How do People Learn from *Not* Being Caught? An Experimental Investigation of a "Non-Occurrence Bias"

Tom Zur[1]

April 2023 draft for Olin submission

Abstract

The law and economics literature has long theorized that one of the goals of law enforcement is specific deterrence, which relies on the conjecture that imperfectly informed offenders learn about the probability of detection from their prior interactions with law enforcement agencies. Surprisingly, however, no empirical study has rigorously tried to identify the learning process that underlies the theory of specific deterrence and, more specifically, whether potential repeat offenders learn from getting caught in the same way as they learn from *not* getting caught. This paper presents novel evidence from a pre-registered randomized controlled trial that sheds new light on these questions. In each of the two stages of the experiment, participants could cheat for an increased monetary payoff at the risk of paying a fine, in the face of an uncertain chance of being audited. Using an incentive-compatible procedure, participants' beliefs regarding the probability of being audited were rigorously elicited both before and after they were either audited or not audited, allowing us to establish the unique rational-Bayesian benchmark and any deviation thereof for each participant. We find that *not* being audited induces a weaker learning effect compared to the learning effect induced by being audited, providing novel evidence for what we call a "non-occurrence bias." These and other findings presented in the paper imply that the specific deterrence benefit from investing in enforcement is lower than predicted by rational choice theory.

JEL CODES: K42, K14, D83, D91

KEYWORDS: Law Enforcement, Learning, Experimental Behavioral Economics, Bayes' rule, Belief updating

[1] John M. Olin Fellow and SJD Candidate, Harvard Law School. For helpful comments and suggestions, I thank Arevik Avedian, Oren Bar-Gill, Alma Cohen, Talia Fisher, Louis Kaplow, Ori Katz, Tamar Kricheli-Katz, Haggai Porat, Omri Porat, Steven Shavell, Holger Spamann, Kathryn Spier, Crystal Yang, seminar participants at Harvard Law School (Law and Economics seminar and the Empirical Legal Studies series), participants of the 39th annual conference of the European Association of Law & Economics and the members of the committee awarding it the Göran Skogh Award for the best paper presented by a young scholar. The hypotheses of this study were pre-registered in the American Economic Association's registry for randomized controlled trials on February 24, 2022, and can be found at https://www.socialscienceregistry.org/trials/8992. The experimental design was reviewed and determined to be exempt from Institutional Review Board (IRB) by Harvard University Area IRB on January 25, 2022 (IRB21-1590). I am also grateful to the John M. Olin Center at Harvard Law School for its financial support. All errors are my own.

Introduction

The most intuitive corner stone of economic theory of deterrence is that an offender will commit an offense if and only if his gain exceeds the expected costs, which is a function of the sanction and the probability of being apprehended and sanctioned (Bentham, 1789; Becker, 1968; Polinsky & Shavell, 1979). Early models of deterrence relied on the assumption that individuals' beliefs regarding the probability of detection are accurate and steady over time. This assumption, however, has long been empirically disproved: beliefs are likely to be inaccurate and change through the offender's experience with criminal activity. Deterrence created through this learning process, which is the focus of this study, is often referred to as *specific deterrence*. Sah (1991) presented the first theoretical model to explicitly account for learning in a law enforcement setting, replacing the static approach of a single, known, and objective probability with a dynamic mechanism where individuals' perceptions regarding the probability of detection evolve through their experience with illegal activity and their interaction (or lack thereof) with law enforcement agencies (similar approaches were later adopted by Shavell, 2004; Maniloff, 2019; Miceli, Segerson, & Earn, 2022; and others). While a static model predicts that individuals who commit an offense once will always find it optimal to commit it again in the future, a dynamic model suggests that the offender's perceived probability of detection may change in response to his experience with crime, and thus might deter him from offending again in the future.

Despite its theoretical importance, the empirical literature that studied how repeat offenders successfully learn about the probability of detection exclusively relies on surveys and self-reported estimations, which are limited in their power to rigorously identify the underlying learning process. More importantly, the empirical literature has only shown that offenders adjust their beliefs in the right *direction*, concluding that this indicates rational learning, whereas a major question is whether offenders adjust their beliefs at the rationally dictated *magnitude*. This is especially surprising given the vast experimental psychological literature that shows that individuals often fail to adjust their beliefs rationally when presented with new information, albeit in the correct direction. Such a simplistic approach overlooks two crucial questions for law enforcement that our study answers. The preliminary question is whether individuals adjust their beliefs to a *magnitude* that is consistent with rational-choice theory. The second and more central question is whether they similarly learn from being caught and from *not* being caught, as expected in a perfectly rational, Bayesian world. An alternative behavioral hypothesis, which is the focus of this study, is

that being caught induces a stronger adjustment of beliefs relative to not being caught – a "non-occurrence bias" – even when both events carry the same informational weight from a purely rational perspective. The hypothesized bias is consistent with the behavioral intuition that not being caught, an event that is framed as lacking a salient consequence, has a weaker effect on the decision-maker.

This study starts filling these gaps by conducting a pre-registered randomized controlled trial that simulates the mechanism of learning about the probability of detection. In the experiment, 350 participants recruited through Amazon Mechanical Turk (Mturk) were faced with two sequential opportunities to cheat to obtain a higher payoff in the face of uncertain risk of being audited. Cheating potentially results in a higher payoff if it was not chosen for inspection, or a lower payoff (due to an imposed sanction) if chosen for inspection. Participants were initially given noisy information about the probability of being audited, and their beliefs were rigorously elicited (using additional monetary incentives) both before and after they were notified whether their first decision had been chosen to be audited or not. This 2×2 between-subjects factorial design allows us to rigorously identify the learning process in a specific deterrence setting. Specifically, participants' elicited prior beliefs regarding the probability of detection were used to construct their unique rational, Bayesian benchmark, which was then compared to their elicited posteriors to identify their unique deviation from rational learning.

The results confirm the existence of a non-occurrence bias. First, we find that subjects who were audited adjust (upwards) their beliefs regarding the probability of being audited in the future to a level that is statistically indistinguishable from the Bayesian estimate, while those who were not audited adjust (downwards) their beliefs to a lesser extent than dictated by rational learning, amounting to 34% under adjustment of beliefs. Second, when focusing on subjects' deviation from their *unique* Bayesian beliefs (based on their reported prior beliefs and observed signal), we find that participants who were not audited exhibit a learning effect that is significantly weaker than the learning effect exhibited by those who were audited, providing further and more rigorous evidence for the hypothesized non-occurrence bias. In further exploratory analysis we find that these findings intensify when focusing on individuals who reported more accurate prior beliefs.

In additional analyses, we explored the effect of one's detection experience on their subsequent decision to cheat. In line with specific deterrence theory, we find that being audited

significantly reduces the likelihood of subsequent cheating, whereas not being audited increases the likelihood of subsequent cheating, and that this is driven by the changes in beliefs. Furthermore, we find that even when controlling for beliefs, those who cheated in the first round were more responsive to the information conveyed by either of the signals – being or not being audited – compared to non-cheaters, as measured by its effect on subsequent behavior.

The findings presented in the paper and the novel "non-occurrence" bias that they uncover have important policy implications. Most notably, they imply that the overall specific deterrence gain from investing in enforcement is lower than predicted by rational choice theory. This is because, holding constant the benefit from the increased deterrence of those who were caught – the loss in deterrence of those who were not caught is milder than predicted by rational choice theory. The clear results obtained by using an incentivized experiment with high internal validity call for further exploration of this phenomenon in the field. If found, these results may provide a novel efficiency-based justification for reducing investment in criminal enforcement, as opposed to many similar claims that are predominantly grounded in fairness concerns or the inefficacy of the US penal system.

The remainder of the paper proceeds as follows. Section 2 reviews the related literature, and section 3 presents a simple economic model to guide the experimental design and flush out its potential policy implications. Section 4 presents the experimental design, section 5 continues to describe the data, and section 6 presents the results of the analysis. Section 7 discusses the results and provides concluding remarks. The complete experimental protocol, and additional robustness checks can be found in the Appendix.

2. Related Literature

Following the influential work by Sah (1991), many empirical studies have been dedicated to studying how offenders' interactions with the criminal justice system and law enforcement agencies affect future criminal behavior through learning. One strand of this literature has focused on offenders' learning about uncertain *criminal sanctions*, either through the severity of the sanction (e.g., Hjalmarsson, 2009; Schargrodsky & Di Tella, 2013) or the length and conditions of imprisonment (e.g., Chen & Shapiro, 2007; Drago, Galbiati, & Vertova, 2011).

A second strand of this literature has focused on offenders' learning about the *probability of detection*. Some use future behavior as a proxy for a change in beliefs. For example, Dušek & Traxler (2022) recently show that drivers who receive speeding tickets tend to drive slower.[2] More closely related to our study are observational survey studies that explore the effect of the offender's prior interactions with law enforcement agencies (or lack thereof) on their beliefs regarding the *probability of detection*. They find that offenders who were arrested more frequently report a higher belief regarding the probability of apprehension, providing suggestive evidence that offenders adjust their beliefs in the Bayesian-rational *direction*. For example, Lochner (2007) shows that respondents who engaged in crime while avoiding arrest revised their perceived probability of future arrest downwards, while those who were arrested adjusted their perceived probability of rearrest upwards. In a similar vein, Huizinga, Matsueda, & Kreage (2006) find that as the number of offenses gone unpunished increases, the perceived probability of arrest decreases monotonically; and Shamena & Loughran (2011) show that the experience of being arrested induces a significant increase in offenders' perceived probability of future arrest, especially for individuals whose prior beliefs were relatively low.

Notwithstanding the importance of these types of studies, they are limited in several crucial ways that this study aims to correct and complement. First, survey studies that rely on self-reported beliefs are inherently prone to bias, which we address by utilizing a rigorous belief elicitation mechanism using additional calibrated monetary incentives. Second, and most importantly, observational studies can only test whether individuals adjust their beliefs in the *right direction*. They are inherently unable to test whether the observed learning is consistent with rational learning as a matter of *degree*, as there does not exist real-world data on the parameters required to construct such a rational benchmark. Consequently, they are also limited in their power to test the hypothesis of a non-occurrence bias, which requires to form the rational benchmark for beliefs both for individuals who were apprehended and those who were not. To the best of our knowledge, this is the first experimental study that is designed in a way that can rigorously identify these learning patterns.

[2] For empirical literature that support the reverse effect, namely – that the experience of punishment might increase recidivism, see, for example Cullen, Jonson & Nagin (2011); Cullen, Jonson & Nagin (2009); Nagin (2013).

More recently, Friehe, Langenbach, & Mungan (forthcoming JLS, 2023) experimentally demonstrated that the severity of the sanction affects the adjustment of beliefs regarding the probability of future apprehension, challenging the assumption that the perceived probability of detection and the magnitude of the sanction are separable. Finally, this paper is also related to the growing experimental psychological literature studying information processing under uncertainty, showing that people systematically deviate from Bayesian predictions (Tversky and Kahneman, 1974; Grether, 1980). Most closely related to the hypothesized non-occurrence bias is the literature on asymmetric learning, and specifically, the so-called "good-news-bad-news" bias, where subjects tend to over-weight events framed as "good news" compared to events framed as "bad news" when adjusting their beliefs. This phenomenon has been documented in various contexts, such as beliefs about financial prospects (Kuhnen, 2014), intellectual abilities (Eil & Rao, 2011; Mobius et al., 2014), life threatening events (Sharot, Korn, & Dolan, 2011), and more.[3] In the context of law enforcement, the hypothesized non-occurrence bias works in the opposite direction of the good-news-bad-news bias, since we hypothesized that the good news (not being caught) induces a weaker response compared to the bad news (being caught), due to the latter being framed as a "non-occurrence," and not the other way around. Identifying the non-occurrence bias is, therefore, an uphill battle, seeing as it is a distinct form of cognitive bias that has yet to be identified in the literature.

3. A Model of Law Enforcement and Learning with a "non-Occurrence Bias"

In this section, we present a simple model of law enforcement where individuals learn about the probability of apprehension, building on the seminal model of Shavell (1991), and adjust it to account for deviation from rational learning in the form of a non-occurrence bias. We use this model to guide the experimental design, described in the following section, and to derive policy implications from the results.

[3] Among the scarce literature that report this asymmetry in the reverse direction, i.e., that "bad news" is weighted more than "good news", see Ertac (2011), and Sunstein, Bobadilla-Suarez & Lazzaro (2016).

Assume that N individuals consider whether to commit an offense that causes a harm of $h > 0$ in each of two sequential periods, $t = 1,2$.[4] Without loss of generality, we will normalize the number of individuals to one. Let e denote the government's expenditures in enforcement, and $p(e)$ denote the actual probability of detection in each period, as a function of e. It is also assumed that there is a decreasing marginal return to investment in enforcement, captured by $p(e)' > 0$ and $p(e)'' < 0$. Let $p_t(e)$ denote the belief of the individual, at the beginning of period t, regarding the probability of detection if they decide to commit an offense in that period. Furthermore, let g denote the gain that the individual derives from committing the offense in each period, which is distributed among individuals by the density function $f(g) > 0$. We assume that $f(g)$ is a single peak distribution, where the gain of individuals who commit the offense in each period exceeds the peak of $f(g)$. let $s > 0$ denote the private disutility from the sanction imposed on the individual.[5] An individual will commit the offense in period t if and only if their gain exceeds the expected sanction, given by:

$$(1)\ g > p_t(e) \times s$$

At $t = 1$, where individuals are imperfectly informed about the actual probability of detection $p(e)$, and has yet to engage in any criminal activity, their prior belief regarding the probability of detection, $p_1(e)$, is based solely on general deterrence efforts.[6] Whether or not an individual chooses to commit the offense at $t = 1$, he acquires new information that he uses to adjust his estimate regarding the probability of apprehension,[7] which informs his decision whether to commit the offense at $t = 2$. This information comes in the form of a partially informative binary signal: being audited or not being audited. When being audited, the individual adjusts his prior belief

[4] We follow the standard economic model of crime approach where the individual is risk-neutral, expected utility maximizer.

[5] For simplicity, we assume that imposing the sanction is socially costless (e.g., a monetary fine), and that the sanction, in terms of both type and magnitude, is objectively known. However, it is plausible that individuals often do not know either the magnitude or the type of the sanction, such that learning is expected not only with respect to the probability of detection, but also with respect to the sanction itself. See, for example, Kaplow (1990), Bebchuk & Kaplow (1992), Ben-Shahar (1997), and more recently Friehe, Langenbach, & Mungan (forthcoming JLS, 2023).

[6] As a result, at $t = 1$, the crime rate equals $1 - F(p_1 s)$.

[7] There are two significant situations where the assumption that individuals learn about the probability of detection even when not committing the offense is realistic. First, when detection is based on *auditing* (e.g., auditing for tax evasion, safety, or other types of regulatory inspection, stop and frisk, sobriety checkpoints for DUI, and so on). Second, when individuals receive information from the experience of *their peers* (as opposed to themselves) with criminal activity. For theoretical models that account for this type of learning, see, e.g., Parker & Grasmick (1979), Sah (1991), and Miceli et al (2022).

regarding the probability of detection *upwards* to $p_2^A = p_1(e) + \Delta_2^A(e)$, creating a socially desirable specific deterrence effect. When the individual is not audited, he adjusts his belief *downwards* to $p_2^{NA} = p_1(e) - \Delta_2^{NA}(e)$, reducing deterrence. $\Delta_2^A(e)$ and $\Delta_2^{NA}(e)$ denote the Bayesian absolute values of the adjustments in the individual's estimate that are induced by observing the signal of either being audited or not being audited, respectively. Given the behavior of the individual in equation (1), the government chooses e to minimize:

$$(2)\ SC = \int_{p_1(e)s}^{\infty} f(g)dg \cdot h + p(e) \cdot \int_{(p_1(e)+\Delta_2^A(e))s}^{\infty} f(g)dg \cdot h +$$

$$\big(1 - p(e)\big) \cdot \int_{(p_1(e)-\Delta_2^{NA}(e))s}^{\infty} f(g)dg \cdot h + 2e$$

The first term in equation (2) represents the aggregate harm from crimes committed at $t = 1$; the second term represents the aggregate harm from crimes committed at $t = 2$ by individuals who were audited at $t = 1$, adjusted their belief upwards and, nonetheless, decided to commit the crime at $t = 2$; the third term represents the aggregate harm from crimes committed at $t = 2$ by individuals who were not audited at $t = 1$ and decided to commit the offense at $t = 2$,[8] and the last term is the government's expenditures in enforcement in the two periods.[9]

The benchmark for the analysis is that individuals are rational-Bayesian decision-makers, i.e., that $\Delta_2^A(e)$ and $\Delta_2^{NA}(e)$ are consistent with Bayesian learning. By taking the partial derivative of the social costs function with respect to enforcement expenditures *e*, we can characterize the optimal level of government expenditures in enforcement in this benchmark scenario, denoted e^*, as the level of investment that solves the following first-order condition:

$$(3)\ \frac{\partial SC}{\partial e}(e^*) = 0 \Rightarrow$$

[8] Note that while this cohort includes both individuals who committed the offense at $t = 1$ and those who did not, by limiting our model to two periods only, consistently with our experimental design, the potential effect of the non-occurrence bias stems only from the latter cohort. The intuition for this result is that if an individual's net benefit from committing the offense at $t = 1$ was sufficiently high under a prior belief of $p_1(e)$ (i.e., $p_1(e) \times s < g$), then a fortiori he will find it worthwhile to do so with a downward revision in the probability of apprehension ($p_2(e) < p_1(e)$), where both g and s being held fixed.

[9] Note that the number of individuals who commit the offense at $t = 2$ is also a function of the actual probability of being audited, $p(e)$, not just their prior belief $p_1(e)$ and the signal they observe at $t = 1$ (which in turn determines whether they adjust their belief upwards by $\Delta_2^A(e)$ or downwards by $\Delta_2^{NA}(e)$), because the actual probability is what determines the number of individuals who get each of these signals at $t = 1$.

$$h \cdot \left[\frac{dp_1(e^*)}{de} \cdot s \cdot f(p_1(e^*)s) + p(e^*) \cdot \frac{d\left(p_1(e^*) + \Delta_2^A(e^*)\right)}{de} \cdot s \cdot f\left(\left(p_1(e^*) + \Delta_2^A(e^*)\right)s\right) \right.$$

$$- \frac{dp(e^*)}{de} \cdot \int_{\left(p_1(e^*)+\Delta_2^A(e^*)\right)s}^{\infty} f(g)dg + \left(1 - p(e^*)\right)$$

$$\cdot \frac{d\left(p_1(e^*) - \Delta_2^{NA}(e^*)\right)}{de} \cdot s \cdot f\left(\left(p_1(e^*) - \Delta_2^{NA}(e^*)\right)s\right) + \frac{dp(e^*)}{de}$$

$$\left. \cdot \int_{(p_1(e^*)-\Delta_2^{NA}(e^*))s}^{\infty} f(g)dg \right] = 2$$

Next, consider the effect of a non-occurrence bias on the optimal *e*. We empirically find, as will be further elaborated in the following sections, that the actual upwards adjustment induced by being audited $\Delta_2^A(e)$ is consistent with rational-Bayesian learning; but that the downward adjustment induced by not being audited $\Delta_2^{NA}(e)$ is smaller than the adjustment predicted by rational learning. To adjust our baseline model to reflect this new information, let substitute $\Delta_2^{NA}(e)$ with $\beta\Delta_2^{NA}(e)$, where $\beta = 1$ if the individual is a perfectly rational-Bayesian decision-maker and $0 < \beta < 1$ if he exhibits a non-occurrence bias.[10] Next, we derive $\Delta_2^A(e)$ and $\Delta_2^{NA}(e)$ according to Bayes' rule. For simplicity, and to remain consistent with our experimental design, assume that individuals' prior beliefs regarding the probability of detection $p_1(e^*)$ have the following structure: at $t = 0$, all individuals believe that there is a 50% chance that the probability of detection is $p^H = p(e) + \Delta_1$ and 50% chance that the probability of detection is $p^L = p(e) - \Delta_1$ (where $p^H < 1, p^L > 0, \Delta_1 > 0$).[11] Deriving $p_2^A(e)$ and $p_2^{NA}(e)$ according to Bayes' rule, we find that $p_2^A(e) = p_1(e) + \frac{\Delta_1^2}{p_1(e)}$ and $p_2^{NA}(e) = p_1(e) - \frac{\beta\Delta_1^2}{1-p_{1(e)}}$.[12] Substituting

[10] $\beta < 0$ means that when not being audited, the individual adjusts his belief upwards (rather than downwards). As elaborated in the previous section, the assumption that offenders adjust their beliefs in the rational-Bayesian direction had long been empirically established.

[11] Note that the assumptions that $p^H < 1$ and $p^L > 0$ follows our assumption that the informational benefit of one's experience with criminal activity takes the form of a *partially* informative binary signal. Where one (or both) of these conditions is violated, one learns the enforcement parameters perfectly (i.e., whether $p(e)$ is p^H or p^L) at the end of $t = 1$.

[12] For individuals who are audited at $t = 1$, $p_2^A(e) = \Pr(p^H|\mathrm{A}) \times p^H + \Pr(p^L|\mathrm{A}) \times p^L$, where: $\Pr(p^H|\mathrm{A}) = \frac{\Pr(\mathrm{A}|p^H)\times\Pr(p^H)}{\Pr(\mathrm{A}|p^H)\times\Pr(p^H)+(\mathrm{A}|p^L)\times\Pr(p^L)} = \frac{0.5p_1(e)+0.5\Delta_1}{p_1(e)}$ and $\Pr(p^L|\mathrm{A}) = 1 - \Pr(p^H|\mathrm{A}) = \frac{0.5p_1(e)-0.5\Delta_1}{p_1(e)}$. Therefore, $p_2^A(e) = \frac{p_1(e)^2 + \Delta_1^2}{p_1(e)} = p_1(e) + \frac{\Delta_1^2}{p_1(e)}$. By the same token, for individuals who are not audited at $t = 1$, $p_2^{NA}(e) =$

$p_1(e) + \frac{\Delta_1^2}{p_1(e)}$ for $p_2^A(e)$ and $p_1(e) - \frac{\beta\Delta_1^2}{1-p_{1(e)}}$ for $p_2^{NA}(e)$ in expression (3), we establish the following proposition:

Proposition: The optimal investment in law enforcement is decreasing in the magnitude of the non-occurrence bias, i.e., $\frac{\partial e^*}{\partial \beta} > 0$.

Remark: The intuition for this result, whose formal proof can be viewed in the Appendix, is as follows. Since individuals who are not audited at $t = 1$ adjust their beliefs downwards to $p_2^{NA}(e) = p_1(e) - \beta\Delta_2^{NA}$ and thus are less deterred at $t = 2$, in the presence of a non-occurrence bias, i.e., whereby $\beta < 1$, the downward adjustment is smaller, hence the loss of deterrence is smaller as well. In essence, this group of individuals discount the signal of not being audited, which in turn leads to a higher expected sanction and consequently a lower willingness to commit the offense at $t = 2$ at any given level of investment in enforcement. The mirror image of this observation is that whenever a non-occurrence bias exists, the benefit from investing in enforcement, which operates to increase deterrence among (also) this group, is lower than predicted by the rational-choice model, implying that enforcement is socially excessive. It follows that at the optimal level of enforcement that is guided by rational theory, e^*, the marginal benefit from investment in enforcement is lower than the marginal cost in the presence of a non-occurrence bias, and hence, the optimal level of investment is lower when β is lower, $\frac{\partial e^*}{\partial \beta} > 0$.

This preliminary analysis elaborates the intuition provided in the introduction that the potential policy implications of the non-occurrence bias is that the socially optimal investment in law enforcement is smaller than dictated by the rational-choice model. The model presented in this section is consistent with the experimental design, to which we turn in the next section, allowing us to derive theoretically driven implications from the empirical results.

$\Pr(p^H|\text{NA}) \times p^H + \Pr(p^L|\text{NA}) \times p^L$, where: $\Pr(p^H|\text{NA}) = \frac{\Pr(\text{NA}|p^H)\times\Pr(p^H)}{\Pr(\text{NA}|p^H)\times\Pr(p^H)+(NA|p^L)\times\Pr(p^L)} = \frac{0.5 - 0.5p_1(e) - 0.5\Delta_1}{1 - p_1(e)}$; and $\Pr(p^L|\text{NA}) = 1 - \Pr(p^H|\text{NA}) = \frac{0.5 - 0.5p_1(e) + 0.5\Delta_1}{1 - p_1(e)}$. Therefore, $p_2^{NA}(e) = \frac{p_1(e) - p_1(e)^2 - \Delta_1^2}{1 - p_1(e)} = p_1(e) - \frac{\Delta_1^2}{1-p_1(e)}$.

4. Experimental Design

Participants were recruited through Mturk for a $1 participation fee and an additional bonus payment based on performance.[13] The average payment was $6 for approximately 10 minutes.[14]

The experiment proceeded in two rounds comprised of the same-identical procedure. In each of the two rounds, participants were asked to roll a virtual fair six-sided dice after reporting their guess of the outcome. Participants were notified that their bonus payment would be based on their *self-reported performance*: $1.5 for a successful guess; and $0.5 for an unsuccessful guess. The participants were informed in advance that some of them would be randomly chosen to be audited after completing the trial, and that in the case of falsely reporting a successful guess, their bonus payment would be reduced to $0.25 ($0.5 for an unsuccessful guess minus a $0.25 fine).

The likelihood of being audited was presented, both by text and visually (as shown in Figure 1 below), as an equally likely chance of being assigned to one of two gumball machines with different proportions of "audit" and "no audit" balls: one was loaded with *seven* red "audit" balls and *three* green "no audit" balls; the other was loaded with *three* red "audit" balls and *seven* green "no audit" balls.[15] Participants were told that the computer had randomly assigned them (by a flip of a coin) either to the seven-audit-balls machine or to the three-audit-balls machine and that

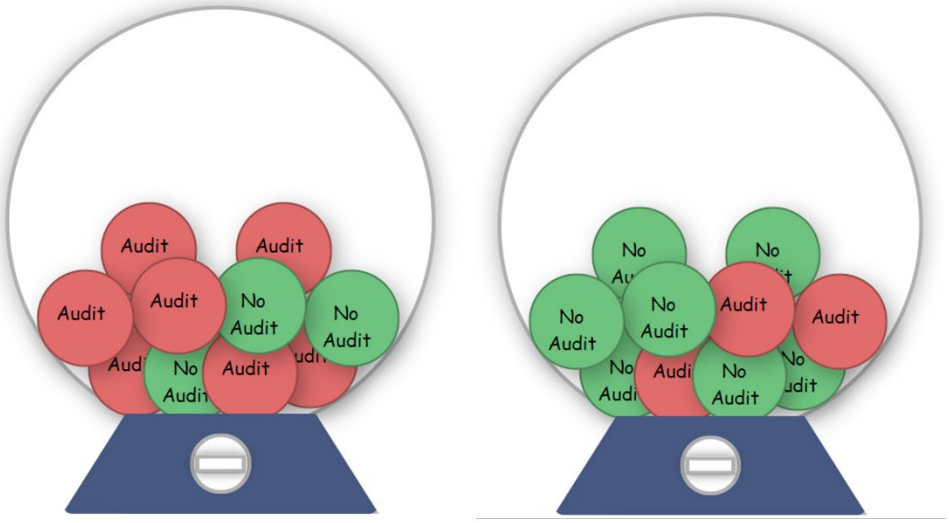
[13] For prior studies on the generalizability of causal relationships from Mturk samples to the U.S broader population, see, e.g., Buhrmester, Kwang, & Gosling, (2011) and Crump, McDonnell, & Gureckis (2013).

[14] Nonetheless, participants' understanding of the various components of the experiment was critical to its credibility and could be achieved only by a careful reading of the instructions. To achieve this, we took two complementary steps: first, we recruited only Mturk workers with a "master" qualification for an additional payment. This designation is awarded to individuals that have "demonstrated a high degree of success in performing a wide range of HITs across a large number of Requesters", as determined by Amazon's algorithm. Second, after reading the instructions and before starting the trial, we asked participants to complete a short simple four-questions comprehension test, of which answering correctly was required to be eligible for the bonus payment. Participants were able to return and read the instructions with no time limit and leave the experiment without penalty. Indeed, our records show that participants allotted 2.5 minutes on average for reading the instructions and that about 1/3 of the participants used the "back" button option (with astonishing average number of "clicks" of 2.5), indicating the effectiveness of this comprehension test altogether.

[15] Numerous empirical studies on color psychology documented the power of colors to affect behavior and cognitive processes. Relevantly to our context, one strand of this literature has shown that red stimuli can have a detrimental effect on performance in high-level cognitive processes (see, e.g, Elliot, Maier, Moller, Friedman, & Meinhardt, 2007). A second strand of literature has shown that red stimuli are often associated with bad outcomes, while green stimuli are typically associated with good outcomes (Gerend & Sias, 2009; Bouhassoun, Naveau, & Delcroi, 2022). Note, however, that both of these effects, like the "good-news bad-news bias", operates in the opposite direction of the hypothesized non-occurance bias, making it harder, rather than easier, to indentify an effect.

their chosen machine would be used in both rounds with replacement, rendering 50% the best guess regarding the probability of being audited at the first round for all participants.

***Figure 1**: The Gumball Machines as Presented to Participants*



To give participants more room for cheating, and to ensure the credibility of the procedure, the following statement appeared below the "Roll!" button: "*Before moving to the next screen, please press the 'Roll!' button a few more times just to make sure the dice is legitimate*" (procedure adapted from Shalvi, Eldar, & Bereby-Meyer, 2012). [16]

After completing the die-rolling task, participants' beliefs regarding the probability of being audited were elicited through the lottery version of the incentive-compatible quadratic loss rule (Mckelvey & Page, 1990), which was designed to ensure that the subject's expected payoff function is maximized by a reported belief that equals her subjective-true belief.[17] Participants

[16] This procedure builds on the well-established behavioral phenomenon according to which observing desired counterfactual information (in our design, the number guessed in irrelevant-for-pay rolls), individuals often feel more comfortable with shuffling the facts in a self-serving way. Indeed, our data reveals that 56% of the respondents rolled the dice several times at least once, 45% out of which were cheaters, which is significantly higher from the overall cheating rate for both the first ($p = 0.021$ in a fisher exact test) and second ($p = 0.002$ in a fisher exact test) rounds. Additional support for this procedure power to facilitate cheating is the observed rolling pattern, where 73% of the cheaters who rolled the dice more than once did so until observing the number they guessed, which is significantly higher than the uniform distribution expected from a fair dice roll ($p < 0.001$).

[17] Specifically, each participant's payoff was determined according to the following scheme: a reported estimate of p is translated to a possibility of winning a lottery that pays \$1 with a probability of $2p - p^2$ if the participant was audited and $1 - p^2$ if he was not audited. After the completion of the trial, we matched each participant with her appropriate probability of winning the \$1 prize in each of the two rounds, $2p - p^2$ if she was audited in that round and $1 - p^2$ if she was not audited. Then, the computer drew a random number r distributed uniformly from 0 to 100: if r is smaller

were asked to provide their "best estimate" regarding the probability of detection, for the opportunity to win an additional $1. Following Eil & Rao (2011), participants were told that "the probability of winning the bonus is higher the closer you are to the correct estimate," with an optional link that refers them to a more thorough explanation and a table of possible payoffs that demonstrate why that is the case.[18] To further facilitate participants' understanding of the elicitation procedure, we added the following intuitive explanation: "*Since your gumball machine contains either 7 red audit balls or 3 red audit balls (out of 10), your estimate is limited to a range between 70% (if you are certain that your assigned machine is the one with 7 audit balls) and 30% (if you are certain that your assigned machine is the one with 3 audit balls). If you are uncertain, the correct estimate lies somewhere in between*". Participants were asked to enter their estimate on a virtual slider, with the lower-end labeled as "I am certain there are 3 audit balls in my machine" and the upper-end labeled as "I am certain there are 7 audit balls in my machine".

After reporting their beliefs, participants were notified whether their report was chosen to be audited (i.e., a ball labeled "audit" was randomly drawn by the computer) or not (i.e., a ball labeled "no audit" was randomly drawn by the computer), accompanied by an animation of the corresponding ball jumping up and down on the screen. Participants' elicited beliefs before and after learning whether they were audited or not were analyzed to identify whether they exhibited a non-occurrence bias. Finally, participants were asked to provide a short explanation of what drove them to adjust their estimation in the second round, followed by a short demographic questionnaire. Participants' written feedback in our post-treatment survey and a short-comprehension test administered after reading the instructions suggested that participants had a good understanding of the various components of the experiment.

Building on prior literature that demonstrated that many individuals suffer from a distaste for cheating (Gneezy, 2005; Fischbacher & Föllmi-Heusi, 2013; Abeler , Becker, & Falk, 2014), we calibrated the payoffs and the fine to be such that for a rational Bayesian risk-neutral individual, cheating will consistently yield a higher expected payoff than not cheating in both rounds, regardless of whether they were audited, to ensure that cheating does not become ill-advised solely

---

or equal to the matched probability, the subject wins the $1 prize in that round. By clicking the link, participants could access a detailed explanation of this mechanism, and the pair probabilities of winning for any given p. This mechanism was later applied by Harrison, Martínez-Correa & Swarthout (2014), Armantier & Treich (2013), among others.

[18] Our records indicates that about 7% of the participants clicked on the link at least once.

on the basis of rational learning. Altogether, the experiment implemented a 2×2 design that varied our two experimental conditions (being audited and not being audited) in two different rounds (first and second rounds),[19] allowing us to causally identify dynamic within-subject effects of learning.

5. Data

350 individuals with a "master" qualification were recruited online via Mturk (164 women, $M_{Age} = 44$, $SD = 10.71$).[20] In the first round (before administering the treatment), participants exhibited an average prior belief of being audited of 49.3% (which is not statistically different from the accurate prior of 50%), and a cheating rate of 37%. 94% of the participants answered at least three of the four comprehension questions correctly, indicating a high level of understanding of the instructions. Table 1 provides summary statistics for our sample. A balance test reported in the last column revealed no significant differences across all pre-treatment variables by experimental condition, suggesting that randomization was effective. Since participants were technically assigned to different gumball machines, which subsequently determined whether or not they were audited, Table 2 provides information on the allocation of participants into the different machines and the corresponding audit rate, consistent with the 30% and 70% probabilities.

[19] Note that there are two dimensions to the randomization in our experiment: being assigned to the 30% condition or the 70% condition (each with a probability of 50%), and then being audited or not being audited (with a probability of either 30% or 70%, depending on the first condition). However, all audited participants experience an identical experimental procedure (and so do all non-audited participants). For this reason, it is more analytically accurate to say that the experiment includes two treatment conditions, and not four.

[20] The choice of the sample size was set based on an a priori power test using the conventional target power of 80% and $\alpha = 0.05$. Four subjects were labeled by Qualtrics as "likely an algorithm" and hence excluded from all analyses. In addition, one participant was excluded due to failing all four questions of the comprehension test.

Table 1: Descriptive statistics

	Not Audited				*Audited*				
	Mean	SD	Min	Max	Mean	SD	Min	Max	Diff.
Prior	0.49	0.1	0.3	0.7	0.49	0.1	0.3	0.7	0.001
Posterior	0.45	0.12	0.3	0.7	0.57	0.11	0.3	0.7	−0.122***
Math literacy	0.3	0.46	0	1	0.28	0.45	0	1	0.025
Cheated in 1st round	0.38	0.49	0	1	0.34	0.48	0	1	0.054
Cheated in 2nd round	0.44	0.5	0	1	0.3	0.46	0	1	0.140**
Female	0.46	0.5	0	1	0.49	0.5	0	1	−0.023
White	0.85	0.36	0	1	0.83	0.38	0	1	0.019
Age	43.34	10.68	23	75	44.6	10.64	24	78	−1.247
College (Bachelor's)	0.64	0.48	0	1	0.7	0.46	0	1	−0.059
Annual income > $50k	0.37	0.48	0	1	0.39	0.49	0	1	−0.022
Total time (in seconds)	494.1	233.3	93	1519	484.1	209.1	195	1322	10.068
Observations	174				171				345

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table 2: Audit rates by rounds and assignment to gumball machines

	3 audit balls machine (n=176)		7 audit balls machine (n=172)	
	1st round	2nd round	1st round	2nd round
Audited	30.7% (54/176)	29.5% (52/176)	69.2% (119/172)	70.4% (121/172)
Not audited	69.3% (122/176)	70.5% (124/176)	30.8% (53/172)	29.6% (51/172)

6. Results

6.1 learning about the probability of detection

This section explores how participants incorporated the signals of "being audited" and "not being audited" into their posterior beliefs. We hypothesized that individuals learn about the probability of detection from *being* caught at a different level of accuracy compared to learning from *not being* caught, challenging the implicit assumption of existing models of specific deterrence that offenders learning conforms with the rational-choice theory.

In our setting, individuals observed a sequence of two signals, whose probability depends on the underlying state of the world: being assigned to the "bad" machine (with seven audit balls) or to the "good" machine (with three audit balls), to be decided by a flip of a coin.[21] This allowed us to construct the accurate prior belief of 50%, such that both signals should rationally induce the *same* change in one's estimate, albeit in a different direction. Concretely, after completing the first round, a rational-Bayesian decision-maker with an accurate prior belief of 50% should adjust her belief eight percentage points upwards or downwards, i.e., a posterior of 58% or 42% (50% $\pm$ 8%), depending on whether they were audited or not, respectively.[22] While symmetric learning is not crucial for identifying the underlying learning process, it allows us to meaningfully compares the differences in absolute terms between participants' actual and Bayesian posterior beliefs across treatments.

A preliminary inquiry reveals that participants' elicited priors (pre-treatment beliefs) were indistinguishable from 50% for both the audited and the non-audited ($M = 49.2\%, t = -0.98, p =$

[21] The choice of the sample size was set based on an a priori power test using the conventional target power of 80% and $\alpha = 0.05$.

[22] To see this, denote the event of being assigned to the 30% condition by $\mathrm{A}_{30\%}$ (with $\Pr(\mathrm{A}_{30\%}) = 0.5$); The event of being assigned to the 70% condition by $\mathrm{A}_{70\%}$ (with $\Pr(\mathrm{A}_{70\%}) = 0.5$); The event of being audited in the first round by C; The event of not being audited in the first round by $\bar{\mathrm{C}}$, and the event of being audited in the second round by B. For participants who *were* audited in the first round, $\Pr(A_{70\%}|C) = \frac{\Pr(C|A_{70\%})\times \Pr(A_{70\%})}{\Pr(C|A_{70\%})\times\Pr(A_{70\%}) + (C|A_{30\%})\times\Pr(A_{30\%})} = \frac{0.7 \times 0.5}{0.7\times0.5 + 0.3 \times 0.5}$ $= 0.7$. Therefore, $\Pr(B|C) = \Pr(A_{70\%}|C) \times 0.7 + \Pr(A_{30\%}|C) \times 0.3 = 0.7 \times 0.7 + 0.3 \times 0.3 = 0.58$. Conversely, for participants who were *not* audited in the first round ($\bar{C}$), we have $\Pr(A_{30\%}|\bar{C}) = \frac{\Pr(\bar{C}|A_{30\%})\times \Pr(A_{30\%})}{\Pr(\bar{C}|A_{70\%})\times\Pr(A_{70\%}) + (\bar{C}|A_{30\%})\times\Pr(A_{30\%})} = \frac{0.7 \times 0.5}{0.3\times0.5 + 0.7 \times 0.5} = 0.7$. Therefore, $\Pr(B|\bar{C}) = \Pr(A_{30\%}|\bar{C}) \times 0.3 + \Pr(A_{70\%}|\bar{C}) \times 0.7 = 0.7 \times 0.3 + 0.3 \times 0.7 = 0.42$.

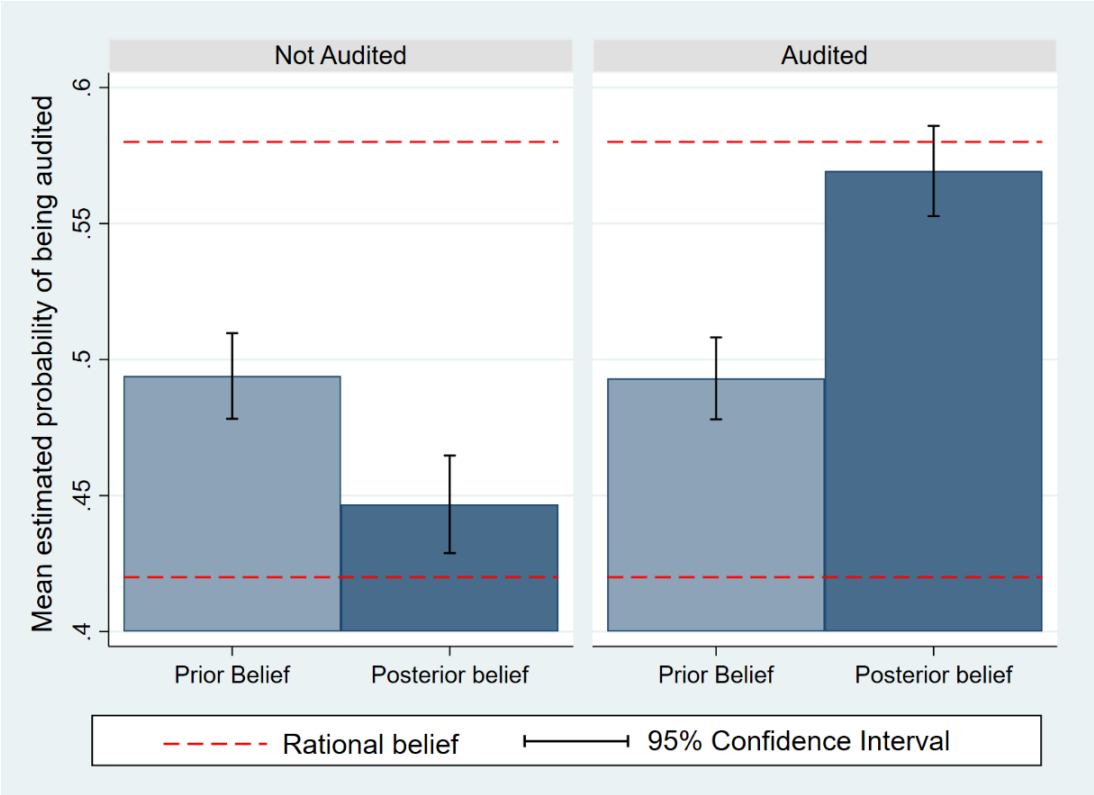
0.33; $M = 49.4\%,\ t = -0.76,\ p = 0.45$, respectively), suggesting that participants understood the likelihood of being audited before experiencing any type of enforcement.[23]

Starting the inquiry with the magnitude of the change in beliefs, the first substantive result is that those who were audited adjusted their belief (upwards) by 7.7 percentage points on average, whereas those who were not audited adjusted their belief (downwards) only by 4.7 percentage points ($t = -2.14,\ p = 0.03$), indicating that not being audited induced stronger adjustment of beliefs than not being audited, as we hypothesized. Turning next to the estimated change in beliefs as compared to rational learning, Figure 2 plots the average reported prior and posterior beliefs across experimental conditions, relative to the correct estimates of 42% and 58%. We find that those who were audited adjusted their beliefs to 57%, which is indistinguishable from the rational-Bayesian benchmark of 58% ($t = -1.25, p = 0.21$), whereas those who were not audited adjusted their beliefs to 44.7%, which is 2.7 percentage points higher from the Bayesian benchmark of 42% ($t = 2.95, p = 0.0037$).[24] This weaker response to the signal of not being audited, which amounts to a 34% under-adjustment of beliefs, provides suggestive evidence of a non-occurrence bias.

[23] Unless noted otherwise, reported values are the result of two-sample t-tests.

[24] As expected, respondents adjusted their beliefs in the right direction, i.e., upwards in the audit condition ($p < 0.001$ in a fisher-exact test), and downwards in the no audit condition ($p < 0.001$ in a fisher-exact test). Indeed, respondents' answers to the open-ended question in our post-treatment survey regarding what drove them to adjust their estimation in the second round in the way they did, revealed that most respondents could provide the intuition behind their decision to adjust their estimation in one direction and not the other, as opposed to just a "gut feeling" (see, for example: "*Since I got no audit the first round, I thought that it would be more likely that I got the gumball machine that only had 3 audit balls in it. I just thought mathematically that it's more likely and chose to adjust my prediction closest to the percentage I thought it would be*"; "*I thought being audited in the first round made it more likely I had the 70% audit condition, though not certain*").

***Figure 2**: Mean Priors and Posteriors across experimental conditions*



However, the assumption that rational learning would result in beliefs of either 42% or 58% assumes that everyone has a prior belief of 50%. While this is true on average, there is a non-negligible variance in participants' prior beliefs, rendering this assumption rather restrictive. Our setting allowed us to use each participant elicited prior to derive her *unique* closed form "Bayesian posterior," as a function of whether she was audited or not. I.e., the posterior belief that each participant would hold had they engaged in rational-Bayesian learning, given their reported prior and observed signal. The more accurate outcome measure of interest is, therefore, the difference between each participant's reported and Bayesian posteriors, which reflects her unique-personal deviation from rational learning.[25] For ease of comparison, the measure for the deviation from

[25] To calculate the "rational" posterior for participants with a prior that is different than 50%, let's denote the reported prior as $m \in [0.3, 0.7]$. We know that $m = p \times 0.3 + (1-p) \times 0.7 \rightarrow p = \frac{0.7 - \mathrm{m}}{0.4}$, and that $1-p = \frac{m-0.3}{0.4}$. For participants who were audited in the first round, $\Pr(A_{70\%}|C) = \frac{\Pr(\mathrm{C}|A_{70\%}) \times \Pr(A_{70\%})}{\Pr(\mathrm{C}|A_{70\%}) \times \Pr(A_{70\%}) + (\mathrm{C}|A_{30\%}) \times \Pr(A_{30\%})} = \frac{0.7 \times \frac{\mathrm{m}-0.3}{0.4}}{0.7 \times \frac{\mathrm{m}-0.3}{0.4} + 0.3 \times \frac{0.7-\mathrm{m}}{0.4}} = 1.75 - 0.525 \times \frac{1}{m}$. Therefore, $\Pr(B|\mathrm{C}) = \Pr(A_{70\%}|\mathrm{C}) \times 0.7 + \Pr(A_{30\%}|\mathrm{C}) \times 0.3 = (1.75 - 0.525 \times \frac{1}{m}) \times 0.7 + (0.525 \times \frac{1}{m} - 0.75) \times 0.3 = 1 - \frac{0.21}{m}$. Conversely, for participants who were *not* audited in

rational learning was normalized in a way that a positive difference means that the individual adjusted their belief in excess of the level dictated by Bayesian learning, while a negative difference means that they adjusted their belief less than they should have. This technical normalization facilitates comparison as it circumvents the fact that being audited calls for an upward adjustment of beliefs while not being audited calls for a downward adjustment of beliefs. The remainder of the analysis will focus on this outcome measure, which we call "*the learning effect*."

When looking at the entire pool of participants ($N = 345$), we find that not being audited induces a learning effect that is 2.79 percentage points *weaker* than the learning effect induced by being audited ($t = -2, p = 0.045$).[26] In other words, being audited creates a stronger signal than not being audited, providing further evidence for a "non-occurrence bias." However, while we find that not being audited induces weaker learning effect, we do not find any evidence that participants' adjusted beliefs are statistically different from the Bayesian benchmark when examining each of the groups separately.

To further explore this issue, recall that while participants reported correct prior beliefs on average (i.e., 50%), many reported incorrect beliefs. Such deviation might be a result of the respondent's mathematical illiteracy (namely, failure to understand that 50% of 70% and 50% of 30% equals 50%), a belief in luck (or lack thereof) that makes one of the machines more plausible than the other;[27] or misunderstanding of the instructions.[28] Notice, for example, that reporting a prior belief of either 30% or 70%, which 14% of participants did, means that the participant is 100% certain which of the gumball machines they were assigned, at a point in the game when the

---

the first round ($\bar{C}$), we have $\Pr(A_{30\%}|\bar{C}) = \frac{\Pr(\bar{C}|A_{30\%})\times \Pr(A_{30\%})}{\Pr(\bar{C}|A_{70\%})\times\Pr(A_{70\%}) + (\bar{C}|A_{30\%})\times\Pr(A_{30\%})} = \frac{0.7 \times \frac{0.7 - \mathrm{m}}{0.4}}{0.3\times\frac{\mathrm{m} - 0.3}{0.4} + 0.7 \times \frac{0.7 - \mathrm{m}}{0.4}} = \frac{0.49 - 0.7m}{0.4 - 0.4m}$.

Therefore, $\Pr(B|\bar{C}) = \Pr(A_{30\%}|\bar{C}) \times 0.3 + \Pr(A_{70\%}|\bar{C}) \times 0.7 = (\frac{0.49 - 0.7m}{0.4 - 0.4m}) \times 0.3 + (\frac{0.3m - 0.09}{0.4 - 0.4m}) \times 0.7 = \frac{0.084}{0.4 - 0.4m}$

[26] Note that the nonparametric equivalent of Mann-Whitney U test is inappropriate for evaluating the differences in mean, given the normal shape of the distributions of the learning effect for the aggregate data (given our failure to reject the null with Shapiro-Wilk test, Shapiro-Francia test, and Skewness and Kurtosis test, despite the relatively large sample size where these tests tend to reject).

[27]Consistently with this conjecture, participants' answers to the open-ended question regarding what drove their decision to adjust their estimation regarding the probability of being audited in the way they did, revealed that some respondents have "priors over priors" due to beliefs in luck (see for example, "*I just felt like the chances were always in favor of being audited, even before the red ball was chosen*"; and "*just a gut feeling*").

[28] Indeed, in our post-treatment survey, we asked respondents to calculate a simple arithmetic ($\frac{(3^3 - 2)}{4}$) without using a calculator. Somewhat surprisingly, only 37% answered this question correctly.

only information they have is that there is an equally likely chance to be assigned to either of the machines (to be decided by a flip of a coin). Anticipating this possible misunderstanding, participants were explicitly told that they should choose either 30% or 70% *only* if they were absolutely certain that their assigned gumball machine was the one with three audit balls or seven audit balls, respectively (and that otherwise, they should set the slider to "somewhere in between").

Seeing as reporting extreme prior beliefs is mostly an indication of one's lack of ability to form accurate estimations, as opposed to the result of a learning process in which one adjusts their beliefs in light of new information, the following exploratory analysis will re-estimate our baseline model for different subgroups of participants who exhibited higher degrees of accuracy in the formation of their prior beliefs.[29] Specifically, we report three sets of results for participants whose prior beliefs are in the ranges of 40%-60%, 45%-55%, and those who accurately reported a prior belief of 50% − being the most common estimates participants reported (218 of the 345 participants reported one of these five estimates).

Focusing on participants whose prior beliefs lie between 40% and 60% ($N = 221$), we find that the learning effect is -0.1 percentage points for those who were audited, which is statistically indistinguishable from the Bayesian benchmark of zero ($t = -0.13, p = 0.9$). In sharp contrast, the learning effect for those who were not audited is -4 percentage points, which is significantly lower than the Bayesian benchmark ($t = 3.96, p < 0.001$), amounting to a striking 62.5% under-adjustment of beliefs. Put differently, not being audited results in a learning effect that is 3.9 percentage points *weaker* than the learning effect induced by being audited ($t = -2.85, p = 0.005$). Turning to participants whose prior beliefs lie between 45% and 55% ($N = 125$) reveals the same general pattern. The learning effect is -1.8 percentage points for those who were audited, which is well approximated by Bayes' rule ($t = -1.65, p = 0.104$), whereas those who were not audited exhibited a significant under-adjustment of beliefs, captured by a learning effect of -5.1 percentage points ($t = -3.95, p < 0.001$). Hence, not being audited results in a learning effect that is 3.3 percentage points weaker than the learning effect induced by being audited ($t = -1.94, p = 0.055$). Finally, narrowing the analysis even further to subjects with a perfectly

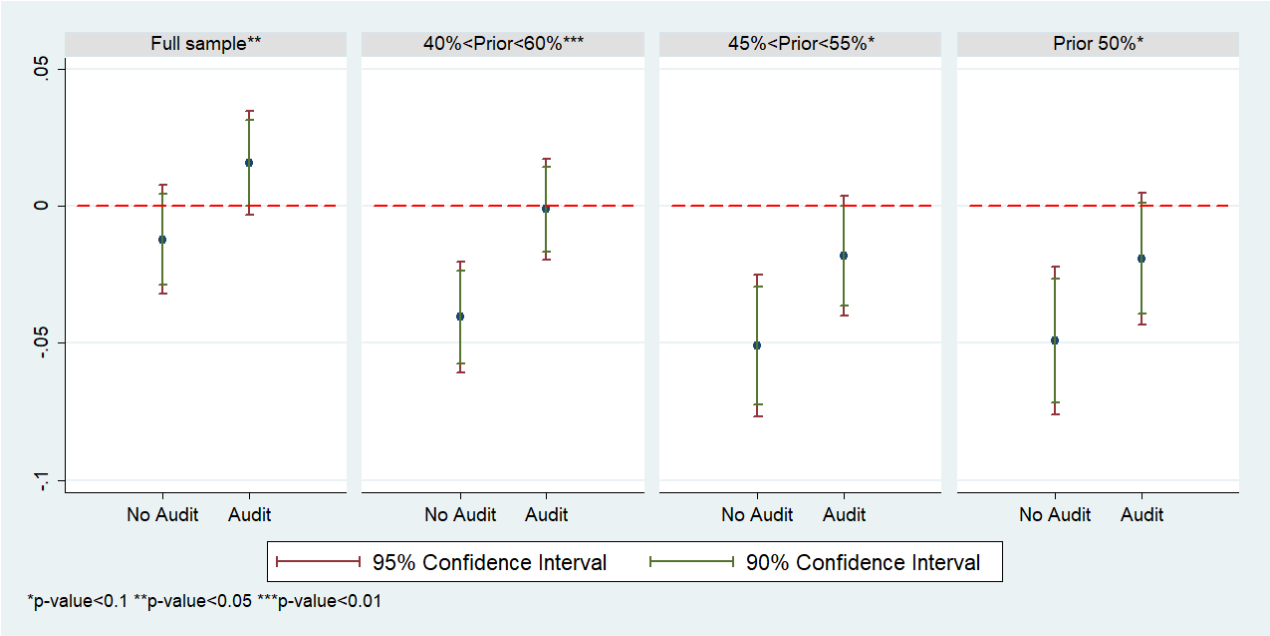
[29] Consistently with using the prior as a proxy for a better understanding of the experimental instructions, the largest cutoff of 40%-60% includes the *entire* pool of participants who answered the two questions in our comprehension test whose understanding is a necessary (and sufficient) condition for the formation of accurate prior: the proportion of "audit" balls in the "bad machine", and the chance of being assigned to either of the machines.

accurate prior belief of 50% ($N = 110$), we find a learning effect of -1.9 percentage points for those who were audited, which is still indistinguishable from rational learning at conventional significance levels ($t = -1.59, p = 0.118$), while those who were not audited exhibit a significant learning effect of -4.9 percentage points ($t = 3.65, p = 0.0006$). Hence, not being audited results in a learning effect that is 3 percentage points weaker than the learning effect induced by being audited, though this is only marginally significant for the parametric student T-test, possibly due to the small size and non-normal distribution of the learning for this cohort ($t = -1.66, p = 0.099$; $p = 0.0021$ in a MWU test).[30]

Figure 3 presents an aggregate view of these results, by plotting the average learning effect for those who were audited and those who were not across the four ranges of reported priors, increasing in accuracy from left to right. As evident from Figure 3 and the forgoing analysis, across all four subgroups, we cannot reject the null that being audited induces Bayesian-rational learning. In sharp contrast, the experience of not being audited consistently yields an under-adjustment of beliefs relative to the rational benchmark in the three subgroups where those reporting extreme prior beliefs are excluded. More importantly, in all subsamples, the learning effect exhibited by the audited compared to the non-audited is significantly larger. Taken together, these results provide strong evidence of a non-occurrence bias.

[30] Consistently with this difference, 50% is the only specification where our normality tests failed, with almost uniform distribution.

Figure 3: *Mean Deviation from Rational Learning ("Learning Effect") across Prior Accuracy*



To further explore whether the finding of a non-occurrence bias is robust to the inclusion of various pre-treatment controls, Table 3 reports the results from a standard ordinary least squares regression of the learning effect on an indicator of whether the individual was audited, controlling for the (pre-treatment) decision to cheat in the first round and various demographic characteristics. Columns 1-3 present the estimates for the full sample, while Columns 4-6 restrict the analysis to participants whose priors lies between 40% and 60% – the largest sub-group with relatively accurate priors. As expected, and consistent with our prior findings, the non-occurrence bias persists across all specifications. To get a sense of the magnitude of this effect, compare the coefficient 0.028 to the size of 0.08, which would be estimated in a hypothetical world with an *extreme* non-occurrence bias, where being audited would induce rational learning while not being audited would not induce *any* learning whatsoever. Appendix Table A.1 re-estimates the baseline results of Table 3 restricting the sample only to active updaters, showing that effect of the non-occurrence bias becomes even stronger.

To further explore whether cheaters differ in their sensitivity to the observed non-occurrence bias, columns 2 and 3 of Table 3 show that first-round cheaters exhibit a stronger learning effect, albeit with only marginal significance. Appendix Table A.2 shows that this effect becomes highly significant once we exclude subjects who, by chance, guessed the dice roll correctly, as they have no reason to cheat, introducing some noise to the analysis. Nonetheless,

using a Welch's t-test (to account for the unequal sample sizes), the non-occurrence bias (i.e., the difference in the learning effect when audited versus not audited) is essentially the same between cheaters and non-cheaters ($t = 0.2758, p = 0.783$). Hence, we cannot rule out the null that cheaters (offenders) are equally likely to exhibit the non-occurrence bias as non-cheaters (non-offenders).

Table 3: the learning effect

	Full sample			40 < Prior < 60		
	(1)	(2)	(3)	(4)	(5)	(6)
Audit	0.028**	0.028**	0.028**	0.039***	0.039***	0.038***
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Cheat 1st round		0.026*	0.026*		0.010	0.015
		(0.01)	(0.02)		(0.02)	(0.02)
Demographics	NO	NO	YES	NO	NO	YES
Constant	-0.012	-0.02*	-0.04*	-0.04***	-0.04***	-0.06**
	(0.01)	(0.01)	(0.24)	(0.01)	(0.01)	(0.25)
Observations	345	345	345	221	221	221
R^2	0.011	0.02	0.029	0.036	0.038	0.05

*Notes: Results from ordinary least squares regressions. The dependent variable is the difference between participant's elicited posteriors and the extrapolated Bayesian posteriors. The reference category for the experimental condition is not being audited. Cheat 1st round is a dummy equal to 1 when the individual falsely reports a successful guess in the first round. Demographics include age, a gender dummy, college dummy, a race dummy, and a dummy for a yearly income of more than $50,000. Robust standard errors are in parentheses. *, **, and *** indicate significance at the 10%, 5%, and 1% level.*

Finally, to complement our analysis with a qualitative measure for the hypothesized bias, we asked subjects in a post-experimental survey to indicate to what extent, if at all, was being notified that they were audited (or not audited) after the first round affected their belief regarding the likelihood of being audited in the second round, on a ten-point Likert-scale ($1 =$ not at all, $10 =$ extremely affected). Consistently with our main findings, participants who were not audited responded with significantly lower scores than their audited counterparts, with an average score of 5.9 and 6.5, respectively ($t = -2.1, p = 0.0352$).

6.2. Deterrence: The effect of learning on behavior

Following the model presented in section 3, the classic economic model of deterrence predicts that an individual will commit an offense if and only if their gain exceeds the expected sanction – the product of the fine and the perceived probability of detection. In our experimental design, where the sanction and the gain remain unchanged, being audited in the first round induces an upward adjustment of beliefs, which *increases* subjects' expected sanction, hence reduces their incentives to cheat in the second round. By the same token, not being audited, which induces a downward adjustment of beliefs, *reduces* subjects' expected sanction, rendering subsequent cheating more attractive. Indeed, we observe this general deterrent effect in the data: while the pre-treatment first-round cheating rate is not statistically different across treatments ($M_A = 31\%, M_{NA} = 32\%, p = 0.818$ in a two-sided fisher-exact test), the second-round cheating rate is 13 percentage points higher among those who were not audited ($M_A = 25\%\,, M_{NA} = 38\%, p = 0.011$ in a two-sided fisher-exact test).

To more closely examine whether the observed pattern is generated through the channel of beliefs, Table 4 presents the results of simple ordinary least squares regressions where the dependent variable is a dummy indicating whether a person cheated in the second round. As expected, we find that the posterior belief regarding the probability of detection has a negative and significant effect on the decision to cheat in the second round, which is robust to the inclusion of various pre-treatment controls.[31]

[31] Appendix Table C.2 re-estimates the baseline results of Table 4 for the pre-treatment analog, i.e., substituting priors for posteriors and first-round cheating for second round cheating, estimating similar effects.

Table 4: the effect of beliefs on the decision to cheat

Dependent Variable: Second-Round Cheating Dummy			
	(1)	(2)	(3)
Belief 2nd round	-0.932***	-0.477**	-0.448**
	(0.20)	(0.19)	(0.18)
Cheat 1st round		0.627***	0.640***
		(0.05)	(0.05)
Demographics	NO	NO	YES
Constant	0.85***	0.38***	0.20
	(0.11)	(0.11)	(0.12)
Observations	292	246	246
R^2	0.07	0.44	0.45

*Notes: Results from ordinary least squares regressions. Estimation samples are restricted to subjects who did not guess correctly in the second round. Belief 2nd round is the participants' reported estimate regarding the probability of being audited in the second round. Cheat 1st round is a dummy equal to 1 when the individual falsely reports a successful guess in the first round. Demographics include age, a gender dummy, college dummy, a race dummy, and a dummy for a yearly income of more than $50,000. Robust standard errors are in parentheses. *, **, and *** indicate significance at the 10%, 5%, and 1% level.*

While our design does not allow us to rigorously isolate the effect of the non-occurrence bias on behavior, a further exploratory analysis reveals two interesting patterns.

First, whereas our baseline model assumes a risk-neutral individual that bears no non-monetary cost from cheating, the theoretical and empirical literature has long emphasized the potential role of individuals' attitude toward risk (e.g, Kaplow & Shavell, 1994; Friehe, Langenbach, & Mungan, 2023), and distaste for cheating (e.g, Gneezy, 2005; Fischbacher & Föllmi-Heusi, 2013; Abeler, Becker, & Falk, 2014) in their decision to engage in criminal or unethical behavior.[32] While variability in risk preferences is less of a problem in our design, as the stakes involved are relatively small and lie in a range in which people are considered to be approximately risk-neutral (Rabin, 2000), distaste for cheating operates in the reverse direction: the smaller the payoff, the higher the potential effect of distaste for cheating on behavior. Table 4 provides suggestive evidence for the power of distaste for cheating to affect behavior, showing that across all specifications, first-round cheating has strong predictive power for second round-cheating, even when beliefs are held constant ($p < 0.001$), increasing the R-squared by 37

[32] Indeed, some responses to our post-treatment survey open-ended question suggest that their choice to report honestly was driven by distaste for cheating (for example: "*I wanted to be fair*"; "*I didn't want to have the bonus based on lying about it*").

percentage points in column 2.[33] Recall that the experiment was designed such that if respondents were risk neutral and have no distaste for cheating, we should expect a cheating rate of 100% in both rounds, regardless of the beliefs regarding the probability of being audited. With this in mind, we derived the expected monetary loss from reporting truthfully by those who did not cheat. We find that this cohort lost \$0.36 on average in the first round ($Min = \$0.125$, $Max = 0.625$), and \$0.33 on average in the second round ($Min = \$0.125$, $Max = 0.625$), which can be interpreted as the lower bound for each participant's monetary equivalence of their distaste for cheating. This significant difference in the sensitivity to economic incentives suggests that a distaste for cheating (offending) has an important role in deterrence. [34]

A second interesting question worth exploring is the extent to which adjustment of beliefs in response to one's detection experience affects subsequent decision to cheat, and whether it differs between those who were audited and those who were not. To answer this question, we estimate the following model:

$$Cheat_{i,2} = \beta_0 + \beta_1 Audit_{i,1} + \beta_2 Change_i + \beta_3 Audit_{i,1} \times Change_i + \beta_4 Prior_i + \beta_5 Cheat_{i,1} + \beta_6 X_i + \varepsilon_i$$

Where $Cheat_{i,1}$ and $Cheat_{i,2}$ are dummies for first- and second-round cheating, respectively; $Change_i$ is 100 times the absolute difference between reported posteriors and prior beliefs; and X_i represents a constant and additional set of covariates. Column (1) of Table 5 shows that a one percentage point (downward) adjustment in reported beliefs in response to not being audited *increases* the likelihood of subsequent cheating by 1.2 percentage points, whereas a similar (upward) adjustment in response to being audited reduces the likelihood of subsequent cheating only by 0.5 percentage points. Namely, the effect of a larger change in beliefs on subsequent cheating is more than *twice* stronger for those who were not audited. However, this difference is

[33] Since the decision to cheat in the first round was made before the treatment was administered, it constitutes a powerful proxy for participants' distaste for cheating.

[34] To further explore the potential effect of distaste for cheating on behavior in our setting, Appendix Figure C.4. compares the second-round cheating rate across treatments separately for those who cheated in the first round and those who did not. Appendix Table D.2 further presents results of an ordinary least squares regressions of first-round cheating behavior on various pre-treatment characteristics that reveal that of the information collected, only gender is predictive of first-round cheating, such that women are 11.6% less likely than men to falsely report correctly guessing the roll of the dice ($p = 0.043$).

only marginally significant at conventional levels ($p = 0.068$) and becomes non-significant when controlling for a first-round cheating dummy in column 2. To complement this finding, Table 5 shows that being audited reduces the likelihood of subsequent cheating, with $\beta_1 < 0$ across all specifications, though these estimates are not statistically significant. This small and insignificant effect implies that detection experience affects subsequent behaviour *only through the channel of beliefs*, further emphasizing its important role in controlling and affecting crime.

Table 5: how adjustment of beliefs affect behavior

Dependent Variable: Second-Round Cheating Dummy			
	(1)	(2)	(3)
Audit (β_1)	-0.046	-0.060	-0.076
	(0.06)	(0.05)	(0.05)
Change (β_2)	0.012***	0.007**	0.006**
	(0.00)	(0.00)	(0.00)
Change × Audit (β_3)	-0.017***	-0.010**	-0.009**
	(0.00)	(0.00)	(0.00)
Prior (β_4)	-0.823***	-0.046	0.011
	(0.31)	(0.23)	(0.24)
Cheat 1st round (β_5)		0.635***	0.650***
		(0.05)	(0.05)
$\mathbb{P}(\beta_2 = \lvert\beta_2 + \beta_3\rvert)$	0.068	0.387	0.371
Demographics (β_6)	NO	NO	YES
Constant	0.79***	0.19	-0.00
	(0.15)	(0.12)	(0.15)
Observations	292	246	246
R^2	0.07	0.45	0.47

*Notes: Results from ordinary least squares regressions. Estimation samples are restricted to subjects who did not guess correctly in the second round. All demographics are described in Table 3. Robust standard errors are in parentheses. *, **, and *** indicate significance at the 10%, 5%, and 1% level.*

Recognizing the debate in the econometric literature concerning the relative merits of various binary dependent variable models (Angrist & Pischke, 2009), in the Online Appendix, we re-estimate all baseline tables in this section using logit and probit models and estimate similar marginal effects.

7. Discussion and Concluding Remarks

This paper is the first to explore whether potential repeat offenders learn about the probability of detection from *being* caught at a different level of accuracy compared to learning from *not being* caught. We find that not being caught induces a *weaker* adjustment of beliefs than the one induced by not being caught, which we call a non-occurrence bias. This is because, in our design, both events – being audited and not being audited – carry the same informational power, but the former is framed as something that has occurred, despite both being presented with identical saliency.

As illustrated in our model, a non-occurrence bias implies that enforcement policy grounded in the rational-choice theory of specific deterrence may result in excessive investment in enforcement. The reason is that the marginal gain from specific deterrence, intuitively embodied in the apprehension of one additional offender, is a function of the wedge between the offender's belief regarding the likelihood of being caught after learning about it from being caught and his belief after learning about it from not being caught. A non-occurrence bias means that this wedge is smaller than predicted by a rational-choice model, and thus that the equilibrium level of investment in enforcement is higher than the optimal level given the non-occurrence bias.

There are two reasons to suspect that the measured non-occurrence bias reflects an underestimation of the magnitude of the phenomenon. First, in our setting, the signals of being audited and not being audited are *equally salient*. In both cases, the participant is actively notified of the drawn ball, accompanied by an animation of a jumping ball with the matching label and color. In non-laboratory settings, however, not being caught (i.e., committing an offense without being arrested) is typically not salient at all, which is likely to further weaken its informational effect (Tversky & Kahneman, 1974). Second, the non-occurrence bias in the context of law enforcement operates in the opposite direction of the established "*good-news-bad-news*" bias, whereby people tend to overweight good news and discount bad news. In the context of specific deterrence, not being caught reflects the "good news," and nonetheless, the signal is found to be weaker. Therefore, one may suspect that the actual magnitude of the non-occurrence bias would likely be larger if it was decoupled from the offsetting "good news" effect, an avenue that should be pursued in future research.

A useful (though not indispensable) feature of our design was that the accurate prior belief regarding the probability of detection is 50%. As mentioned earlier, this feature simplified the analysis because it meant that both treatments – being audited or not being audited – should induce the *same* absolute average change in one's estimate, to one direction or the other. However, in most circumstances, the probability of apprehension is significantly lower than 50% (Shavell 1993), and hence priors, through investment in general deterrence, are expected to be significantly lower than 50% on average. Formally, it means that not being caught should rationally induce a smaller adjustment of beliefs compared to being caught. Further research should test whether the non-occurrence bias changes in form or magnitude under such a richer setting, in the light of some evidence that systematic deviations from Bayes' rule may vary across different values of individuals' priors (see, e.g., Holt and Smith 2009; Coutts 2018).

References

Abeler , J., Becker, A., & Falk, A. (2014). Representative evidence on lying costs. *Journal of Public Economics, 113*, 96-104.

Angrist, J. D., & Pischke, J.-S. (2009). *Mostly Harmless Econometrics: An Empiricist's Companion.* New Jersey: Princton University Press.

Armantier, O., & Treich, N. (2013). Eliciting beliefs: Proper scoring rules, incentives, stakes and hedging. *European Economic Review, 62*, 17-40.

Bebchuk, L., & Kaplow, L. (1992). Optimal Sanctions When Individuals Are imperfectly Informed about the Probability of Apprehension. *Journal of legal studies, 21*.

Becker, G. S. (1968). Crime and Punishment: An Economic Approach. *Journal of Political Economy, 76*, 169-217.

Ben-Shahar, O. (1997). Playing without a rulebook: Optimal enforcement when individuals learn the penalty only by committing the crime. *International Review of Law and Economics, 17*, 409-421.

Bentham, J. (1789). *An Introduction to the Principles of Morals and Legislation.* Garden City, New York: Anchor Books.

Bouhassoun, S., Naveau, M., & Delcroi, N. (2022). Approach in green, avoid in red? Examining interindividual variabilities and personal color preferences through continuous measures of specific meaning associations. *Psychological Research*.

Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's Mechanical Turk: A New Source of Inexpensive, Yet High-Quality, Data? *Perspectives on Psychological Science, 6*, 3-5.

Chen, K. M., & Shapiro, J. M. (2007). Do Harsher Prison Conditions Reduce Recidivism – A Discontinuity-Based Approach. *American Law and Economics Review, 9*, 1-29.

Coutts, A. (2018). Good news and bad news are still news: experimental evidence on belief updating. *Experimental Economics, 22*.

Crump, M. J., McDonnell, J. V., & Gureckis, T. M. (2013). Evaluating Amazon's Mechanical Turk as a Tool for Experimental Behavioral Research. *Plos One, 8*, 1-18.

Cullen, F. T., Cheryl , L., & Nagin, D. S. (2011). Prisons Do Not Reduce Recidivism: The High Cost of Ignoring Science. *The Prison Journal, 91*, 48-65.

Drago, F., Galbiati, R., & Vertova, P. (2011). Prison Conditions and Recidivism. *American Law and Economics Review, 13*, 103-130.

Dušek, L., & Traxler, C. (2022). Learning from Law Enforcement. *Journal of the European Economic Association, 2*, 739-777.

Eeckhout, J., Persico, N., & Todd, P. E. (2010). A Theory of Optimal Random Crackdowns. *American Economic Review, 100*, 1104-1135.

Eil, D., & Rao, J. M. (2011). The Good News-Bad News Effect: Asymmetric Processing of Objective Information about Yourself. *American Economic Journal: Microeconomics, 3*, 114-138.

Elliot, A. J., Maier, M. A., Moller, A. C., Friedman, R., & Meinhardt, J. (2007). Color and psychological functioning: The effect of red on performance attainment. *Journal of Experimental Psychology: General, 136*, 154-168.

Fischbacher, U., & Föllmi-Heusi , F. (2013). Lies in Disguise - An Experimental Study on Cheating. *Journal of the European Economic Association, 11*, 525-547.
Fischbacher, U., & Föllmi-Heusi, F. (2013). Lies in Disguise - An Experimental Study on Cheating . *Journal of the European Economic Association, 11*, 525-547.
Friehe, T., Langenbach, P., & Mungan, M. C. (2023). Sanction Severity Influences Learning About Enforcement Policy: Experimental Evidence. *Journal of Legal studies (forthcoming)*.
Gerend, M. A., & Sias, T. (2009). Message framing and color priming: How subtle threat cues affect persuasion. *Journal of Experimental Social Psychology, 45*, 999-1002.
Gneezy, U. (2005). Deception: The Role of Consequences. *American Economic Review, 95*, 384-394.
Grether, D. M. (1980). Bayes Rule as a Descriptive Model: The Representativeness Heuristic. *Querterly Journal Economics, 95*.
Harrison, G., Martínez-Correa , J., & Swarthout, T. (2014). Eliciting Subjective Probabilities with Binary Lotteries. *Journal of Economic Behavior and Organization, 101*, 128-140.
Hjalmarsson, R. (2009). Juvenile Jails: A Path to the Straight and Narrow or to Hardened Criminality. *Journal of Law and Economics, 52*, 779-806.
Huizinga, D., Matsueda, R. L., & Kreage, D. A. (2006). Deterring Delinquents: A Rational Choice Model of Theft and Violence. *American Sociological Review, 71*, 95-122.
Kaplow, L. (1990). Optimal Deterrence, Uninformed Individuals, and Acquiring Information about Whether Acts Are Subject to Sanctions. *Journal of Law, Economics & Organization, 6*, 93-128.
Kaplow, L., & Shavell, S. (1994). Accuracy in the Determination of Liability. *Journal of Public Economics, 37*, 1-15.
Kuhnen, C. M. (2014). Asymmetric Learning from Financial Information. *The Journal of Finance, 70*, 2029-2062.
Lochner, L. (2007). Individual Perceptions of the Criminal Justice System. *American Economic Review, 97*, 444-460.
M.Smith, C. A. (2009). An update on Bayesian updating. *Journal of Economic Behavior & Organizations, 69*.
Mckelvey, R. D., & Page, T. (1990). Public and Private Information: An Experimental Study of Information Pooling. *Econometrica, 58*, 1331-1339.
Miceli, T. J., Segerson, K., & Earn, D. (2022). The role of experience in deterring crime: A theory of specific versus general deterrence. *Economic Inquiry, 60*, 1833-1853.
Nagin, D. S. (2013). Deterrence: A Review of the Evidence by a Criminologist for Economists. *Annual Review of Economics, 5*, 83-105.
Nagin, D. S., Cullen, F. T., & Jonson, C. L. (2009). Imprisonment and Reoffending. *Crime & Justice, 38*, 115-200.
Polinsky, M. M., & Shavell, S. (1979). The optimal tradeoff between the probability and magnitude of fines. *American Economic Review, 69*, 880-891.
Rabin, M. (2000). Risk Aversion and Expected-Utility Theory: A Calibration Theorem. *Econometrica, 68*, 1281-1292.
Sah, R. K. (1991). Social Osmosis and Patterns of Crime. *Journal of Political Economy, 99*, 1272-1295.
Schargrodsky, E., & Di Tella, R. (2013). Criminal Recidivism after Prison and Electronic Monitoring. *Journal of Political Economy, 121*, 28-69.

Shalvi, S., Eldar, O., & Bereby-Meyer, Y. (2012). Honesty Requires Time (and Lack of Justifications). *Psychological Science, 23*, 1264-1270.
Shamena, A., & Loughran, T. (2011). Testing a Bayesian Learning Theory of Deterrence among Serious Juvenile Offenders . *Criminology, 49*, 667-694.
Sharot, T., Korn, C. K., & Dolan, R. J. (2011). How unrealistic optimism is maintained in the face of reality. *Nature Neuroscience, 14*, 1475-1479.
Shavell, M. P. (1999). On the Disutility and Discounting of Imprisonment and the Theory of Deterrence. *Journal of Legal Studies, 2*.
Shavell, S. (1991). Specific versus general enforcement of law. *Journal of Political Economy, 99*, 1088-1108.
Shavell, S. (1993). The Optimal Structure of Law Enforcement. *Journal of Law and Economics, 36*, 255-287.
Shavell, S. (2004). *Foundations of Economic Analysis of Law.* Cambridge, Massachusetts: Harvard University Press.
Sunstein, C., Bobadilla-Suarez, S., & Lazzaro, S. C. (2016). How People Update Beliefs about Climate Change: Good News and Bad News. *Cornell Law Review, 102*, 1431-1443.
Tversky, A., & Kahneman, D. (1974). Judgment under Uncertainty: Heuristics and Biases. *Science, 185*.

Appendix

A. Proof of proposition

Let R denote the partial derivative of the social costs function with respect to enforcement expenditures *e*, as given by expression (3). Using the general formula for the derivative of an implicit function, we can now demonstrate that the optimal investment in law enforcement is decreasing in the magnitude of the non-occurrence bias, i.e., $\frac{\partial e^*}{\partial \beta} = -\frac{R_\beta}{R_{e^*}} > 0$.

First, taking the derivative of R with respect to β (omitting e^* for expositional purposes):

$$(4)\ R_\beta = \left[f'\left(\left(p_1 - \frac{\beta {\Delta_1}^2}{1-p_1} \right) s \right) \cdot \left(\beta \frac{\Delta_1^2}{(1-p_1)^2} - 1 \right) {p_1}' \Delta_1^2 s^2 \beta \right]$$

Note that the first term in expression (4), $f'\left(\left(p_1 - \frac{\beta {\Delta_1}^2}{1-p_1} \right) s \right)$, is negative by virtue of the assumption that the gain of individuals who choose to commit the offense is larger than the peak of $f(g)$; and that the second term in brackets, $\left(\beta \frac{\Delta_1^2}{(1-p_1)^2} - 1 \right)$, is also negative, given that $p^H = p + \Delta_1 < 1$ by design (hence $1 - p > \Delta_1 \leftrightarrow \frac{{\Delta_1}^2}{(1-p_1)^2} < 1 \leftrightarrow \beta \frac{{\Delta_1}^2}{(1-p_1)^2} < 1 \leftrightarrow \beta \frac{{\Delta_1}^2}{(1-p_1)^2} - 1 < 0$). Further note that p_1' is positive by virtue of the assumption that $p(e)' > 0$ (as $p_1 = p$ by design). The rest of the elements in (4) $-$ Δ_1, s, β $-$ are positive by design, hence R_β is positive.

Next, taking the derivative of R with respect to e^*:

$$(5)\ R_{e^*} = h \cdot \Bigg[\overbrace{s \cdot \left\{ p_1'' \cdot f(p_1 s) + (p_1')^2 \cdot f'(p_1 s) \cdot s \right\}}^{1} + \overbrace{\left\{ \begin{matrix} S\left[\left(2{p_1}'^2 + p p_1'' \left(1 - \frac{{\Delta_1}^2}{{p_1}^2} \right) \right) \cdot f\left(\left(p_1 + \frac{{\Delta_1}^2}{p_1} \right) s \right) \right] + \\ S^2 \left[p \cdot \left({p_1}' - \frac{{p_1}' {\Delta_1}^2}{{p_1}^2} \right)^2 \cdot f'\left(\left(p_1 + \frac{{\Delta_1}^2}{p_1} \right) s \right) \right] + \left[-p'' \cdot \int_{\left(p_1 + \frac{{\Delta_1}^2}{p_1} \right) s}^{\infty} f(g) dg \right] \end{matrix} \right\}}^{2} + \overbrace{S \left\{ \begin{matrix} S\left[\left(-2{p_1}'^2 + (1-p) p_1'' \left(1 - \beta \frac{{\Delta_1}^2}{(1-p_1)^2} \right) \right) \cdot f\left(\left(p_1 - \beta \frac{{\Delta_1}^2}{(1-p_1)^2} \right) s \right) \right] \\ + S^2 \left[(1-p) \cdot \left({p_1}' - \frac{\beta {p_1}' {\Delta_1}^2}{(1-p_1)^2} \right)^2 \cdot f'\left(\left(p_1 - \frac{\beta {\Delta_1}^2}{1-p_1} \right) s \right) \right] + \left[p'' \cdot \int_{\left(p_1 - \frac{\beta {\Delta_1}^2}{1-p_1} \right) s}^{\infty} f(g) dg \right] \end{matrix} \right\}}^{3} \Bigg]$$

Starting with the first term in brackets in (5), $p_1'' \cdot f(p_1 s) + \left(p_1{}'\right)^2 \cdot f'(p_1 s) \cdot s$, notice that p_1'' is negative under the assumption that there is a decreasing marginal return to investment in enforcement, hence $p'' < 0$, and that $f'(p_1 s)$ is negative by virtue of the assumption that the gain of individuals who choose to commit the offense is larger than the peak of $f(g)$. The rest of the elements in the first term are positive, given the aforementioned assumptions. Hence the first term is negative.

Moving next to the second and third terms in (5), observe that each of these terms is the summation of three parallel components in brackets. Starting with the first component of the second term, $\left[\left(2{p_1}'^2 + p p_1'' \left(1 - \frac{{\Delta_1}^2}{{p_1}^2} \right) \right) \cdot f\left(\left(p_1 + \frac{{\Delta_1}^2}{p_1} \right) s \right) \right]$, while this component can be either positive or negative, its summation with its parallel in the third term, $\left[\left(-2{p_1}'^2 + (1-p) p_1'' \left(1 - \beta \frac{{\Delta_1}^2}{(1-p_1)^2} \right) \right) \cdot f\left(\left(p_1 - \beta \frac{{\Delta_1}^2}{(1-p_1)^2} \right) s \right) \right]$, is clearly negative. To see this, note that the summation of $2{p_1}'^2 f\left(\left(p_1 + \frac{{\Delta_1}^2}{p_1} \right) s \right)$ and $-2{p_1}'^2 f\left(\left(p_1 - \beta \frac{{\Delta_1}^2}{(1-p_1)^2} \right) s \right)$ is negative, as $f\left(\left(p_1 + \frac{{\Delta_1}^2}{p_1} \right) s \right) < f\left(\left(p_1 - \frac{\beta {\Delta_1}^2}{1-p_1} \right) s \right)$ given

the assumption that $f(g) > 0$ and the fact that $\left(p_1 - \frac{\beta\Delta_1^2}{1-p_1}\right) < \left(p_1 + \frac{\Delta_1^2}{p_1}\right)$ by design. Also notice that $pp_1''\left(1 - \frac{\Delta_1^2}{p_1^2}\right) f\left(\left(p_1 + \frac{\Delta_1^2}{p_1}\right)s\right)$ and $p_1''(1-p)\left(1 - \beta\frac{\Delta_1^2}{(1-p_1)^2}\right) f\left(\left(p_1 - \beta\frac{\Delta_1^2}{(1-p_1)^2}\right)s\right)$ are both negative under the assumptions above (i.e. $p_1'' < 0$ and $p, f(g) > 0$); the fact that $p^L = p - \Delta_1 > 0$ by design (hence $p > \Delta_1 \leftrightarrow \frac{\Delta_1^2}{p_1^2} < 1 \leftrightarrow 1 - \frac{\Delta_1^2}{p_1^2} > 1$); and the fact that $p^H = p + \Delta_1 < 1$ by design (hence $1 - p > \Delta_1 \leftrightarrow \frac{\Delta_1^2}{(1-p_1)^2} < 1 \leftrightarrow \beta\frac{\Delta_1^2}{(1-p_1)^2} < 1 \leftrightarrow 1 - \beta\frac{\Delta_1^2}{(1-p_1)^2} > 0$).

The second component of both the second and third terms, $\left[p \cdot \left(p_1' - \frac{p_1'\Delta_1^2}{p_1^2}\right)^2 \cdot f'\left(\left(p_1 + \frac{\Delta_1^2}{p_1}\right)s\right)\right]$ and $\left[(1-p) \cdot \left(p_1' - \frac{\beta p_1'\Delta_1^2}{(1-p_1)^2}\right)^2 \cdot f'\left(\left(p_1 - \frac{\beta\Delta_1^2}{1-p_1}\right)s\right)\right]$ respectively, is also negative given the aforementioned assumptions.

Finally, the summation of the last components of the second and third terms, $\left[-p'' \cdot \int_{\left(p_1 + \frac{\Delta_1^2}{p_1}\right)s}^{\infty} f(g)dg\right]$ and $\left[p'' \cdot \int_{\left(p_1 - \frac{\beta\Delta_1^2}{1-p_1}\right)s}^{\infty} f(g)dg\right]$ is also negative, as the perceived probability of individuals who were audited $\left(p_1 + \frac{\Delta_1^2}{p_1}\right)$, is higher than that of individuals who were not audited $\left(p_1 - \frac{\beta\Delta_1^2}{1-p_1}\right)$ by design, hence $\int_{\left(p_1 + \frac{\Delta_1^2}{p_1}\right)s}^{\infty} f(g)dg < \int_{\left(p_1 - \frac{\beta\Delta_1^2}{1-p_1}\right)s}^{\infty} f(g)dg$.

Since we demonstrated that the first term in (5) is negative, and that the summation of the second and third terms is negative, we demonstrated that R_{e^*} is negative as well. Therefore, as $R_\beta > 0$ and $R_{e^*} < 0$ it follows that $\forall e^*, \beta$: $\frac{\partial e^*}{\partial \beta} = -\frac{R_\beta}{R_{e^*}} > 0$. Q.E.D.

B. Experimental protocol

Consent form

Description: You are invited to participate in a research study on decision-making and economic behavior. You will be asked to play a game and to fill out a demographic survey. You must be at least 18 years of age to participate. Your participation will take about 10 minutes. There are no risks associated with this study, and your identity will be kept confidential.

Payment: In addition to the $1 payment for taking this HIT, you will receive a bonus payment of up to $5, based partially on your performance, partially on your decisions, and partially on chance.

Participant's rights: If you decide to participate in this experiment, please note that your participation is voluntary and that you may withdraw your consent or discontinue participation at any time without penalty. Your privacy will be maintained in all published and written data resulting from the study. Your name will never be connected to any decision you make. For scientific reasons, you may be unaware of the study hypotheses and the research questions being tested.

Contact Information: If you have any questions, concerns, or complaints about this research, its procedures, risks or benefits, contact the Protocol Director, at zurlab975@gmail.com.

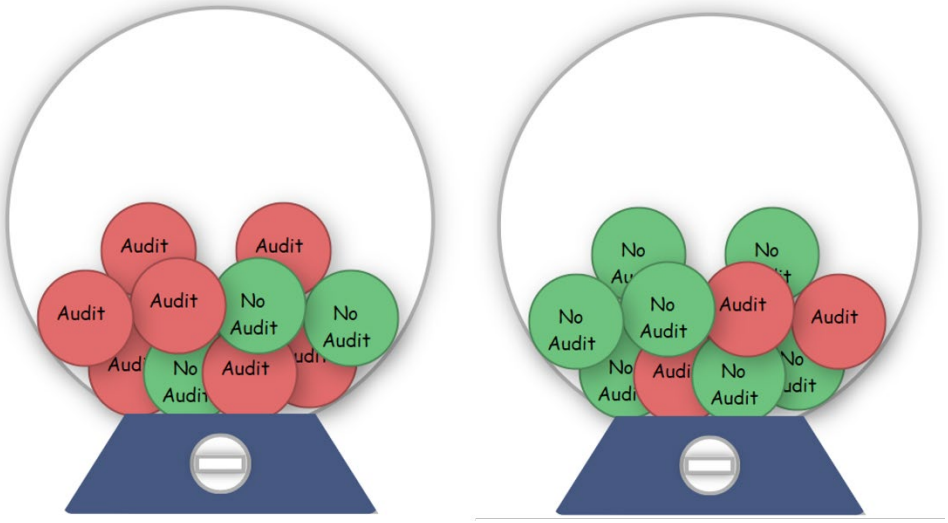
By clicking on the link below, you confirm that you have read the consent form, are at least 18 years old, and agree to participate in the research.

Introductory Instructions

This experiment includes two identical rounds. In each round, we will ask you to perform a simple task of guessing the outcome of a dice roll. You will be asked to report whether your guess was correct, and the payment will be based on your report: **$1.50** for a correct guess and **$0.50** otherwise. However, after each round, some participants, **randomly** chosen, will be notified that their report will be audited. If your report is audited, falsely reporting a correct guess will reduce your payment to **$0.25**.

The decision whether to inspect your report in each round will be determined by the computer randomly drawing a ball from a gumball machine that was loaded with **10** balls labeled either "Audit" or "No Audit." The computer has randomly assigned you, by a flip of a coin, either to a gumball machine with **7** red balls labeled "Audit" and **3** green balls labeled "No Audit"; or to a gumball machine with **3** red balls labeled "Audit" and **7** green balls labeled "No Audit". Once the

computer has assigned you to one of the gumball machines, **that machine will be used in both rounds**, but you will not know for certain which one it is.



You will also have a chance to receive an additional bonus of **$1.00** in each round by estimating the probability of being audited. In the first round, you will not have any concrete information about which gumball machine was assigned to you. In the second round, however, you will know whether or not you were audited in the first round, which will allow you to improve your estimate.

Final notes: the outlined procedure is completely true and will be strictly followed by the experimenter. **Remember:** your identity will never be connected to any decision you make and will not affect you negatively in any manner or shared with Amazon Mechanical Turk or anyone else.

Comprehension test

To ensure that you understand the instructions, please answer the following questions. You must

answer these questions correctly to be eligible to receive the bonuses in the following sections.

What is the color of the ball that will result in an audit?

What are the odds that you will be assigned to a gumball machine with 3 green balls?

How many times you will be asked to guess the outcome of a dice roll?

One of the gumball machines has more green balls than the other. How many green balls does it have?

Back

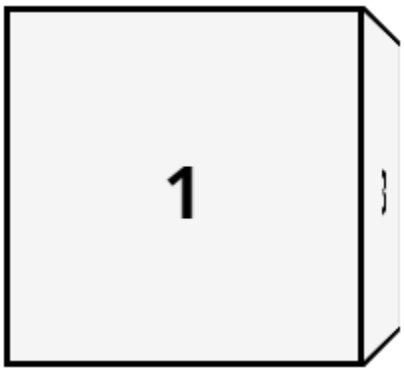
Next

Please guess a number from 1 to 6:

- 1
- 2
- 3
- 4
- 5
- 6

Next

To roll the dice, please click on the “Roll!” button below, and make a mental note of whether you guessed correctly



Roll!

Before moving to the next screen, feel free to press the 'Roll!' button a few more times to see that the dice is fair.

Next

Did your guess match the number in the first roll?

- Yes
- No

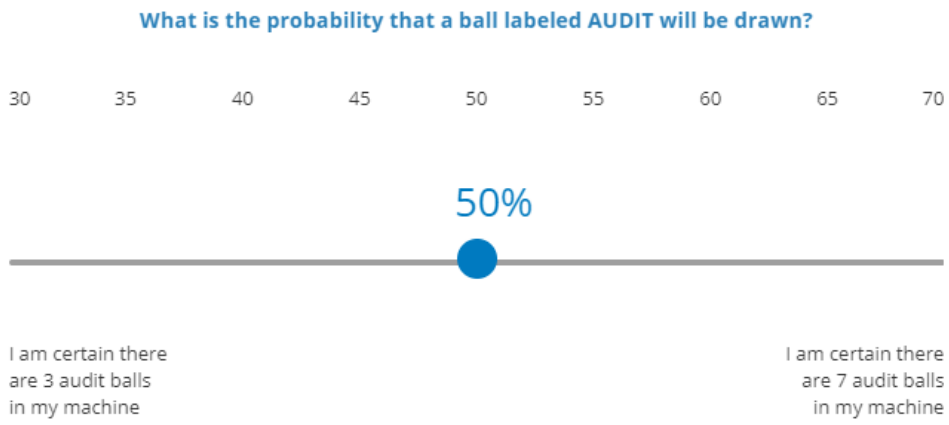
Remember: your bonus payment will be either $0.50 or $1.50, based on this report only, unless you will be chosen to be audited based on the ball drawn from your gumball machine, in which case falsely reporting a correct guess will reduce your payment to $0.25.

Next

Before the computer draws a ball from your assigned gumball machine to determine whether your report in this round will be audited, you have a chance to win an **additional** bonus payment of **$1.00** by reporting your best estimate regarding the likelihood that you will be audited, given the information you have.

You will enter your estimate by adjusting the slider at the bottom of the screen to the number that best reflects your belief regarding the chance you will be audited. Since your gumball machine contains either 7 red audit balls or 3 red audit balls (out of 10), your estimate is limited to a range between 70% (if you are certain that your assigned machine is the one with 7 audit balls) and 30% (if you are certain that your assigned machine is the one with 3 audit balls). If you are uncertain, the correct estimate lies somewhere in between.

Your estimate will be used to determine whether you win the additional $1.00, by means of a lottery conducted by the computer at the end of the experiment. **The probability of winning the bonus is higher the closer you are to the correct estimate.** If you want, you can click here to see how the accuracy of your estimation increases the chance of earning the $1.00 bonus.



Once you are ready, click "Next" to draw a ball from the gumball machine.

Next

The computer randomly drew an "**audit**" ball. Therefore, you will be audited, and your bonus payment in this round will be based on whether your guess matches the roll of the dice, which will be verified after completing the experiment.

Next

Now, we will ask you to perform the same task again, for additional bonuses. As before, your payment will be based on your report, unless it will be randomly chosen for inspection. Your assigned gumball machine from the first round is the same one that will be used in the second round, and the ball that was drawn in the first round was returned to the gumball machine.

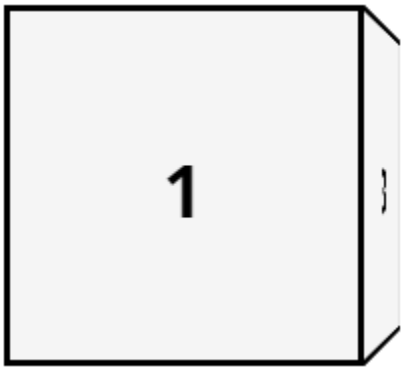
Please click on the "Next" button below to start the second round.

Please guess a number from 1 to 6:

- 1
- 2
- 3
- 4
- 5
- 6

Next

To roll the dice, please click on the "Roll!" button below, and make a mental note of whether you guessed correctly



Roll!

Before moving to the next screen, feel free to press the 'Roll!' button a few more times to see that the dice is fair.

Did your guess match the number in the first roll?

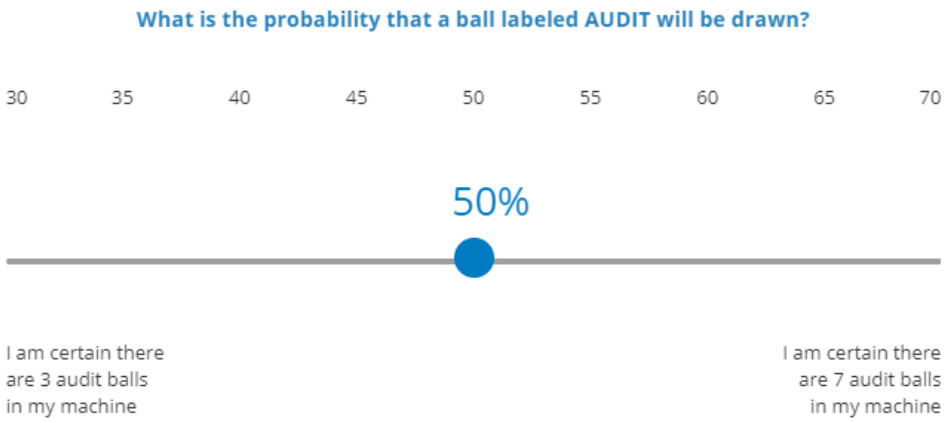
- Yes
- No

Remember: your bonus payment will be either $0.50 or $1.50, based on this report only, unless you will be chosen to be audited based on the ball drawn from your gumball machine, in which case falsely reporting a correct guess will reduce your payment to $0.25.

Again, we will ask you to report your estimate regarding the likelihood that you will be audited, for a chance to win an **additional** bonus payment of $1.00, by adjusting a slider at the bottom of the screen. The slider is currently set to your previous estimate, and you are free to change as you see fit.

Note: since a red "audit" ball was drawn in the first round, you can use this information to improve your estimation regarding which gumball machine was assigned to you and, consequently, what are your chances of being audited.

As before, **your chance of winning the additional $1.00 is higher the closer you are to the correct estimate.** If you want, you can click here to see how the accuracy of your estimation increases the chance of earning the bonus.



Once you are ready, click "Next" to draw a ball from the gumball machine.

Next

The computer randomly drew a "**no audit**" ball. Therefore, you will not be audited, and your bonus payment in this round will be based on your report.

Next

Please answer the following questions. **Remember:** Your identity will remain confidential.

To what extent, if at all, was being notified that you were audited after the first round affected your belief regarding the likelihood of being audited in the second round? Please answer on a scale of 1 to 10 [1=not at all; 10=extremely affected]

1	2	3	4	5	6	7	8	9	10
○	○	○	○	○	○	○	○	○	○

Please provide a short explanation of what drove you to update your estimation the way you did, or not to update it, if you did not change it. Feel free to provide any other feedback on the instructions, user interface, etc.

In which state do you currently reside?

What is your age?

What is the highest level of school you have completed or the highest degree you have received?

- Less than high school degree
- High school graduate (high school diploma or equivalent including GED)
- Some college but no degree
- Associate degree in college (2-year)
- Bachelor's degree in college (4-year)
- Master's degree
- Doctoral degree
- Professional degree (JD, MD)

Choose one or more races that you consider yourself to be:

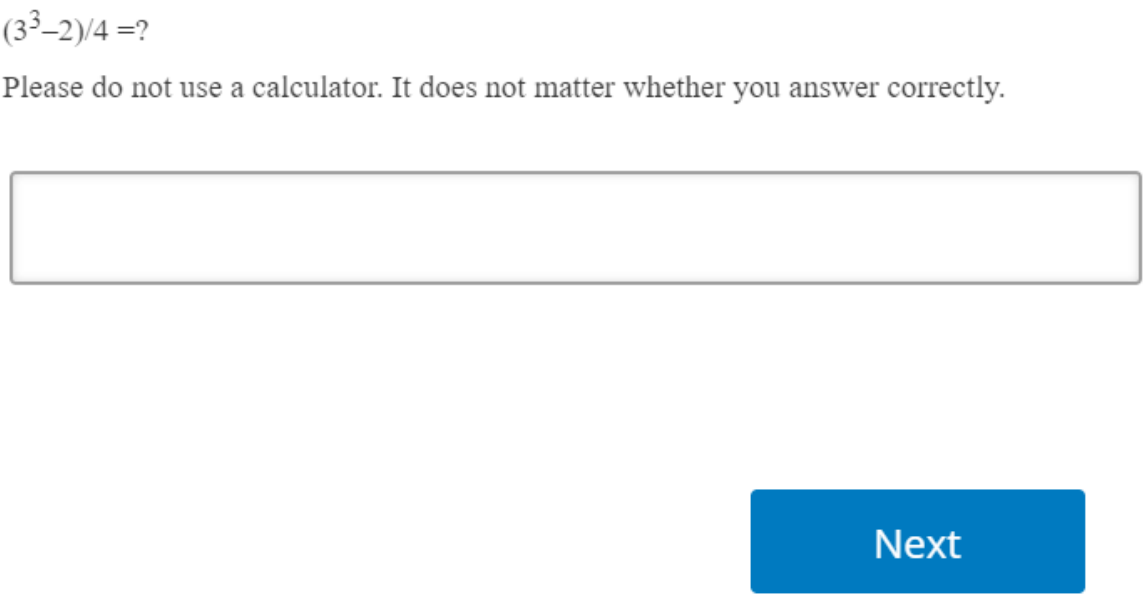
White/Caucasian ☐	Asian ☐
African American ☐	Native Hawaiian/Pacific Islander ☐
Native American ☐	Other ☐

What is your gender?

- Male ○
- Female ○
- Transgender Male ○
- Transgender Female ○
- Non-binary ○
- Other ○

Which of these describes your annual income last year?

- $0
- $1 to 9,999$
- $10,000 to 24,999$
- $25,000 to 49,999$
- $50,000 to 74,999$
- $75,000 or more
- prefer not to say



A. Robustness: Beliefs

A.1. Restricting to Active Updates

An interesting question pertains to whether the observed asymmetry extends to the decision of *whether* to adjust one's belief or not in response to the observed signal, as opposed to the magnitude of the adjustment once it occurs. Indeed, our aggregate date included a non-negligible number of non-updates, where 25% of reported posteriors were identical to reported priors. Our analysis reveals, however, that the number of individuals who did not revise their belief in response to their detection experience does not differ across treatments ($M_A = 26.3\%, M_{NA} = 25.8\%, p = 0.710$ in a fisher-exact test).[35] Seeing as the decision to not revise one's belief implies that the drawn ball is not informative for whether one was assigned to the "bad machine" or the "good machine", the finding of no difference between these two groups is rather intuitive.[36]

[35] Indeed, the degree to which the signal of being audited or not being audited is informative highly depends on priors. To complement this analysis, we regress a dummy of whether one revised his prior (i.e., whether $prior \neq posterior$) on audit, controlling for reported prior, which derived similar marginal effects ($p = 0.678$).

[36] Consistently with this conjecture, non-updaters' answers to the open-ended question in our post-treatment survey regarding what drove them to adjust their estimation in the second round in the way they did, revealed that most respondents thought that one-single ball, if not informative for their assigned machine (see, for example: "*I did not*

With this in mind, Table A.1 re-estimates the baseline results of Table 3, restricting the sample only to active updaters ($N = 259$), showing that the effect of the non-occurrence bias becomes even stronger: Not being audited results in a learning effect that is 4 percentage points *weaker* than the learning effect induced by being audited ($t = -2.24, p = 0.0258$).

Table A.1: OLS Regressions predicting the learning effect for active updates

	Full sample			40 < Prior < 60		
	(1)	(2)	(3)	(4)	(5)	(6)
Audit	0.028**	0.039**	0.038**	0.05***	0.051***	0.05***
	(0.01)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
Cheat 1st round		0.037*	0.026*		0.010	0.022
		(0.02)	(0.02)		(0.02)	(0.02)
Demographics	NO	NO	YES	NO	NO	YES
Constant	-0.002	-0.01	-0.04	-0.03**	-0.04**	-0.05*
	(0.02)	(0.01)	(0.24)	(0.01)	(0.01)	(0.32)
Observations	259	259	259	174	174	174
R^2	0.01	0.03	0.05	0.05	0.05	0.08

*Notes: Results from ordinary least squares regressions. The dependent variable in all regressions is the difference between participant's elicited posteriors and the extrapolated Bayesian posteriors. The reference category for the experimental condition is not being audited. Cheat 1st round is a dummy equal to 1 when the individual falsely reports a successful guess in the first round. All demographics are described in Table 3. Robust standard errors are in parentheses. *, **, and *** indicate significance at the 10%, 5%, and 1% level.*

have enough data to determine which machine I was drawing from"; "*I didn't update it because a single point wasn't enough for me to make any meaningful change*").

Table A.2: OLS Regressions predicting the learning effect excluding those who guesses correctly

	Full sample			40 < Prior < 60		
	(1)	(2)	(3)	(4)	(5)	(6)
Audit	0.032**	0.033**	0.033**	0.043***	0.043***	0.041***
	(0.01)	(0.01)	(0.01)	(0.02)	(0.02)	(0.02)
Cheat 1st round		0.032**	0.033**		0.015	0.021
		(0.02)	(0.02)		(0.02)	(0.02)
Demographics	NO	NO	YES	NO	NO	YES
Constant	-0.02	-0.03**	-0.06**	-0.04***	-0.05***	-0.07***
	(0.01)	(0.01)	(0.25)	(0.01)	(0.01)	(0.25)
Observations	294	294	294	187	187	187
R^2	0.012	0.03	0.05	0.04	0.05	0.08

*Notes: This table re-estimates the baseline results presented in Table 3 for subjects who did not guess correctly in the first round. The dependent variable in all regressions is the difference between participant's elicited posteriors and the extrapolated Bayesian posteriors. All demographics are described in Table 3. Robust standard errors are in parentheses. *, **, and *** indicate significance at the 10%, 5%, and 1% level.*

B. Extensions: Beliefs

B.1. Time spent as a potential deriving mechanism

We had no a priori hypothesis regarding how (and whether) the time spent on contemplating whether to cheat or not, or on estimating the probability of detection, should differ across treatments. As a further exploratory analysis, Table B.1 estimates the effects of receiving the signal of "being audited" on response time and log response (in seconds) using OLS regressions (columns (1) and (2)) and quantile regressions for the 25th, 50th, and 75th percentiles (columns (1)-(4)).[37] We find that time allotted throughout the experiment is virtually identical across treatments, with two interesting exceptions. First is the part of the experiment where the treatment was administered, where participants were notified whether their report was chosen for inspection (i.e., a ball labeled "audit" was randomly drawn by the computer) or not (i.e., a ball labeled "no audit" was randomly drawn by the computer), accompanied by an animation of the corresponding

[37] As common in economics and mathematical psychology, we rely on log response time, $\ln(1+T)$, because of the oftentimes skewed nature of response time data. Using instead the raw response time (T) delivers the same qualitative results.

ball jumping up and down on the screen. The average time spent by participants who were audited is close to 1.5 times as long as the time spent by those who were not audited, an effect which remain highly significant across all specifications ($p < 0.001$). Second is where participants made their decision of whether to cheat for the second time, i.e., after treatment was administered, where those who were audited in the first-round spent more time contemplating whether to cheat or not. However, this result was highly significant only for the 25th percentile. While exploratory in nature, both results are in line with the behavioral intuition that not getting caught induces a stronger response relative to not being caught – even when both events carry the same informational weight from a purely rational perspective.

Table B.1: Time spent across experimental conditions

	OLS		Quantile regression		
	(1)	(2)	(3)	(4)	(5)
Total time	-9.916	-0.012	-7.000	-13.000	48.000
	(23.91)	(0.04)	(18.86)	(22.99)	(39.62)
Instructions	-13.711	-0.077	-6.096	-10.313	-8.509
	(12.39)	(0.09)	(5.55)	(7.61)	(9.95)
Attention test	-5.304	-0.003	2.325	8.197	-3.031
	(8.16)	(0.08)	(4.01)	(6.01)	(11.51)
Cheat 1st round	1.180	0.097	0.600	0.798	-0.235
	(0.89)	(0.07)	(0.40)	(0.53)	(1.23)
Cheat 2nd round	0.759*	0.133*	0.490***	0.299	0.088
	(0.43)	(0.07)	(0.18)	(0.29)	(0.53)
Report 1st round	0.747	0.028	0.978	-3.031	1.137
	(3.03)	(0.07)	(2.55)	(3.29)	(3.77)
Report 2nd round	-2.758	0.018	0.774	-0.281	-0.217
	(3.02)	(0.08)	(0.73)	(0.98)	(2.17)
Notify 1st round	2.189***	0.344***	1.329***	1.986***	2.672***
	(0.36)	(0.05)	(0.29)	(0.37)	(0.53)
Notify 2nd round	0.213	0.053	0.038	0.118	0.344
	(0.26)	(0.07)	(0.19)	(0.27)	(0.37)
Observations	345	345	345	345	345

Notes: Each column represents a separate specification regressing each dependent variable on an audit indicator. The dependent variable is the number of seconds passed between the moment that a choice task is displayed on the screen until the moment that the participant presses the Next button. Robust standard errors clustered at the individual level are in parentheses. The reference category for the experimental condition is not being audited. *, **, and *** indicate significance at the 10%, 5%, and 1% level.

C. Robustness: Behavior

Table C.1. Logit and Probit regressions of second-round cheating

Panel A: Logit Regressions

Dependent Variable: First-Round Cheating Dummy			
	(1)	(2)	(3)
Belief 2nd round		-0.610***	-0.590***
		(0.16)	(0.16)
Cheat 1st round	0.373***	0.356***	0.354***
	(0.02)	(0.03)	(0.03)
Demographics	NO	NO	YES
Constant	-1.68***	0.28	-0.05
	(0.02)	(0.56)	(0.75)
Observations	345	345	345
Pseudo R^2	0.19	0.22	0.27

Panel B: Probit Regressions

Dependent Variable: First-Round Cheating Dummy			
	(1)	(2)	(3)
Belief 2nd round		-0.620***	-0.600***
		(0.16)	(0.16)
Cheat 1st round	0.392***	0.370***	0.368***
	(0.03)	(0.03)	(0.03)
Demographics	NO	NO	YES
Constant	-1.0***	0.15	-0.06
	(0.09)	(0.32)	(0.42)
Observations	345	345	345
Pseudo R^2	0.19	0.22	0.23

*Notes: Table C.1 re-estimates the baseline results presented in Table 4, using logit (Panel A) or probit (Panel B) regressions. All reported coefficients represent marginal effects. Cheat 1st round is a dummy equal to 1 when the individual falsely reports a successful guess in the first round. Demographics include age, a gender dummy, college dummy, a race dummy, and a dummy for a yearly income of more than $50,000. Robust Standard errors are in parentheses. *, **, and *** indicate significance at the 10%, 5%, and 1% level.*

Table C.2: the effect of beliefs on the decision to cheat

Dependent Variable: First-Round Cheating Dummy		
	(1)	(2)
Belief 1st round	-0.816***	-0.800***
	(0.28)	(0.28)
Demographics	NO	YES
Constant	0.77***	0.87***
	(0.14)	(0.16)
Observations	294	294
R^2	0.03	0.5

*Notes: Table C.2 re-estimates the baseline results presented in Table 4, for the parallel pre-treatment measures. Belief 1st round is the participants' elicited prior beliefs. All demographics are described in Table 3. Robust standard errors are in parentheses. *, **, and *** indicate significance at the 10%, 5%, and 1% level.*

Table C.3: Logit and Probit regressions of how change in beliefs affects behavior

Panel A: Logit Regressions

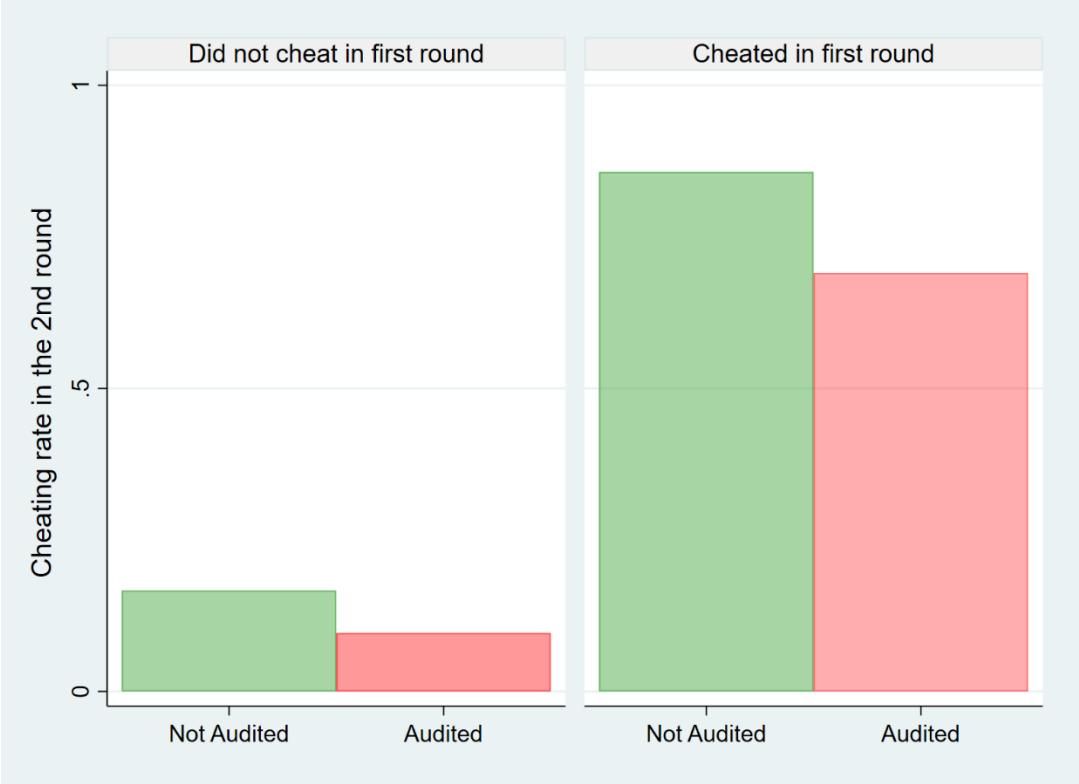
Dependent Variable: Second-Round Cheating Dummy			
	(1)	(2)	(3)
Audit (β_1)	-0.035	-0.047	-0.066
	(0.06)	(0.05)	(0.05)
Change (β_2)	0.012***	0.007**	0.006**
	(0.00)	(0.00)	(0.00)
Change × Audit (β_3)	-0.017***	-0.010**	-0.009**
	(0.00)	(0.00)	(0.00)
Prior	-0.799***	-0.017	0.059
	(0.30)	(0.22)	(0.22)
Cheat 1st round		0.423***	0.440***
		(0.01)	(0.02)
$\mathbb{P}(\beta_2 = \lvert\beta_2 + \beta_3\rvert)$	0.068	0.387	0.371
Demographics	NO	NO	YES
Constant	1.29*	-1.78*	-3.58***
	(0.69)	(0.92)	(1.27)
Observations	292	246	246
R^2	0.06	0.38	0.40

Panel B: Probit Regressions

Dependent Variable: Second-Round Cheating Dummy			
	(1)	(2)	(3)
Audit (β_1)	-0.038	-0.040	-0.057
	(0.06)	(0.05)	(0.05)
Change (β_2)	0.012***	0.007**	0.006**
	(0.00)	(0.00)	(0.00)
Change × Audit (β_3)	-0.017***	-0.010**	-0.009**
	(0.00)	(0.00)	(0.00)
Prior (β_4)	-0.796***	-0.036	0.014
	(0.30)	(0.23)	(0.23)
Cheat 1st round (β_5)		0.446***	0.459***
		(0.02)	(0.02)
$\mathbb{P}(\beta_2 = \lvert\beta_2 + \beta_3\rvert)$	0.068	0.387	0.371
Demographics (β_6)	NO	NO	YES
Constant	0.79*	-1.02*	-1.87***
	(0.42)	(0.52)	(0.68)
Observations	292	246	246
R^2	0.06	0.38	0.39

*Notes: Table C.3 re-estimates the baseline results presented in Table 5, using logit (Panel A) or probit (Panel B) regressions. Estimation samples are restricted to subjects who did not guess correctly in the second round. All reported coefficients represent marginal effects. Change reflects the (absolute) difference between posterior and prior beliefs. Cheat 1st round is a dummy equal to 1 when the individual falsely reports a successful guess in the first round. All demographics are described in Table 3. Robust standard errors are in parentheses. *, **, and *** indicate significance at the 10%, 5%, and 1% level.*

Figure C.4. Differences in cheating rates across experimental conditions and first-round cheating



D. Extensions: Behavior

Table D.1: Time spent by first-round cheaters

	OLS		Quantile regression		
	(1)	(2)	(3)	(4)	(5)
Total time	-52.791**	-0.080	3.000	-30.000	-79.000*
	(24.66)	(0.05)	(21.19)	(26.52)	(47.34)
Instructions	-10.134	-0.004	-0.690	-12.385	-5.465
	(13.20)	(0.10)	(5.92)	(7.87)	(12.08)
Attention test	-4.719	-0.077	-2.503	-6.929	-10.037
	(9.68)	(0.09)	(4.42)	(6.92)	(10.15)
Cheat 1st round	1.629	0.163*	0.674	1.460*	2.891**
	(1.05)	(0.08)	(0.52)	(0.79)	(1.37)
Cheat 2nd round	-0.694	-0.137*	-0.207	-0.417	-0.870
	(0.45)	(0.08)	(0.22)	(0.31)	(0.60)
Report 1st round	-6.440**	-0.114	-0.256	0.328	-5.176
	(3.04)	(0.08)	(2.95)	(3.68)	(3.94)
Report 2nd round	-5.743*	-0.307***	-2.036**	-2.964***	-5.414***
	(2.99)	(0.09)	(0.87)	(0.92)	(1.97)
Notify 1st round	-1.298***	-0.237***	-0.924***	-1.489***	-2.064***
	(0.42)	(0.07)	(0.33)	(0.41)	(0.60)
Notify 2nd round	-0.375	-0.147*	-0.279	-0.664**	-0.493
	(0.32)	(0.08)	(0.21)	(0.30)	(0.41)
Observations	294	294	294	294	294

*Notes: Each column represents a separate specification regressing each dependent variable on an audit indicator. The dependent variable is the number of seconds passed between the moment that a choice task is displayed on the screen until the moment that the participant presses the Next button. Estimation samples are restricted to subjects who did not guess correctly in the first round. The reference category for the experimental condition is not being audited. Standard errors, clustered at subject level, are in parentheses. *, **, and *** indicate significance at the 10%, 5%, and 1% level.*

Table D.2: Heterogeneity in first-round cheating

	OLS	PROBIT	LOGIT
	(1)	(2)	(3)
age_40	-0.028	-0.027	-0.028
	(0.06)	(0.06)	(0.06)
Female	-0.119**	-0.117**	-0.118**
	(0.06)	(0.06)	(0.06)
College	0.021	0.020	0.020
	(0.06)	(0.06)	(0.06)
White	-0.072	-0.069	-0.069
	(0.08)	(0.07)	(0.07)
Rich	0.016	0.015	0.015
	(0.06)	(0.06)	(0.06)
mathQ_correct2	-0.006	-0.004	-0.005
	(0.06)	(0.06)	(0.06)
Constant	0.48***	-0.03	-0.04
	(0.09)	(0.25)	(0.4)
Observations	294	294	294
R^2	0.02	0.02	0.02

*Notes: Each column represents a separate specification regressing a dummy equal to 1 when the individual falsely reports a successful guess in the first round on various pre-treatment characteristics. Estimation samples are restricted to subjects who did not guess correctly in the first round. All reported coefficients in models (2) and (3) represent marginal effects. Robust standard errors are in parentheses. *, **, and *** indicate significance at the 10%, 5%, and 1% level.*