http://pubsonline.informs.org/journal/mnsc

MANAGEMENT SCIENCE
Articles in Advance, pp. 1–11
ISSN 0025-1909 (print), ISSN 1526-5501 (online)

Mis-Nudging Morality

Coby Morvinski,[a,*] **Silvia Saccardo,**[b] **On Amir**[c]

[a] Ben-Gurion University, Be'er Sheva 8410501, Israel; [b] Carnegie Mellon University, Pittsburgh, Pennsylvania 15213; [c] University of California San Diego, La Jolla, California 92093
*Corresponding author

Contact: cobym@bgu.ac.il, https://orcid.org/0000-0001-8081-7085 (CM); ssaccard@andrew.cmu.edu, https://orcid.org/0000-0001-7560-8262 (SS); oamir@ucsd.edu, https://orcid.org/0000-0002-0364-8027 (OA)

Received: November 5, 2020
Revised: June 17, 2021; August 28, 2021
Accepted: October 20, 2021
Published Online in Articles in Advance: March 24, 2022

https://doi.org/10.1287/mnsc.2022.4344

Abstract. Morals constrain self-serving behavior. Yet, self-regulation failures in the face of monetary temptation are common at the workplace. To limit such failures, organizations can design environments that limit the temptation to behave self-servingly, nudging workers to uphold their morals. In a series of experiments where participants may be tempted to take excessive pay after exerting effort, we study whether a simple intervention—asking individuals to state the wage they believe should be paid ex ante, before facing the temptation to take excessive compensation—prevents self-serving behavior. In contrast to lay beliefs and the predictions from prior work, we find that such an intervention is not effective, leading to self-serving behavior. However, a more realistic elicitation procedure of the appropriate wage mitigates this effect. These findings contribute to work on the malleability of moral behavior showing that simple interventions thought to effectively mitigate self-serving behavior can prompt individuals to stretch their moral boundaries. They also stress the importance of properly testing interventions that might seem intuitive.

History: Accepted by Yan Chen, behavioral economics and decision analysis.

Funding: Financial support from the Israel Science Foundation [Grant 766/19] is gratefully acknowledged.
Supplemental Material: The online appendix and data are available at https://doi.org/10.1287/mnsc.2022.4344.

Keywords: nudges • self-serving behavior • morals

> The road to hell is paved with good intentions.
> —Henry G. Bohn (1855)

1. Introduction

Individuals strive to perceive themselves as fair, moral, and ethical (Bénabou and Tirole 2004, Mazar et al. 2008), which helps constrain the temptation to misbehave (e.g., Abeler et al. 2019). These self-constraints are important for the effective functioning of organizations, where monitoring and standard enforcement are not always feasible (Greenberg 2002). On top of mitigating blatant misconduct, such as stealing resources from the workplace, these motives can also mitigate subtler transgressions that can be detrimental to organizations' goals, such as claiming undeserved credit or exhibiting excessive entitlement to rewards. Yet, research shows that what people construe as fair or moral is flexible: Individuals find ways to justify self-serving behavior that violates their own moral standards (e.g., Babcock et al. 1995, Dana et al. 2007, and Gneezy et al. 2019). Finding ways to design work environments that promote ethicality, nudging individuals to uphold their own morals or fairness standards, is therefore an important challenge for organizations.

A growing body of work shows that interventions that leverage behavioral science insights, such as sending moral reminders (e.g., Bursztyn et al. 2019), or restructuring the environment where individuals make decision by, for example, leveraging defaults (Mazar and Hawkins 2015) or altering the order in which they receive information (e.g., Gneezy et al. 2020 and Saccardo and Serra-Garcia 2021) can effectively reduce self-serving behavior. Yet, not all behaviorally informed interventions are successful in every domain (Kristal and Whillans 2020), and some may even backfire (Monin and Miller 2001, Schultz et al. 2007, Sachdeva et al. 2009, Beshears et al. 2015, Damgaard and Gravert 2018, Pierce et al. 2020). On top of that, individuals often fail to anticipate what policies or interventions are needed or whether they will be effective (e.g., Cain et al. 2005, Goswami and Urminsky 2016,

DellaVigna and Pope 2018, Daniels and Zlatev 2019, DellaVigna and Linos 2020, and Reiff et al. 2020). In this paper, we document the unintended effect of a behavioral intervention that both laypeople and experienced managers expect to mitigate anticipated self-serving behavior.

Focusing on workplace entitlement, the tendency to demand or expect excessive rewards relative to effort or performance (e.g., Harvey and Martinko 2009), we conduct a set of experiments where individuals face a temptation to behave self-servingly. In the experiments, individuals are hired to work on a real-effort task and choose their own payment after being informed about the "standard" compensation rate. Specifically, upon completing the task, participants privately take as much compensation as they want from a bowl full of money. In this setting, temptation may lead individuals to behave self-servingly, taking more money than they deserve. This take-what-you-want setting is inspired by Greenberg (2002), who used a similar setup to investigate overpayment as a form of stealing. Based on work suggesting that moral transgressions become more likely when financial considerations are made salient (Gino and Pierce 2009, Gino et al. 2011, Kouchaki et al. 2013, Gneezy et al. 2019), individuals asked to pay themselves might be tempted to stretch the boundaries of their moral standards by taking more than appropriate. An incentive-compatible norm-elicitation experiment confirms that individuals consider taking more money than the standard wage as a norm violation (Online Appendix A).

In this setting, we study how a simple intervention—asking individuals to make a nonbinding statement about the payment they should receive ex ante, before facing the temptation of taking excessive pay from the bowl—affects self-serving behavior. Such an intervention is potentially effective in preventing individuals from violating their own moral/fairness standards for two main reasons: First, prior work has suggested that making norms of what is fair or ethical top of mind could raise attention to moral standards (Mazar et al. 2008, Epley and Tannenbaum 2017). Asking people to consider their actions in advance, before facing temptation, may increase the salience of fairness/ethical considerations over self-interest (Tenbrunsel et al. 2010), helping them resist "want" choices (Milkman et al. 2008). Second, research on goal attainment has shown that adherence to goals increases when the target behavior is specified ex ante (Gollwitzer 1999, O'Hora and Maglieri 2006) because of people's intrinsic need for consistency (Cialdini et al. 1995). Further, asking people to state their intentions before they face temptation to misbehave could work as a nonbinding precommitment nudge, encouraging follow-through (Tenbrunsel et al. 2010, Rogers et al. 2013, Zhang et al. 2014).[1] Similarly, nudges asking participants to state what the moral behavior should be ex ante have been shown to increase generosity in economic games (Capraro et al. 2019). In line with the predictions from this work, two pilot studies that describe the experimental setting and ask individuals to predict the effect of this intervention (relative to a setting without the intervention) find that both lay participants and experienced managers expect that stating the appropriate compensation in advance, prior to enacting this decision, should successfully mitigate the temptation to take excessive pay.

In contrast with these predictions and lay beliefs, four incentive-compatible experiments, including a high-powered preregistered study, show that introducing this intervention enables, rather than constrains, moral flexibility, leading individuals to take more money than they would have otherwise taken. We note that although individuals expect self-serving behavior in this context, baseline behavior is not self-serving, as, on average, individuals take a fair compensation. This result is in line with other work documenting how morals often constrain the temptation to misbehave for private gains (e.g., Abeler et al. 2019). However, introducing the intervention substantially increases the average amount taken, enabling individuals to be more self-serving. This finding is robust to different real-effort tasks and elicitation procedures. Our results further suggest that the unintended effect of expressing an appropriate wage ex ante can be explained by research on the hypothetical bias, a systematic discrepancy between hypothetical statements and actual behavior that leads people to underestimate the emotional impact that "being in the situation" would have on their behavior (List and Gallet 2001, Murphy et al. 2005). Indeed, we show that reducing the hypothetical nature of the elicitation procedure and making it more realistic (Morales et al. 2017) mitigate the effect.

Past work has highlighted the malleability of norms of fairness and ethical conduct. In this paper, we document one new way in which these norms can be unexpectedly stretched. Our findings show that an intervention expected to effectively mitigate anticipated self-serving behavior does not work as intended, prompting individuals to take more, rather than less, money. Further, our findings contribute to recent work on predicting the effects of nudges, which has documented how predictions from experts and policymakers are sometimes biased (e.g., Daniels and Zlatev 2019 and DellaVigna and Linos 2020). In our experiments, those making predictions may fail to anticipate the extent to which morals prevent self-serving behavior in the absence of the intervention or may not expect hypothetical statements about appropriate behavior to backfire, ultimately recommending the deployment of an intervention that has unintended

effects. These results underscore the importance of testing how choice architecture and behavioral interventions affect actual behavior, rather than merely relying on intuition. Finally, these findings have implications for the design of interventions intended to prevent norm transgressions, stressing the importance of designing interventions that closely resemble the decision environment where individuals might be tempted to "misbehave," so as not to suffer unintended consequences.

2. Experimental Setting and Pilot Studies

2.1. Setting

We design incentive-compatible experiments in which individuals work on a task and take their desired compensation from a bowl full of money. In the experiments, we study how an intervention asking individuals to express the appropriate wage before having a chance to take the money affects self-serving behavior. This setting is designed to mimic workplace situations in which employees might demand or expect rewards disproportionate to their effort or performance.

In our experiments, participants are first asked to complete a short version of a task (about one minute) for a fixed wage (norm-setting stage). Upon completing the first task, participants then complete a longer version of the task (about five minutes) without specific compensation information. Instead, they learn that they will receive their compensation at the end of the study. The first round serves as a calibration to reduce variability in the perceived difficulty of the task and anchor all participants to the same reference wage (i.e., norm). Upon completing the second task, participants proceed to a different room, where they take their compensation in private (in one online study, privacy was naturally part of the setting). That is, participants are asked to take as much money as they think they deserve from a bowl filled with $10 in quarters. This setting creates a tension between maximizing payoffs and taking a fair wage because individuals may be tempted to behave self-servingly—that is, to take more than they deserve. As described above, a norm-elicitation experiment, where we describe this setting and ask participants to indicate the moral appropriateness of taking any given amount, confirms that individuals consider taking more money than the standard wage as significantly less appropriate than taking the standard wage (see Online Appendix A).

Using this setting, we study how making an ex ante explicit statement about the appropriate wage affects subsequent behavior. First, we present two pilot studies demonstrating that laypeople expect expressing an appropriate wage before self-payment to be an effective intervention for mitigating subsequent self-serving behavior. In the next section, we document the actual effect of making such a statement on behavior.

2.2. Pilot Study 1—Prediction

We asked 200 Amazon Mechanical Turk (MTurk) participants (M_{age} = 36.2; 39.5% females) to imagine a hypothetical laboratory participant, John, participating in an experiment with the setting described above. We then described two potential scenarios: one in which John expressed, in writing, how much money he should have been paid before heading to the payment room (Ex Ante Statement) and another in which John did not state anything in advance (Baseline). Participants indicated which of the two courses of action, if any, would be more effective at mitigating John "taking more money than he deserves" and provided reasons for their choice. For generalizability, we presented the same scenario (preregistered; https://aspredicted.org/y8cb5.pdf) to 40 MBA students from a southwestern university who had an average work experience of 9.38 years.

2.2.1. Results. A majority of participants (75%, n = 149 in the MTurk sample; and 92.5%, n = 37 in the MBA sample) indicated that the scenario in which John would write down the wage he believed he should be paid ex ante would be the most effective at mitigating self-serving behavior. These proportions are significantly higher than chance ($p < 0.001$). The remaining participants were divided nearly equally between having John take the money directly (N_{MTurk} = 24 and N_{MBA} = 2) or indicating both courses of action as equally effective (N_{MTurk} = 26 and N_{MBA} = 1). In the MBA sample, the results are virtually the same if we restrict the analysis to participants with managerial experience or supervisory responsibilities (Online Appendix C). Taken together, these results suggest that both lay individuals and experienced managers expect that asking workers to make an ex ante statement about the appropriate wage to be a desirable approach for mitigating anticipated self-serving behavior.

2.3. Pilot Study 2—Preferences

A second pilot study investigates whether participants would choose themselves to make an explicit statement about the appropriate compensation before taking the money. Ninety undergraduate students completed the study in the laboratory for course credit (M_{age} = 21.3; 61% females). To make sure that participants experienced working on the task themselves, all participants completed short and long versions of a real effort task (listening to an annoying sound) before indicating whether they would prefer to make a statement about the appropriate compensation before taking the money. Participants imagined that the compensation for the short task was $0.25. Upon completing the longer version of the task, they indicated whether, hypothetically, they would prefer to directly proceed to the

payment room or to first write down how much they should be paid (the order of choices was counterbalanced). Participants did not receive the explicit goal of avoiding self-serving behavior as in the first pilot study because we were interested in observing the reasons participants naturally give for choosing to make an ex ante statement (versus not). Following their choice, participants were asked to indicate the reason for their preferences in an open-ended response. Finally, participants indicated the amount of money that they believed they should be (hypothetically) paid for the second task.

2.3.1. Results. Participants were nearly equally divided in their choices, with 44 participants (48.9%) showing a preference for the ex ante statement. Most participants indicated that this course of action would help them stay moral, honest, be less greedy, or be less tempted to take money. We coded each entry for whether it reflected (a) a desire to limit the amount of money taken, (b) a desire not to limit the amount taken, (c) indifference, or (d) none of the above (see coding methods and data in Online Appendix B). On average, 54% of participants choosing to express the appropriate amount provided reasons that reflected an intent to avoid taking more money than needed, as opposed to only 13% of those who preferred taking the money directly ($t(88) = 7.54$, $p < 0.001$). Conversely, the main reasons for taking money from the bowl directly were that participants did not want to be limited in how much money they would take (31%) and that they were indifferent between the courses of actions (30%); these values were 16% ($t(88) = 3.38$, $p = 0.001$) and 11% for participants who favored expressing the appropriate wage beforehand. Taken together, these results suggest that the vast majority of participants expected expressing the appropriate wage ex ante to limit self-serving behavior. Finally, the amount of money participants believed they should be paid did not differ, regardless of the chosen course of action (Med_{State} = \$2.00, $Mean_{State}$ = 3.131 [SD_{State} = 2.546]; $Med_{NotState}$ = \$2.00, $Mean_{NotState}$ = 2.377 [$SD_{NotState}$ = 1.505]; p = 0.328, Mann-Whitney[2]), suggesting that differential feelings of entitlement did not drive preferences for ex ante statement.

Together, Pilot Studies 1 and 2 demonstrate that people believe that expressing an appropriate wage ex ante would mitigate anticipated self-serving behavior. Next, we present three consequential studies showing that people's intuition is, in fact, in stark contrast to their behavior and attempt to shed light on the reasons for this counterintuitive effect. Study 1 documents the effect of expressing an appropriate wage ex ante on subsequent behavior. Studies 2a and 2b replicate the results from Study 1 and document a moderating factor. Complete experimental instructions are in the online appendix.

3. Incentive-Compatible Experiments

3.1. Study 1—Do Explicit Statements Prior to Actual Behavior Mitigate Self-Serving Tendencies?

Study 1 investigates how asking workers to make an explicit statement about the wage they believe they should be paid *before* they have a chance to take the money affects the chosen compensation.

3.1.1. Design. Participants from a southwestern university (n = 82, M_{age} = 21.6, 58.5% females) completed the study for course credit at the end of a one-hour laboratory session and were informed that they would receive payment for this portion of the study.[3] Participants completed two rounds of a task requiring them to listen to an unpleasant noise and count the number of embedded bell rings. Participants completed a short (67 seconds, four rings) version of the task for \$0.25 (norm-setting stage), followed by a long (285 seconds, 11 rings) version of the same task. After completing the task, the experimenter led participants, one by one, to a small office containing a table with a bowl filled with \$10 in quarters. A note placed on the table read, "Please take \$0.25 for the first listening task, plus what you believe you should be paid for the second listening task." After taking their money, participants left the laboratory. For each participant, the experimenter counted how much money was left in the bowl.[4]

In the *Baseline* treatment, immediately after completing the final task, participants proceeded to the payment room, chose their payment, and left. In the *Statement* treatment, after completing the final task, we asked participants to indicate in an open-ended question how much they "believed [they] should be paid for the second listening task." This nonbinding statement may be interpreted as simple "cheap talk" and thus be irrelevant to subsequent behavior; alternatively, it may induce a sense of commitment and influence subsequent behavior (Milkman et al. 2011). After writing the amount, participants proceeded to the payment room, where they picked their desired payment from the bowl.

In both conditions, participants were informed that they would be paid at the end of the study, but the payment procedure was not revealed to them in advance. All participants also completed a demographic survey, an adapted scale of dispositional greed (Seuntjens et al. 2015), and a measure of self-perceived social status (Anderson et al. 2012). We collected these measures about 45 minutes prior to our main study.

3.1.2. Results. Does making an explicit statement prior to actual behavior mitigate self-serving behavior? It does not. On average, participants in the *Statement* condition took *more than twice* as much money as those in the *Baseline* condition ($Mean_{Statement}$ = \$2.35 [$SD_{Statement}$ = 2.728], $Mean_{Baseline}$ = \$1.03 [$SD_{Baseline}$ = 0.998], $t(80)=2.96$, $p = 0.004$). Figure 1 shows the distribution of payments. Given that these distributions are largely skewed, we use nonparametric tests hereafter. A Mann-Whitney *U*-test leads to a similar conclusion ($Med_{Statement}$ = \$1.25, $Med_{Baseline}$ = \$0.75, $p = 0.005$).

We regress the amount taken on a treatment dummy (Table 1). Columns (A) and (B) report the results of quantile regressions and confirm that the median amount taken is higher in the *Statement* treatment ($p = 0.040$), even when controlling for dispositional greed, perceived sociometric status, gender, age, and native language ($p = 0.012$). Columns (C) and (D) replicate the results with ordinary least-squares (OLS) regressions, where we log-transformed the amount taken.[5] None of the other measures significantly predicted the amount taken. In the online appendices, we present additional robustness analyses for all consequential studies.

By looking at the stated amounts, we investigate why an ex ante statement leads individuals to take more, rather than less, money. Both the average and median stated amounts were higher than the actual amount participants in this treatment took (Med_{Stated} = \$1.75, $Mean_{Stated}$ = \$3.65 [SD_{Stated} = 3.60] versus Med_{Taken} = \$1.25, $Mean_{Taken}$ = \$2.35 [SD_{Taken} = 2.73], $p = 0.001$, Wilcoxon). We find evidence that stated amounts correlate with the amounts of money participants take, suggesting that these amounts might serve as a reference value when determining actual wages. The higher the stated amount, the higher the taken amount ($\beta = 0.728$, $t(33) = 4.28$, $p < 0.001$, OLS regression). Online Appendix C shows that participants either take the amount they stated or adjust downward.

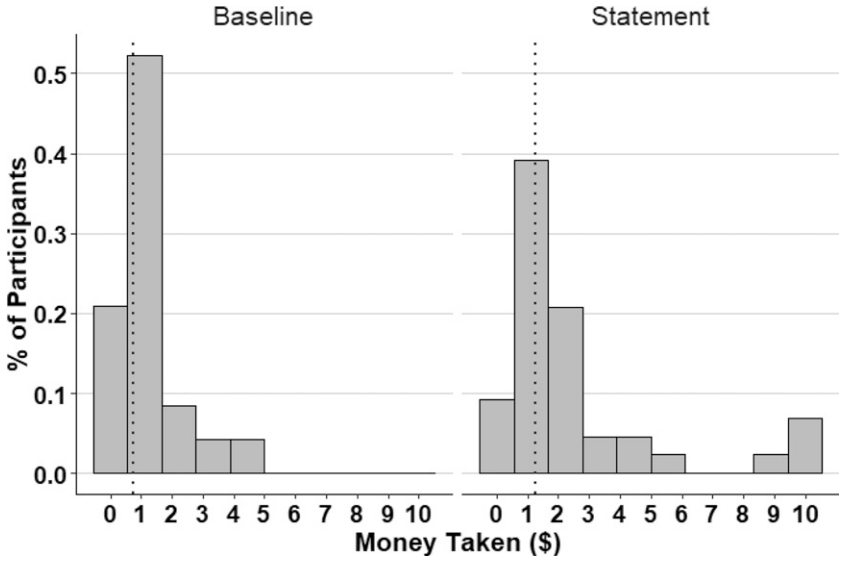
Table 1. Study 1 Regressions Results

Variable	(A)	(B)	(C)	(D)
Statement	0.500*	0.460*	0.732*	0.697†
	(0.239)	(0.229)	(0.358)	(0.382)
Dispositional Greed		0.232		0.087
		(0.182)		(0.269)
Sociometric Status		0.190		0.096
		(0.135)		(0.233)
Female		−0.381†		−0.418
		(0.206)		(0.367)
Age		−0.129		0.181
		(0.082)		(0.113)
English		0.190		0.240
		(0.210)		(0.426)
Constant	0.750***	−3.399	−0.609*	−5.161†
	(0.086)	(2.15)	(0.247)	(2.928)
N	82	82	82	82
R^2	0.027	0.070	0.038	0.033

Notes. Columns (A) and (B) use quantile regressions, and columns (C) and (D) use log-transformed OLS regressions. *Statement* is a dummy variable coded as one for participants in treatment Statement, and zero otherwise. *Dispositional Greed* is constructed by averaging the adapted Dispositional Greed Scale items. *Sociometric Status* is constructed by averaging the perceived sociometric status items. *Female* is a dummy variable coded as one for female participants, and zero otherwise. *English* is a dummy variable coded as one for native speakers, and zero otherwise. Pseudo R^2 based on Koenker and Machado (1999) is shown for median regressions, and Adjusted R^2 is shown for OLS regressions. Standard errors are presented in parentheses below parameter estimates.
†$p < 0.1$; *$p < 0.05$; ***$p < 0.001$.

3.1.2.1. Discussion. Study 1 shows that, contrary to lay belief that a nonbinding statement may constrain later temptation, participants who first made a

Figure 1. Study 1 Distribution of Money Taken by Treatment



nonbinding statement about how much they thought they should be paid left the laboratory with more money than those who did not. This result is driven by participants *stating* larger amounts than what they would otherwise take. The observed gap between stated and taken amounts suggests that individuals adjust downward when taking their money, but do it insufficiently, and thus end up taking more than those who do not state any amount. Importantly, *Baseline* participants conform to the norm, suggesting that *Statement* participants are violating the established moral standard, behaving more self-servingly.

In Study A2, fully reported in Online Appendix D, we replicate this finding using a different real-effort task and explore four additional variations of the treatments to address potential confounds, as well as document a boundary condition. Other than changing the task, we followed the procedures of Study 1, where participants first work on a short task for \$0.25 and then choose their own compensation for a longer task. Study A2's results replicate those of Study 1, even when controlling for actual performance in the task. This result persists when bounding statements to the actual amount of money in the bowl (\$10), but not when individuals are asked to simply "think about the value" as opposed to writing it down. This latter result is in line with Falk and Zimmermann (2018), who only find consistency when individuals state an initial amount in writing, but not when asked to merely think about it. The study also shows that taking more money in the *Statement* treatment is not a way to compensate for the extra time people spent in the laboratory in this treatment, as compared with the *Baseline* treatment. Finally, it shows that changing the payment medium to include both coins and dollar bills does not significantly increase the amount taken.

3.2. Study 2—Aligning Elicitation Procedure with the Actual Experience

Why might, counter to people's intuitions, asking individuals to state the appropriate wage ex ante backfire? The results from Studies 1 and A2 suggest that participants state high amounts, but adjust downward when actually taking the money. This finding suggests that the unintended effect of prestating an appropriate wage may stem from misalignment between the abstract statement and the actual money-taking experience. For example, when writing the wage participants should be paid ex ante, they may fail to consider how guilty they may feel when taking the money. The hypothetical nature of the elicitation procedure may lead participants to state an inflated compensation. A large body of literature has documented such hypothetical bias in valuations (Cummings and Taylor 1999, List 2001, Champ et al. 2009). After coming up with an estimate, participants use this constructed value as a reference for their subsequent behavior (the amount they take). This idea was also proposed by Murphy et al. (2005, p. 320) as an explanation for the hypothetical bias, the discrepancy between hypothetical and incentive-compatible decisions, suggesting that "participants might try to maintain some consistency between their hypothetical and actual values."

One way that policymakers can bridge the gap between hypothetical and real valuations is to better align the hypothetical procedure with the natural "look and feel" of a field domain (Bateman et al. 2009, Fiore et al. 2009). Building on this work, we expect that making the elicitation procedure resemble the actual situation in which individuals take money should reduce their overvaluation, leading them to state a lower ex ante wage than the one stated in the standard *Statement* condition, thus reducing self-serving behavior. That is, we predict that a more realistic appropriate-wage elicitation procedure will mitigate the previously observed unintended effect of stating the appropriate wage on subsequent behavior. In two different experimental environments, Study 2 replicates the effect documented in Study 1 and tests one debiasing technique—aligning the elicitation procedure with the actual experience.

3.2.1. Design.

3.2.1.1. Study 2a—Laboratory Environment. We replicate the design and procedures from Study 1 with participants from the same subject pool ($n = 220$, $M_{age} = 21.2$, 37.1% females) and test whether making the appropriate-wage elicitation procedure more similar to the actual money-taking experience attenuates self-serving behavior. In addition to the *Baseline* and *Statement* treatments, we add a treatment in which the elicitation procedure more closely resembles the act of taking money from the bowl. In the *High-Realism Statement* treatment (*Statement-HR*), instead of asking participants to enter an amount in writing, participants selected virtual quarters on the screen, one by one, until they reached the desired amount. Whereas stating an amount in writing is abstract, doing so by clicking on coins on the screen is more similar to taking money from the bowl and may evoke a similar response (Fiore et al. 2009). As such, we expected participants in this treatment to express a desire for a lower wage and to subsequently take less money than those in the (abstract) *Statement* condition. Because in this condition money is more vivid, a counterhypothesis could be that this treatment may not mitigate, and might even increase, stated amounts and subsequent self-serving behavior. As part of this study, we also included a treatment in which we attempted to mitigate the stated amount by setting a norm, where we informed participants that people in a different session took a median amount of \$1.00 (*Statement-Norm* treatment).

3.2.1.2. Study 2b—Virtual Environment. We replicate the *Baseline* and *Statement* as well as the more realistic *Statement-HR* treatments in an online environment, due to the restriction on in-person studies posed by the 2020 COVID-19 pandemic. The study design, sample size, and analysis plans were preregistered (https://aspredicted.org/bb76a.pdf). Participants from the same subject pool of Study 1 (n = 300, M_{age} = 20.91, 45.7% females) took part in the study for class credit and completed the tasks following the same instructions of Study 2a. We simulated taking money from a virtual bowl in an online environment (for a demonstration, see https://tinyurl.com/2p9ba39r). At the end of the study, the amount "taken" from the bowl was transferred to each participant via Venmo.

3.2.1.3. Additional Measures. After participants took the money from the virtual bowl, we collected additional secondary measures to better understand what guided their decisions. Specifically, we administrated an exploratory thought-listing protocol in which we asked participants to list all the thoughts they had when they first considered how much money to take. All participants also indicated how guilty they felt taking the amount they took (1 = *not at all*, 7 = *very much*). Finally, using a multiple-price list, participants indicated how guilty they thought they would feel had they taken amounts ranging from \$1 to \$10. Consistent with moral regulation theory, we predicted that participants would expect to feel guiltier the more money they took for themselves. However, we did not expect those in the *Statement* treatment to feel guiltier after taking more money than those in the *Baseline* condition because taking larger amounts in this condition is assumed to be intrinsically justified. Further, we used the thought-listing results to shed light on participants' intrinsic motivations to take larger amounts in the Statement treatment and investigated whether participants in this treatment indicated reasons that differed from those of the *Baseline* condition. Particularly, we anticipated those who already stated a relatively high nonbinding appropriate wage (versus baseline) to be less motivated to adhere to the normative wage set by the experimental design. Finally, participants in the two *Statement* conditions also indicated whether they thought the ex ante statement was intended to influence their behavior and, if so, in which direction (results of this measure can be found in the online appendix).

3.2.2. Results. The results are depicted in Figure 2 and Table 2, and regression analyses are reported in Table 3.

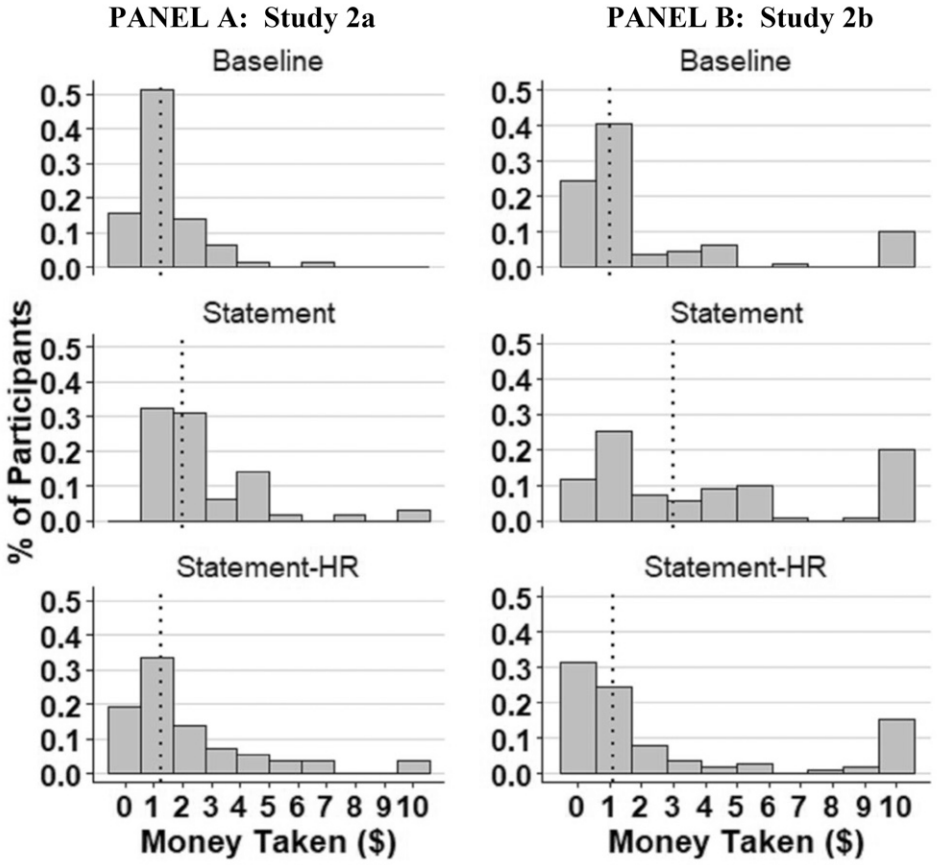
3.2.2.1. Study 2a. As in Study 1, participants in the *Statement* treatment took significantly more money than those in *Baseline* ($Med_{Baseline}$ = \$1.00, $Mean_{Baseline}$ = 1.20 [$SD_{Baseline}$ = 1.12]; versus $Med_{Statement}$ = \$1.75, $Mean_{Statement}$ = \$2.46 [$SD_{Statement}$ = 1.97], $p < 0.001$, Mann-Whitney). More importantly, as expected, our intervention achieved the intended goal of mitigating the unintended effect of stating an appropriate wage ex ante. Participants in *Statement-HR* left the study with significantly *less* money than those in the *Statement* treatment (Med_{HR} = \$1.00, $Mean_{HR}$ = \$1.99 [SD_{HR} = 2.24]; $p = 0.012$); the amount taken did not differ from the *Baseline* condition ($p = 0.114$; Mann-Whitney). Quantile regressions as well as log-transformed OLS regressions with and without controls confirm this result (Table 3, Panel A). As hypothesized, the difference between the two *Statement* treatments is driven by participants in the more realistic *Statement-HR* treatment stating lower amounts than those who stated an appropriate wage in writing ($p = 0.011$). Finally, the *Statement-Norm* condition did not reduce the amount taken, potentially because it failed to convey a norm of taking an amount close to the standard wage (Online Appendix C).

3.2.2.2. Study 2b. As in Studies 1 and 2a, participants in the *Statement* treatment ($Med_{Statement}$ = \$2.75, $Mean_{Statement}$ = \$3.89 [$SD_{Statement}$ = 3.57]) took significantly more money than those in *Baseline* ($Med_{Baseline}$ = \$0.75, $Mean_{Baseline}$ = 2.13 [$SD_{Baseline}$ = 2.99], $p < 0.001$, Mann-Whitney). Importantly, as in Study 2a, participants in the *Statement-HR* treatment (Med_{HR} = \$0.88; $Mean_{HR}$ = \$2.67 [SD_{HR} = 3.63]) did not take more money than those in *Baseline* ($p = 0.979$; Mann-Whitney) and took significantly less than those in the *Statement* treatment ($p < 0.001$; Mann-Whitney). Making the appropriate-wage elicitation procedure more realistic (*Statement-HR*) led participants to initially state smaller amounts (Med_{HR} = \$0.875; $Mean_{HR}$ = \$2.54 [SD_{HR} = 3.58] versus $Med_{Statement}$ = \$2.75; $Mean_{Statement}$ = \$3.88 [$SD_{Statement}$ = 3.35], $p < 0.001$; Mann-Whitney). In both treatments, the stated amount was highly correlated with the actual amount taken (Table 2 and Online Appendix C).

Quantile regressions and log-transformed OLS regressions confirm the treatment differences with respect to the amount taken (Table 3, Panel B). The table shows that compared with the *Baseline* treatment, the median (columns (A) and (B)) and average (columns (C) and (D)) amount taken were significantly larger in the *Statement* treatment. Importantly, participants who stated an appropriate wage by selecting quarters on the screen (*Statement-HR*) took similar amounts to those who did not state this wage.

3.2.2.3. Thought-Listing. Two coders blind to the hypotheses and treatments independently coded participants' thoughts following our preregistered coding scheme (see Online Appendix C). First, they indicated whether participants discussed reasons for taking less than, equal to, or more than the standard wage of

Figure 2. Study 2 Distribution of Money Taken by Treatment



Notes. (a) Study 2a. (b) Study 2b.

$0.25. Second, they indicated whether participants expressed an intention to adhere to acceptable norms of conduct (e.g., being fair or ethical) and whether they were trying to minimize negative feelings (e.g., guilt). Answers were not mutually exclusive.

With respect to the first measure (kappa = 0.60), the majority of participants (82%) referred to the standard wage in their answers. Participants in the *Baseline* treatment were more likely to mention reasons for taking less than the standard wage than participants in the *Statement* treatment (33% versus 14%, $t(198) = 3.63$, $p < 0.001$) and similarly likely to those in *Statement-HR* (29%, not significant, ns). Conversely, participants in the *Statement* treatment were more likely to mention reasons for taking more than the standard wage (39.5%) than those in *Baseline* (24.5%, $t(198) = 2.49$, $p = 0.014$) and as likely as those in the *Statement-HR* treatment (38.5%, ns).

In addition, the effect of treatment (*Baseline* versus *Statement*) on the actual amount taken is mediated by whether participants mention reasons for taking less than the standard wage (95% confidence interval (CI) [0.26, 1.05]) or more than the standard wage (95% CI [0.10, 1.08]). These results further support the hypothesis that writing an appropriate wage in advance steers participants away from what they would otherwise consider acceptable (see Online Appendix C for full analyses).

With respect to the second measure, we see that, on average, participants in the *Statement* condition were marginally less likely to state that they were trying to adhere to acceptable norms of conduct ($M_{\text{Baseline}} = 0.45$ versus $M_{\text{Statement}} = 0.36$, $t(198) = 1.71$, $p = 0.088$), but as likely as those in the *Statement-HR* treatment ($M = 0.39$, $p = 0.590$). Finally, we observed no difference across treatments in whether participants indicated a desire to minimize negative feelings ($F(2, 297) = 0.51$, $p = 0.601$, $\eta^2 = 0.003$).

3.2.2.4. Guilt. The analysis of guilt reveals that participants expected feelings of guilt to increase with amount taken ($b = 0.404$, $t(298) = 58.93$, $p < 0.001$; mixed effect). However, when assessing how guilty participants actually felt after taking the money, we

Table 2. Study 2 Summary Statistics

Condition	*N*	Percentile amount taken ($) 25th	50th	75th	Mean Taken ($)	Median stated ($)	Mean stated ($)	Deserved amount vs. money taken	Difference from baseline
Panel A: Study 2a									
Baseline	58	0.50	1.00	1.44	1.21	—	—	—	—
Statement	58	1.25	1.75	3.38	2.46	2.00	3.44	$p < 0.001$	$p < 0.001$
Statement-HR	51	0.75	1.00	2.50	1.99	1.25	2.69	$p = 0.015$	$p = 0.012$
Panel B: Study 2b									
Baseline	100	0.25	0.75	2.06	2.13	—	—	—	—
Statement	100	1.00	2.75	5.06	3.89	2.75	3.88	$p = 0.011$	$p < 0.001$
Statement-HR	100	0.25	0.88	3.38	2.70	0.875	2.54	$p = 0.122$	$p = 0.862$

Notes. The "deserved amount vs. money taken" column compares prestated amounts and actual money taken within the same subject group using a Wilcoxon signed-rank test. The "Difference from baseline" column compares prestated amounts of those in the relevant condition and actual money taken by those in the baseline condition using a Mann-Whitney *U*-test. Deserved amounts of four participants in Study 2a and three participants in Study 2b exceeded $10 and were winsorized to $10.

see no difference across treatments ($F(2, 297) = 0.505$, $p = 0.604$, $\eta^2 = 0.003$). This result suggests that although participants in the *Statement* treatment took significantly higher amounts, they did not feel guilty about doing so, presumably because adhering to the stated amount implicitly increased the compensation they felt was right. Additionally, communicating the appropriate wage to others (i.e., the experimenter) may facilitate justifying taking a similar amount without feeling guiltier, as merely thinking about the appropriate wage without communicating it did not increase the amount taken (Study A2).

Taken together, these findings suggest that stating an appropriate wage ex ante can increase self-serving behavior, altering normative ethics without increasing feelings of guilt. However, making this nonbinding statement and temptation more realistic can be effective in mitigating this effect. We note that in Study 2b, participants in the *Baseline* condition paid themselves immediately after they completed working on the tasks because they did not need to walk to a payment room. This design makes a "cooling off" explanation less likely: The notion that participants in the *Statement* treatment are in a "hot state," whereas those who walk to the payment room (*Baseline* treatment) are given a chance to cool off and therefore take less money, does not apply to this design.

Table 3. Study 2: Regressions Results

	Panel A: Study 2a (A)	(B)	(C)	(D)	Panel B: Study 2b (A)	(B)	(C)	(D)
Statement	0.750***	1.000***	0.991***	1.024***	2.00**	1.750**	0.980**	0.983**
	(0.184)	(0.249)	(0.251)	(0.248)	(0.609)	(0.658)	(0.311)	(0.310)
Statement-HR	0.000	0.250	0.250	0.366	0.250	0.000	−0.164	−0.181
	(0.192)	(0.270)	(0.259)	(0.263)	(0.193)	(0.243)	(0.311)	(0.311)
Statement-Norm	0.750***	0.750***	0.517*	0.511*				
	(0.189)	(0.163)	(0.256)	(0.256)				
Female		−0.250†		−0.110		−0.250		−0.335
		(0.146)		(0.194)		(0.256)		(0.257)
Age		0.000		0.056		0.000		−0.072
		(0.047)		(0.047)		(0.082)		(0.077)
Other payments		0.000		−0.025				
		(0.024)		(0.028)				
Constant	1.00***	1.00	−0.325†	−1.387	1.00***	1.25	−0.510*	1.145
	(0.102)	(0.963)	(0.177)	(1.048)	(0.086)	(1.756)	(0.220)	(1.639)
N	220	195	220	195	300	300	300	300
R^2	0.056	0.198	0.060	0.069	0.028	0.030	0.044	0.045

Notes. Dependent variables are money taken from the bowl for the second task. Columns (A) and (B) use median regressions performed with R and quantreg package using nid standard error estimation. Columns (C) and (D) use log-transformed OLS regressions. *Statement* and *Statement-HR* are dummy variables coded as one for the respective treatments. *Female* is a dummy variable coded as one for female participants, and zero otherwise. *Age* is a continuous variable indicating participants' age. Pseudo R^2 based on Koenker and Machado (1999) is shown for median regressions, and Adjusted R^2 is shown for OLS regressions. Standard errors are presented in parentheses below parameter estimates.
†$p < 0.1$; *$p < 0.05$; **$p < 0.01$; ***$p < 0.001$.

4. Conclusion

Self-serving behavior is often a consequence of self-regulation failures. Interventions such as precommitment nudges that ask individuals to state their preferences ex ante—before facing temptation to misbehave—might limit self-serving behavior by making future temptations costlier. But does this intervention always work as intended? In a series of consequential studies in which individuals pay themselves after performing work, we show that, contrary to lay beliefs, stating beliefs about the appropriate wage one should be paid ex ante can cause one to be *more* rather than less self-serving (Study 1). These findings are in line with work showing that a hypothetical statement of one's preference can unconsciously influence subsequent behavior (Fitzsimons and Shiv 2001). The above unintended effect is robust to several manipulations, but mitigated by a more realistic elicitation procedure that mimics the real experience of taking money (Studies 2a and 2b). Taken together, these findings contribute to a growing body of knowledge about the unintended consequences of behavioral nudges (e.g., Reiff et al. 2020) and to research documenting biased predictions about the effectiveness of behavioral interventions (DellaVigna and Pope 2018, Daniels and Zlatev 2019).

Our findings have direct implications for practitioners who seek to use nonbinding pledges, or, as in our case, statements about appropriate wages, to mitigate undesirable behaviors. In particular, they highlight the importance of carefully designing interventions that consider features of the environment in which individuals make decisions. As we demonstrate, different elicitation procedures can yield different outcomes. In some situations, abstract elicitations may produce unintended effects, as in the case of the self-serving behavior we documented. However, in other situations, such misalignment may be beneficial. For example, in the domain of charitable giving, individuals who pledge to donate a given amount in the future may underestimate the costs of giving. In such situations, abstract prestatement elicitations may be effective at raising larger donations. Future research could further our understanding of when different types of nonbinding statements about appropriate behavior are effective and when they may influence behavior in unintended ways.

Endnotes

[1] It is nonbinding because individuals ex ante state an intention to exhibit a given behavior, but are not prevented from behaving differently later.

[2] We exclude six participants who indicated demanded compensations that exceeded \$10; including all of them but one who indicated \$1 million did not change the results.

[3] Subjects completed all experiments in the session in the same order, with our experiment being last.

[4] New bowls were set out for each participant.

[5] Shapiro-Wilk tests show that the residuals are not normally distributed ($p < 0.001$).

References

Abeler J, Nosenzo D, Raymond C (2019) Preferences for truth-telling. *Econometrica* 87(4):1115–1153.

Anderson C, Kraus MW, Galinsky AD, Keltner D (2012) The local-ladder effect: Social status and subjective well-being. *Psych. Sci.* 23(7):764–771.

Babcock L, Loewenstein G, Issacharoff S, Camerer C (1995) Biased judgments of fairness in bargaining. *Amer. Econom. Rev.* 85(5): 1337–1343.

Bateman IJ, Day BH, Jones AP, Jude S (2009) Reducing gain–loss asymmetry: A virtual reality choice experiment valuing land use change. *J. Environ. Econom. Management* 58(1):106–118.

Bénabou R, Tirole J (2004) Willpower and personal rules. *J. Polit. Econom.* 112(4):848–886.

Beshears J, Choi JJ, Laibson D, Madrian BC, Milkman KL (2015) The effect of providing peer information on retirement savings decisions. *J. Finance* 70(3):1161–1201.

Bohn HG (1855) *A Hand-book of Proverbs: Comprising Ray's Collection of English Proverbs, with His Additions from Foreign Languages. And a Complete Alphabetical Index* (G. Bell).

Bursztyn L, Fiorin S, Gottlieb D, Kanz M (2019) Moral incentives in credit card debt repayment: Evidence from a field experiment. *J. Polit. Econom.* 127(4):1641–1683.

Cain DM, Loewenstein G, Moore DA (2005) The dirt on coming clean: Perverse effects of disclosing conflicts of interest. *J. Legal Stud.* 34(1):1–25.

Capraro V, Jagfeld G, Klein R, Mul M, Van De Pol I (2019) Increasing altruistic and cooperative behaviour with simple moral nudges. *Sci. Rep.* 9(1):1–11.

Champ PA, Moore R, Bishop RC (2009) A comparison of approaches to mitigate hypothetical bias. *Agricultural Resource Econom. Rev.* 38(2):166-180.

Cialdini RB, Trost MR, Newsom JT (1995) Preference for consistency: The development of a valid measure and the discovery of surprising behavioral implications. *J. Personality Soc. Psych.* 69(2):318–328.

Cummings RG, Taylor LO (1999) Unbiased value estimates for environmental goods: A cheap talk design for the contingent valuation method. *Amer. Econom. Rev.* 89(3):649–665.

Damgaard MT, Gravert C (2018) The hidden costs of nudging: Experimental evidence from reminders in fundraising. *J. Public Econom.* 157:15–26.

Dana J, Weber RA, Kuang JX (2007) Exploiting moral wiggle room: Experiments demonstrating an illusory preference for fairness. *Econom. Theory* 33(1):67–80.

Daniels DP, Zlatev JJ (2019) Choice architects reveal a bias toward positivity and certainty. *Organ. Behav. Human Decision Processes* 151:132–149.

DellaVigna S, Linos E (2022) RCTs to scale: Comprehensive evidence from two nudge units. *Econometrica* 90(1):81–116.

DellaVigna S, Pope D (2018) Predicting experimental results: who knows what? *J. Polit. Econom.* 126(6):2410–2456.

Epley N, Tannenbaum D (2017) Treating ethics as a design problem. *Behav. Sci. Policy* 3(2):72–84.

Falk A, Zimmermann F (2018) Information processing and commitment. *Econom. J.* 128(613):1983–2002.

Fiore SM, Harrison GW, Hughes CE, Rutström EE (2009) Virtual experiments and environmental policy. *J. Environ. Econom. Management* 57(1):65–86.

Fitzsimons GJ, Shiv B (2001) Nonconscious and contaminative effects of hypothetical questions on subsequent decision making. *J. Consumer Res.* 28(2):224–238.

Gino F, Pierce L (2009) The abundance effect: Unethical behavior in the presence of wealth. *Organ. Behav. Human Decision Processes* 109(2):142–155.

Gino F, Schweitzer ME, Mead NL, Ariely D (2011) Unable to resist temptation: How self-control depletion promotes unethical behavior. *Organ. Behav. Human Decision Processes* 115(2):191–203.

Gneezy U, Saccardo S, Van Veldhuizen R (2019) Bribery: Behavioral drivers of distorted decisions. *J. Eur. Econom. Assoc.* 17(3):917–946.

Gneezy U, Saccardo S, Serra-Garcia M, van Veldhuizen R (2020) Bribing the self. *Games Econom. Behav.* 120:311–324.

Gollwitzer PM (1999) Implementation intentions: Strong effects of simple plans. *Amer. Psych.* 54(7):493–503.

Goswami I, Urminsky O (2016) When should the ask be a nudge? The effect of default amounts on charitable donations. *J. Marketing Res.* 53(5):829–846.

Greenberg J (2002) Who stole the money, and when? Individual and situational determinants of employee theft. *Organ. Behav. Human Decision Processes* 89(1):985–1003.

Harvey P, Martinko MJ (2009) An empirical examination of the role of attributions in psychological entitlement and its outcomes. *J. Organ. Behav.* 30(4):459–476.

Koenker R, Machado JA (1999) Goodness of fit and related inference processes for quantile regression. *J. Amer. Statist. Assoc.* 94(448):1296–1310.

Kouchaki M, Smith-Crowe K, Brief AP, Sousa C (2013) Seeing green: Mere exposure to money triggers a business decision frame and unethical outcomes. *Organ. Behav. Human Decision Processes* 121(1):53–61.

Kristal AS, Whillans AV (2020) What we can learn from five naturalistic field experiments that failed to shift commuter behaviour. *Nature Human Behav.* 4(2):169–176.

List JA (2001) Do explicit warnings eliminate the hypothetical bias in elicitation procedures? Evidence from field auctions for sportscards. *Amer. Econom. Rev.* 91(5):1498–1507.

List JA, Gallet CA (2001) What experimental protocol influence disparities between actual and hypothetical stated values? *Environ. Resource Econom.* 20(3):241–254.

Mazar N, Hawkins SA (2015) Choice architecture in conflicts of interest: Defaults as physical and psychological barriers to (dis)honesty. *J. Experiment. Soc. Psych.* 59:113–117.

Mazar N, Amir O, Ariely D (2008) The dishonesty of honest people: A theory of self-concept maintenance. *J. Marketing Res.* 45(6):633–644.

Milkman KL, Rogers T, Bazerman MH (2008) Harnessing our inner angels and demons: What we have learned about want/should conflicts and how that knowledge can help us reduce short-sighted decision making. *Perspect. Psych. Sci.* 3(4):324–338.

Milkman KL, Beshears J, Choi JJ, Laibson D, Madrian BC (2011) Using implementation intentions prompts to enhance influenza vaccination rates. *Proc. Natl. Acad. Sci. USA* 108(26): 10415–10420.

Monin B, Miller DT (2001) Moral credentials and the expression of prejudice. *J. Personality Soc. Psych.* 81(1):33–43.

Morales AC, Amir O, Lee L (2017) Keeping it real in experimental research—Understanding when, where, and how to enhance realism and measure consumer behavior. *J. Consumer Res.* 44(2): 465–476.

Murphy JJ, Allen PG, Stevens TH, Weatherhead D (2005) A meta-analysis of hypothetical bias in stated preference valuation. *Environ. Resource Econom.* 30(3):313–325.

O'Hora D, Maglieri KA (2006) Goal statements and goal-directed behavior: A relational frame account of goal setting in organizations. *J. Organ. Behav. Management* 26(1-2):131–170.

Pierce L, Rees-Jones A, Blank C (2020) The negative consequences of loss-framed performance incentives. NBER Working Paper 26619, National Bureau of Economic Research, Cambridge, MA.

Reiff J, Dai H, Beshears J, Milkman KL (2020) Save more today or tomorrow: The role of urgency and present bias in nudging pre-commitment. Preprint, submitted June 11, https://dx.doi.org/10.2139/ssrn.3625338.

Rogers T, Milkman KL, John L, Norton MI (2013) Making the best-laid plans better: How plan making increases follow-through. Working Paper, Harvard University, Cambridge, MA.

Saccardo S, Serra-Garcia M (2021) Cognitive flexibility or moral commitment? Evidence of anticipated beliefs distortion. Preprint, submitted August 18, https://dx.doi.org/10.2139/ssrn.3676711.

Sachdeva S, Iliev R, Medin DL (2009) Sinning saints and saintly sinners: The paradox of moral self-regulation. *Psych. Sci.* 20(4): 523–528.

Schultz PW, Nolan JM, Cialdini RB, Goldstein NJ, Griskevicius V (2007) The constructive, destructive, and reconstructive power of social norms. *Psych. Sci.* 18(5):429–434.

Seuntjens TG, Zeelenberg M, Van de Ven N, Breugelmans SM (2015) Dispositional greed. *J. Personality Soc. Psych.* 108(6):917–933.

Tenbrunsel AE, Diekmann KA, Wade-Benzoni KA, Bazerman MH (2010) The ethical mirage: A temporal explanation as to why we are not as ethical as we think we are. *Res. Organ. Behav.* 30:153–173.

Zhang T, Gino F, Bazerman MH (2014) Morality rebooted: Exploring simple fixes to our moral bugs. *Res. Organ. Behav.* 34:63–79.