incentivised Control Treatment. The bars represent 95% confidence intervals.

Fig. 2 Effect of Incentives on Dishonesty

the Control treatment only for those questions for which, on average, subjects report below average levels of confidence and we find that subjects distort their answers to increase their bonus by $11.10 and move 16 percentage points closer to the maximal possible bonus.

In columns (3) and (4), we estimate the same model as in columns (1) and (2) but only for those questions which subjects reported above average levels of confidence. Now we find minimal and insignificant effects of the Incentive Treatments.

In columns (5) (6), we pool together all 11 questions and interact the dummy for the incentive treatment with the dummy for the questions that have below average levels of confidence. Consistent with columns (1)-(4), subjects significantly distort their responses in the presence of financial incentives when they have below average confidence in the objectively correct response compared to when they have above average confidence.

Result 2 Dishonesty in this experiment is driven by self-serving dishonesty; that is, when subjects are confident in the objective truth, there is very little evidence of dishonesty: in only two of the nine survey questions where subjects are confident in the objective truth do we detect lying, and this only accounts for 24% of the total gains from dishonesty. On the other hand, when subjects were less confident in the objective truth, we detect significantly more lying than in any of the questions where they knew the objective truth, and these two questions alone (just 18% of the questions) account for over three-fourths of all the financial gains received from dishonesty. Thus, when subjects do not know the objective truth, they are more likely to significantly and substantially distort their responses in a financially self-serving way than when they are certain of the objective truth.

Table 4 Self-Serving Dishonesty

	Total Lie in $	% of Maximal Bonus	Total Lie in $	% of Maximal Bonus	Total Lie in $	% of Maximal Bonus
Incentive	10.62***	0.15***	0.91**	0.01	0.91**	0.01
	(1.13)	(0.02)	(0.43)	(0.01)	(0.45)	(0.01)
Incentive × Below Average Conf	.	.	.	.	9.71***	0.14***
					(1.05)	(0.02)
Below Average Conf	.	.	.	.	-33.93***	-0.13***
					(0.91)	(0.02)
Constant	-30.50***	0.56***	3.44***	0.7***	3.44***	0.7***
	(0.98)	(0.01)	(0.37)	(0.009)	(0.39)	(0.008)
Observations	1220	1220	5490	5490	6710	6710
	Below Average Confidence Questions		Above Average Confidence Questions		All Questions	

Random effects model using observations from the Incentive Treatments and Control Treatment. Robust standard errors clustered at the subject-level in parentheses and *, ** and *** indicate statistical significance at the 10%, 5% and 1% levels, respectively

3.3 Self-serving dishonesty is non-strategic

Last, we briefly examine whether financial incentives motivated subjects to distort their confidence in the correctness of their answers in order to self-deceive; that is, we analyze whether subjects report lower confidence on their responses to questions in the presence of financial incentives in order to ex-post justify their lying. If this is the case, then we would expect to see less confidence in the Incentive Treatments than the Control Treatment. To test this, we compare the levels of confidence for each question between the Control and the Incentive conditions.

Table A1 reports the test statistic from a Mann-Whitney two-sample test for each of the 11 questions. We find that there are no significant differences in the confidence responses for 10 of the 11 questions, including, critically, both of the two questions with below average confidence (Miles and Value), between the Control Treatment (Fig. 1a) and the Incentive Treatments (Fig. 1b). Thus, we find no evidence that subjects were distorting their beliefs about the accuracy of their responses in a self-serving manner to justify their lying.

Result 3 We find no evidence that self-serving dishonesty is strategic. Subjects do not report higher levels of uncertainty about the truthfulness of their responses when there is a financial incentive to lie.

In sum, Results 1 and 2 provide evidence that when it is pay-off favorable, subjects distort their answers to the survey, but only for those questions in which they are significantly *less* certain of the accuracy of their response; that is, subjects display self-serving dishonesty. However, we find no evidence consistent with *strategic* self-serving dishonesty—subjects' confidence is unaffected by the presence of incentives to lie. This suggests that subjects respond to their exogenous level of uncertainty by "filling-in" their uncertainty in a self-serving way.

3.4 Do honor codes mitigate self-serving dishonesty?

Self-serving dishonesty suggests that subjects distort their answers in a self-serving way when they are uncertain about the truth, but it is also possible for dishonesty to arise when individuals are uncertain about what is the morally appropriate behaviour. For example, individuals may argue that it is not clear what behavior is expected from them in a particular context. To mitigate dishonesty stemming from moral ambiguity with respect to rule clarity, we implement a set of Treatments including an honor statement that explicitly informs subjects that dishonesty is inappropriate and asks them to agree to behave honestly. We focus on the "Miles" and "Value" questions for which subjects are significantly less certain of their answers, since the vast majority of dishonesty stems from these two questions.[4]

[4] Table A2 replicates Table 5 for the other 9 questions in the survey.

Result 4 The interventions targeting moral ambiguity reduce total dishonesty in the Miles question by $2.20 (16%) and are ineffective at mitigating dishonesty in the "Values" question. Of the honor code interventions, the Honor Code with the Check Box and the Honor Code with the Signature on the same screen were effective at decreasing dishonesty in the "Miles" question.

In Table 5 we report coefficients from OLS regressions in which we regressed responses to the "Value" and "Miles" question on a dummy for assignment to the Honor Code Interventions and the Control Treatment using assignment to the Incentive Treatments as the omitted category. The coefficients for each variable represent the additional money added to their bonus (deducted from their overall premium) by their reports *relative to* the Incentive Treatments. In column (1) we find that the Honor Code interventions, on average, do not significantly change the responses to the Value question relative to the Incentive Treatments. Similarly, in column (2) we disaggregate each of the Honor Code interventions and find that each of the individual interventions added approximately $6.80-$7.30 to their bonus relative to the Control Treatment, but also that none of the individual interventions mitigated the dishonesty motivated by financial incentives.

However, columns (3) and (4) show that the Honor Code interventions had a significant impact on mitigating dishonesty motivated by financial incentives in the "Miles" question. Column (3) reports that dishonesty due to financial incentives is mitigated by $2.20 or 16% ($2.20/$13.68 of the total change in the bonuses due to the financial incentives) and column (4) shows that this was

Table 5 SELF-SERVING DISHONESTY & HONOR CODES

	Value	Value	Miles	Miles
All Honor Code Interventions	-0.65	.	-2.20**	.
	(0.96)		(1.04)	
Honor Code with Check Box	.	-0.24	.	-2.50
		(1.41)		(1.54)
Honor Code with Signature on Prev Screen	.	-0.74	.	-0.87
		(1.26)		(1.40)
Honor Code with Signature on Same Screen	.	-0.97	.	-3.23**
		(1.40)		(1.57)
Control Treatment	-7.55***	-7.55***	-13.68***	-13.68***
	(1.48)	(1.49)	(1.51)	(1.51)
Constant	-13.38***	-13.38***	-26.38***	-26.38***
	(0.69)	(0.69)	(0.71)	(0.71)
Observations	1069	1069	1069	1069
R^2	0.03	0.03	0.07	0.08
Omitted Group	Incentive Treatments			

OLS regression estimates. Columns (1) and (3) pool the honor code interventions, while columns (2) and (4) consider each of the three honor code interventions separately. Robust standard errors in parentheses and *, ** and *** indicate statistical significance at the 10%, 5% and 1% levels, respectively

driven by the Honor Code with the Check Box Treatment and the Honor Code with a Signature Treatment. Our results in column (4) are contrary to Shu et al. (2012) who find that those who sign such a veracity statement at the beginning of a document, rather than at the end, are more likely to act honestly.

4 Conclusion

We provide experimental evidence that questionnaire items asking for information that respondents may not be sure about are prone to trigger responses that are biased in respondents' monetary favour. In particular, we ask a sample of U.S. car owners to respond to a questionnaire that resembles an auto insurance underwriting questionnaire. When facing monetary incentives to indicate particular responses (reflecting the calculation of insurance premiums based on the information provided), we observe dishonesty, but almost entirely arising for the questions where participants are not sure about the correctness of the response they indicated. Participants are thus behaving consistently with "self-serving dishonesty", meaning they respond to their uncertainty about the objectively correct response in a self-serving manner and choose responses that are more financially beneficial for them. Interventions that target commonly used strategies to justify dishonest behaviour like exploiting moral ambiguity are only minimally effective in reducing dishonesty.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11166-022-09380-1.

Acknowledgements We acknowledge funding from University of Sydney Faculty of Arts and Social Sciences and helpful comments and conversations.

Funding Open Access funding enabled and organized by CAUL and its Member Institutions. This project was funded by the University of Sydney Faculty of Arts and Social Sciences internal grant scheme.

Data availability All data, software for analyses and experimental instructions are available upon request.

Declarations

Conflict of interest There are no conflicts of interests for any of the authors.

Ethical approval The experiment received ethics approval from the University of Sydney Human Research Ethics Council.

Consent participate All participants read a consent statement and agreed to participate before beginning the experimental study.

References

Abeler, J., Nosenzo, D., & Raymond, C. (2019). Preferences for truth-telling. *Econometrica, 87,* 1115–1153.

Ayal, S., Gino, F., Barkan, R., & Ariely, D. (2015). Three principles to revise people's unethical behavior. *Perspectives on Psychological Science, 10,* 738–741.

Bandura, A. (1999). Moral disengagement in the perpetration of inhumanities. *Personality and social psychology review, 3,* 193–209.

Becker, G. S. (1968). Crime and punishment: An economic approach. In *The economic dimensions of crime* (pp. 13–68). Springer.

Bellé, N., & Cantarelli, P. (2017). What causes unethical behavior? a meta-analysis to set an agenda for public administration research. *Public Administration Review, 77,* 327–339.

Bénabou, R., & Tirole, J. (2016). Mindful economics: The production, consumption, and value of beliefs. *Journal of Economic Perspectives, 30,* 141–64.

Chou, E. Y. (2015a). Paperless and soulless: E-signatures diminish the signer's presence and decrease acceptance. *Social Psychological and Personality Science, 6,* 343–351.

Chou, E. Y. (2015b). What's in a name? the toll e-signatures take on individual honesty. *Journal of Experimental Social Psychology, 61,* 84–95.

Cressey, D. R. (1986). Why managers commit fraud. *Australian & New Zealand Journal of Criminology, 19,* 195–209.

Dana, J., Weber, R. A., & Kuang, J. X. (2007). Exploiting moral wiggle room: experiments demonstrating an illusory preference for fairness. *Economic Theory, 33,* 67–80.

Exley, C. L. (2016). Excusing selfishness in charitable giving: The role of risk. *The Review of Economic Studies, 83,* 587–628.

Exley, C. L., & Kessler, J. B. (2019). *Motivated errors*. Technical Report National Bureau of Economic Research.

Fukukawa, K. (2002). Developing a framework for ethicallyquestionable behavior in consumption. *Journal of Business Ethics, 41,* 99–119.

Gino, F., Norton, M. I., & Weber, R. A. (2016). Motivated bayesians: Feeling moral while acting egoistically. *Journal of Economic Perspectives, 30,* 189–212.

Gneezy, U., Saccardo, S., Serra-Garcia, M., & van Veldhuizen, R. (2020). Bribing the self. *Games and Economic Behavior, 120,* 311–324.

Golman, R., Hagmann, D., & Loewenstein, G. (2017). Information avoidance. *Journal of Economic Literature, 55,* 96–135.

Grossman, Z., & Van Der Weele, J. J. (2017). Self-image and willful ignorance in social decisions. *Journal of the European Economic Association, 15,* 173–217.

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

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological science, 23,* 524–532.

Köneke, V., Müller-Peters, H., Fetchenhauer, D., et al. (2015). *Versicherungsbetrug verstehen und verhindern*. Springer.

Konow, J. (2000). Fair shares: Accountability and cognitive dissonance in allocation decisions. *American economic review, 90,* 1072–1091.

Kristal, A. S., Whillans, A. V., Bazerman, M. H., Gino, F., Shu, L. L., Mazar, N., & Ariely, D. (2020). Signing at the beginning versus at the end does not decrease dishonesty. *Proceedings of the National Academy of Sciences, 117,* 7103–7107.

Kunda, Z. (1990). The case for motivated reasoning. *Psychological bulletin, 108,* 480.

List, J. A., Sadoff, S., & Wagner, M. (2011). So you want to run an experiment, now what? some simple rules of thumb for optimal experimental design. *Experimental Economics, 14,* 439.

Maggian, V., & Villeval, M. C. (2016). Social preferences and lying aversion in children. *Experimental Economics, 19,* 663–685.

Mazar, N., Amir, O., & Ariely, D. (2008). The dishonesty of honest people: A theory of self-concept maintenance. *Journal of marketing research, 45,* 633–644.

Oster, E., Shoulson, I., & Dorsey, E. (2013). Optimal expectations and limited medical testing: evidence from huntington disease. *American Economic Review, 103,* 804–30.

Pittarello, A., Leib, M., Gordon-Hecker, T., & Shalvi, S. (2015). Justifications shape ethical blind spots. *Psychological science, 26,* 794–804.

Rabin, M. (1995). Moral preferences, moral constraints, and self-serving biases.

Schwardmann, P., & Van der Weele, J. (2019). Deception and self-deception. *Nature human behaviour, 3,* 1055–1061.

Schweitzer, M. E., & Hsee, C. K. (2002). Stretching the truth: Elastic justification and motivated communication of uncertain information. *Journal of Risk and Uncertainty, 25,* 185–201.

Shalvi, S., Gino, F., Barkan, R., & Ayal, S. (2015). Self-serving justifications: Doing wrong and feeling moral. *Current Directions in Psychological Science, 24,* 125–130.

Shu, L. L., Gino, F., & Bazerman, M. H. (2011). Dishonest deed, clear conscience: When cheating leads to moral disengagement and motivated forgetting. *Personality and social psychology bulletin, 37,* 330–349.

Shu, L. L., Mazar, N., Gino, F., Ariely, D., & Bazerman, M. H. (2012). Signing at the beginning makes ethics salient and decreases dishonest self-reports in comparison to signing at the end. *Proceedings of the National Academy of Sciences, 109,* 15197–15200.

Sykes, G. M., & Matza, D. (1957). Techniques of neutralization: A theory of delinquency. *American sociological review, 22,* 664–670.

Thunström, L., Nordström, J., Shogren, J. F., Ehmke, M., & van't Veld, K. (2016). Strategic self-ignorance. *Journal of Risk and Uncertainty, 52,* 117–136.