Breaking bad:
Malfunctioning control institutions erode good behavior in a cheating game

Rustamdjan Hakimov and Agne Kajackaite*

September 9, 2024

Abstract

This paper studies whether malfunctioning (or unenforced) institutions erode good behavior. We use a large-scale online experiment, in which participants play a repeated observed cheating game. When we ask participants to report honestly and promise no control, we find low cheating rates. When control of truthful reporting is introduced, low cheating rates remain. In our main treatment with a malfunctioning institution, participants do not know whether they are in the treatment with or without control. In this treatment, participants who do not face control for some rounds start cheating significantly more often, reaching highest cheating rates. That is, a malfunctioning institution leads to more cheating than no institution at all, which indicates that the development of cheating behavior is endogenous to the institutions. Our findings suggest a novel negative effect of unenforced laws.

Key Words: Lying, Cheating, Institutions, Control, Crowding-Out, Experiment.
JEL Classification: C90, D81, D82, D91.

* Hakimov: University of Lausanne, Lausanne, CH-1015 Switzerland (email: rustamdjan.hakimov@unil.ch); Kajackaite: University of Milan, Milan, 20122, Italy (email: agne.kajackaite@unimi.it). We thank Tilman Fries, Kai Barron, Fabio Galeotti, Uri Gneezy, Dorothea Kübler, Dirk Sliwka, Joel Sobel, Gary Charness, two anonymous reviewers, and seminar participants at WEBEAS, CREST Ecole Polytechnic Paris, University of Kiel, NYU Abu Dhabi, Humboldt University, McMaster University, University of Milan und University of Vienna for their valuable comments. The paper is part of a pre-registered project #49192 on AsPredicted registry (https://aspredicted.org/gy47f.pdf); The Ethics Committee of LABEX, University of Lausanne, approved the experimental design. Rustamdjan Hakimov acknowledges financial support from the Swiss National Science Foundation project #100018_207722.

1 Introduction

The importance of institutions for the functioning of society is enormous. The economics literature has devoted considerable attention to institutions that help maintain desirable outcomes for society. For instance, punishment institutions have been found to be useful for maintaining cooperation (e.g., Fehr and Gächter, 2000; Gürerk et al., 2006; Nikiforakis and Normann, 2008; Dai et al., 2015). The other stream of research emphasizes the possible detrimental effect of audit/control institutions, because they crowd out intrinsic motivation and thus change people's behavior (e.g., Frey and Oberholzer-Gee, 1997; Gneezy and Rustichini, 2000; Frey and Jegen, 2001; Fehr and Rockenbach, 2003; Dickinson and Villeval, 2008). In this paper, we also concentrate on audit/control institutions, but the main novelty of this paper comes from analyzing the effect of an audit/control institution that is announced and not enforced.

Indeed, in reality, especially in context of developing countries, some laws are not enforced, i.e. institutions are malfunctioning. The unenforced laws are also common in developed countries, and their effects are debated among law scholars.[1] The focus of this paper is to study the effects of malfunctioning institutions, that is, institutions that are supposedly in place but are likely not functioning. We hypothesize that the potential positive impact of strong

[1] The law scholars' views range widely. Some scholars argue that the unenforced laws suggest the norm of good behavior and thus have a positive effect despite no enforcement (e.g. Cooter, 1998). Other scholars claim that the effect of unenforced rules is detrimental, either because the violators might feel that they are being perceived as criminals (Leslie, 2000) or because unenforced rules impose an externality of discomfort for non-violators who observe the violations (Depoorter and Tontrup, 2016). Finally, some scholars argue that unenforced laws are problematic, because they undermine the power of law. This argument was captured by Justice Brandeis in Olmstead v. United States, 227 U.S. 438, 485, 48 S.Ct. 564, 575, 72 L.Ed. 944 (1928): "Our government… teaches the whole people by its example. If the government becomes the lawbreaker, it breeds contempt for law; it invites every man to become a law unto himself; it invites anarchy." Also, there is a debate in financial regulations literature on whether it is worth having regulations that authorities cannot enforce. For instance, laws against insider trading are not enforced in 70% of developing countries (Bhattacharya and Daouk, 2002).

control institutions[2] on "good" behavior could be reversed if uncertainty exists regarding whether the institutions are functioning.

We study the effect of malfunctioning intitutions on good behavior in the context of a cheating game.[3] In an online experiment, we use a version of the observed cheating game (Gneezy et al., 2018) repeated over 20 rounds. Experimental participants observe a number between 1 and 10 and are asked to report it to the experimenter. In the baseline treatment (NoControl), participants' monetary payoff is equal to the number they report in euros. Participants are told they are expected to report truthfully and they will not be controlled. By contrast, in the treatment with a strong institution (Control30), participants are informed that in each round, they have a 30% probability of being controlled. If they are controlled, participants receive the number of euros corresponding to the number they have reported, if they told the truth; and if they have misreported the number, they receive zero. Participants learn after each round whether a control was in place. We expect the strong control institution to work well in monitoring people's truthfulness and the controlling effect to outweigh possible crowding-out effects. These two treatments serve as important benchmarks for the malfunctioning institution treatment.

In many situations, strong institutions cannot be implemented. Despite the presence de jure of robust and modern institutions in many cases, they malfunction de facto. Examples of malfunctioning institutions range widely. For instance, in Indonesia, drug users could be

[2] We call a control institution "strong," when the control probability and the penalty are high enough to change the behavior of utility maximizers.

[3] Honesty is crucial for the functioning of societies and economies (e.g., Mauro, 1995; Pranab, 1997; Olken and Pande, 2012). From a homo-economicus perspective, a person will cheat for a monetary benefit, if both the possibility of being caught and the penalty are low enough (see Becker, 1968). However, the literature on lying in economics shows people will often forego an opportunity to lie, even when there is no monetary penalty for lying (see Gneezy, 2005; Shalvi et al., 2011; Fischbacher and Föllmi-Heusi, 2013; Abeler et al., 2014; Gächter and Schulz, 2016; Kajackaite and Gneezy, 2017; Gneezy et al., 2018; Abeler et al., 2019, among many others).

punished by a prison sentence, but in practice, their punishment is often restricted to a bribe paid to a police officer. Corruption by public officials is one of the most widespread crimes in many developing countries, because the punishment of the officials is often not enforced in practice. Furthermore, even in developed economies, tax returns are often only checked superficially and waved through, if no noticeable irregularities are detected.[4] Finally, due to the COVID-19 pandemic, many countries introduced quarantine laws and rules. Some of these rules, however, could not be enforced.[5] Taken together, in all these examples, even though institutions are used as a threat, they are not enforced.

What all these malfunctioning institutions have in common is that, when introduced, uncertainty exists regarding their enforcement. Over time, with experience, people develop endogenous beliefs about the probability that the law can be enforced. The longer they observe unpunished violations, the greater the likelihood that the institution is malfunctioning.[6] We introduce this uncertainty about the institution that is in place in the third treatment. In this treatment, the chance of being in the NoControl or Control30 treatment is 50-50. After each round, independently of their truthfulness, subjects learn whether they have been controlled. We distinguish between two sub-treatments: 50-50NoControl and 50-50Control30. Participants in the 50-50Control30 sub-treatment learn they are in this sub-treatment after the first round in which control is implemented. Those in the 50-50NoControl sub-treatment never

[4] In developed countries, there is a high heterogeneity in level of enforcement of laws. For instance, there is a negative correlation between the strictness of labor laws and level of their enforcement (Kanbur and Ronconi 2016).

[5] For instance, Switzerland required people coming from risk regions to isolate themselves from other household members within their place of residence; violations could lead to fines up to 10,000 CHF. Such a law is hard to implement, because verifying whether family members are remaining separate within the household is difficult.

[6] For example, when submitting a tax return in a new country, we do not know how thoroughly tax filings are being checked, but will find out over time. If we do not get audited and observe others not being audited for a long time, we might believe the audits are not being performed despite the promise. The same goes for the quarantine rules—we do not know whether we will actually be inspected, but we learn what kind of institution is in place over time.

deterministically learn they are in this sub-treatment,[7] and we refer to this sub-treatment as a malfunctioning-institution treatment. Our central hypothesis is that over time, subjects will lie more in the malfunctioning-institution sub-treatment than in the treatment with no institution.[8]

The results from our online experiment are in line with our central hypothesis. Participants break bad[9] and cheat the most when the institution is malfunctioning—the 50-50NoControl treatment has the lowest truthfulness rates. We also find that, contrary to our hypotheses, a strong control institution does not lead to more overall truthfulness than NoControl.

Further analyses of lying on the intensive and extensive margins show treatment effects that are masked by the overall truthfulness rates. Although Control30 does not lead to different truthfulness rates than NoControl, more subjects lie at least once in the course of the experiment in Control30 than in NoControl. This detrimental effect is counteracted by subjects in Control30 lying less on the intensive margin than in NoControl—conditional on lying at least once, paricipants in Control30 lie less often than in NoControl.

The malfunctioning institution, on the other hand, combines the worst of the two worlds and erodes good behavior. The proportion of subjects who lie at least once in this treatment is similar to the proportion of subjects who lied at least once in Control30, and, conditional on lying at least once, with experience, subjects in 50-50NoControl lie as often as subjects who lied at least once in NoControl. These results suggest that although the potential presence of a strong control institution crowds out some participants' honesty, learning over time that one is

[7] However, note that after 10 rounds without control, the probability that the sub-treatment is 50-50NoControl is 97.25%.
[8] An ideal test for our hypothesis would be running a treatment, in which we announce the Control30 treatment, but never control and allow subjects to communicate with others whether they were controlled. This design would involve deception, so we believe our design captures the main feature of the malfunctioning institution environment respecting the no-deception rule.
[9] By "breaking bad," we refer to increased lying rates over time.

almost certainly in NoControl does not prompt honesty, but allows the individual to lie (almost) without the risk of a punisment.

What can explain our main result that the malfunctioning institution leads to the lowest level of honesty? One possibility is that the result is driven by the, in expectation, very small control probability that crowds out intrinsic motivation for honesty while posing no risk of being controlled. This effect would be in line with Gneezy and Rustichini (2000), who show that very small incentives might negatively affect behavior. We run an additional treatment to control for this small chance of a fine effect: in this treatment, subjects know the probability of control is 0.03%, the exact control probability of the Bayesian updater in the last round of the malfunctioning-institution treatment. However, we find no significant difference between NoControl and Control0.03 in the proportion of misreports. Thus, the effect of the malfunctioning institution cannot be explained simply by the crowding out due to the small probability of control. The effect goes beyond a "pay-enough-or-don't-pay-at-all" type of behavior.

Another explanation for our main result is that the *threat* of the *strong* control institution is what crowds out honest behavior in the malfunctioning-institution sub-treatment. To scrutinize this conclusion thoroughly, and control for the potential influence of institutional uncertainty, we implemented an additional experimental treatment. The treatment is identical to the 50-50 treatment, but with a twist: participants learn about their assigned treatment before the first round commences. Participants assigned to NoControl, herein referred to as Known50-50NoControl, encounter the threat, and promptly learn about its irrelevance to their payment. Participants assigned to Control (Known50-50Control), on the other hand, immediately learn that they are in the sub-treatment with control. In the Known50-50NoControl, we found the lowest rates of truthful reporting, evident from the earliest rounds. The difference is highly significant compared to all other treatments.

That is, the mere announcement of a threat scenario has a crowding-out effect, while informing that there will be no control does not instigate a crowding-in of motivation to behave honestly. We conclude that intimidating individuals with a strong, albeit unimplementable, institution generates more pervasive dishonesty than maintaining a weak yet functional institution (Control0.03), or having no institutional control whatsoever (NoControl). This detrimental effect persists even when the threat is retracted immediately after its announcement.

Thus, we find that overall truthful reporting is considerably lower in Known50-50NoControl than in 50-50NoControl. This result is not predicted by our theory. While our paper focuses on the performance of malfunctioning institutions compared to other institutions, the variance of lying rates within the malfunctioning institutions under different levels of uncertainty is intriguing and warrants further investigation. It is plausible that participants in 50-50NoControl do not update their beliefs correctly about the possibility of being in 50-50NoControl after each round in which they observe no control. If they overestimate the chance of being in 50-50Control, it could lead to less lying than if they had accurate beliefs. However, we do not know whether these incorrect beliefs are the mechanism responsible for the large differences in lying rates between Known50-50NoControl and 50-50NoControl. Pinpointing the precise mechanism that leads to more lying when the threat of control is announced and then immediately revoked, compared to when the threat of control is not revoked, is left for future research.

Summing up, our findings present a new reason for the potential erosion of good behavior - the threat of strong control.[10] This insight extends beyond the predictions of well-

[10] We use the threat of strong control to represent the concept of malfunctioning institutions. One might question whether the "strong" element is essential for the observed adverse effects. This remains an open question for future research, as we did not implement a threat of weak control in our treatment, partly due to the practical cost of introducing such a condition.

established crowding-out theories (see, for example, Frey and Jegen, 2001). These theories do not predict the erosion of good behavior in malfunctioning institutions, as they do not account for incentives that are announced but not fulfilled in practice. They also do not anticipate this decline when uncertainty is resolved with no control imposed. We believe that this detrimental effect goes beyond the context of cheating behavior.

Related literature. Our paper relates to several streams of research in economics and behavioral economics. First, it closely relates to the above-discussed experiments on lying behavior (e.g., Gneezy et al., 2018; Abeler et al., 2019). A notable contribution to the lying literature is our finding from the Known50-50NoControl treatment, in which the average truthfulness rate amounts to 59.8% compared to 82.1% in NoControl. The reaction to this treatment variation appears more pronounced compared to, for example, the response observed in Kajackaite and Gneezy (2017), when monetary incentives to lie in a mind game were increased fiftyfold. It is surprising how impactful it can be to announce a control possibility and then retract it immediately.

Second, our study relates to experiments on cheating, which include control institutions. To our knowledge, only two studies look at how control affects lying. In a deception-game setting (as in Gneezy, 2005), Laske et al. (2018) explore how the size of fines and the probability of being caught affect decisions to deceive by sending a wrong message to the receiver. In the repeated version of the experiment, they find that participants are sensitive to both the size of the fine and the probability of being monitored (5% vs. 50% chance). The higher and more likely the fine, the less participants cheated. Thus, unlike in our experiment, both weak and strong institutions worked to reduce lying.

The other experiment that looks at the effect of control on lying was conducted by Galeotti et al. (2021). The researchers were specifically interested in seeing the spillover effects

of control. In the first stage of the quasi-experiment, some people had their ticket checked on public transportation and some did not. In the second stage, an actor followed the participants, then acted as if they were picking up a 5 euro banknote and asked the person whether it was theirs. The authors found that observing ticket checks on public transport led to more cheating in the form of individuals claiming the banknote was theirs. Interestingly, observing ticket checks led to more lying about the banknote for both groups—for those who cheated in the public transport and those who did not. Following Sliwka (2007), the authors argue the ticket checks signal a social norm of dishonesty in society and hence crowd out intrinsic motivation among some participants. Our paper shows that just announcing a possibility of control might crowd out intrinsic motivation to be honest.

The other stream of research that relates to our paper is the tax-evasion literature. Most of the experiments show that increasing detection probabilities leads to more truthful reporting (see, e.g., Webley, 1997; Beck et al., 1991; Alm et al., 1992). Similar results have also been found when analyzing the effect of the probability of being caught stealing, with stealing decreasing with the probability of being caught (e.g., Harbaugh et al., 2013). Our paper shows that the relation between the probability of control and cheating might not be linear and that the level of laws enforcement is a crucial variable to consider when designing institutions.

Our paper relates to the intrinsic-motivation crowding-out literature in economics and psychology (e.g., Frey and Oberholzer-Gee, 1997; Deci et al., 1999; Gneezy and Rustichini, 2000; Frey and Jegen, 2001; Fehr and Rockenbach, 2003; Benabou and Tirole, 2003, 2006; Sliwka, 2007; Dickinson and Villeval, 2008; Mellström and Johannesson, 2008; Gneezy et al., 2011). We contribute to this literature by showing that the crowding out effect of control is persistent even if the control is not enforced.

Finally, our paper contributes to the financial regulation literature. Bae and Goyal (2009) show that variation in the enforceability of contracts matters for the way banks structure and price loans. Bhattacharya and Daouk (2009) argue that no regulation is often better than an unenforced regulation in the context of the ban on insider trading Our paper provides empirical evidence in line with the authors' argument.

The rest of the paper is organized as follows. In section 2, we introduce the experimental design and procedure. In section 3, we state the main hypotheses for our experiment. In section 4.1, we provide the overall effects of control institutions on truthfulness, and in sections 4.2 and 4.3, we discuss lying behavior on extensive and intensive margins, respectively. Section 4.4 looks at how observing control affects lying, and section 4.5 discusses the weak-institution treatment. Section 4.6 discusses the treatment when uncertainty is lifted. Section 5 concludes the paper.

2 Experimental design

2.1 Experimental procedure

In the experiment, subjects play a repeated-cheating game (see Appendix B for instructions). The setup for the experiment is an individual decision-making situation, with no interactions between subjects. We use the observed-cheating game (as in Gneezy et al., 2018). In each round, on their screens, participants see 10 boxes with hidden outcomes behind them. The outcomes behind the boxes are numbers between 1 and 10, placed in a random order, where each box has a different number. After clicking on one of the boxes and seeing the number, participants are asked to truthfully report the number to the experimenter. The higher the number reported, the higher the payoff, which creates an incentive to cheat by over-reporting. In the absence of a control, the participant's payoff for the round is the number she reports in

euros. If a control is present, the payoff is equal to the number reported if the report is truthful, and 0 otherwise.[11]

Each experimental session had 15–40 subjects, and each subject played the cheating game for 20 rounds. At the end of the experiment, one round was picked at random to be payoff relevant. The experiment was run online using oTree software (Chen et al., 2016). The participants logged in to the experiment online and were observed through Zoom until the cheating game started. The instructions were read aloud at the beginning of the session, and subjects could ask questions in private in a chat.

Before the start of the game, subjects had to pass a quiz about the game's rules. If the subjects gave a wrong answer to at least on question, they had to take the quiz again. A maximum of 10 attempts was allowed, and failing subjects (around 2%) were prohibited from participating in the experiment.[12]

Before the start of the first round, subjects had to turn off their cameras, so the subjects could not interact or observe the reactions of other participants.[13] We treated each participant as one independent observation.

We ran experiments with members of the general population, whom the laboratory of the University of Valencia (LINEEX) recruited for the study using online advertisements. In total, we had 1014 participants:143 in NoControl, 145 in Control30, 168 in 50-50NoControl, 159 in 50-50Control30, 141 in Control0.03 (see section 4.5 for this condition), and 131 in Known50-50NoControl and 127 in Known50-50Control30 (see section 4.6 for this condition) treatments.

[11] In our case, in all treatments, the experimenter observes the lies, which might lead to less lying. However, the design choice does not affect the treatment differences. Alternative designs might include mind and wheel tasks (e.g., Galeotti et al., 2020).

[12]We consider a failing rate of 2% to be high, likely because we use a general population and not a laboratory subject pool.

[13]We asked participants to turn on their cameras at the beginning of the experiment and during the introduction stage to ensure that participants take the participation seriously and to make the environment closer to an "offline" lab.

The average duration of a session was about one hour, and the average payoff was 11.70 euros, including a show-up fee of 5 euros.

2.2 Treatments

To answer our research questions, we ran the following three treatments in a between-subjects design:

1. **No-institution treatment (NoControl).** This treatment served as a baseline and entailed no control institution. Subjects received a payoff equal to the number reported in euros. Subjects were told they were expected to report truthfully and would not be controlled.
2. **Strong-institution treatment (Control30).** In this treatment, subjects knew that in each round, they might be controlled with a 30% probability. In the case of control, the subjects received no payoff if they misreported the number, or they received the number reported in euros if they were truthful. After each round, subjects learned whether they had been controlled in that round, independently of their truthfulness.
3. **50% chance of being in either No-institution or in Strong-institution treatment. (50-50).** Subjects received instructions for both NoControl and Control30 treatments and were told that with a 50% chance, they were in one of the treatments. After each round, subjects learned whether they had been controlled in that round, independently of their truthfulness.[14] We distinguish between two sub-treatments:
    a. **50-50Control30**. This sub-treatment consists of participants who were in the Control30 treatment. They had learned deterministically that they were in the Control30 treatment after the first round in which control was implemented.

[14] This design feature is important. In all treatments, one finds out whether a control was in place, *independently* of their truthfulness. Thus, "experimenting" through lies in order to see whether a control was in place is not a rational strategy.

b. **Malfunctioning-institution treatment (50-50NoControl)**. This sub-treatment consists of participants who were in the NoControl treatment. These participants had never deterministically learned they were in the NoControl treatment. However, after 10 rounds of absent control, the probability that the treatment was NoControl was 97.25%, and after 19 rounds of absent control, the probability that the treatment was NoControl was 99.89%.

3 Hypotheses

We start with a simple theoretical framework to outline the predictions of the theory. We base our framework on Gneezy et al. (2018). Our theory differs from theirs in that we do not differentiate between the intrinsic and social-identity costs of lying, and instead consider the fixed psychological cost of lying. Furthermore, we introduce to the theoretical framework a crowding out of intrinsic motivation for truth-telling, as a coefficient that reduces the psychological cost of lying, if a possibility of a control institution is announced.

An agent's type consists of a tuple (i, t, θ), where $i = \{1,2 \dots ,10\}, t \in [0, T]$, and $\theta \in [0,1]$. The value i represents the agent's actual observation of the random draw; t is the fixed psychological cost of lying; and θ is the parameter of crowding out due to the announcement of the control institution, with $\theta = 1$ in NoControl. The agent can report any number $j = \{1,2, \dots ,10\}$. Given the type (i, t, θ) and report j, the utility of the agent (assuming a linear utility of money) is

$$U = \begin{cases} j & if\ i = j \\ j - \theta t & if\ i \neq j, no\ control. \\ -\theta t & if\ i \neq j, control \end{cases}$$

Because we assume a fixed cost of lying and the cost therefore does not depend on the size of the lie, agents who lie, lie to the maximal extent and report a 10. We derive the optimal reporting in Proposition 1.

Proposition 1:

Under NoControl, $\theta = 1$. *Then, the optimal report is the following:* $j = \begin{cases} i & if\ i > 10 - t \\ 10 & otherwise \end{cases}$.

Under Control30, the optimal report is the following: $j = \begin{cases} i & if\ i > 7 - \theta t \\ 10 & otherwise \end{cases}$.

Under 50-50, the optimal report in round r *is the following:*

$$j = \begin{cases} i & if\ i > (10 - \theta t)p_r + (7 - \theta t)(1 - p_r) \\ 10 & otherwise \end{cases},$$

where p_r *is the probability of being in the 50-50NoControl sub-treatment in round* r.

Our central assumption is that crowding out happens based on the institution's announcement, even if the institution ends up being malfunctioning. That is, for those whose intrinsic cost of lying was crowded out, based on the assumption that the law might be enforced, learning that the law is most likely unenforced does not crowd in their motivation to behave well.

We formulate our theoretical predictions of treatment differences assuming that the presence of a control institution crowds out intrinsic motivation for honesty by reducing the fixed cost of lying, namely, $\theta < 1$.[15] We look at the truthfulness rates in the last 10 rounds. In these rounds, 50-50Control30 and Contro30 are virtually the same because the chance of not facing control in the first 10 rounds is only 2.82%, and thus, almost all agents in the 50-50Control30 treatment learn deterministically that they are in 50-50Control30. For 50-50NoControl, after 10 rounds of absent control, the probability of being in NoControl is $p_r >$

[15] In absence of crowding out the comparison of truthful rates in the last 10 rounds is predicted to be $Control30 \approx 50 - 50Control30 > 50 - 50NoControl \approx NoControl$.

97% for a Bayesian updater. Therefore, 50-50NoControl and NoControl are virtually the same treatment, with respect to the control probability.

Prediction 1: In the last 10 rounds,

for any $\theta < \frac{t-0.09}{t}$*, truthful rates under* $50 - 50NoControl < NoControl.$[16]

for any θ *, truthful rates under* $50 - 50NoControl < Control30.$

Prediction 1 states the main predictions of the paper.

The comparison of NoControl and Control30 treatments depends on the level of crowding out, and whether the controlling effect outweighs the effect of crowding out. For $\theta > \frac{t-3}{t}$ (low crowding out),[17] Control30 leads to higher level of truthful reporting than NoControl, while for $\theta < \frac{t-3}{t}$ (high crowding out), the relation is reversed.

Before running the study, we based our hypotheses on assumption of relatively low crowding out effects, such that $\theta > \frac{t-3}{t}$. This resulted in the following pre-registered hypothesis concerning the order of the proportions of honest reports in the last 10 rounds between the (sub-)treatments: [18]

[16] In the eleventh round, more truthful reporting will occur in NoControl than in 50-50NoControl, if $10 - t < (10 - \theta t)0.97 + (7 - \theta t)0.03$, which leads to $\theta < \frac{t-0.09}{t}$. That is, truthful reporting will be lower in 50-50NoControl as long as there is *even very mild* crowding out.

[17] In Control30, participants will report more truthfully than in NoControl if the crowding out is low enough. More truthful reporting will occur in Control30 than in NoControl, if $7 - \theta t < 10 - t$, which leads to $\theta > \frac{t-3}{t}$.

[18] We pre-registered the design and hypotheses at aspredicted.org: https://aspredicted.org/gy47f.pdf. Note we have changed the design of the weak-institution treatment (see section IV.V) compared to the pre-registration. The previous design of decreasing the control probability in each round (see the pre-registration for more details) had a clear demand effect (decreasing the probability in each round and showing this decrease before each round to the participants is like asking them to cheat more and more in each round). The treatment was poorly designed, and we changed the design of this treatment after running one session. Furthermore, we pre-registered too few observations for the Control30 treatment because we were expecting to pool the data from Control30 and 50-50Control30 treatments (we expected and pre-registered that the lying behavior will be the same in both treatments). We could not pool the data, due to treatment differences; thus, we collected additional observations for Control30 to achieve a similar number of observations in each (sub-)treatment. Finally, we pre-registered that we would run the treatments with a dictator game too, but we did not conduct the dictator-game experiments, because we felt adding an additional game is beyond the scope of one paper (however, we are open to running it for our next study). The Ethics Committee of LABEX, University of Lausanne, approved the experimental design.

Control30=50-50Control30>NoControl>50-50NoControl.

As mentioned above, our hypothesis that the malfunctioning control institution—50-50NoControl—leads to the most substantial adverse effect on truthfulness rates is based on the assumption that crowding out of motivation happens at the announcement of the institution stage. Intrinsic motivation is not crowded in even when agents learn that the institution is malfunctioning.

4 Results

We test the hypothesis about the detrimental effect of the malfunctioning institution in an online experiment. In what follows, we first present the overall truthfulness rates over the (sub-)treatments. We then report the results regarding the extensive margin of lies, after which, we present the results regarding the intensive margin of lies. Then, we move to a further analysis of how observing control affects lying. Then, we present the results from two additional treatments.

Throughout the results section, we call a result significant if it is significant at least at the 5% level. We use ">" to communicate "significantly more" at the 5% level, whereas we use "=" to communicate no significance at the 5% level. For all the results, we use the significance of the coefficient of interest in regression analyses with standard errors clustered at the individual level. Importantly, in all main analyses, as pre-registered, we analyze the last 10 rounds of the experiment.

4.1 Truthfulness rates

First, we test our main (pre-registered) hypothesis and compare the rates of truthful reporting in the last 10 rounds between treatments.

Result 1: *The comparison of truthful reporting between treatments in the last 10 rounds shows that truthfulness rates are significantly lower in the malfunctioning-institution treatment than in all other (sub-)treatments.*[19] *We find no significant differences between truthfulness rates in the remaining (sub-)treatments:*

Control30=50-50Control30=NoControl>50-50NoControl.

Support: Figure 1 presents the proportion of truthful reports by rounds and treatments.[20] As can be seen in the figure, in the first rounds, all treatments lead to a similar rate of truthful reports.[21] However, over time, the rate of truthful reporting becomes significantly lower in 50-50NoControl than in other treatments. In the sub-sample of the first seven rounds, the truthfulness rate in 50-50NoControl is significantly lower than in NoControl, Control30, or 50-50Control30.[22]

[19] The tests are based on regressions with controls for age, gender, draw, and rounds with the sample restricted to two treatments of interest for the comparison.
[20] See Figure A1 in the appendix for the truthful reporting by the actual observed draw.
[21] Furthermore, note the average truthful reporting in our experiment is higher than in the comparable experimental literature using the observed cheating game, such as Gneezy et al. (2018). For instance, in the NoControl condition, over all rounds, the average truthful reporting amounts to 82.1%. Such high truthfulness rates are likely driven by the subject pool, which consists of a general Spanish population (see Abeler et al., 2014, who show the general population lies less than a student subject pool). Furthermore, the majority of our subject pool—68%—are women, who tend to lie less than men (see Table 1). Finally, in the NoControl treatment, we explicitly state that participants will not be controlled, which signals trust, and therefore might lead to more truthfulness. In Gneezy et al. (2018), no such statement is made.
[22] These results are based on pairwise regression comparisons, including controls.

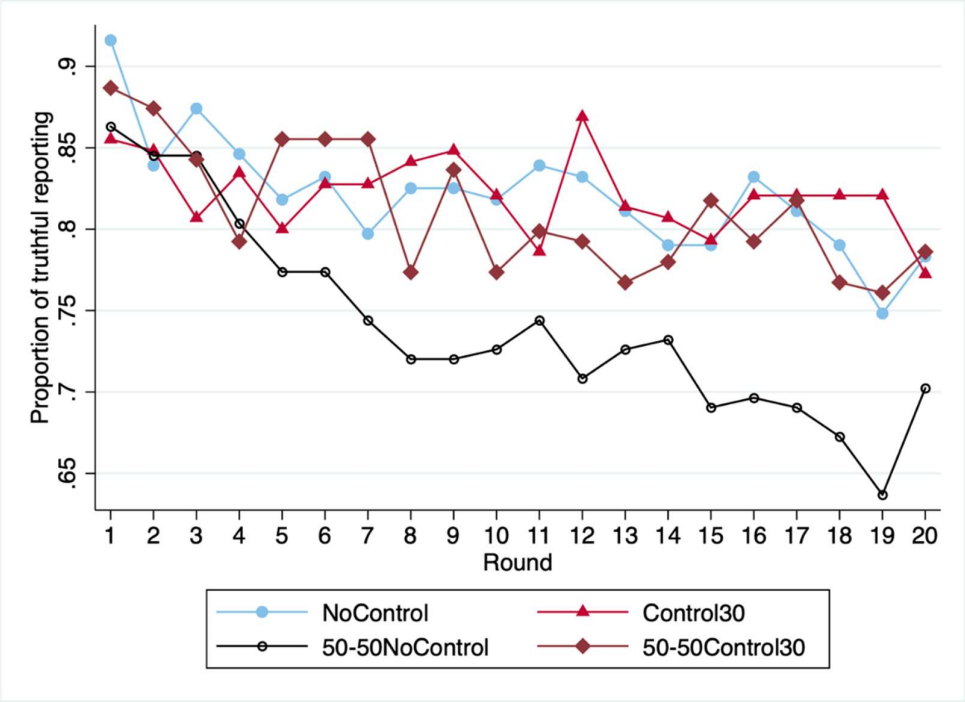


Figure 1. Proportion of truthful reporting by treatments and rounds

Table 1 presents the marginal effects of probit regressions of truthful reporting in the last 10 rounds.[23] Model (1) contains only treatment dummies as explanatory variables, and Model (2) contains additional controls. Both models support the intuition from Figure 1 that in the later rounds, the 50-50NoControl treatment leads to significantly lower rates of truthful reporting than in the NoControl condition. In the last 10 rounds, the truthful reporting is around 11 percentage points lower in the 50-50NoControl than in the NoControl condition (see Model (2)). This detrimental effect of the malfunctioning institution is highly significant at the 1% level. Furthermore, we conduct pairwise comparisons using regression analyses and find that in the last 10 rounds, truthfulness in 50-50NoControl is significantly lower than in any other treatment.

[23] See Table A1 in the Appendix for the average truthfulness rates in the last 10 rounds by treatment.

	Truthful (1)	Truthful (2)
Control30	0.010	0.016
	(0.037)	(0.035)
50-50NoControl	-0.097***	-0.111***
	(0.036)	(0.035)
50-50Control30	-0.015	-0.008
	(0.035)	(0.033)
Female		0.073***
		(0.023)
Age		0.010***
		(0.002)
Draw		0.052***
		(0.002)
Round		-0.005***
		(0.001)
Observations	6150	6150
Number of clusters	615	615
Sample	Last 10 rounds	Last 10 rounds
Pseudo *R*-squared	0.039	0.173

Notes: Standard errors in parentheses, * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Standard errors are clustered at the individual level.

Table 1. Marginal effects of probit regressions of truthful reporting in the last 10 rounds

Overall, our central hypothesis that the malfunctioning institution erodes honest behavior is strongly supported by the data. However, some of our additional predictions find no support in the data: we do not observe a positive overall effect of Control30 and 50-50Control30 on the lying rates. The results on the overall truthful reporting might be masking some underlying shifts in behavior on the extensive (lying at least once) and intensive margins (how often one lies conditional on lying at least once and what one reports conditional on misreporting). We investigate the extensive and intensive margins separately in the next sections to detect the monitoring and crowding-out effects of our treatments.

4.2 Extensive margin: Lying at least once

As discussed in the hypotheses section, announcing control can lead to two effects. Firstly, there is the monitoring effect, as lying becomes financially costly. Secondly, there is the

crowding-out effect of the psychological cost of lying. To identify the crowding-out of lying costs, we can examine a participant's decisions across rounds. Note that the cost affects the likelihood to lie during the experiment, while the decision to lie in a particular round also depends on the payoff from lying, i.e., the draw. Thus, if the crowding-out of cost of lying is present, we should observe more participants willing to lie at least once in the treatments where control was announced—specifically, in the Control 30 and 50-50 treatments, compared to the NoControl treatment. Our data indeed confirms these predictions:

Result 2: *Participants are more likely to lie at least once when they are in either the strong institution or the 50-50 condition than when no institution is in place. The comparison of proportions of subjects who lied at least once for the course of the experiment leads to the following result:*

$$Control30=50\text{-}50 Control30=50\text{-}50 NoControl > NoControl.$$

Support: Figure 2 presents the proportion of participants who lied at least once over all rounds by treatments and rounds.[24] As can be seen in the Figure, NoControl has the lowest proportion of those who lied at least once. Table 2 presents the results of probit regressions for the dummy of lying at least once over all rounds. The regressions show that on the extensive margin, Control30, 50-50Control30, and 50-50NoControl significantly crowd out honesty relative to the NoControl. We also conduct pairwise regressions, and they show no significant differences between Control30, 50-50Control30, and 50-50NoControl.

[24] Unlike in other analyses, here we look not only at the last 10 rounds, but at all the rounds. The reason is that if somebody has not lied in the last 10 rounds but did in the first 10 rounds, categorizing him as a non-liar would be odd. All these results hold if we consider only the last 10 rounds of the experiment.

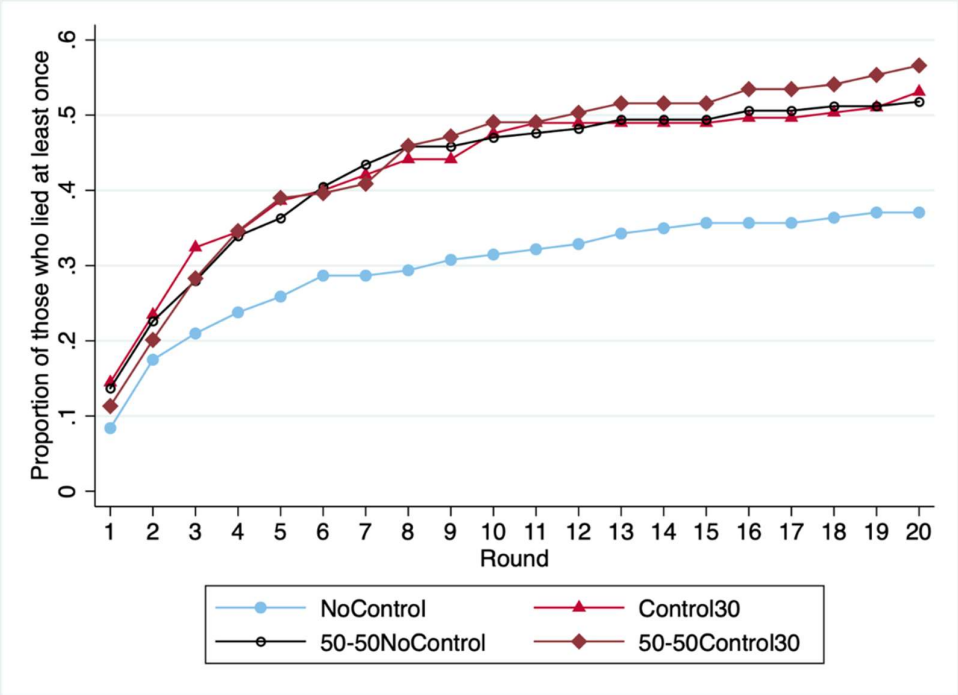


Figure 2. Proportion of participants who lied at least once from the start of the experiment

	Lied at least once (1)	Lied at least once (2)
Control30	0.160***	0.154***
	(0.057)	(0.055)
50-50NoControl	0.147***	0.169***
	(0.056)	(0.054)
50-50Control30	0.195***	0.197***
	(0.056)	(0.053)
Female		-0.176***
		(0.039)
Age		-0.019***
		(0.003)
Average draw		-0.044
		(0.028)
Observations	615	615
Sample	All rounds	All rounds
Pseudo *R*-squared	0.016	0.080

Notes: Standard errors in parentheses $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$. Standard errors are clustered at the individual level.

Table 2. Marginal effects of probit regressions of a dummy of lying at least once for the duration of the experiment

That is, the overall null effect of Control30 and 50-50Control30 on the truthfulness rates masks the crowding-out effect of Control30 and 50-50Control30 that is highly significant on the extensive margin.

Finally, note the crowding out on the extensive margin starts early in the experiment. Using regression analyses, we find that in the sub-sample of the first four rounds, in Control30, 50-50Control30 and 50-50NoControl, lying at least once is significantly more observed than in NoControl.

Next, we analyze the effects of control institutions on the intensive margin of lying behavior.

4.3 Intensive margin: Lying pattern of those who lied at least once, and average reporting, conditional on lying

As discussed in previous sections, the announcement and implementation of control may result in the monitoring effect, making lying financially costly. This implies that, given a fixed cost of lying, a participant is less likely to lie when presented with a draw. If the monitoring effect is present, we should observe fewer lies told by liars in the treatments, in which control was realized – that is, in Control30 and 50-50Control30. This is indeed what we observe in our data:

Result 3: *In the last 10 rounds, participants lie less often in Control30 and 50-50Control30 than in other conditions, conditional on lying at least once during the whole experiment. The comparison of the proportions of truthful reporting in the last 10 rounds for subjects who lied at least once in the experiment leads to the following result:*

Control30=50-50Control30>NoControl=50-50NoControl.

Support: Figure 3 presents the proportions of truthful reporting by round and treatment for those who lied at least once in the experiment.[25] The figure shows that participants in Control30 and 50-50Control30 lie less often than participants in other treatments, conditional on lying at least once. Probit models in Table 3 show that in the last 10 rounds, participants in Control30 and 50-50Control30 are highly significantly less likely to misreport, conditional on lying at least once in the experiment, than the participants in NoControl. We also conduct pairwise regressions and find Control30 and 50-50Control30 do not differ statistically from each other, conditional on lying at least once.

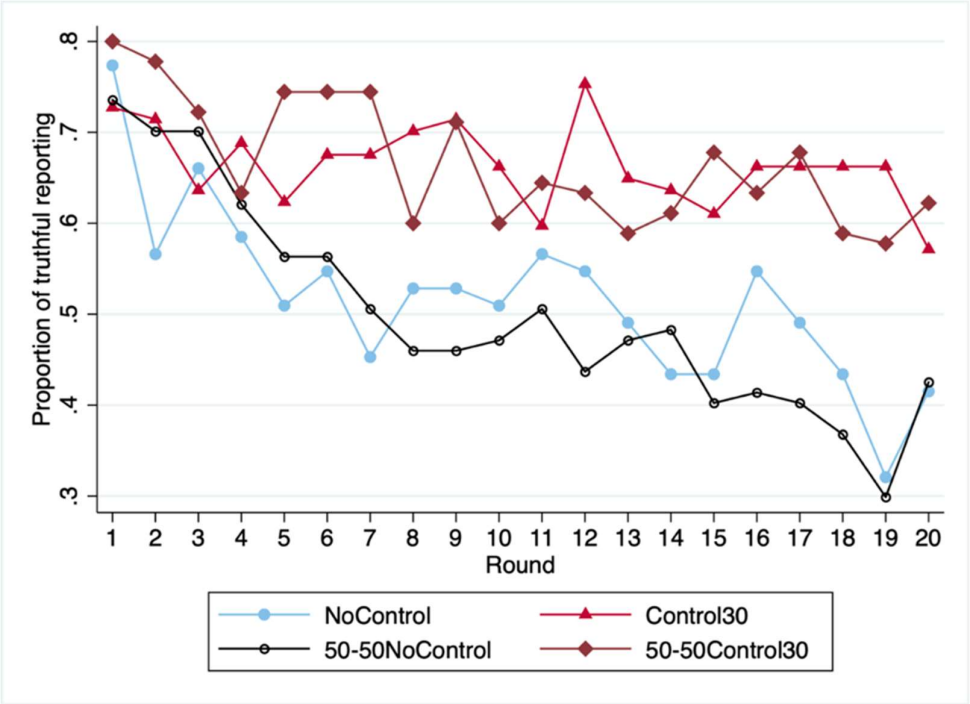


Figure 3. Proportion of truthful reporting by treatments and rounds for those who lied at least once during the experiment

That is, the monitoring of a strong institution *does* work to prevent frequent lies. As a result, the negative effect of crowding out due to the presence of control (extensive margin) is

[25] See Figure A2 in the Appendix for the truthful reporting, conditional on lying at least once, by the actual observed draw.

compensated by less frequent lies, conditional on lying at least once (intensive margin), which at the end leads to the aggregate absence of a difference in overall truthfulness between NoControl, Control30, and 50-50Control30.

In the next step, we turn to the dynamics of the breaking-bad effect on the intensive margin. Model (1) of Table 4 shows that in the first five rounds of the experiment, the rate of truthful reporting among those who lied at least once is significantly lower in NoControl than in all other treatments. Model (2) of Table 4 shows the pattern persists in Control30 and 50-50Control30 for the last five rounds of the experiment but not in 50-50NoControl. Thus, when learning that one is almost certainly in the 50-50NoControl sub-treatment, participants who lied at least once start to lie more often, which suggests they are aware no control will be in place.

	Truthful (1)	Truthful (2)
Control30	0.175***	0.174***
	(0.049)	(0.048)
50-50NoControl	-0.046	-0.051
	(0.049)	(0.049)
50-50Control30	0.154***	0.156***
	(0.046)	(0.045)
Female		0.015
		(0.029)
Age		0.003
		(0.004)
Draw		0.088***
		(0.002)
Round		-0.008***
		(0.002)
Observations	3070	3070
Number of clusters	307	307
Sample	Lied at least once in the experiment; last 10 rounds	Lied at least once in the experiment; last 10 rounds
Pseudo *R*-squared	0.029	0.341

Notes: Standard errors in parentheses *$p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Standard errors are clustered at the individual level.

Table 3. Marginal effects of probit regressions of truthful reporting in the last 10 rounds for subjects who lied at least once during the experiment

	Truthful (1)	Truthful (2)
Control30	0.194***	0.173***
	(0.053)	(0.050)
50-50NoControl	0.141***	-0.073
	(0.052)	(0.051)
50-50Control30	0.221***	0.165***
	(0.050)	(0.046)
Female	0.011	0.008
	(0.034)	(0.031)
Age	0.000	0.001
	(0.004)	(0.004)
Draw	0.089***	0.088***
	(0.002)	(0.002)
Observations	784	1491
Number of clusters	216	307
Sample	Those who lied at least once in rounds 1 to 5; first 5 rounds	Those who lied at least once in rounds 1 to 20; last 5 rounds
Pseudo *R*-squared	0.400	0.349

Notes: Standard errors in parentheses, $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$. Standard errors are clustered at the individual level.

Table 4. Marginal effects of probit regressions of truthful reporting in the first five and last five rounds for subjects who lied at least once

Together, the 50-50NoControl treatment combines the worst of the two worlds—it makes one more likely to lie at least once than in NoControl (extensive margin) and it makes one lie as frequently, conditional on lying at least once, as in the NoControl (intensive margin).

In further analyses on the lying patterns on the intensive margin, we compare the average numbers reported, given a lie, over the treatments.

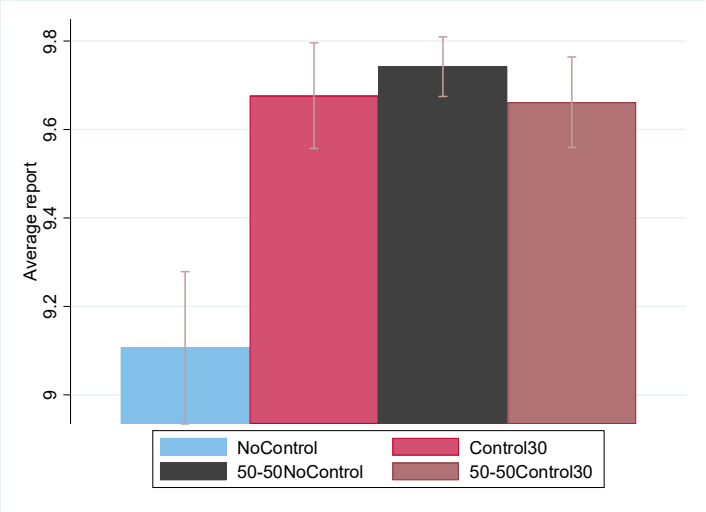
Result 4: *The average reporting, given a lie, is significantly higher in all (sub-)treatments than in the NoControl condition. The comparison of average reports in the last 10 rounds, given a lie, leads to the following result:*

Control30=50-50Control30=50-50NoControl>NoControl.

Support: Figure 4 presents the average reports by treatment in the last 10 rounds, conditional on lying. In NoControl, the average reported number given a lie is 9.11, which is significantly

lower than in all other treatments. We find no significant difference in average reports between all other treatments. OLS regressions in Table 5 confirm that in all the treatments, the average number reported in the last 10 rounds, conditional on lying, is significantly higher than in NoControl.

A closer look at the distributions of numbers reported when lying reveals the lowest average reporting when lying in NoControl is driven by the lowest proportion of reporting a 10 for those who lie. In the last 10 rounds, only 62% of the participants who lie report a 10 in the NoControl condition, whereas the fraction amounts to 82% in Control30, 81% in 50-50NoControl, and 82% in 50-50Control30.[26]



Notes: Gray bars represent 95% confidence intervals.

Figure 4. Average report, conditional on lying in the last 10 rounds

[26] In the NoControl condition, lying to the full extent is similar to the one observed by Gneezy et al. (2018). In Gneezy et al. (2018), in the observed game (treatment "Numbers"), 68% of the participants who lie, lie to the full extent.

	Report (1)	Report (2)
Control30	0.570***	0.630***
	(0.187)	(0.184)
50-50NoControl	0.636***	0.648***
	(0.175)	(0.169)
50-50Control30	0.555***	0.631***
	(0.184)	(0.179)
Age		-0.002
		(0.010)
Female		-0.085
		(0.096)
Draw		0.075***
		(0.015)
Round		0.014
		(0.009)
Constant	9.106***	8.673***
	(0.163)	(0.311)
Observations	1395	1395
Number of clusters	289	289
Sample	Dishonest; last 10 rounds	Dishonest; last 10 rounds
R-squared	0.052	0.076

Notes: Standard errors in parentheses $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$. Standard errors are clustered at the individual level.

Table 5. OLS of the reported number in the last 10 rounds, conditional on lying

4.4 Does observing control affect lying differently in Control30 and 50-50?

In Control30, observing control should not affect a rational agent, as he understands that the control probability is independent of the previous round. However, in the 50-50 treatment, observing control shows that one is in 50-50Control30, whereas not observing control yet makes one surer over time that one is in 50-50NoControl. Thus, not observing any control yet should lead to more lying in the 50-50treatment and should have no effect in Control30. We test in probit analyses whether participants use this reasoning.

	Truthful (1)	Truthful (2)	Truthful (3)	Truthful (4)
No control so far	-0.007	0.000	-0.048**	-0.053**
	(0.026)	(0.026)	(0.022)	(0.023)
Female	0.090**	0.091**	0.072***	0.067**
	(0.036)	(0.036)	(0.027)	(0.028)
Age	0.004	0.005	0.009***	0.010***
	(0.003)	(0.003)	(0.003)	(0.003)
Draw	0.049***	0.050***	0.054***	0.056***
	(0.005)	(0.005)	(0.003)	(0.003)
Observations	1450	2175	3270	4905
Number of clusters	145	145	327	327
Sample	First 10 Control30	First 15 Control30	First 10 50-50	First 15 50-50
Pseudo R-squared	0.20	0.20	0.21	0.20

Notes: Standard errors in parentheses, *$p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Standard errors are clustered at the individual level. Models (1) and (3) look at the first 10 rounds. Models (2) and (4) look at the first 15 rounds.

Table 6. Marginal effects of probit regressions of truthful reporting for 50-50 and Control30 treatments in the first 10 and 15 rounds

Table 6 presents marginal effects of Probit regressions of truthful reporting for the first 10 (Models (1) and (3)) and the first 15 (Models (2) and (4)) rounds, with the sample restricted to participants in Control30 and 50-50 treatments. The main goal of this analysis is to see the effect of the absence of control on truthfulness rates in Control30 and in 50-50 treatments. We find the variable "No control so far" is not significant for Control30, meaning the absence of control in Control30 does not influence the behavior of subjects. This behavior is rational, because the presence of control with a probability of 30% is deterministic in this treatment. The variable "No control so far" for the 50-50 treatment, on the other hand, is negative and significant, meaning subjects lie more if they do not observe control in the 50-50 treatment, which is consistent with updating beliefs in favor of being in the environment without a control institution.

4.5 Weak institution

One might argue the detrimental effect of the 50-50NoControl treatment is driven by, in expectation, a low probability of control, and subjects perceiving it as a weak institution. We

argue, however, that the detrimental effect of the malfunctioning institution goes beyond the effect of a small probability of control. We argue the *threat* of the *strong* control institution and no monitoring possibility are what drive the detrimental result.

In an additional treatment, we aim to distinguish the effect of weak institutions from malfunctioning ones. We introduce a treatment with a weak institution—Control0.03. In this treatment, subjects know that in each round, they might be controlled with a 0.03% probability. In the case of control, the subject receives no payoff if she misreported the number, or she receives the number reported in euros if she was truthful. After each round, subjects learn whether they were controlled in that round, independently of their truthfulness. Note the 0.03% probability corresponds to the Bayesian belief about the probability of control in the 50-50 treatment in round 20, after not being controlled in any previous round.

We expect Control 0.03 to lead to more cheating than NoControl, because of the crowding out of intrinsic motivation for truthtelling and no substantial possibility of being controlled.

From a rational perspective, the 0.03% probability of control does not affect the misreporting rate under risk neutrality, because misreporting under all draws below ten is beneficial. Therefore, such a weak control institution should have no positive effect on truthfulness for payoff maximizers. At the same time, we expect the crowding-out effect from the presence of the control institution for the participants with a psychological cost of lying. However, we expect this crowding out to be weaker with a weak institution than with the malfunctioning institution, because the threat of control is low under a weak institution. In terms of our model, we assume in what follows that θ depends on the size of the control- the

higher the threat the stronger is the crowding out.[27] We hypothesize the following order of the proportions of honest reports in the last 10 rounds between the (sub-) treatments:

Control30=50-50Control30>NoControl>Control0.03>50-50NoControl.

In the experiment, however, we find no overall effect of weak institutions on truthfulness. Figure 5 shows that, in aggregate, lying in Control0.03 is similar to that in NoControl, 50-50Control30, and Control30, and shows significantly lower levels of lying than 50-50NoControl.

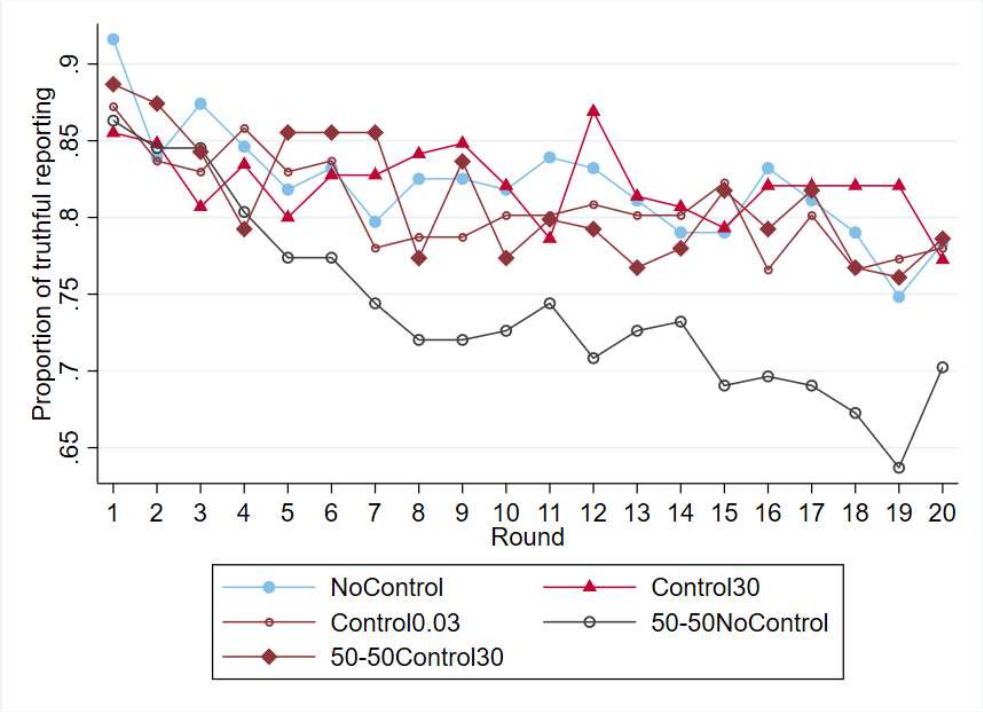


Figure 5. Average proportion of truthful reporting by treatments and rounds, including Control0.03

[27] Note that if θ is independent of the size of control, prediction would be that Control0.03 leads to the same level of misreporting as 50-50NoContol.

Although we observe minimal evidence of crowding out on the intensive margin (in Control0.03, the average reported number given a lie is significantly higher than in NoControl), the aggregate effect is not significant either economically or statistically.

This result suggests θ might be endogenous to the probability of control. Although a threat of a strong institution crowds out intrinsic motivation for honesty, a certain weak institution has no such effect. That is, having no institution or a weak institution is less detrimental than having a malfunctioning one.

4.6 Is it about the threat?

The previous section demonstrates that the impact of a malfunctioning institution is not determined solely by a small expected probability of control. In summary, the results from the earlier sections indicate that uncertain control institution leads to crowding out. This crowding-out effect is supported by an increased number of subjects who lie at least once (suggesting an early crowding out upon the announcement of the possibility of control) and a higher percentage of rounds in which they lie (based on learning that the control institution is likely malfunctioning) in 50-50NoControl than in NoControl. A crucial question arises as to whether the crowding-out effect is driven by uncertainty about the control institution, allowing subjects to attribute the responsibility of lying to the potential presence of control, or if the announcement itself undermines the motivation to report honestly, regardless of the uncertainty of the control institution.

To explore this question, we conducted an additional treatment Known50-50. The design and procedures are the same as in the 50-50 treatment, with one key difference: participants are informed of their assignment to either the Control30 or NoControl treatment just before the first round. We refer to these sub-treatments as Known50-50Control30 and

Known50-50NoControl. This treatment allows us to determine whether the negative impact of the threat announcement persists even when uncertainty about the control institution is absent.

Recall that under 50-50, the optimal report in round r is the following:

$$j = \begin{cases} i \ \ if \ i > (10 - \theta t)p_r + (7 - \theta t)(1 - p_r) \\ 10 \quad otherwise \end{cases},$$

where p_r is the probability of being in the NoControl sub-treatment in round r

In Known50-50NoControl, $p_r = 1$ starting from round 1, thus

$$j = \begin{cases} i \ \ if \ i > 10 - \theta t \\ 10 \quad otherwise \end{cases}.$$

Thus, we have the following predictions for the truthfulness rates between treatments:

Control30=50-50Control30=Known50-50Control30>NoControl>50-50NoControl>Known50-50NoControl.

In other words, the announcement of a threat followed by an immediate resolution of the uncertainty is expected to result in the highest rate of misreporting. It is important to note, however, that this prediction assumes that motivation does not crowd in. Our understanding of when and for whom crowding in might occur is still developing, and this treatment could offer a valuable opportunity to explore the relevance of the crowding-in phenomenon.

In the experiment, we observe a significant impact of an announced but *with certainty* not implemented control institution on truthfulness. Figure 6 clearly illustrates that, in aggregate, truthfulness in the Known50-50NoControl treatment is consistently the lowest among all the treatments right from the start of the experiment. In contrast, truthfulness in the Known50-50Control30 treatment is comparable to that in the NoControl, 50-50Control30, Control30, and Control0.03 treatments.

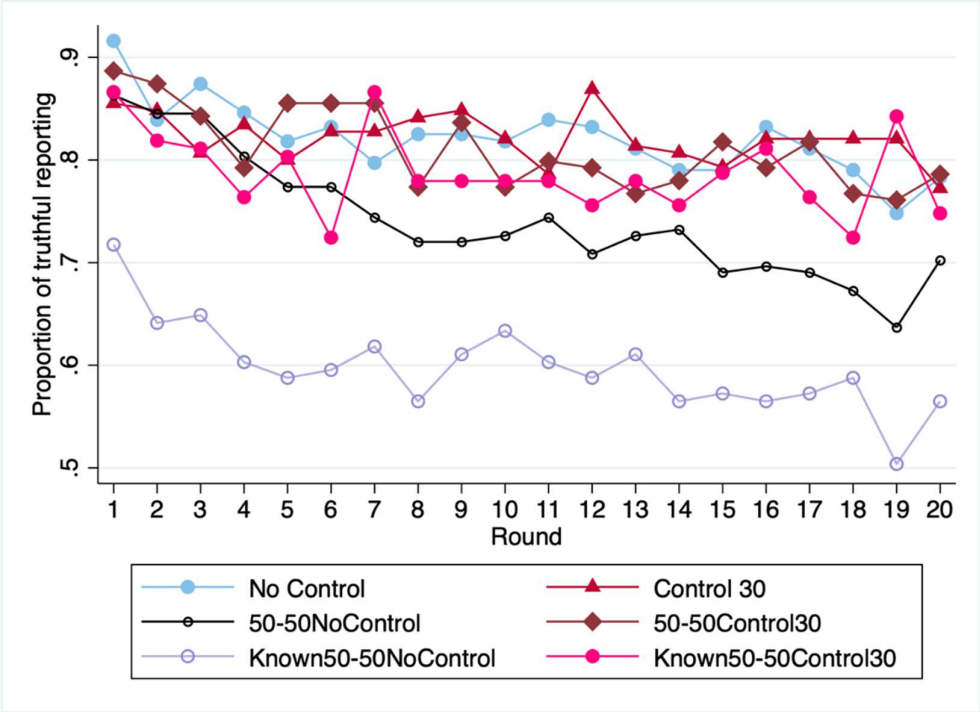


Figure 6. Proportion of truthful reporting by treatments and rounds for all treatments

Table 7 presents the marginal effects of probit regressions of truthful reporting in the last 10 rounds. Model (1) contains only treatment dummies as explanatory variables, and Model (2) contains additional controls. Both models support the intuition from Figure 6 that in the later rounds, the Known50-50NoControl treatment leads to significantly lower rates of truthful reporting than in the NoControl condition. Furthermore, we conduct pairwise comparisons using regression analyses and find that in the last 10 rounds, truthfulness in Known50-50NoControl is significantly lower than in any other treatment (including 50-50NoControl).

	Truthful (1)	Truthful (2)
Control30	0.017	0.031
	(0.031)	(0.030)
50-50NoControl	-0.095***	-0.097***
	(0.031)	(0.030)
50-50Control30	-0.010	0.003
	(0.029)	(0.028)
Known50-50NoControl	-0.200***	-0.182***
	(0.033)	(0.032)
Known50-50Control30	-0.024	-0.022
	(0.029)	(0.029)
Female		0.083***
		(0.019)
Age		0.006***
		(0.002)
Draw		0.053***
		(0.002)
Round		-0.005***
		(0.001)
Observations	10140	10140
Number of clusters	991	991
Sample	Last 10 rounds	Last 10 rounds
Pseudo *R*-squared	0.026	0.27

Notes: Standard errors in parentheses, * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Standard errors are clustered at the individual level.

Table 7. Marginal effects of probit regressions of truthful reporting in the last 10 rounds

Overall, the results from the Known50-50NoControl sub-treatment reveal that the mere announcement of the possibility of control is sufficient to crowd out good behavior. Even when we subsequently *retract the threat* by assuring subjects that they will definitely not be subjected to control, it still results in negative effects.

At this point, it is important to note that the theory does not fully account for the substantial differences in lying behavior between 50-50NoControl and Known50-50NoControl. According to the theory, lying rates in 50-50NoControl should be somewhat lower than in Known50-50NoControl, but only slightly. After several rounds of observing no control in 50-50NoControl, the lying rates should approach those observed in Known50-50NoControl.

Although our experimental data cannot pinpoint the mechanism behind the stark differences in lying behavior between 50-50NoControl and Known50-50NoControl, several potential explanations could account for these results. As noted in the introduction, it's possible that participants in 50-50NoControl may not accurately update their beliefs about the likelihood of being in that condition. If they overestimate the probability of being in 50-50Control, this could lead them to lie less than they would if their beliefs were accurate.

Another possibility is that in Known50-50NoControl, some participants might feel compelled to lie after being explicitly told that there was a possibility of being in the treatment with control, but that they specifically ended up in the treatment without control. In contrast, since no such statement is made in 50-50NoControl, participants may feel less pressure to lie. There are undoubtedly several other explanations for the differences between treatments, and identifying the precise mechanisms will be a task for future research.

5 Conclusion

We present empirical evidence that malfunctioning institutions lead to more lying compared to trust-based institutions. This detrimental effect of malfunctioning institutions aligns with our theory, which suggests that the threat of a strong control institution can crowd out some individuals' intrinsic motivation to tell the truth. Learning almost with certainty that no strong institution is in place does not crowd in intrinsic motivation, but rather leads to these individuals lying more on the intensive margin, because no monitoring exists, unlike in the strong institution. The effect of the strong threat is established in comparison to weak institutions, strong institutions, and no institution at all. Whether a weak threat would produce similar effects remains an open question for future research.

The main contribution of this paper is the discovery of a new behavioral regularity. To our knowledge, none of the previous theories or experimental investigations on a crowding-out effect have analyzed uncertain incentives, which are common in real life. We show that an uncertain threat of punishment leads to severe detrimental effects and high costs for an institution designer. Importantly, we also find that a threat that is unequivocally retracted leads to the highest levels of dishonesty. In other words, announcing distrust and subsequently "taking it back" represents the most detrimental course of action for an institution designer.

The detrimental effect of the threats can be rationalized by the theory of Sliwka (2007), which argues that control institutions signal social norms. Adopting Sliwka's (2007) theory to our context, the population consists of three types: unconditional truthful types, utility maximizers, and conditional truthful types, with the latter reporting truthfully if they expect others to report truthfully. The mere possibility of being in Control30 signals to conditional truthful types that there is a high proportion of lying in the population, causing them to adopt a utility-maximizing behavior. While Control30 can counterbalance this crowding-out effect with their control institution, the 50-50NoControl and Known50-50NoControl conditions lack monitoring capabilities. Consequently, these institutions combine the worst aspects of both worlds: they lead to a crowding-out of conditional honest types into utility maximizers, while lacking the ability to effectively monitor and enforce truth-telling.

We are also the first to show the crowding-out effect exists in cheating games—an effect that has not been studied, even given the vast lying literature in behavioral economics.

We believe the insights of our paper generalize to other contexts, like altruistic behavior, public good contributions, and other contexts with where intrinsic motivation matters for behavior. A policy implication from our study is that having no institutions in place is better than threats of strong institutions that will not be implemented. For instance, it might be more

efficient to let people quarantine away from their families on a trust basis than threatening them with large fines. It could also be more efficient to explicitly state to low income tax payers (who are in any case not controlled extensively), that their reports will not be audited. Such trust institutions might work better than threatened with but not implemented ones. Beyond economics, we believe our paper sheds light on the potential effects of unenforced laws. We provide empirical evidence that law might crowd out intrinsic motivation to behave as the law prescribes before it is evident that it is not enforced.

Lastly, our paper has certain limitations. As mentioned earlier, the theory does not fully explain the stark differences in lying rates between Known50-50NoControl and 50-50NoControl. Future research should focus on understanding the mechanism that leads to stronger detrimental effects when the possibility of control is announced and immediately revoked, compared to when the possibility of control is announced but not revoked. Additionally, future studies could examine the relative importance of the "threat" versus the "strength of the threat" of control in contributing to the observed detrimental effects, compared to the absence of institutions. Further research could also explore the effects of unenforced institutions in other decision-making situations and games beyond cheating games.

References

Abeler, J., Becker, A., Falk, A. (2014). Representative Evidence on Lying Costs. *Journal of Public Economics*, 113, 96–104.

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

Alm, J., McClelland, G. H., Schulze, W. D. (1992). Why do people pay taxes? *Journal of Public Economics*, 48, 21-38.

Bae, K. H., Goyal, V. K. (2009). Creditor rights, enforcement, and bank loans. *The Journal of Finance,* 64(2), 823-860.

Bhattacharya, U., Daouk, H. (2002). The world price of insider trading. *The Journal of Finance,* 57 (1), 75-108.

Bhattacharya, U., Daouk, H. (2009). When no law is better than a good law. *Review of Finance,* 13(4), 577-627.

Beck, P.J., Davis, J.S, Jung, W.O. (1991). Experimental Evidence on Taxpayer Reporting under Uncertainty. *The Accounting Review*, 66, 535-558.

Becker, G. S. (1968). Crime and punishment: an economic approach. *Journal of Political Economy*, 76(2), 169–217.

Bénabou, R., Tirole, J. (2003). Intrinsic and extrinsic motivation. *Review of Economic Studies*, 70, 489–520.

Bénabou, R., Tirole, J. (2006). Incentives and prosocial behavior. *American Economic Review*, 96 (5), 1652–1678.

Chen, D. L., Schonger, M., Wickens, C. (2016). oTree—An open-source platform for laboratory, online, and field experiments. *Journal of Behavioral and Experimental Finance*, 9, 88-97.

Cooter, R. D. (1998). Expressive Law and Economics. *Journal of Legal Studies*, 27(2), 585-607.

Dai, Z., Hogarth, R. M., Villeval, M. C. (2015). Ambiguity on audits and cooperation in a public goods game. *European Economic Review*, 74, 146-162.

Deci, E. L., Koestner, R., Ryan, R. M. (1999). A meta-analytic review of experiments examining the effects of extrinsic rewards on intrinsic motivation. *Psychological Bulletin*, 125(6), 627–668.

Depoorter, B., Tontrup, S. (2016). Costly Rights - A Field Study on Symbolic Laws. *SSRN Electronic Journal.*

Dickinson, D. L. and Villeval, M. C. (2008). Does monitoring decrease work effort? The complementarity between agency and crowding-out theories', *Games and Economic Behavior*, 63(1), 56–76.

Falk, A., Kosfeld, M. (2006). The hidden costs of control. *American Economic Review*, 96 (5), 1611–1630.

Fehr, E., Gächter, S. (2000). Cooperation and punishment in public goods experiments. *American Economic Review*, 90(4), 980–994.

Fehr, E., List, J. A. (2004). The hidden costs and returns of incentives – Trust and trustworthiness among CEOs. *Journal of the European Economic Association*, 2(5), 743–771.

Fehr, E., Rockenbach, B. (2003). Detrimental effects of sanctions on human altruism. *Nature*, 422, 137–140.

Fischbacher, U., Föllmi-Heusi, F. (2013). Lies in Disguise: An Experimental Study on Cheating. *Journal of the European Economic Association*, 11(3): 525–547.

Frey, B. S., Jegen, R. (2001). Motivation crowding theory: A survey of empirical evidence. *Journal of Economic Surveys*, 15, 589–611.

Frey, B. S., Oberholzer-Gee, F. (1997). The cost of price incentives: An empirical analysis of motivation crowding-out. *American Economic Review*, 87 (4), 746–755.

Galeotti, Fabio, Charlotte Saucet, and Marie Claire Villeval. "Unethical amnesia responds more to instrumental than to hedonic motives." Proceedings of the National Academy of Sciences 117.41 (2020): 25423-25428.

Galeotti, F., Maggian, V., Villeval, M. C. (2021). Fraud deterrence institutions reduce intrinsic honesty. *The Economic Journal*, 131, 2508–2528.

Gächter, S., Schulz, J.F. (2016). Intrinsic honesty and the prevalence of rule violations across societies. *Nature*, 531(7595), 496–499.

Gneezy, U. (2005). Deception: The Role of Consequences. *American Economic Review*, 95(1): 384–394.

Gneezy, U., Meier, S., Rey-Biel, P. (2011). When and Why Incentives (Don't) Work to Modify Behavior. *Journal of Economic Perspectives*, 25 (4): 191-210.

Gneezy, U., Kajackaite, A., Sobel, J. (2018). Lying Aversion and the Size of the Lie. *American Economic Review*, 108 (2), 419-53.

Gneezy, U., Rustichini, A. (2000). A fine is a price. *Journal of Legal Studies*, 29(1), 1–18.

Gürerk, O., Irlenbusch, B., Rockenbach, B. (2006). The competitive advantage of sanctioning institutions. *Science*, 312(5770), 108-111.

Harbaugh, W. T., Mocan, N., and Visser, M. S. (2013). Theft and Deterrence. *Journal of Labor Researc*h, 34, 389-407.

Kajackaite, A., Gneezy, U. (2017). Incentives and Cheating. *Games and Economic Behavior*, 102, 433–444.

Laske, K., Saccardo, S., Gneezy, U. (2018). Do Fines Deter Unethical Behavior? The Effect of Systematically Varying the Size and Probability of Punishment. *Working Paper.*

Leslie, C. (2000). Creating criminals: The injuries inflicted by "unenforced" sodomy laws. *Harvard Civil Right – Civil Liberties Law Review*, 35(1), 103-181.

Mauro, P. (1995). Corruption and growth. *Quarterly Journal of Economics*,110, 681–712.

Mellström, C., Johannesson, M. (2008). Crowding out in blood donation: Was titmuss right? *Journal of the European Economic Association*, 6(4), 845–863.

Nikiforakis, N., Normann, H. T. (2008) A comparative statics analysis of punishment in public-good experiments. *Experimental Economics*, 11, 358–369.

Olken, B. A., Pande, R. (2012). Corruption in developing countries. *Annual Review of Economics*, 4, 479–509.

Pranab, B. (1997). Corruption and development: a review of issues. *Journal of Economic Literature, 35*, 1320–1346.

Shalvi, S., Dana, J, Handgraaf, M. J. J., De Dreu, C. K. W. (2011). Justified Ethicality: Observing Desired Counterfactuals Modifies Ethical Perceptions and Behavior." *Organizational Behavior and Human Decision Processes*, 115(2), 181–190.

Sliwka, D. (2007). Trust as a Signal of a Social Norm and the Hidden Costs of Incentive Schemes. *American Economic Review*, 97 (3), 999-1012.

Webley, P. (1987). Audit probabilities and tax evasion in a business simulation. *Economics Letters*, 25, 267-270.