Fraud Deterrence Institutions Reduce Intrinsic Honesty

Fabio Galeotti, Valeria Maggian, Marie Claire Villeval

GATE LYON SAINT-ÉTIENNE
UMR 5824

93, chemin des Mouilles
69130 Ecully - France

Maison de l'Université, Bâtiment B
10, rue Tréfilerie
42023 Saint-Etienne cedex 02 - France

http://www.gate.cnrs.fr
gate@gate.cnrs.fr

WP 2039 – December 2020

Fraud Deterrence Institutions Reduce Intrinsic Honesty

Fabio Galeottia, Valeria Maggian, Marie Claire Villeval

Abstract:

This paper studies the effect of transportation networks on spatial inequality within metropolitan areas. It uses a spatial equilibrium model featuring nonhomotheticities and worker heterogeneity, allowing to capture rich patterns of workers sorting on commuting costs and amenities. The model is calibrated for the Paris urban area. Counterfactual simulations study the effects of a) the Regional Express Rail and b) restricting car use in the city center. Despite a strong contribution to suburbanization and reducing welfare inequality, the public transport network plays no role in reducing income segregation. The effects of banning cars depends critically on the response of residential amenities in the city. If it is low enough, it reduces income disparities between Paris and its suburbs at the cost of a substantial welfare loss. If it is high enough, the policy creates welfare gains but steepens the income gradient.

Keywords:

Deterrence Institutions, Intrinsic Honesty, Spillovers, Quasi-Experiment

JEL codes:
C93, K42, D02, D91

Fraud Deterrence Institutions Reduce Intrinsic Honesty

Fabio Galeotti[a], Valeria Maggian[b] and Marie Claire Villeval[c,d]

Abstract: Deterrence institutions are widely used in modern societies to discourage rule violations but whether they have an impact beyond their immediate scope of application is usually ignored. Using a quasi-experiment, we found evidence of spillover effects across contexts. We identified fraudsters and non-fraudsters on public transport who were or not exposed to ticket inspections by the transport company. We then measured the intrinsic honesty of the same persons in a new, unrelated context where they could misappropriate money. Instead of having an expected educative effect across contexts, the exposure to deterrence practices increased unethical behavior of fraudsters but also, strikingly, of non-fraudsters, especially when inspection teams were larger. Learning about the prevailing norm is the most likely channel of this spillover effect.

JEL-Classification: C93, K42, D02, D91

Keywords: Deterrence Institutions, Intrinsic Honesty, Spillovers, Quasi-Experiment.

[a] Univ Lyon, CNRS, GATE, UMR 5824, F-69130 Ecully, France. galeotti@gate.cnrs.fr

[b] Cà Foscari University of Venice, Department of Economics. Cannaregio 873, Fondamenta San Giobbe, 30121 Venice, Italy. valeria.maggian@unive.it

[c] Corresponding author: Univ Lyon, CNRS, GATE, UMR 5824, F-69130 Ecully, France. villeval@gate.cnrs.fr

[d] IZA, Schaumburg-Lippe-Strasse 5-9, 53113 Bonn, Germany

Acknowledgments: We thank J. Andreoni, M. Bigoni, G. Charness, J. Van de Ven, and audiences at seminars (at the Universities of Aarhus, Ancona, Bilkent, Birmingham, Innsbruck, Maastricht, Montpellier, Paris 1, Venice) and conferences (AFSE conference in Paris, ASFEE conference in Nice, 1st BEEN meeting at the University of Bologna, CBT Research day of EMLyon, ESA North-American meeting in Antigua, ESA World meeting in Berlin, EWEBE meeting in Tilburg, GATE-Lab-CORTEX workshop in Lyon, Organizations and Markets workshop in Dijon, Workshop on Social Economy for Young Economists in Bologna, 10th Birthday Conference of the Collegium of Lyon) for useful comments. We are grateful to R. Sauter and the TCL company in Lyon for facilitating our access to the transport network. This research has benefited from financial support from the LABEX CORTEX (ANR-11-LABX-0042) of Université de Lyon, within the program Investissements d'Avenir (ANR-11-IDEX-007) operated by Agence Nationale de la Recherche (ANR), from the INDEPTH program of IDEXLYON (No 183634), and from a grant of the French National Research Agency (ANR, DECISION project, ANR-19-CE26-0019). This project has also received funding from the European Union's Horizon 2020 research and innovation program under the Marie Sklodowska-Curie grant agreement No 661645. The IRB of INSERM (IRB00003888) has approved the project (decision n°17-416). The AEARCT identification number of the present study is AEARCTR-0004656.

1. Introduction

Honesty and norm compliance are fundamental for the maintenance of trust and the development of prosperous societies (Mauro, 1995; Knack and Keefer, 1997). Norms can be sustained by an internalization mechanism that induces individuals to comply even in absence of any threat of punishment (Axelrod, 1986; Gintis, 2003). However, intrinsic honesty is not sufficient to prevent violations, and varies widely across cultures (Fisman and Miguel, 2007; Gächter and Schulz, 2016; Cohn *et al*., 2019). While peer punishment of violations facilitates compliance (Fehr and Gächter, 2000; Masclet *et al*., 2003), inspections and sanctions by centralized authorities are the most common institutional practices adopted in modern societies to deter deviant behavior. When they raise the costs of breaking the rule above its benefits, these institutions can discourage the targeted misbehavior (Becker, 1968; Di Tella and Schargrodsky, 2004; Fisman and Miguel, 2007; Baldassarri and Grossman, 2011; Ariely, 2012; Chalfin and McCrary, 2017). However, they sometimes crowd out the intrinsic motivation to comply (Gneezy and Rustichini, 2000; Frey and Jegen, 2001; Falk and Kosfeld, 2006; Dickinson and Villeval, 2008), with potential spillovers into adjacent activities (Belot and Schröder, 2016). This results from control-averse individuals who directly reciprocate against a distrustful authority that reduces their freedom of choice.

While past estimations of deterrence effects focus almost exclusively on the targeted misbehavior, we contend that indirect effects may expand across contexts and impact both compliers and non-compliers. For example, it has been found that past exposure to institutions fostering prosocial norms can improve future pro-sociality even when the institution is no longer enforced (*e.g.*, Cassar *et al.,* 2014; Peysakhovich and Rand, 2016; Engl *et al*., 2017; Galbiati *et al.,* 2018). Here, we look at possible spillover effects across contexts from inspecting and sanctioning people for rule violations on one of the most fundamental traits of human beings: intrinsic honesty. Investigating whether these spillovers exist is essential to better understand the overall effectiveness of these institutions, which crucially depends on whether they also affect socially desirable behavior beyond their immediate scope of intervention. Yet, to the best of our knowledge, no one has ever shed light on this issue.[1]

[1] The literature identifying negative effects of monitoring people in different productivity dimensions (*e.g*., Gneezy and Rustichini, 2000; Falk and Kosfeld, 2006; Dickinson and Villeval, 2008) evaluated these effects only in the institution's direct operational context. For example, Belot and Schröder (2016) show that controlling employees'

Why should we expect spillover effects of deterrence institutions on individuals' intrinsic honesty across contexts? The traditional economic approach to crime (Becker, 1968) is silent on their existence.[2] Alternatively, psychological and behavioral economics theories could account for these effects – which may be positive or negative. Focusing on dishonest individuals, if past experience of a deterrence institution recalls what society expects from individuals, it may serve as an educative tool for the future and foster intrinsic honesty (on the socio-pedagogical effect of punishment see, *e.g.*, Hawkins, 1969; Andenaes, 1974; Hampton, 1984). At the same time, individuals who are caught breaking the rule are usually fined. In a subsequent, unrelated context, they may be tempted to misbehave again in order to recover their financial loss (Sharma *et al.*, 2013). Intrinsic honesty may also decrease if dishonest individuals evaluate their moral activities dynamically (Nisan, 1991; Effron and Conway, 2015) and consider that the sanction has cleansed their past immoral actions ("*I paid for my sin*"), reducing the discrepancy between one's perceived self-image and the desired moral self.

Exposure to a deterrence institution may also have spillover effects on the intrinsic honesty of norm-compliers. On the one hand, the educative effect of deterrence institutions can act as a positive reinforcer. On the other hand, signaling theories (*e.g.*, Benabou and Tirole, 2003) deliver mixed predictions: an inspection may signal to compliers that they are honest, reinforcing their intrinsic motivation; but it may also remind some people that their intrinsic motivation for compliance is avoiding a fine, and crowds out their honesty in subsequent contexts where they know the deterrence institution is not in place anymore. Also, because of social learning, the enforcement of the deterrence institution may affect compliers' beliefs about the spread of norm violation in society. Observing many violators being punished or large police teams, may reveal

performance may reduce their punctuality. These spillover effects within the same context are usually explained by direct reciprocity. However, this literature ignores whether these effects spill over to other contexts that are not regulated by the institution – where direct reciprocity is ruled out – by affecting individuals' intrinsic motivation.

[2] Most of the literature on deterrence in the Beckerian tradition examines whether variations in the probability of detection *vs*. severity of sanctions affect criminality (see Chalfrin and McCrary, 2017, for a survey). The only spillovers considered are those related to crime displacement following a sudden increase in the intensity of policing (see review in Weisburd *et al*., 2006), or those related to the incidence of more serious crimes following an increase in the intensity of arrest for small crimes (*e.g.*, Wilson and Kelling 1982). These studies – mostly conducted at the aggregate level – tend to be afflicted by simultaneity bias, omitted variables, and identification problems (Chalfrin and McCrary, 2017). In addition, they do not inform on spillover effects of the enforcement of the institution on intrinsic honesty. They consider only whether offenders reduce their criminal activities or relocate somewhere else after they update their perceived risk of apprehension in response to an increase in policing.

that misconduct is socially widespread and has become the norm, which may lead compliers to behave accordingly (Keizer *et al*., 2008; Dickinson *et al*., 2015).

In sum, there exist several mechanisms that could lead to spillover effects of deterrence institutions across contexts at the individual level. However, whether these spillover effects truly exist in the real world, whether they are positive or negative, and whether they affect rule compliers and non-compliers alike remain open questions. To shed light on these speculations, we ran a quasi-experiment in public transport and on the streets in Lyon, France, with 708 passengers. We collected a direct and unbiased measure of dishonest behavior (*i.e*., fare evasion). Individuals were observed in a daily-life situation and were not aware of their taking part in a study. Our quasi-experiment allows us to overcome the limitations associated with laboratory experiments (Levitt and List, 2007; List, 2011) especially when investigating dishonesty, since possible scrutiny by the experimenter can considerably influence unethical behavior (Gneezy *et al*., 2018). Moreover, eliminating any experimenter demand effect is of utmost importance to rule out direct reciprocity as a possible explanation of our results, especially when formal monitoring can be perceived as a form of distrust. Finally, besides contributing to the analysis of the dynamics of unethical behavior (but in a different sense to Welsh, 2015, and Garrett *et al.,* 2016, who look at escalation effects), we focus on both compliers and non-compliers.

We chose to conduct our quasi-experiment in public transport because in France all socio-demographic categories use public transport and fare evasion is relatively widespread (Cour des Comptes, 2016; Dai *et al*., 2018).[3] This means that when we study fraudsters, we are not looking at a tiny minority of people. Moreover, in this setting dishonest behavior is publicly identifiable with almost no measurement error, since every passenger must validate a ticket or a pass every time they enter a public vehicle.

Our quasi-experiment consisted of two stages. The first stage took place on board buses and trams and produced two main natural conditions. In the *Inspection condition*, the targeted passenger was controlled by ticket inspectors from the transport company during his or her journey whereas

[3] A 2011 survey conducted by OpinionWay in Lyon for the local public transport company revealed that 55% of the participants sometimes travel without a valid ticket (Keolis, 2014). The company estimates that around 1 out of 7 trips on the tram or bus is irregular (http://www.sytral.fr/uploads/Externe/9d/310_765_CP_CS_02_02_2018.pdf, accessed 23.09.2020).

in the *No-Inspection condition*, no ticket inspection occurred.[4] The second stage took place when the passenger exited the vehicle, on the street. A professional actor who was part of the research team walked behind the targeted passenger and suddenly bent down to seemingly pick up a banknote on the ground. The actor then called the passenger's attention by asking whether they had lost the banknote. We measured intrinsic honesty by recording whether the passenger took or not the banknote, and tested whether this correlates with their compliance on public transport. To identify the causal effect of the deterrence institution on compliers and non-compliers, we contrasted intrinsic honesty in the Inspection *vs*. No Inspection conditions. We explored the possible mechanisms behind our findings by means of an additional survey conducted in public transport, and two laboratory experiments.

We found that instead of having a positive immediate educative effect in the new context, the direct exposure to a deterrence institution in public transport increased the misconduct of fraudsters on the street. More strikingly, it also significantly increased the unethical behavior of non-fraudsters. The effect was highly significant and was of the same magnitude for both groups (between 14% and 19% of the base level). This rejects a general explanation of the spillover in terms of monetary loss recovery. Interestingly, the effect increased with the size of the ticket inspection team, especially for non-fraudsters. Without rejecting the role of emotions in affecting our results, this suggests that one mechanism behind such effect may be a normative channel: larger inspection teams may signal more widespread dishonesty and a weaker social norm of honesty in the society. Overall, our findings show that to optimize the design of deterrence institutions and evaluate their full efficiency, policy makers should also consider the spillover effects of these institutions on intrinsic honesty beyond the context where these institutions directly apply.

[4] Note that our identification strategy is not based on exogenous shocks in the deterrence policy but on natural variations in its implementation. This implementation, we believe, acts as a reminder of the existence of the institution and, thus, as a proxy of a change in the institution. Since inspections result from the transport company policy, and not from a random intervention by the researchers, our study can be defined as a quasi-experiment. We address the potential issues regarding randomization in section 2.

2. Experimental Design

Our quasi-experiment consisted of two stages and was conducted by teams composed of a research assistant and a professional actor, both blind to the hypotheses of the study.[5] [6] The first stage aimed at identifying dishonesty in a natural setting where formal deterrence institutions could be enforced. It took place on public transport in Lyon (France) where the identification of fraudsters and non-fraudsters is direct: in order not to incur a fine, all passengers need to validate their ticket or pass at fareboxes located on board public vehicles each time they enter a new vehicle. Re-validation is compulsory even if the ticket is valid for one hour from the first validation and has been already validated in a previous vehicle. Thus, someone who does not validate a pass or a ticket is in an irregular situation, and he or she is classified as a fraudster in our analysis.[7] Details about the public transport network in Lyon and ticket inspections can be found in Appendix 1.1.

First stage. In the first stage of the experiment, the research assistant and the actor traveled on board a bus or tram. The former had to stay next to a validating farebox and focus attention on the first four of five passengers boarding and validating or not their ticket. This was done for logistical reasons: first, it was easier to recognize those who validated and those who did not by targeting the first passengers entering the public vehicle; second, it was easier to recall who stamped the ticket and who did not if only a few passengers at a time were targeted. The actor waited on board the public vehicle without giving any impression of travelling with the research assistant. Once the first of these passengers got off the vehicle, both the research assistant and the actor also got off behind the targeted passenger. This procedure avoided subjectivity in the choice of the target

[5] We used professional actors to ensure that the scene was played as similarly as possible across conditions. Four actors (two males, two females) were selected after a recorded casting with 18 candidates in a professional acting school. 21 subjects from the subject pool of the GATE-Lab in Lyon were recruited via Hroot (Bock *et al*., 2014) and paid €15 to evaluate the actors in terms of performance, apparent honesty, trustworthiness, attractiveness, credibility, seriousness, and friendliness, based on the videos. We selected those actors with similar high scores in performance and credibility (see Figure A1 in Appendix 4). Before us, only a few studies have used professional actors (Fischer *et al*., 2006; Swami *et al*., 2008; Gino *et al*., 2009; Wang *et al.,* 2012; Antonakis *et al*., 2015; Sands, 2017; Winter and Zhan, 2018).

[6] Note that 30% of the observations were collected in the presence of an experimentalist who was not blind to the research questions, due to the unavailability of the assistant or because, in the Audience condition – presented below – we needed two assistants (one to walk by the actor and another one collecting observable characteristics). As shown in the next section, this presence did not affect the results.

[7] After scanning a ticket or a pass, the farebox emits a clear beeping sound, which makes forgetting to validate unlikely if other people are boarding at the same time. In buses, front door entry is compulsory but drivers have no responsibility for checking validation and they actually do not inspect. The only possible measurement error is when a passenger validates a ticket with a special tariff (*e.g.,* tariffs for seniors or unemployed) he or she is not entitled to.

passenger and was cost-effective, preventing the research team from spending too much time on board.

There were two conditions that occurred naturally. In the Inspection condition (I, hereafter), the targeted passenger was controlled by a team of ticket inspectors from the transport company during or at the end of the ride, whereas in the No-Inspection condition (NI, hereafter), no ticket inspection occurred. We address the question of randomization in a separate subsection below.

Second stage. The second stage of the experiment was conducted on the street, where we measured the intrinsic honesty of the same targeted passengers in a context where no formal institution applies. The actor, while having a fake phone conversation to minimize interactions, suddenly bent down to seemingly pick up a 5-Euro banknote on the ground, just behind the targeted passenger. The actor then called the attention of the targeted passenger by asking whether they had lost the banknote. Accepting or not the banknote is our measure of intrinsic honesty. Meanwhile, the research assistant observed the scene and collected data on a tablet, regarding the decision to accept or not the banknote, any observable characteristics of the passenger (*e.g.*, apparent wealth and age, gender, emotional reaction to an inspection) and the environment (*e.g.*, approximate number of people on board, number of ticket inspectors, payment of a fine). The actor was instructed to play the scene with no audience within hearing distance. As a robustness check of the effect of an audience on intrinsic honesty we ran an additional condition, the No-Inspection-Audience condition (NI-A, hereafter). Here, the assistant walked by the actor and explicitly observed the scene. This allowed us to isolate the possible role played by an observer in influencing individuals' unethical behavior.

The actors were asked to use their mobile phone as an audio recording device when playing the scene on the street.[8] In order to make sure that (*i*) the actors played the scene similarly across conditions, and (*ii*) one's false claim of ownership of a banknote found by someone else violates an injunctive ethical norm, we conducted a laboratory experiment (called "Laboratory Experiment 1" that is presented in detail in Appendix 2.1 (see instructions in Appendix 3.1)).

[8] These recordings were used to verify that the actors played the scene according to the protocol (see below), and as a robustness check to ensure that any minimal deviation from this protocol did not affect the internal validity of our results. We thank James Andreoni for suggesting this to us. The script given to the actors is available in Appendix 1.2.

Recorded variables. The research assistant had to record several pieces of information on a tablet. He recorded the name of the actor playing the scene, the time of day, the weather (sunny, cloudy or rainy), the bus/tram line, whether the subject validated a ticket or a monthly pass or nothing, the bus/tram stop where the subject got off the public vehicle, the approximate number of people on board (almost empty, quite crowded but everyone could sit, crowded), whether someone could notice the scene played in the street, whether the subject took or not the banknote, the gender, estimated age (18-24, 25-34, 35-44, 45-59, 60 or more), estimated economic status based on appearance (poor, average, wealthy), and ethnicity (Caucasian, Arab, African, Asian, other) of the targeted passenger, and whether the subject wore religious symbols. In the Inspection condition, the research assistant also recorded the number of ticket inspectors, whether the inspection was conducted at the tram/bus stop or on board, whether the ticket inspectors wore uniform or plain clothes, the gender of the controller who inspected the targeted passenger, and, only for inspected fraudsters, whether the passenger paid the fine immediately, and whether he or she had an emotional or aggressive reaction to being fined.

Tables A1 and A2 in Appendix 5 present descriptive statistics of the targeted passengers' individual background variables in each condition, for the whole sample, and for the sub-groups of fraudsters and non-fraudsters, respectively.

Sample size and power analysis. The experiment was run on weekdays in 2017. On a typical weekday, we collected on average 21 observations between 9:00AM and 6:15PM, avoiding rush hours because passengers may anticipate that the risk of ticket inspection is lower during these hours. In total, our study involved 708 passengers: 358 non-fraudsters (104 in the I condition, 140 in NI and 114 in NI-A) and 350 fraudsters (100 in the I condition, 140 in NI, and 110 in NI-A). When collecting data, we excluded vulnerable persons, minors and tourists (based on subjective judgment), persons accompanied by children, friends, colleagues or partners. Details about the sample distribution across lines and at different times of the day and locations are reported in Table A3 in Appendix 5, and Figure A4 in Appendix 4 displays the frequency of inspections on a map.

To determine the sample size for both the NI and I conditions, we conducted an a priori power analysis. For the NI condition, we built on the results of Dai *et al*. (2018) to achieve a sample size of 92 subjects per group (fraudsters and non-fraudsters), which we rounded to a more conservative 100 (see details in Appendix 1.3). For the I condition, it was too speculative to allow any prediction about the direction and the effect size by comparing NI to I. We thus set the sample size at 100

observations (*i.e.*, the optimal sample size for the NI condition) for each treated group (fraudsters and non-fraudsters) and computed the minimum detectable effect size for $\alpha = 0.05$ and power = 0.8. The minimum detectable effect size was 0.19 for fraudsters and 0.20 for non-fraudsters. This corresponded to a Cohen's *h* of approximately 0.4. Hence, a sample size of 100 was large enough to detect a small-medium treatment effect. In running the quasi-experiment, we thus decided to stop collecting data once we reached (roughly) 100 observations per group in the I condition.

Collecting data in both conditions was time-consuming, especially in the I condition, which depended on the natural occurrence of ticket inspections.[9] Therefore, we instructed the research assistant and the actor to primarily focus on searching for ticket inspectors by travelling up and down a random line and switching to another if unsuccessful. They were asked to start from a different main line every day in a direction chosen at random, with the overall objective of keeping a roughly constant proportion of I and NI observations throughout the day. While we endeavoring to reach the target of 100 for the I condition, we collected, however, more data in the NI condition so as not to waste the actors' time (they were paid per hour). It is important to note that every day we sampled roughly three or four random observations in the NI condition for each observation sampled in the I condition, to account for the different sampling costs, and tried to maintain this ratio throughout the experiment.[10] The higher number of observations in the NI condition does not reflect any problem with the first hundred observations in this condition.

Identification. Our experimental design, combining the three conditions described above (I, NI, NI-A) with the regular or irregular condition of the passenger on board the bus or tram, allows us to achieve a twofold objective: first, to investigate whether there is a correlation between the honesty of passengers in the bus/tram and on the street and, second, to identify the causal effect of ticket

[9] By choosing a target of 100 in each condition, we obviously over-represented the population of inspected passengers. In fact, the probability of being inspected is quite low in the field. Egu and Bonnel (2020) estimate that in 2017 in Lyon, the ratio between the number of ticket inspections and the number of boardings was 0.017 for the tram and 0.012 for the bus. Precisely, boardings amounted to 95,2M and 166,1M for the tram and bus network, respectively, as measured by the counting system placed at each vehicle door; controllers inspected 1,6M and 1,9M people in the tram and bus, respectively. Thus, respecting this proportion with the constraint of 100 observations in the I condition would have required collecting between 1200 and 1700 observations in the NI condition.

[10] Imposing this 3:1 or 4:1 ratio allowed us to account for the substantially different costs of sampling in the two conditions. Indeed, it has been shown that the optimal sample sizes should be inversely proportional to the square root of the relative sampling costs (see Pentico, 1981; List et al. 2011). Since the sampling costs are much higher in the I condition, the NI sample should be larger than the I sample.

inspections on the latter. Our identifying strategy relies on the assumptions that our sampling of participants is random and that ticket inspections are orthogonal to intrinsic honesty.

Regarding the first assumption and the first dimension (fraudsters *vs*. non-fraudsters), there is no reason to believe that the order in which passengers board a public vehicle correlates with their intrinsic honesty. By focusing on the first four or five passengers boarding we could randomly observe fraudsters or non-fraudsters. But one could question whether the rule pertaining to targeting the first pre-identified passenger getting off the vehicle generated a lack of randomization by focusing on short trips. In fact, we do not necessarily over-represent short trips. According to Egu and Bonnel (2020), passengers in Lyon change between 1.30 and 1.45 vehicles per trip. So, when a passenger enters a new vehicle it may be the final leg of a longer trip. Moreover, while the motivation to defraud may be different if one considers a long, rather than short trip, it is not clear how this would affect people's reaction to a ticket inspection in terms of acceptance of the banknote and how long a trip need be to observe a different response.

Regarding the other dimension (inspected *vs*. non-inspected passengers), imposing a ratio between observations in the I and NI conditions in the data collection, as explained above, circumvented collecting all the NI observations immediately, and the I observations later, which could have raised selection issues. Moreover, to verify that our randomization strategy worked, we checked with the transport company the consistency between the frequency of inspections observed in our data and those reflected in the inspection plans of the company. Figure A6 in Appendix 4 depicts the distribution of these inspections over time; the inverted U-shape pattern is analogous to that observed in our field data (see Figure A5). A regression analysis of the occurrence of a planned inspection by the company at the time of our experiment is reported in Table A4 in Appendix 5. Overall, consistency was high, suggesting that there was no bias in the method of our collecting the I condition observations.

The second identification assumption (orthogonality of ticket inspections to intrinsic honesty) would automatically hold if ticket inspections were completely random across lines and times of day. However, the randomization of inspections is not perfect since they result from the company policy. While this, in itself, does not pose a threat to our identification strategy as long as ticket inspections are not systematically conducted in areas where or at times when intrinsic honesty is particularly low, it might be a source of concern were there an asymmetric selection in the samples

of inspected and non-inspected people. A first important point is that ticket inspections are organized by the company such that they are difficult to predict by passengers, a major source of randomness. As explained by the company, inspections are largely random in order to maintain uncertainty and prevent fare-dodgers learning where and when inspections could happen. Some randomness also results from the fact that if inspection teams receive an inspection plan at the beginning of their shift, they often change that plan to adapt to the occurrence of incidents. Indeed, they are also in charge of the security of the system: in the case of an incident on a line, they may switch tasks and lines independently of the initial plan.

Table A3 in Appendix 5 shows that ticket inspections were more frequent in the Center Metropolitan Lyon, on certain lines and in the early afternoon. Therefore, we include, in the parametric models reported in the results section, fixed-effect variables (time of day, geolocation and transport line) that might be a source of selection for inspections, as detailed in the next section. To assess the robustness of our results, a series of additional steps is then implemented. In Tables A9 and A10 of Appendix 5, we consider additional regression models where we include finer control which takes into account the environmental variables as well as interactions between each transport line and the time of day, and interactions between the geolocation and the time of day. In Appendix 1.4 we also report a propensity score matching analysis to account for possible selection on observables. Indeed, if ticket inspectors act in a non-random way, their selection criteria should be mainly based on observable characteristics of a transport (a specific area, line or time), which limits the risk of selection on unobservables. They are unlikely to base selection on observable individual characteristics, since all passengers are checked in the case of an inspection. For example, we tested for correlation between the size of the inspection teams and the individual characteristics of the subjects in terms of gender, age, and ethnicity and found none (available upon request). The results of our robustness checks confirm our main analysis, suggesting that differences in timing and location of the controls are unlikely to be responsible for our results.

3. Results

Our first result shows that not validating the ticket on public transport is associated with a lower intrinsic honesty.

In the absence of ticket inspection in the first stage, passengers without a validated ticket or pass (*i.e.*, fraudsters) were more likely than passengers with a validated ticket or pass (*i.e.*, non-fraudsters) to claim ownership of the banknote on the street in the second stage. Figure 1 displays the percentage of fraudsters and non-fraudsters who took the banknote, depending on whether a ticket inspection occurred (I) or not (NI). In NI, 52.86% of the fraudsters took the banknote compared to 32.14% of the non-fraudsters. This difference is significant (Chi-squared test: $\chi^2(1) = 12.29$, $p < 0.001$),[11] revealing that the disparity in ethical behavior correlates across the two contexts. The observed pattern of cross-context unethical behavior was not affected by the presence of an observer in the second stage (see Figure A2 in Appendix 4). Although fraudsters were slightly less likely to take the banknote when being observed by a third person (45.45% in NI-A *vs*. 52.86% in NI), the difference is not significant (χ^2 (1) = 1.35, p = 0.245). Similarly, there was little difference in the percentage of non-fraudsters who took the banknote in NI-A and NI (33.33% *vs*. 32.14% in NI; χ^2 (1) = 0.04, p = 0.841).

Next, we investigated the effect of the enforcement of the deterrence institution in the first stage of the quasi-experiment on the intrinsic honesty of fraudsters in the second stage (*i.e.*, behavior on the street). This spillover effect was negative: the percentage of fraudsters who took the banknote increased significantly from 52.86% to 67% after an inspection ($\chi^2(1) = 4.81$, p = 0.028). This reveals that inspections and sanctions had no immediate educative effect on the intrinsic honesty of fraudsters.

[11] All the reported non-parametric statistics are two-tailed and take each individual as one independent observation.

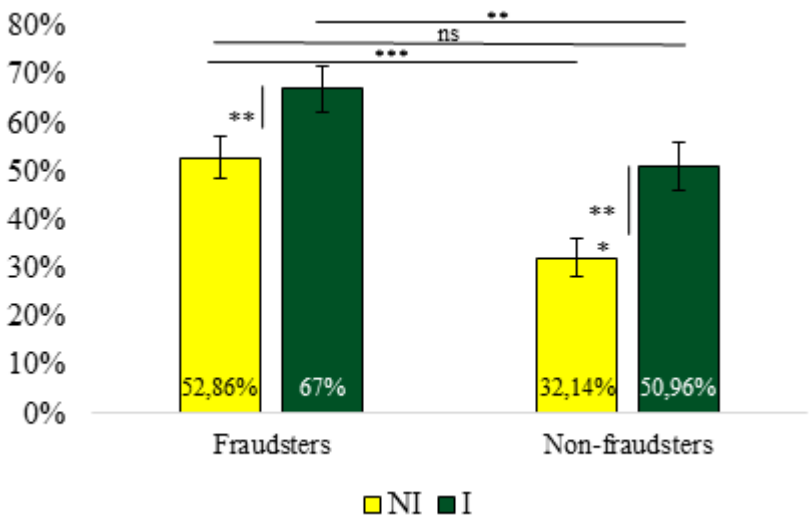

Figure 1. Percentage of fraudsters and non-fraudsters accepting the banknote in the conditions with (I) and without (NI) ticket inspection.

Notes: The light bars are for the NI condition and the dark bars for the I condition. N = 140 (NI, fraudsters), 100 (I, fraudsters), 140 (NI, non-fraudsters), and 104 (I, non-fraudsters). Error bars, mean ± SEM. Significance levels: *** $p < 0.01$, ** $p < 0.05$, *ns* not significant, Chi-squared tests.

Fraudsters caught travelling irregularly had to pay a fine. Hence, the mechanism behind this negative spillover could be that fraudsters try to partially recover the loss incurred by the fine (Sharma *et al.,* 2013). However, if loss recovery solely explained the spillover, we should observe no spillover effect for non-fraudsters. Strikingly, the percentage of non-fraudsters accepting the banknote increased from 32.14% to 50.96% after a ticket inspection ($\chi^2(1) = 8.79$, $p = 0.003$). Thus, the enforcement of the deterrence institution also reduced the intrinsic honesty of law-abiding passengers. The percentage of passengers who took the banknote following an inspection was still significantly higher for fraudsters than for non-fraudsters ($\chi^2(1) = 5.41$, $p = 0.020$), but between non-fraudsters after an inspection and non-inspected fraudsters it was no longer different, even without an audience ($\chi^2(1) = 0.08$, $p = 0.769$).

We now turn to a regression analysis to control for the environment and for the individuals' socio-demographic characteristics. The coefficients from four linear probability regressions in which the dependent binary variable is the decision to take or not the banknote are reported in Table 1.[12] In Model (1), the effect of inspections is investigated: the independent variable

[12] Since the estimated coefficients on interactions in ordered models are difficult to interpret with standard marginal effects (Ai and Norton, 2003), we run the whole analysis with linear probability models.

Inspection is equal to 1 when a control occurred on the bus/tram and is equal to 0 otherwise; additionally, *Audience* takes value 1 when the corresponding experimental condition applies and 0 otherwise. Finally, *Fraudster* identifies a dummy variable that is equal to 1 when the individual does not hold a validated ticket or pass and 0 otherwise. In Model (2), we add the *Inspection*Fraudster* and *Audience*Fraudster* variables, representing interaction terms between *Fraudster* and the *Inspection* condition.

In Model (3), the following control variables are also included in the analysis. First, based on the actors' evaluation given by the experimental subjects in the laboratory during the casting phase (see footnote 5), we categorize actors and actresses depending on their relative score (high or low) and include them as dummies in the regression, with the high score actress taken as the baseline category. Second, with respect to passengers' socio-demographics, we control for apparent age (coded as a continuous variable), and we include dummy variables for gender, ethnicity (identifying Caucasian – the baseline group –, Arab, African, Asian, or any other ethnic group), apparent wealth (poor – the baseline group –, average, and rich), and whether religious signifiers were visible or not. We also controlled for environmental conditions. The geolocation is captured through three dummy variables (Center Metropolitan Lyon – the baseline category –, North-East Metropolitan Lyon, and South-East Metropolitan Lyon). Fixed effects are added for the transport line and for the time of day. More specifically, we included a set of dummy variables for each of the main tram lines in our sample (T1, T2 and T4), with the remaining, minor lines representing the omitted reference category, as well as dummy variables for each time interval (morning from 9:00AM to 11:59AM – the baseline category –, early afternoon from 12:00PM to 2:59PM, and late afternoon from 3:00PM to 6:15PM).[13] We also included dummies for whether the public vehicle was crowded and for the noticeability of the scene on the street. Weather was coded as a set of binary variables (sunny – the baseline category –, cloudy, and rainy). Finally, Model (4) is similar to Model (3), except that the subject's gender dummy and the actor dummies are replaced with three indicator variables capturing the gender composition of the actor-passenger pair (with female pairs as the baseline category). This aimed to test whether passengers reacted differently with someone more similar to them.

[13] The information about the time of day, the geolocation and the transport line was grouped into categories to preserve the information without over-parameterizing the model. This avoided singleton dummies and too-sparse data in certain categories, and allowed a reasonable amount of variation among our key variables within each category.

Model (1) shows that the average probability of accepting the banknote increases by 16 percentage points after an Inspection and by 17 percentage points when not holding a ticket. In Model (2), the positive coefficient of the variable Inspection indicates that, when the passenger holds a ticket, being inspected by controllers significantly increases the probability of accepting the banknote. Additionally, the interaction term between Inspection and Fraudster in Model (2) was not significant ($p = 0.602$), confirming that ticket inspections increased the unethical behavior of both fraudsters and non-fraudsters on the street and thus, loss recovery cannot be the only explanation of these cross-context spillover effects. Models (3) and (4) confirmed these findings with very minor changes in the coefficients of the variables of interest. This was after controlling notably for time of day, geolocation and public transport line, which had no significant effect on the likelihood of taking the banknote.[14] A few socio-demographic matters were noted: older subjects were more likely to take the banknote (possibly driven by a selection effect as, on average, wealthier older people use public transport less) while people of poorer appearance were also significantly more likely to violate the norm. Finally, individuals with visible religious signs exhibited a lower propensity to behave unethically in the second stage of our quasi-experiment.

[14] As already mentioned, 30% of the observations were collected in the presence of an experimenter who was not blind to the research questions. Controlling directly for this presence in the regression analysis does not change the results (see Table A8 in Appendix 5).

Table 1. Determinants of the decision to take the banknote.

Dependent variable: Decision to take the banknote	Model (1) Coeff.	se	Model (2) Coeff.		Model (3) Coeff.	se	Model (4) Coeff.	se
Inspection (baseline = No Inspection)	0.165***	0.045	0.188***	0.063	0.150**	0.065	0.159**	0.066
Audience (baseline = no audience)	-0.031	0.043	0.012	0.059	0.051	0.061	0.034	0.061
Fraudster (baseline = no fraudster)	0.166***	0.036	0.207***	0.058	0.177***	0.057	0.171***	0.058
Inspection*Fraudster			-0.047	0.089	-0.012	0.087	-0.021	0.089
Audience*Fraudster			-0.086	0.087	-0.095	0.086	-0.074	0.087
Constant	0.342***	0.034	0.321***	0.040	0.431***	0.122	0.545***	0.120
Actors/Actress (baseline = Higher−score actress)								
Lower-score actress					0.189***	0.049		
Higher-score actor					-0.016	0.071		
Lower-score actor					0.073	0.051		
Gender interaction (baseline = Female actress, Female passenger)								
Female actress, Male passenger							0.029	0.046
Male actor, Female passenger							-0.049	0.057
Male actor, Male passenger							-0.027	0.050
Male passenger (baseline = female passenger)					0.023	0.037		
Age [e]					0.028*	0.014	0.242*	0.014
Time of day (baseline = 9:00AM – 11:59AM)								
12:00PM – 2:59PM					-0.070	0.044	-0.059	0.045
3:00PM – 6:15PM					-0.074	0.048	-0.056	0.048
Geolocation (baseline = Center Metropolitan Lyon)								
North-East Metropolitan Lyon					-0.030	0.047	-0.027	0.048
South-East Metropolitan Lyon					-0.070	0.051	-0.056	0.051
Line public Transport (baseline = other)								
T1					0.028	0.071	0.034	0.073
T2					-0.044	0.085	-0.094	0.085
T4					0.010	0.087	0.031	0.089
Ethnicity (baseline = Caucasian) [e]								
Arab					0.062	0.055	0.038	0.054
African					0.063	0.051	0.051	0.051
Asian					0.079	0.117	0.078	0.125
Other					0.090	0.126	0.102	0.126
Social appearance (baseline = poor) [e]								
Average					-0.204***	0.050	-0.199***	0.050
Rich					-0.254***	0.076	-0.255***	0.078
Religious signs (baseline = no religious signs)					-.0.227**	0.110	-0.215**	0.107
Crowded (baseline = No Crowded)					-0.048	0.038	-0.061	0.038
Weather (baseline = sunny)								
Cloudy					-0.017	0.052	-0.041	0.052
Rainy					-0.086	0.080	-0.151	0.082
Someone could notice the scene (baseline = no one)					-0.041	0.039	-0.045	0.039
Obs	708		708		708		708	
R2	0.0546		0.0559		0.1483		0.1277	
Prob > F	0.0000		0.0000		0.0000		0.0000	

Notes: Table 1 reports the coefficients from linear probability estimates as well as robust Standard Errors. [e] estimated by the research assistant. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$ (Wald tests).

To further check the robustness of our results, we replicated the analysis provided in Model (3) and included either interaction terms between the time of day and each transport line category, or between geolocation and time of day (see Model (1) and Model (2), respectively, in Table A9 in Appendix 5). In Table A10, we instead included hourly dummies for the time of day, a dummy for each minor line, and local district fixed effects for geolocation. The magnitude and significance of the coefficients of interest remained almost the same as in Table 1. Furthermore, in Appendix 1.4, we report the analysis based on propensity score matching, where we matched controlled and not controlled passengers on some key observable environmental variables. This undertook to account for the fact that ticket controls might not be entirely random but rather might vary depending on the time of day, geolocation and public transport line. The matching procedure allowed us to restrict the analysis to observations for which the controlled and not controlled subjects were more similar. The results of this analysis (reported in Tables A11 and A12 in Appendix 5) confirmed the results reported in Table 1, suggesting that differences in timing and location of the inspections were unlikely to be responsible for our results.[15]

Overall, this analysis confirms that the enforcement of the deterrence institution reduces the intrinsic honesty not only of fare dodgers, but also of law-abiding passengers. To dig deeper into the mechanisms that could explain this result, we investigated whether the strength of the deterrence institution matters beyond the occurrence of a ticket inspection. This strength can be proxied by the size of the inspection teams.[16] In Table 2, we focus on the Inspection condition only, and isolate the impact of the number of ticket inspectors during a control on the decision to take the banknote, estimating a set of linear probability models. In Model (1), the independent variables include one that indicates whether the subject is a fraudster, and a set of dummies for the inspection team size grouped into three categories (2-5 – the baseline category –, 6-10, and 11 to 20). We also controlled for whether the inspection occurred at the bus stop or on board the public vehicle. In Model (2), we add interaction terms between the set of dummies identifying the number of

[15] Time of day, geolocation and public transport line may only be imperfect proxies of how ticket inspections are conducted in the field by the transport company. Selection on unobservables may still be present. Using the method developed by Oster (2019), we assessed whether unobserved characteristics that drive ticket controls could bias our estimates. If anything, we found that our results underestimate the true effect of ticket controls on the probability of accepting the 5-Euro banknote.

[16] Figure A3 in Appendix 4 shows a histogram with the distribution of the number of ticket inspectors per inspection in our quasi-experiment.

inspectors and the Fraudster variable. Finally, in Model (3) we add the same individual and environmental controls, fixed effects for time of day, geolocation, and transport line as in Table 1.

Model (1) shows that the probability of taking the banknote is sensitive to the size of the inspection team, since the coefficients associated with a team of 6 to 10 inspectors and with a team larger than 10 inspectors are both positive and significant. In Models (2) and (3), the two coefficients are also significant, suggesting that non-fraudsters, the reference category, were more likely to accept the banknote when inspected by a medium-sized or a large team of controllers rather than a small one. Moreover, the coefficient associated with a team of more than 10 inspectors is larger than the coefficient associated with a team of 6 to 10 inspectors, suggesting that non-fraudsters were more likely to accept the banknote as the team of inspectors increased in size (p = 0.028 in Model (2), p = 0.004 in Model (3)). It is interesting to note that in both Models (2) and (3), the coefficient of the interaction term between being a fraudster and being inspected by a medium-sized team is not significant, indicating no difference between fraudsters and non-fraudsters with respect to the impact of a team with 2 to 5 inspectors *vs.* a team with 6 to 10 inspectors. In contrast, the interaction term between being a fraudster and being inspected by a large size team is significant and negative. This indicates that the non-fraudsters had the greater reaction to an increase in the size of the inspection teams. This finding can help to discern the various possible mechanisms triggering the spillover effects of the deterrence institution on intrinsic honesty, which we discuss in the next section. Finally, being controlled on board, rather than at the bus/tram stop, significantly increased the probability of passengers' accepting the banknote.

This analysis supports our third result: the size of the inspection team increases the spillover effects of the deterrence institution on intrinsic honesty, especially for non-fraudsters.

Table 2. Effect of the Number of Ticket Inspectors on the Decision to Take the Banknote

Dependent variable: Decision to take the banknote	Model (1) Coeff.	se	Model (2) Coeff.	se	Model (3) Coeff.	se
Fraudster (baseline = no fraudster)	0.188***	0.071	0.413***	0.123	0.509***	0.114
Number of controllers (baseline = 1 - 5)						
6 – 10	0.466**	0.211	0.669***	0.268	0.688***	0.214
11 – 20	0.581***	0.215	0.918***	0.267	1.090***	0.247
*Fraudster * Number of controllers*						
Fraudster * 6 - 10 Controllers			-0.210	0.162	-0.226	0.155
Fraudster *11 - 20 Controllers			-0.542***	0.190	-0.612***	0.219
Control on board (baseline = control on the bus)	0.384*	0.214	0.450*	0.233	0.658***	0.204
Constant	0.032	0.209	-0.177	0.252	-0.691**	0.277
Actors/Actress (baseline = Higher−score actress)						
Lower-score actress					0.305***	0.093
Higher-score actor					-0.001	0.192
Lower-score actor					0.047	0.097
Additional controls for passengers (gender, age, ethnicity, social appearance, religious signs)	No		No		Yes	
Additional controls for the environment (weather, audience in the vehicle, audience on the street)	No		No		Yes	
Time of day dummies	No		No		Yes	
Geolocation dummies	No		No		Yes	
Line Public Transport dummies	No		No		Yes	
Obs	199[§]		199[§]		199[§]	
R2	0.0600		0.0973		0.3386	
Prob > F	0.0144		0.0019		0.0000	

Notes: Table 2 reports the coefficients and robust standard errors from linear probability estimates on the Inspection condition. [§] Five observations were excluded because the information about the number of inspectors was missing. Significance levels: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$ (Wald tests).

4. Discussion

Our results provide strong evidence for cross-context spillover effects of inspections and sanctions on intrinsic honesty. Strikingly, these effects equally applied to fraudsters and to non-fraudsters. In what follows, we discuss which mechanisms could explain these spillover effects. We begin with

the mechanisms that receive less support from our data and move towards those more consistent with our findings.

Indirect reciprocity - Negative direct reciprocity against the authority that signals distrust by enforcing inspections is ruled out by design, since behavior on the street cannot affect the transport company. However, people may still want to harm a stranger because of indirect reciprocity (Nowak and Sigmund, 2005). While we cannot exclude this possibility, it seems unlikely for a number of reasons. First, it is unclear why an inspected passenger would like to exploit a 'kind' third party who has just offered them money. Second, indirect reciprocity often arises for strategic motives (*e.g.*, Engelman and Fischbacher, 2009; Stanca, 2009) that are absent in our setting.

Loss recovery - Fraudsters' willingness to recover a loss after being fined might conceivably explain their subsequent unethical conduct. However, it cannot explain the negative effect of inspections on non-fraudsters' intrinsic honesty across contexts. Moreover, we found no difference in the banknote acceptance rate between fraudsters who paid their fine on the spot (N = 30/41 = 73.17%) and those who did not (N = 33/52 = 63.46%; Chi2 test, $\chi^2(1) = 0.989$, p=0.320). Since almost two thirds of the fines that are not paid immediately are never recovered by the company, we know that a significant proportion of those who did not pay their fine on the spot will not actually suffer a monetary loss. Thus, loss recovery is unlikely the main mechanism behind our findings.

Signaling - An alternative mechanism could be that ticket inspections prompt people to update their belief about the intrinsic cost of honesty. For example, in their self-signaling model, Benabou and Tirole (2003) assume that people have an imperfect self-knowledge. In our context, this may be an imperfect knowledge of their intrinsic honesty (how intrinsically costly it is to not validate the ticket). A passenger might interpret a ticket inspection as a signal that their intrinsic honesty is low ("I am inspected because I am suspected of being dishonest"), and revise their beliefs about the intrinsic cost to them. This might, in turn, affect a subsequent moral decision, inducing the person to accept the banknote more often in the following context. Benabou and Tirole's (2003) theory hinges on the assumption that the principal (the public transport company in our setting) possesses certain relevant information regarding the unknown characteristic of the agent or task at hand. However, this is unlikely in our setting (inspectors target everyone in a vehicle, including law-

abiding passengers). Moreover, it is difficult to believe that fraudsters are not aware of their moral type.

A related signaling interpretation is that the inspection makes salient to a proportion of non-fraudsters that they paid for their ticket not because they are using a public service, but because they otherwise fear being fined. The inspection may signal their mainly extrinsic motivation, whose higher saliency might make them more willing to take the banknote in the new context where money can be earned unethically, but without the risk of sanction. However, it is unclear how and why this signaling mechanism would depend on the size of the inspection team, given our finding that the spillover effects on non-fraudsters were particularly reactive to the number of inspectors.

Moral balancing - A psychological explanation in terms of moral licensing (Nissan, 1991) could apply to non-fraudsters if, after a ticket inspection that reinforced their positive self-image, they loosen their moral standards while maintaining their self-concept of honesty (Benabou and Tirole, 2006; Shalvi *et al.*, 2011). Symmetrically for fraudsters, paying a fine may lead to moral cleansing if the sanction reduces the dissonance between the individual's self-image and his desired moral self. However, as already noted, in acceptance of the banknote, there was no difference between the fraudsters who paid their fine on the spot – leading to possible moral cleansing – and those who did not. And again, there is no obvious link between such moral balancing strategies and the sensitivity to the size of the inspection team.

Emotions - Inspections might trigger negative emotions in both fraudsters and non-fraudsters; the former may feel anger or shame at being fined and the latter, sadness due to the experience of distrust, or anger when they observe rule violations. Cross-context spillovers might then emerge as a consequence of the passenger's mood that leads to their punishing whoever can be associated with the transport company (*e.g.*, another passenger) or society in general (see, in other contexts, *e.g.*, Card and Dahl, 2011; Munyo and Rossi, 2013). To investigate whether emotions arising during an inspection might affect passengers' subsequent misbehavior, we explored the impact of a visible emotional reaction (crying, screaming) expressed during an inspection. This analysis revealed that the banknote acceptance rate among detected fraudsters did not depend on their expressing a strong emotion (N=11/16, 68.75%) or not (N=50/74, 67.57%) after being fined ($\chi^2(1) = 0.01$, $p = 0.927$). However, people may feel angry or sad without any overt expression.

Therefore, we explored whether ticket inspections made non-fraudsters susceptible to emotional responses by means of a new study. Indeed, this would be a precondition for emotions to be the general mechanism behind spillovers. Several weeks after the main experiment, we conducted a survey of 160 passengers who validated a ticket or a pass on public transport in Lyon, either after a ticket inspection (51 subjects) or without a ticket inspection (109 subjects). Following the same identification procedure as in our quasi-experiment, we asked passengers to self-report their feelings of happiness and nervousness using Self-Assessment Manikins (SAM) (Lang, 1980) (see Figure A9 in Appendix 4). Self-reported happiness and nervousness after a ticket inspection (N = 51, mean = 3.88, S.D. = 1.16 and mean = 2.24, S.D. = 1.36, respectively, on a scale from 1 to 5) and when no inspection occurred (N = 109, mean = 4.14, S.D. = 0.92 and mean = 1.93, S.D. = 1.08, respectively) revealed that the inspected non-fraudsters were less happy and more nervous compared to uninspected ones. However, while reported emotions for the non-inspected tended to be closer to the mean than were those for the inspected passengers, the observed difference between these two groups was not statistically significant (Mann-Whitney test, z = 1.13, p = 0.257 for happiness; z = −1.149, p = 0.251 for nervousness).

Therefore, our data provides little evidence that emotions triggered the spillover effects observed in our main experiment.[17] We acknowledge, however, the limitation that the sample size of the inspected non-fraudsters in the survey is smaller, and the confidence band larger than that of the non-inspected non-fraudsters. Moreover, measurement errors could have played a role since we had only one question to measure each of the two emotions. Also, there was insufficient statistical power to detect any effect on emotions of the size of inspection teams. The occurrence of an inspection may also have generated a selection bias in the willingness to participate in our survey. Therefore, we cannot reject the role of emotions, especially for fraudsters, and suggest that more systematic investigations of this mechanism should be conducted.

Social norms - Finally, the fact that individuals, and in particular non-fraudsters, reacted to the size of the inspection team, suggests that the spillover effect might be driven not so much by the

[17] Note that even when emotions are measured by physiological responses, there is no consensus in the (limited) experimental literature on the relationship between emotions and unethical behavior. In a cheating game with no risk of detection, Pittarello *et al*. (2018) found a correlation between a higher emotional arousal and a lower likelihood of cheating. In contrast, in a tax evasion game where fraud could be detected and fined, Coricelli *et al*. (2010) showed that cheaters tend to be more emotionally aroused than non-cheaters, both at the time of deciding whether to evade and in reaction to an audit; compliers were not more emotionally aroused after an audit than when not audited.

inspection itself, but rather by the information that is conveyed when many people are observed being inspected. The visibility of inspections may affect people's perception of the injunctive norm (what one ought to do or not do) or the descriptive norm (what most people do), decreasing one's intrinsic honesty as a consequence (Gino *et al*., 2009). We investigated the first hypothesis by means of a second laboratory experiment ("Laboratory Experiment 2"). In this experiment, we elicited the injunctive social norm following the same procedure as in Experiment 1, but *after* new subjects (N = 96) had played a simplified version of the public transport game of Dai *et al.* (2017) and received feedback on their payoff in this game (see details in Appendix 2 and instructions in Appendix 3). In this incentivized game, subjects had to decide whether to purchase or not a ticket, being uninformed of the exact probability of a control (50%). The results reject that perceptions of the injunctive norm differ between inspected and non-inspected non-fraudsters in this game.[18]

Nevertheless, inspections might still inform people about the descriptive norm in the field (Sliwka, 2007; Dickinson *et al*., 2015). If ticket inspections signal the prevalence of rule violations, people may revise downwards their perception of society's descriptive norm after an inspection. This may particularly affect those who had initially higher beliefs about the honesty of citizens (thus, presumably more the non-fraudsters than the fraudsters) and as a result, it may weaken their own moral stance. This effect might be stronger if a larger inspection team is perceived as a signal of a higher prevalence of violations. Our results in the field are consistent with such a normative channel: non-fraudsters were significantly more likely to accept the banknote when they had been inspected by a medium-sized team of inspectors compared to a small team, and even more so when inspected by a team composed of more than 10 individuals than when inspected by a medium-sized team, whereas fraudsters reacted less to the size of the inspection team.[19] While we cannot

[18] Claiming ownership of a banknote found on the ground by oneself was considered as "somewhat or very socially appropriate" by 93.54% of the non-inspected non-fraudsters in the game and by 88% of the inspected non-fraudsters (Mann-Whitney tests, $p = 0.823$). Claiming ownership when the banknote has been found by another person was considered as "somewhat or very socially inappropriate" by 96.78% of the non-inspected non-fraudsters in the game and by 92% of the inspected non-fraudsters ($p = 0.816$). See Table A13 in Appendix 5.

[19] In the survey that we conducted several weeks after our experiment, we also elicited the beliefs of the respondents about the percentage of passengers travelling without a valid transport ticket/pass on the transport network in Lyon. We did not observe that a ticket inspection changes the beliefs of non-fraudsters in the field about the prevalence of fare evasion on public transport in Lyon: Non-inspected non-fraudsters: N = 109, mean belief about the percentage of fraudsters = 30.93%; inspected non-fraudsters: N = 50, mean belief = 29.28%. Mann-Whitney test, z = 0.441, p = 0.659. Insufficient variation in the data did not allow us to test an effect of the inspection team size. It is still possible that information about others' norm violations, conveyed by the number of inspectors, receives more attention only when it is relevant to the individual's goals than if it is not, so that the cross-context spillover spreads once people actually have the opportunity to behave unethically.

unambiguously isolate a unique explanation for the observed spillover effects, our evidence points in the direction that (the strength of) inspections might act as a proxy for the descriptive norm, spreading unethical behavior across contexts.

5. Conclusion

Modern societies have developed centralized institutions to protect citizens and assets against dishonesty. Since the honesty norms prevailing in the environment, *i.e.,* the frequency of violations, can compromise intrinsic honesty in a society (Gächter and Schulz, 2016), one might expect that these institutions contribute to the elevation of intrinsic honesty. However, solely focusing on the impact of such institutions in their context of application cannot isolate their pure effect on intrinsic honesty, since this is confounded by other factors such as material cost-benefit considerations (*e.g.*, avoiding a sanction) or direct reciprocity. By studying their effect outside their scope of application, our quasi-experiment reveals that the relationship between deterrence institutions and intrinsic honesty is more complex than might be expected.

Deterrence institutions create incentives to behave honestly to avoid a sanction but, at the same time, as our evidence has shown, may also effect a reduction in intrinsic honesty. Instead of observing an educative effect across contexts, we found that following a ticket inspection not only evaders, but also those who abided by the law behaved unethically in a setting where the institution does not apply. The enforcement device, when made visible to individuals, might act as a proxy of the (otherwise less salient) prevailing descriptive norm, spreading unethical behavior in contexts other than that directly targeted by the institution. Our results do not mean that such institutional enforcement is detrimental to compliance — our study is silent about the impact of ticket inspection on the willingness of passengers to pay for their next fares. Building on the contribution of Becker (1968), a huge theoretical and empirical literature has shown the positive effects of deterrence on compliance (see the review of Chalfin and McCrary, 2017). But our results point to the existence of a negative externality of this deterrence institution on intrinsic honesty, something that has largely been ignored both in the literature and by policy makers.

Teasing the mechanisms behind the negative spillovers from inspections on the level of intrinsic honesty of fraudsters and non-fraudsters requires additional investigation. Indeed, these

mechanisms are not necessarily the same for fraudsters and non-fraudsters and they may not be unique. In particular, it might be useful to elicit in a large-scale study how passengers' empirical norms about compliance varies with ticket inspections, with the size of inspection teams, and with the number of non-compliers fined. This would help identify how the deterrence institution affects the perceived social norm and whether spillovers vary with such normative views. Since we cannot refute the role of emotions, perhaps especially of fraudsters who are publicly exposed as cheaters in the case of an institutional control, it would be important to induce emotions exogenously to measure the extent to which their variations affected spillovers. Another extension could be to introduce rewards (such as symbolic thank-you cards or bonuses on loyalty cards) given by inspectors to compliers to determine whether this would reduce the spillover for this group, which could be anticipated if self-signaling plays a role in the spillover.

Teasing out these mechanisms would help to refine the policy implications raised by our study. The negative spillover of making a deterrence policy implementation visible to individuals suggests that crackdown interventions should be used with parsimony if there is a willingness to limit negative externalities. If large inspection teams signal a high fraud rate and contribute to weakening rather than strengthening the compliance norm, inspections conducted by small teams of inspectors in plain clothes might generate less spillover effects across contexts, at least for non-fraudsters. If a self-signaling mechanism plays a role, the negative reactions of incentive-sensitive compliers to inspections might be counteracted by the introduction of positive incentives associated with inspections, such as loyalty card-type bonuses or any such expressions of approval. More generally, our study invites policy-makers to adopt a broader view in evaluating the efficacy of an institution. A social welfare perspective requires ensuring that, in the aggregate, the positive effects of a deterrence institution are not cancelled out by spillovers into contexts beyond its direct target.

References

Ai, Chunrong, and Edward C. Norton 2003. "Interaction terms in logit and probit models." *Economics Letters*, 80(1):123-129.

Andenaes, Johannes. 1974. "*Punishment and Deterrence*." Michigan University Press.

Antonakis, John, Giovanna d'Adda, Roberto Weber, and Christian Zehnder. 2015. ""Just words? Just speeches?" On the economic value of charismatic leadership." *Working Paper Department of Organizational Behavior,* University of Lausanne.

Ariely, Dan. 2012. "*The honest truth about dishonesty. How we lie to everyone---especially ourselves*." New York: Harper Collins.

Axelrod, Robert. 1986. "An Evolutionary Approach to Norms." *American Political Science Review,* 80:1095-1011. DOI: https://doi.org/10.1017/S0003055400185016

Baldassarri, Delia, and Guy Grossman. 2011. "Centralized sanctioning and legitimate authority promote cooperation in humans." *Proceedings of the National Academy of Sciences, 108(27):* 11023-11027. DOI: 10.1073/pnas.1105456108

Becker, Gary S. 1968. "Crime and punishment: an economic approach." *Journal of Political Economy,* 76(2): 169-217. DOI: https://doi.org/10.1086/259394

Belot, Michele, and Marina Schröder. 2016. "The spillover effects of monitoring: A field experiment." *Management Science,* 62(1):37-45. http://dx.doi.org/10.1287/mnsc.2014.2089

Benabou, Roland, and Jean Tirole. 2003. "Intrinsic and extrinsic motivation." *Review of Economic Studies,* 70(3): 489-520. DOI: https://doi.org/10.1111/1467-937X.00253

Benabou, Roland, and Jean Tirole. 2006. "Incentives and prosocial behavior." *American Economic Review,* 96(5): 1652-1678. DOI: 10.1257/aer.96.5.1652

Card, David, and Gordon B. Dahl. 2011. "Family violence and football: The effect of unexpected emotional cues on violent behavior." *The Quarterly Journal of Economics*, 126 (1): 103-143. DOI: https://doi.org/10.1093/qje/qjr001

Cassar, Alessandra, Giovanna d'Adda, and Pauline Grosjean. 2014. "Institutional Quality, Culture, and Norms of Cooperation: Evidence from Behavioral Field Experiments." *Journal of Law and Economics,* 57(3): 821-863. DOI: http://dx.doi.org/10.1086/678331

Chalfin, Aaron, and Justin McCrary. 2017. Criminal Deterrence: A Review of the Literature. *Journal of Economic Literature*, 55(1): 5-48. DOI: https://doi.org/10.1257/jel.20141147

Cohn, Alain, Michel-André Maréchal, David Tannenbaum and Christian Lukas Zünd. 2019. "Civic honesty around the globe." *Science*, 70-73. DOI: 10.1126/science.aau8712

Coricelli, Giorgio, Mateus Joffily, Claude Montmarquette, and Marie-Claire Villeval. 2010. "Cheating, Emotions, and Rationality: An Experiment on Tax Evasion." *Experimental Economics*, 13, 226-247. DOI: 10.1007/s10683-010-9237-5.

Cour des Comptes. 2016. "La lutte contre la fraude dans les transports urbains en Île-de-France: un échec collectif. " *Rapport Public Annuel*, 537-577.

Dai, Zhixin, Fabio Galeotti, and Marie Claire Villeval. 2017. "The efficiency of crackdowns: a lab-in-the-field experiment in public transportations." *Theory and Decision,* 82(2): 249-271. DOI: https://doi.org/10.1007/s11238-016-9561-0

Dai, Zhixin, Fabio Galeotti, and Marie Claire Villeval. 2018. "Cheating in the lab predicts fraud in the field. An experiment in public transportations." *Management Science,* 64(3): 1081-1100. DOI: 10.1287/mnsc.2016.2616

Di Tella, Rafael, and Ernesto Schargrodsky. 2004. "Do police reduce crime? Estimates using the allocation of police forces after a terrorist attack." *American Economic Review,* 94(1): 115-133. DOI: 10.1257/000282804322970733

Dickinson, David L., E. Glenn Dutcher, and Cortney S. Rodet. 2015. "Observed punishment spillover effects: a laboratory investigation of behavior in a social dilemma." *Experimental Economics,* 18: 136-153. DOI: 10.1016/j.jpubeco.2010.11.021

Dickinson, David L., and Marie Claire Villeval. 2008. "Does monitoring decrease work effort? The complementarity between agency and crowding-out theories." *Games and Economic Behavior,* 63: 56-76. DOI: https://doi.org/10.1016/j.geb.2007.08.004

Diekmann, A., W. Przepiorka, and H. Rauhut. 2015. "Lifting the Veil of Ignorance: An Experiment on the Contagiousness of Norm Violations," Rationality and Society, 27: 309–333.

Effron, Daniel A., and Paul Conway. 2015. "When virtue leads to villainy: advances in research on moral self-licensing." *Current Opinion in Psychology,* 6: 32-35. DOI: https://doi.org/10.1016/j.copsyc.2015.03.017

Engelman, Dirk, and Urs Fichbacher. 2009. "Indirect reciprocity and strategic reputation building in an experimental helping game." *Games and Economic Behavior*, 67(2): 399-407. DOI: https://doi.org/10.1016/j.geb.2008.12.006

Engl, Florian, Arno Riedl, and Roberto A. Weber. 2017. "Spillover Effects of Institutions on Cooperative Behavior, Preferences, and Beliefs." *CESifo Working Paper No6504*, Munich (2017). Retrieved from http://www.cesifo-group.de/DocDL/cesifo1_wp6504.pdf on 3 August 2019.

Egu, Oskar, and Patrick Bonnel 2020. "Can we estimate accurately fare evasion without a survey? Results from a data comparison approach in Lyon using fare collection data, fare inspection data and counting data." *Public Transport*, 12(1): 1-26. DOI: https ://doi.org/10.1007/s12469-019-00224-x

Falk, Armin, and Michael Kosfeld. 2006. "The hidden costs of control." *American Economic Review,* 96(5): 1611-1630. DOI: 10.1257/aer.96.5.1611

Fehr, Ernst, and Simon Gächter. 2000. "Cooperation and punishment in public goods experiments." *American Economic Review,* 90(4): 980-994. DOI: 10.1257/aer.90.4.980

Frey, Bruno S., and Reto Jegen. 2001. "Motivation crowding theory." *Journal of Economic Surveys,* 15: 589-611. DOI: https://doi.org/10.1111/1467-6419.00150

Fischer, Peter, Tobias Greitemeyer, Fabian Pollozek, and Dieter Frey. 2006. "The unresponsive bystander: Are bystanders more responsive in dangerous emergencies?" *European Journal of Social Psychology,* 36: 267-278. DOI: https://doi.org/10.1002/ejsp.297

Fisman, Raymond, and Edward Miguel. 2007. "Corruption, norms, and legal enforcement: Evidence from diplomatic parking tickets." *Journal of Political Economy,* 115(6): 1020-1048. DOI: https://doi.org/10.1086/527495

Gächter, Simon, and Jonathan F. Schulz. 2016. "Intrinsic honesty and the prevalence of rule violations across societies." *Nature,* 531: 496-499. DOI: 10.1038/nature17160

Galbiati, Roberto, Emeric Henry, and Nicolas Jacquemet. 2018. "Dynamic effects of enforcement on cooperation." *Proceedings of the National Academy of Sciences,* 115(49): 12425-12428. DOI: https://doi.org/10.1016/j.jempfin.2018.07.008

Garrett, Neill, Stephanie Lazzaro, Dan Ariely, and Tal Sharot. 2016. "The brain adapts to dishonesty." *Nature Neuroscience* 19(12):1727-1732. DOI: 10.1038/nn.4426

Gino, Francesca, Shahar Ayal, and Dan Ariely. 2009. "Contagion and Differentiation in Unethical Behavior: The Effect of One Bad Apple on the Barrel." *Psychological Science,* 20(3): 393-398. DOI: https://doi.org/10.1111/j.1467-9280.2009.02306.x

Gintis, Herbert. 2003. "The hitchhiker's guide to altruism: Gene-culture coevolution and the internalization of norms." *Journal of Theorical Biology,* 220(4): 407-418. DOI: https://doi.org/10.1006/jtbi.2003.3104

Gneezy, Uri, Agne Kajackaite, and Joel Sobel. 2018. "Lying Aversion and the Size of the Lie." *American Economic Review,* 108(2): 419-453. DOI: https://doi.org/10.1257/aer.20161553

Gneezy, Uri, and Aldo Rustichini. 2000. "Pay enough or don't pay at all." *Quarterly Journal of Economics,* 115(3): 791-810. DOI: https://doi.org/10.1162/003355300554917

Hampton, Jean. 1984. "The moral education theory of punishment." *Philosophy and Public Affairs,* 13, 208-238.

Hawkins, Gordon. 1969. "Punishment and Deterrence: The Educative, Moralizing, and Habituative Effects." *Wisconsin Law Review* 550.

Keolis. 2014. "Fraude: comment lutter?" Available online at https://fr.slideshare.net/Keolis/keo-fvrier-2014 (accessed 23 September 2020).

Knack, Stephen, and Philip Keefer. 1997. "Does social capital have an economic payoff? A cross-country investigation." *Quarterly Journal of Economics,* 112(4): 1251-1288. DOI: https://doi.org/10.1162/003355300555475

Lang, Peter. 1980. "Behavioral treatment and bio-behavioral assessment: computer applications." in Sidowski J.B., J.H. Johnson, and T.A. Williams, eds. *Technology in mental health care delivery systems*. Norwood: Ablex, 119-l37.

Levitt, Steven D., and John A. List. 2007. "What do laboratory experiments measuring social preferences tell us about the real world?" *Journal of Economic Perspectives,* 21(2): 153-174. DOI: 10.1257/jep.21.2.153

List, John A. 2011. "Why economists should conduct field experiments and 14 tips for pulling one off." *Journal of Economic Perspectives, 25*(3): 3-16. DOI: 10.1257/jep.21.2.153

Masclet, David, Charles Noussair, Steven Tucker, and Marie Claire Villeval. 2003. "Monetary and non-monetary punishment in the voluntary contributions mechanism." *American Economic Review,* 93: 366-380. DOI: 10.1257/000282803321455359**Mauro, Paulo**. 1995. "Corruption and Growth." *Quarterly Journal of Economics,* 110(3): 681-712. DOI: https://doi.org/10.2307/2946696

Munyo, Ignacio and Martín A. Rossi. 2013. "Frustration, euphoria, and violent crime." *Journal of Economic Behavior & Organization*, 89: 136-142. DOI: https://doi.org/10.1016/j.jebo.

Nisan, Mordechai. 1991. "The moral balance model: Theory and research extending our understanding of moral choice and deviation." in Kurtine, W.M., and J.L. Gewirtz, eds. *Handbook of Moral Behavior and Development,* Lawrence Erlbaum Associates, Inc., 3-213.

Nowak, Martin A., and Karl Sigmund. 2005. "Evolution of Indirect Reciprocity." *Nature*, 437(7063): 1291-1298. DOI: 10.1038/nature04131

Oster, Emily. (2019). "Unobservable Selection and Coefficient Stability: Theory and Evidence." *Journal of Business & Economic Statistics,* 37(2): 187-204.

Peysakhovich, Alexander, and David G. Rand. 2016. "Habits of virtue: creating norms of cooperation and defection in the laboratory." *Management Science,* 62(3): 631-647. DOI: http://dx.doi.org/10.1287/mnsc.2015.2168

Pittarello, Andrea, Beatrice Conte, Marta Caserotti, Sara Scriminand Enrico Rubaltelli. 2018. "Emotional intelligence buffers the effect of physiological arousal on dishonesty." *Psychonomic Bulletin Review* 25: 440–446. DOI: https://doi.org/10.3758/s13423-017-1285-9

Rand, David G., Alexander Peysakhovich, Gordon T. Kraft-Todd, George E. Newman, Owen Wurzbacher, Martin A. Nowak, and Joshua D. Greene. 2014. "Social heuristics shape intuitive cooperation." *Nature Communications,* 5: 3677. DOI: 10.1038/ncomms4677

Sands, Melissa L. 2017. "Exposure to inequality affects support for redistribution." *Proceedings of the National Academy of Sciences,* 114(4): 663-668. DOI: 10.1073/pnas.1615010113.

Shalvi, Shaul, Jason Dana, Michel J.J. Handgraaf, and Carsten K.W. De Dreu. 2011. "Justified ethicality: observing desired counterfactuals modifies ethical perceptions and behavior." *Organizational Behavior and Human Decision Processes,* 115: 181-190. DOI:10.1016/j.obhdp.2011.02.001

Sharma, Eesha, Nina Mazar, Adam L. Alter, and Dan Ariely. 2014. "Financial deprivation selectively shifts moral standards and compromises moral decisions." *Organizational Behavior and Human Decision Processes,* 123(2): 90-100. DOI: http://dx.doi.org/10.1016/j.obhdp.2013.09.001

Sliwka, Dirk. 2007. "Trust as a signal of a social norm and the hidden costs of incentive schemes." *American Economic Review,* 97(3): 999-1012. DOI: 10.1257/aer.97.3.999

Stanca, Luca. 2009. "Measuring indirect reciprocity: Whose back do we scratch?" *Journal of Economic Psychology,* 30(2): 190-202. DOI: https://doi.org/10.1016/j.joep.2008.07.010

Swami, Viren, Flora Chan, Vivien Wong, Adrian Furnham, and Martin J. Tovée. 2008. "Weight-based discrimination in occupational hiring and helping behavior." *Journal of Applied Social Psychology,* 38: 968-981. DOI: https://doi.org/10.1111/j.1559-1816.2008.00334.x

Wang, Lu, Gregory B. Northcraft, and Gerben A. Van Kleef. 2012. "Beyond negotiated outcomes: The hidden costs of anger expression in dyadic negotiation". *Organizational Behavior and Human Decision Processes,* 119(1): 54-63. DOI: 10.1016/j.obhdp.2012.05.002

Weisburd, David, Laura A. Wyckoff, Justin Ready, John E. Eck, Joshua C. Hinkle, and Frank Gajewski. 2006. "Does Crime Just Move around the Corner? A Controlled Study of Spatial Displacement and Diffusion of Crime Control Benefits." *Criminology,* 44(3): 549–592. DOI: https://doi.org/10.1111/j.1745-9125.2006.00057.x

Welsh, David, Lisa D. Ordóñez, Deirdre G. Snyder, and Michael S. Christian. 2015. "The slippery slope: how small ethical transgressions pave the way for larger future transgressions." *Journal of Applied Psycholo*gy, 100: 114-127. DOI: 10.1037/a0036950

Wilson, James Q., and George L. Kelling. 1982. "Broken Windows: The Police and Neighborhood Safety." *Atlantic Monthly*, 249(3): 29-38.

Winter, Fabian, and Nan Zhan. 2018. "Social norm enforcement in ethnically diverse communities." *Proceedings of the National Academy of Sciences,* 115(11): 2722-2727. DOI: 10.1073/pnas.1718309115.

APPENDIX 1: QUASI-EXPERIMENT - ADDITIONAL INFORMATION ON THE FIELD SETTING, THE PROTOCOL AND ROBUSTNESS CHECKS

1.1. TICKET INSPECTIONS IN THE PUBLIC TRANSPORT NETWORK IN LYON

The experiment was run in the main tram and bus lines of Lyon and we informed the transport company about our experiment.

The public transport network in Lyon comprises 4 metro lines, 5 tramway lines, 2 funicular lines and over 130 bus lines. Tickets can be purchased from vending machines located at each tram stop and metro entrance, from dedicated agencies or from the bus driver on board the public vehicle. In 2017, when the experiment was run, a single ticket cost €1.80 at vending machines and €2 on board. The single ticket enables passengers to use the public transport network for an unlimited number of times and any distance during one hour. Passes require the client to buy a smartcard at a cost of €5 and then a pass which is held on the smartcard. In 2017, the monthly pass cost €63.20 and the yearly pass €60.10 per month, with a discount for people less than 21 years old and half of it being reimbursed by the employer, according to the labor law.

To avoid a fine, passengers must validate their ticket or pass every time they board a new public vehicle even if they have already validated it in a previous journey. The fine amounts to €60 if paid on the spot, while it increases to €80 (€110) if paid with a maximum delay of 7 days (2 months). If a person did not validate a pass, this is also considered as an infraction but the amount of the fine is reduced to €5.

The transport company conducts ticket inspections every day. Most inspectors wear official uniforms but some of them wear plain clothes. Ticket inspectors work in teams of different sizes (typically of 4) and they can be as many as 20 during a hot-spot inspection. Figure A3 in Appendix 4 shows a histogram with the distribution of the number of ticket inspectors per inspection that we encountered in our quasi-experiment. The teams can be on board, changing line and direction as they wish, or waiting for the public vehicle to arrive at a stop. In the first case, the inspection is conducted during the ride. In the second case, it is done at the stop. In both cases, everyone on board is controlled. The inspector scans the client's ticket or pass in a device. If a passenger is caught fare dodging, the controller issues a fine. The enforcement of a fine takes several minutes (sometimes more if the passenger tries to find excuses or confront the inspector). As a result, everyone is able to see a person who gets a fine. Ticket inspectors are paid a flat wage and receive no incentives for the number of fare-dodgers they catch or people they control.

According to the company, ticket inspections are mostly random and irregular in order to maintain uncertainty and prevent fare-dodgers to learn where inspections could be. The inspection plans change every day and are subject to unexpected changes within the day. Since inspectors usually have a large discretion regarding where to go within a predefined area, it is very difficult to localize them. We also checked whether there existed apps for smartphones able to signal the presence of inspectors in the public transport network. We did not find any that worked during the realization of our experiment. All this largely explains why we could not collect more than 21 observations per day on average.

Of course, ticket inspections might occur more frequently in certain lines or areas for logistical reasons (*e.g.,* accessibility of the zone, shift work organization, number of people using a line) (see Table A3 in Appendix 5). We can check for this by looking at the frequency of inspections and their geolocation observed in our quasi-experiment. We covered three main areas of Metropolitan Lyon: Center Metropolitan

Lyon, North-East Metropolitan Lyon and South-East Lyon. These are areas that can be easily accessed by metro and tram. We did not cover West Metropolitan Lyon since it is a hilly area with no metro or tram stops. For logistical reasons, we also did not visit far-away neighborhoods in the East or South of Lyon. Figure A4 in Appendix 4 plots the frequency of inspections on a map covering the area of Metropolitan Lyon which can be reached by tram or metro. Inspections were more frequent in Center Metropolitan Lyon (darkest blue shaded area) than North-East Metropolitan Lyon (medium blue shaded area; χ^2 test, $p = 0.029$) and South-East Metropolitan Lyon (light blue shaded area; $p = 0.002$). North-East and South-East Metropolitan Lyon present similar frequencies of inspections ($p = 0.294$). This is not surprising given that more lines pass through Center Metropolitan Lyon and it is visited by more people.

1.2. **SCRIPT GIVEN TO THE ACTORS ABOUT THE SCENE TO PLAY** (Translated from French)

The actor/actress is on the phone. He/she follows a passenger, indicated by the RA, at the exit of the bus/tram. After 20/30 meters, he/she catches up and pretends to pick up a 5-Euro banknote from the floor. The actor/actress calls the attention of the target passenger, interrupting for a moment his/her phone call, to ask if the person has lost the banknote. The banknote must be clearly visible: it must be immediately clear that the banknote has just been picked up and that it is a 5-Euro banknote. The sentence to say is: "Sir/Madam, did you lose this?", in a neutral tone.

The actor/actress pretends not to pay too much attention to the targeted passenger's response and remains focused on the phone call he/she is having. If the passenger responds affirmatively, the actor/actress gives the banknote to the person without showing any signs of surprise. If the passenger responds negatively, the actor/actress puts the banknote in his/her pocket. In both cases, after the interaction, the actor/actress resumes his telephone call.

In the event that the passenger interacts with the actor/actress, for example by asking him/her if he/she has seen the money falling from the passenger's pocket, the actor/actress must avoid initiating a conversation. In the example just given, the actor/actress will simply answer "No idea" or "I did not pay attention".

1.3. **SAMPLE SIZE AND POWER ANALYSIS**

To form reasonable predictions about the behavior of fraudsters and non-fraudsters in the control group (*i.e.*, NI condition), we built on the results of Dai *et al.* (2018) who ran an artefactual field experiment in public transport in Lyon using a similar subject pool as ours. Dai *et al*. (2018) estimated the proportion of dishonest individuals among fare-dodgers and non-fare-dodgers in a die-under-the-cup task. The estimated proportion of fully (partially) dishonest subjects was between 0% and 19% (41% and 60%) for non-fraudsters, and between 9% and 46% (37% and 74%) for fraudsters. Assuming similar proportions of full and partial liars in our quasi-experiment, and assuming that full (partial) liars accept the banknote all (half of) the times, we predicted between 46% and 64.5% (30% and 39.5%) of fare-dodgers (non-fare-dodgers) taking the banknote. Using the midpoints of these intervals and assuming a type-I error rate of $\alpha = 0.05$ and a power level of 0.8, we computed a sample size of 92 subjects per group (fraudsters and non-fraudsters), which we rounded to a more conservative 100. Sample sizes are computed for two-sample proportions tests.

1.4. **PROPENSITY SCORE MATCHING ANALYSIS**

As we explain in the main text, ticket controls are organized in such a way that they are difficult to predict by the users. This implies some randomness in their occurrence. This randomness may however not be perfect. Controls might be more frequent during certain times of the day or in certain areas or lines where intrinsic honesty is lower (or where it is logistically easier to conduct them). An inspection of our data reveals that observations on controlled individuals are more concentrated in the Center Metropolitan Lyon, in certain lines (e.g., tram T4 and buses lines), and in the early afternoon (see Table A3). This means that the observations on controlled and not controlled passengers (henceforth, treated and untreated units) are not balanced on these dimensions. This imbalance could bias our estimate of the effect of ticket controls on intrinsic honesty if these characteristics are also associated with the likelihood of accepting the banknote. To alleviate this problem, we include, in the parametric models reported in the main text, fixed-effect variables (time of day, geolocation and transport line) that are hypothesized to be associated with both the probability of being controlled and, possibly, accepting the banknote. As a further robustness check, in Table A9 and A10, we also consider additional models where we include interactions between each transport line and the time of day, and interactions between the geolocation and the time of day, as well as hourly dummies for the time of day, a dummy for each minor line, and local district fixed effects for geolocation. The results replicate those reported in the main text.

In this section, we consider an alternative approach based on propensity score matching to correct for the potential sample selection bias due to observable differences between treated and untreated units. The objective is to identify and compare treated and untreated units that are similar in terms of certain key environmental characteristics that are associated with the probability of being controlled. We match inspected and non-inspected passengers on some key observable environmental variables. The matching procedure allows us to restrict the analysis to observations for which the inspected and non-inspected subjects are more similar. We describe below how we implemented this approach.

To estimate the propensity score, we estimate a logit model with the dummy for being controlled on the left-hand side. We first consider a parsimonious specification by only including the covariates (related to the time of day, the geolocation and the transport line) that seem *a priori* important to explain the probability of being controlled and, possibly, to accept the banknote.[20] We then explore different specifications by iteratively adding variables to the specification, including interactive combinations of the linear terms. We follow the algorithm proposed by Imbens and Rubin (2015) to select the final set of predictors to be included in the propensity score model.[21] We then estimate this model and calculate the

[20] We group the information about the time of day, the geolocation and the transport line, respectively, into categories that preserve the information without over-parameterizing the model. This is to reduce the variance in the propensity scores and not to exacerbate the support problem (see Caliendo and Kopeinig, 2008). In particular, we create three dummies for the time of day (morning from 9AM to 11:59AM, early afternoon from 12PM to 2:59PM, and late afternoon from 3PM to 6:15PM), three dummies for the geolocation (Center Metropolitan Lyon, North-East Metropolitan Lyon, South-East Metropolitan Lyon), and four dummies for the transport line (T1, T2, T4 and a residual category for all the other lines). In the initial specification of the propensity score model, we included a dummy for early afternoon, one for Center Metropolitan Lyon, and two dummies for the line T4 and other minor lines, respectively.

[21] The procedure consists in conducting several logistic regressions, one for each new covariate, and retains the model with the highest likelihood ratio statistic (the null hypothesis is that the parameter estimate of the additional covariate is zero), if this statistic is greater than a pre-set constant. This is repeated for each new model until no statistic is greater than the pre-set constant. Following Imbens and Rubin (2005), we set the constant to 1 for linear terms and 2.71 for

predicted probabilities (*i.e.*, the propensity score) of being controlled. Figure A7 in Appendix 4 shows the degree of overlap in the propensity scores between treated and untreated units in our data.

We then match treated and untreated units on the propensity score, using different matching strategies, and estimate weighted regressions using the weights obtained from the different matching methods.[22] We first consider one-to-one single-nearest neighbor matching (NN) where each treated unit is matched with the closest untreated unit in terms of the propensity score. To increase the quality of the matching, we allow replacement so that an untreated unit can be used multiple times as a match.[23] Second, we consider a caliper matching (Caliper) where each treated unit is matched to the closest comparable untreated unit within a given caliper radius.[24] As a third method, we perform a radius matching (Radius) which uses all the untreated units within the caliper. The advantage of this method is that it minimizes the possibility of bad matches since it uses more or less untreated units depending on how many good matches are available within the caliper (Caliendo and Kopeinig, 2008). Finally, we employ both a Kernel matching (KM, hereafter) and a local linear matching (LLM, hereafter). They both use weighted averages of the untreated units to construct the counterfactual outcome of each treated unit (Caliendo and Kopeinig, 2008). KM includes only an intercept in the weighting function, while LLM also adds a linear term.[25] Both KM and LLM are more efficient than previous approaches since they use all the untreated units,[26] with the risk, however, of including bad matches. For each matching method, we also impose a common support by removing all treated units with a propensity score larger than the maximum or smaller than the minimum propensity score of the untreated units (Leuven and Sianesi, 2003).[27]

At the end of each matching procedure, we evaluate the quality of the matching by comparing how balanced the relevant variables are before and after the matching. In particular, for each covariate, we compute two-sample t-tests for equality of means between treated and untreated units (before and after the matching),[28] the standardized percentage bias (before and after the matching) and the achieved percentage reduction in the bias.[29] We also compute some overall measures of imbalance. In particular, we re-estimate the propensity score using the matched data, compute the corresponding pseudo-R^2, and perform a likelihood ratio test on the joint significance of all coefficients in the model. If the matching worked well, the pseudo-R^2 (a measure

interactive terms. As the interaction between the dummy for other lines and Center Metropolitan Lyon perfectly predict a control, we drop these observations from the propensity score estimation.

[22] All the different matching algorithms were implemented in Stata via the command psmatch2 by Leuven and Sianesi (2003).

[23] Since high propensity scores are mostly associated with treated units, the replacement option prevents bad matches between high and low propensity score units. It also avoids the problem that the estimates would depend on how observations are ordered for the matching (Caliendo and Kopeinig, 2008). In case of ties (multiple observations with the same propensity score), the algorithm matches all tied observations.

[24] Following Cochran and Rubin (1973) and Rosenbaum and Rubin (1985), we use a caliper width equal to 0.2 of $\sqrt{(s^2_{treat} + s^2_{untreat})/2}$ where s^2_{treat} and $s^2_{untreat}$ are the point estimates of the variance of the log odds for treated and untreated units, respectively. Such caliper size eliminates 99% of the bias in the observed covariates.

[25] For both KM and LLM, we impose a bandwidth of 0.06 (see Heckman, Ichimura, and Todd, 1997).

[26] With NN, Caliper and Radius, unmatched units are dropped.

[27] It turns out that all observations are on support. So, effectively, we do not drop any observations.

[28] The test is constructed by regressing the variable on the ticket control dummy (= 1 for treated units and 0 for untreated units). Before the matching, the regression is unweighted, while after matching, the regression is weighted using the matching weights.

[29] The standardized percentage bias measures the mean difference between the treated and untreated units and it is computed as a percentage of the average standard deviation (see Rosenbaum and Rubin, 1985).

of how well the relevant covariates explain the probability of being inspected) should be low and the likelihood ratio test should not reject the null hypothesis after matching. Finally, we compute the overall mean and median bias (from the distribution of the individual biases), and the Rubin's B (the standardized difference in the means of the propensity scores in the treated and untreated units) and the Rubin's R (the ratio of the variances of the propensity score in treated and untreated units) (see Rubin, 2001).[30]

Figure A8 in Appendix 4 graphs, for each matching method, the extent of covariate imbalance in terms of standardized percentage bias before and after matching using dot charts. Table A11 in Appendix 5 summarizes the main measures of overall imbalance. In general, the estimated propensity score balances very well the matched treated and untreated units. With NN, Caliper and LLM, all treated units are perfectly matched to their untreated counterparts, and we completely remove all bias. This is not surprising since there is a perfect overlap in propensity scores between treated and untreated units (see Figure A7). By allowing replacement, a treated unit can always find a match with the same propensity score. With Radius and Kernel, some small bias persists but it is drastically reduced compared to pre-matching.

Table A12 in Appendix 5 displays the results of weighted regressions based on the matching procedures described above. The resulting estimated effects are analogous to those reported in the paper, if not, a bit stronger. This suggests that differences in key observable characteristics between treated and untreated units are unlikely to be responsible for our results.

[30] A sample is sufficiently balanced if $B < 25$ and $0.5 < R < 2$ (Rubin, 2001).

APPENDIX 2: DETAILS OF LABORATORY EXPERIMENTS 1 AND 2

2.1. **LABORATORY EXPERIMENT 1**

Our Laboratory experiment 1 is divided in two parts.[31] In the first part we test whether one's false claim of ownership of a banknote found by someone else indeed violates an injunctive ethical norm. More specifically, in the first part of the laboratory experiment we employed the norm-elicitation procedure introduced by Krupka and Weber (2013) in order to identify the shared normative judgment about the actions undertaken in the second stage of our quasi-experiment. In the second part of our laboratory experiment we verified that the actors played the scene similarly across conditions.

The entire experiment was computerized using the z-Tree software (Fischbacher, 2007), and conducted at GATE-Lab, Lyon (France). We recruited 45 subjects (46.67% males, 55.56% students, mean age = 28.18, S.D. = 12.20) with the Hroot software (Bock *et al*., 2014).

Experimental Design

First Part: Social norm elicitation. In the first part of the experiment, subjects were presented with two scenarios, one at the time and in random order, and asked to evaluate, on a four-point scale and for each scenario, whether the action taken by a person A was "very socially inappropriate", "somewhat socially inappropriate", "somewhat socially appropriate" or "very socially appropriate". The incentives provided to the subjects were not to report their own normative view but to match the response of the majority of subjects participating in the same session. In one scenario, person A walks on the street, with no one around. She picks up a €5 banknote on the floor knowing that it does not belong to her. In the second scenario, two strangers (A and B) walk on the street, with no one around. Person B picks up a €5 banknote on the floor and calls Person A's attention, asking whether she has lost it. Person A takes the banknote knowing that it does not belong to her. In both scenarios, subjects were asked to judge person A's decision to take the banknote according to the majority. At the end of the experiment, one of the two scenarios was randomly drawn. For this scenario, if a subject's answer coincided with the answer given by the majority of all participants in the session, the subject earned €5.

Table A5 in Appendix 5 shows that 80% of the subjects believe that for the majority, claiming ownership of a banknote found on the ground by oneself is "somewhat or very socially appropriate". The mean score is positive and statistically different from zero ($p < 0.001$) and the modal response is "somewhat socially appropriate", with 67% of subjects agreeing on that response. In contrast, 100% of them believe that for the majority, taking the banknote when someone else has found it is "somewhat or very socially inappropriate". The modal response is "very socially inappropriate" with 63% of subjects agreeing on that response. The difference in scores between the two scenarios is highly significant (rank-sum test: $z = -4.84$, two-sided $p < 0.001$). We can thus safely conclude that the decision to take the €5 banknote is collectively perceived as socially inappropriate and thus considered as a violation of a social norm.[32]

[31] 30 subjects completed both parts; 15 subjects received only the second part. A translation of the instructions is provided in Appendix 3. On average, subjects who completed both parts earned €17.50 and those who completed only part 2 earned €14.07, including a €5 show-up fee.

[32] Similar results are obtained if we only consider the first scenario encountered by each participant and compare the scores of Scenarios 1 and 2 between subjects (Mann-Whitney test, $p < 0.001$). We can do that since each within-subject scenario was presented on a different computer screen and participants did not know about the content of the second scenario when they were responding to the first.

Second Part: Guessing task. In the second part of the laboratory experiment, we verified that the actors, who recorded their fake phone conversation during the second stage of the quasi-experiment, played the scene similarly across conditions. . In the laboratory we explained to the subjects the context where the recordings were made but we did not tell them anything about the first stage of the quasi - experiment. Then, the subjects were incentivized to guess whether or not the targeted passenger took the banknote after listening to 48 randomly selected audio files (12 for each main condition, I and NI, and each category of passengers, fraudsters and non-fraudsters).[33] At the end of the experiment, we drew 5 guesses for each subject and paid him or her €4 for each correct guess.

To determine whether the lab participants could predict the behavior of the subjects in the field, we constructed the following measure of the guessing ability (Belot and van de Ven, 2017):

$$A = F(T \mid T) - F(T \mid NT)$$

where $F(T \mid T)$ is the proportion of lab participants that guessed "took" when the field subject indeed took the banknote, while $F(T \mid NT)$ is the proportion of lab participants that guessed "took" when the field subject did not take the banknote. The advantage of this measure is that it is independent of the number of times the field subjects took the banknote. Depending on the value of A, we can make the following claims:

- If $A \leq 0$, the lab participants are not able to predict the behavior of the field subjects (*i.e*., the probability of guessing that the person took the banknote is independent of the actual behavior of the person) or are worse than chance.
- If $0 < A < 1$, the lab participants can to some extent predict the behavior of the field subjects.
- If $A = 1$, the lab participants can perfectly predict the behavior of the field subjects.

We found that lab subjects were not able to predict the behavior of the field subjects. Table A6 in Appendix 5 shows that in all conditions and groups of field subjects, our measure of the ability to predict field behavior is either not significantly different from zero or weakly negative (in the I condition for fraudsters), meaning that lab subjects were, if anything, worse than chance in predicting field behavior. We found no significant differences in the ability to predict field behavior across conditions and groups of subjects (Wilcoxon signed-rank tests on pairwise comparisons, $p > 0.1$ for all comparisons). The Logit regressions on the probability of guessing (correctly or not) that the person took the banknote presented in Table A7 in Appendix 5 confirms that subjects did not assign a different probability of taking the banknote across conditions and groups. We can therefore safely conclude that the actors played the scene similarly across conditions and with fraudsters and non-fraudsters.

2.2 **LABORATORY EXPERIMENT 2**

The negative spillover effect of the deterrence institution that we observe in our quasi-experiment in the field may be the result of a change in the perception of the injunctive norm. Both fraudsters and non-

[33] For each audio file, participants listened to the voice of the actor asking whether the targeted person has lost the €5 banknote but not the answer of the person. They were allowed to replay each audio file as many times as they wanted before reporting their guess. The test controls for the actors' tone of voice and actual words spoken, but not for body language since it was forbidden to film the scenes. We did not record the audio of the scenes in the first experimental sessions. Also, due to some technical or environmental problems (*e.g.,* the actor forgot to press the record button or the quality of the audio was too poor), we failed to record the audio of a few other scenes.

fraudsters may, after an inspection, revise downward their beliefs about what ought to be done when the €5 banknote is proposed. To test this conjecture, we conducted Lab Experiment 2 with 96 participants from our subject-pool at GATE-Lab, Lyon (France). Subjects were recruited via the online software Hroot (Bock *et al*., 2014). The experiment was programmed and conducted using z-Tree (Fischbacher, 2007). Subjects were mostly students (92.71%), 56.25% were males, and the average age was 21.82 (S.D. = 6.71).

Experimental Design. There were two parts in the experiment. In part 1, subjects played a simplified version of the public transport game (Dai *et al*., 2017). In this game, subjects had to make a risky choice which was described as the decision to buy or not a ticket for using a (fictional) bus, knowing that there was a risk of inspection. The ticket cost €1.8 (which was equivalent to the price of a ticket in Lyon when our quasi-experiment was run). Each subject was inspected with 50% probability (this was randomly determined by the computer and it was independent for each subject). Subjects were not told about the precise probability of inspection (they only knew that there could be one). If a subject was inspected, the computer informed the subject about the inspection and displayed pictures and a video of real ticket inspectors in action to increase the salience of the event. An inspected subject who did not buy the ticket had to pay €4.80 (a fine of €3 plus the price of the unpaid ticket). There were no financial consequences for those who did not buy the ticket and were not inspected. Those who bought the ticket paid €1.8 both in the event of an inspection or no inspection. Any loss was deducted from the show-up fee which was purposely increased to €10 to make sure that subjects did not earn less than a minimal participation fee of €5.20.

In part 2, we employed the same norm-elicitation task that we used in Laboratory Experiment 1 (see instructions in Appendix 3). Subjects earned on average €10.61.

Results are presented in Table A13.

APPENDIX 3: INSTRUCTIONS OF LABORATORY EXPERIMENTS

3.1. INSTRUCTIONS OF LABORATORY EXPERIMENT 1

Instructions are translated in English from French.

Hello. Thank you for participating in this study. Please turn off your mobile phone. It is forbidden to communicate with other participants for the duration of the session.
If you have any questions at any time, please press the red button on the side of your desk and an assistant will come to answer your questions in private.
The experience is divided into two parts. At the end of the session, you will receive your earnings from parts 1 and 2 as well as a show-up fee of €5. Your earnings will be paid to you privately in a separate room to maintain confidentiality.
Please press OK to see the rest of the instructions.
{OK}

Part 1

Your task

The following screens will describe two situations in which a person "A" makes a choice. After you read the description of the situation, we will ask you to evaluate the choice made by person A and to indicate whether this choice is "socially appropriate" and "consistent with moral or proper social behavior in society" or "socially inappropriate" and "inconsistent with moral or proper social behavior in society". By "socially appropriate", we mean a behavior considered correct and ethical by the majority of people.
For each of your responses, we would like you to answer as truthfully as possible, based on your opinions of what constitutes socially appropriate or socially inappropriate behavior. To enter your response, you will have to click on one of the following options.

Person A's choice	**Very socially inappropriate**	**Somewhat socially inappropriate**	**Somewhat socially appropriate**	**Very socially appropriate**
	O	O	O	O

{OK}

Your earnings

At the end of the experiment, we will randomly select one of the two situations. For the situation selected, we will determine which response was selected by the largest number of participants in this session. If you

give the same response as that most frequently given by the other participants, then you will receive an additional €5 which will be paid to you at the end of the session.
For instance, suppose that in the situation selected for the payment, your response had been "somewhat socially inappropriate", then you would receive €5 if this was the response selected by the largest number of participants in today's session.
If you have any questions, please press the red button on the side of your desk. Otherwise, press OK to start the task.

{OK}

Situation 1

Description: two persons (A and B) who do not know each other are walking on the street, with no one around. Person B walks behind person A. Person B picks up a €5 banknote from the ground and calls person A to ask if she has lost it. Person A takes the banknote knowing that it does not belong to her.
Please indicate whether you think that person A's choice is very socially inappropriate, somewhat socially inappropriate, somewhat socially appropriate or very socially appropriate. To indicate your answer, click one of the options below. Remember that if this question is selected for the payment you will earn €5 if your response is the same as the most common response given by the other participants in today's session.

Person A's choice	**Very Socially Inappropriate**	**Somewhat socially inappropriate**	**Somewhat socially appropriate**	**Very socially appropriate**
Take the banknote	O	O	O	O

{OK}

Situation 2

Description: One person (A) is walking on the street, with no one around. She picks up a €5 banknote from the ground knowing that it does not belong to her.
Please indicate whether you think that person A's choice is very socially inappropriate, somewhat socially inappropriate, somewhat socially appropriate or very socially appropriate. To indicate your answer, click one of the options below. Remember that if this question is selected for the payment you will earn €5 if your response is the same as the most common response given by the other participants in today's session.

Person A's choice	**Very Socially Inappropriate**	**Somewhat socially inappropriate**	**Somewhat socially appropriate**	**Very socially appropriate**
Take the banknote	O	O	O	O

{OK}

Part 2

Your task

Your task is to listen to 48 audio files. These audio files correspond to recordings made during a study conducted on the streets of Lyon in the following context.

For each recording, an actor or actress is on the phone. He/she is following a person on the street. After 20/30 meters, he/she catches up with the person and pretends to pick up a €5 banknote on the ground. The actor/actress calls the attention of the targeted person, holding for a moment his/her phone conversation, to ask if the person has lost the banknote. If the targeted person responds affirmatively, the actor/actress gives the banknote to the person. If the targeted person responds negatively, the actor/actress puts the banknote in his/her pocket. In both cases, the interaction between the person and the actor/actress stops, the actor/actress resumes his/her phone conversion and leaves.

Each recording was made by the actor/actress. The voice that you will hear is that of the actor/actress when he/she asks the targeted person if the €5 banknote belongs to him/her. The targeted person is not aware of the existence of the recording. You will not hear the answer of the targeted person. Some sentences may be different from each other but the context is always the same.

Your task is to guess, for each audio file, whether or not the targeted person took or not the €5 banknote.

You can replay each audio file multiple times before making your guess. An example of the situation is accessible by clicking here:
{VIDEO}
{OK}

Your earnings

At the end of the session, the program will select at random five audio files. You will be paid for your guesses in these five audio files. For each selected audio file:

- you will earn €4 if your prediction is correct (i.e., you have correctly guessed whether the person took or not the €5 banknote);
- you will earn €0 if your guess is incorrect.

These earnings will be added to your other earnings of the session.
If you have any questions, please press the red button on the side of your desk. Otherwise, press OK to start the task.
{OK}

Audio file 1 of 48
Please click on "Listen" to play the audio file.
{Listen}
Your prediction:

O The person takes the €5 banknote
O The person does not take the €5 banknote

To what extent are you sure of your prediction, on a scale of 1 (totally uncertain) to 5 (totally certain)?
1 2 3 4 5

3.2. INSTRUCTIONS OF LABORATORY EXPERIMENT 2

Instructions are translated in English from French. We only report the instructions of Part 1. The instructions of Part 2 are similar to those used in Part 1 of Laboratory Experiment 1.

Hello. Thank you for participating in this study. Please turn off your mobile phone. It is forbidden to communicate with other participants for the duration of the session.
If you have any questions at any time, please press the red button on the side of your desk and an assistant will come to answer your questions in private.
The experience is divided into two parts. At the end of the session, you will receive your earnings from parts 1 and 2 as well as a show-up fee of €10. Your earnings will be paid to you privately in a separate room to maintain confidentiality.
Please press OK to see the rest of the instructions.

{OK}

Part 1

Your task

Imagine that you take a bus to reach a certain destination. Taking the bus requires you to buy a ticket that costs €1.8. There could be a ticket inspection on the bus. This inspection is determined by the computer program with a certain probability that you do not know. If you are not inspected or if you are inspected and you have bought a ticket, there is no consequence. If you are inspected and you have not bought a ticket, you will have to pay a fine of €3 and the price of the ticket (€4.8 in total).
Your task consists of deciding whether you want to buy the ticket or not. After your decision, there are four possible scenarios:
You have not bought the ticket and you are not inspected: your loss is €0.
You have not bought the ticket and you are inspected: your loss is €4.8.
You have bought the ticket and you are not inspected: your loss is €1.8.
You have bought the ticket and you are inspected: your loss is €1.8.
The losses of this part will be deducted from the show-up fee of €5.
Please click "OK" to make your decision.
{OK}

Decision

Click on "Ticket €1.8" if you want to buy the ticket or "No ticket" if you do not want to buy it.
You will know immediately if you are inspected. If you are not inspected, you will go directly to the next part.

{Ticket € 1.8} {No ticket}

Inspection!

You are inspected!

You have bought the ticket: there is no consequence [You have not bought the ticket: you pay a fine of €3 and the price of the ticket].
{OK}

In the on-screen original instructions, the picture in the middle is a video.

APPENDIX 4: FIGURES

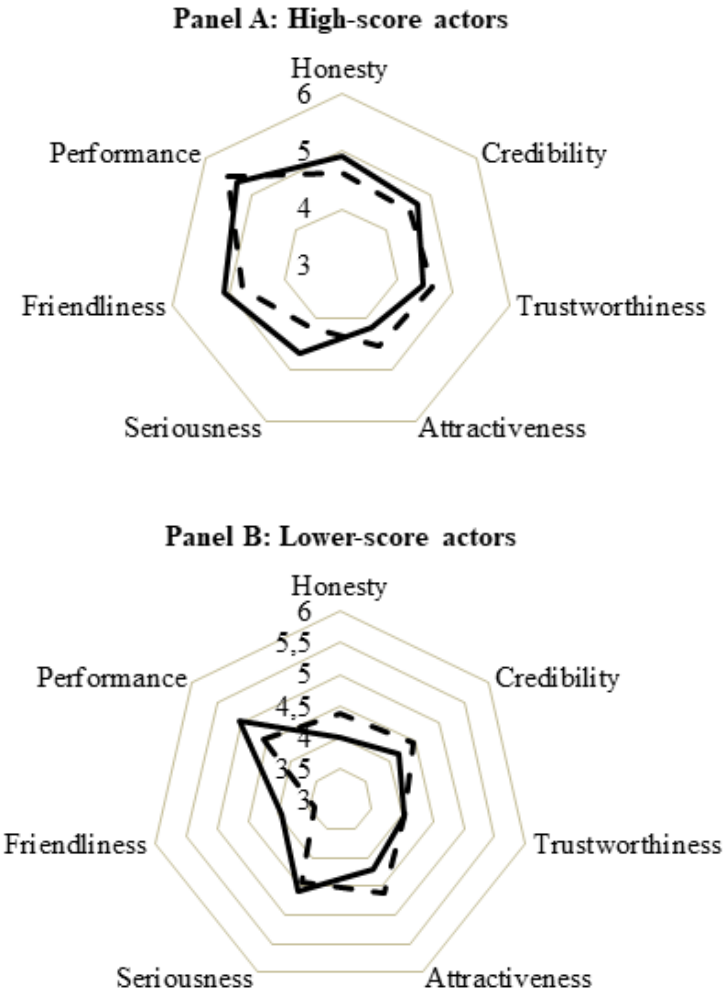

Figure A1. Average scores of the four selected actors on the different characteristics.

The figure depicts the average score given to each of the four selected actors by the 21 subjects from the subject pool of GATE-LAB that were recruited to evaluate the actors. After watching the videos of the 18 actors, these subjects rated each candidate in terms of performance, honesty, trustworthiness, attractiveness, credibility, seriousness, and friendliness. We selected two actors and two actresses with similar high scores in performance and credibility, and similar scores in the other dimensions (one pair made of one actor and one actress with high scores in all the other dimensions, and another pair with lower scores). Panel A refers to the two actors with higher scores and Panel B refers to the two actors with lower scores. Each characteristic is measured on a scale from 1 to 7. The dashed line identifies an actress and the solid line an actor. This procedure ensured a neutral selection of the actors. In the main econometric analysis we introduced individual fixed effects to control for the different observable and unobservable characteristics of the actors.

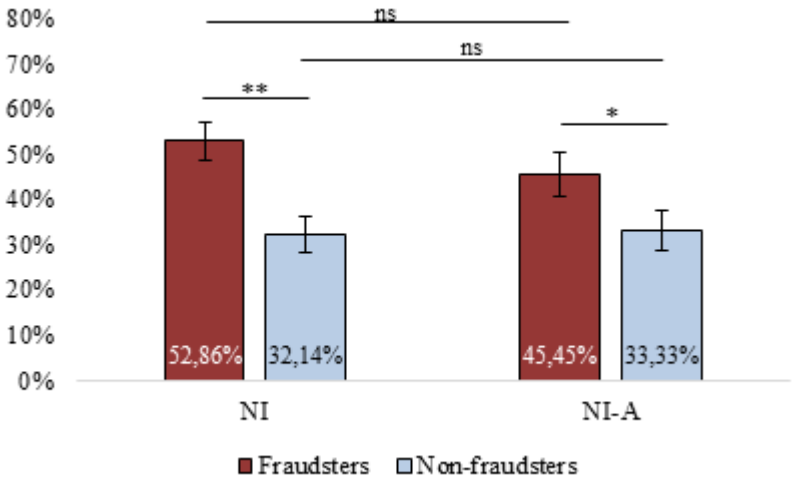

Figure A2. Percentage of fraudsters and non-fraudsters accepting the banknote in the conditions with no ticket inspection.

The figure presents the percentage of fraudsters and non-fraudsters who took the banknote in the NI and NI-A conditions. The light bars are for non-fraudsters and the dark bars for fraudsters in the two conditions without prior ticket inspection (NI and NI-A). N = 140 (NI, fraudsters), 140 (NI, non-fraudsters), 110 (NI-A, fraudsters), and 114 (NI-A, non-fraudsters). Error bars, mean ± SEM. Significance levels: *** $p < 0.01$, * $p < 0.1$, *ns* not significant, Chi-squared tests.

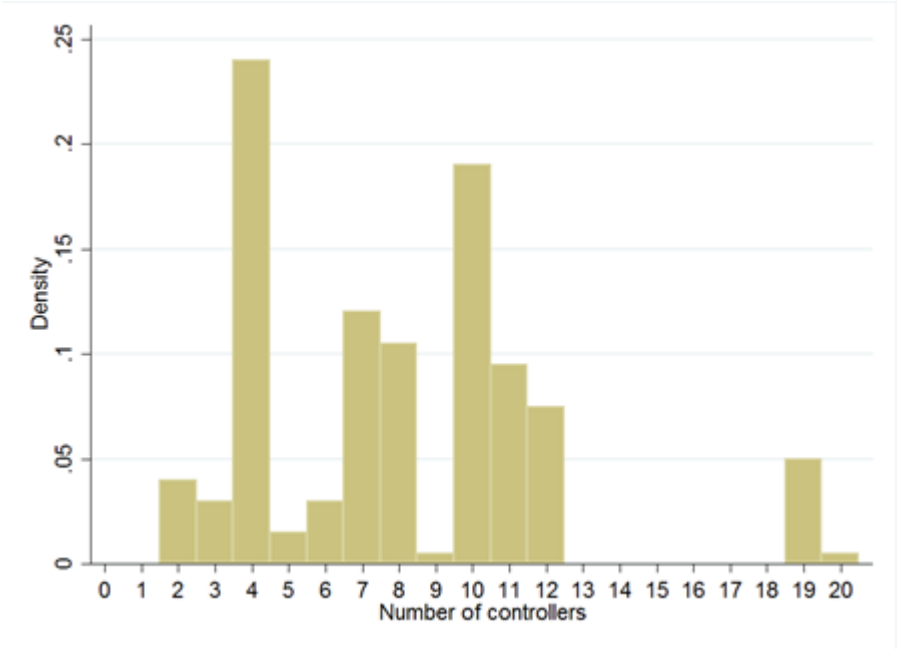

Figure A3. Histogram of the number of ticket inspectors in the quasi-experiment in the field.

The figure shows the histogram of the number of ticket inspectors (per inspection) that we encountered in the experiment in the field (N = 200).

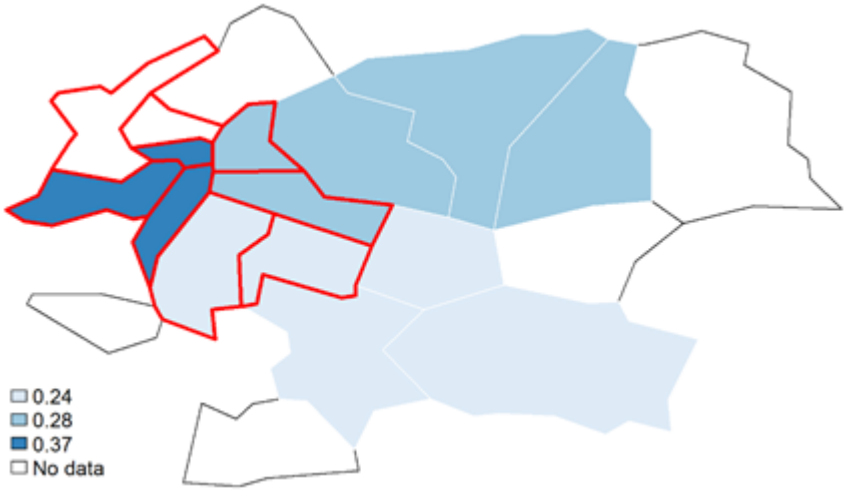

Figure A4. Frequency of ticket inspections in the bus-tram area of Metropolitan Lyon, as measured in our experiment in the field.

The figure plots the frequency of observed ticket inspections on a map covering the area of Metropolitan Lyon that can be reached by tram or bus. The darkest blue shaded area identifies Center Metropolitan Lyon; the medium blue shaded area corresponds to North-East Metropolitan Lyon; while the light blue shaded area represents South-East Metropolitan Lyon. White segments are areas which we did not cover for logistical reasons. The area with red contours identifies the city of Lyon.

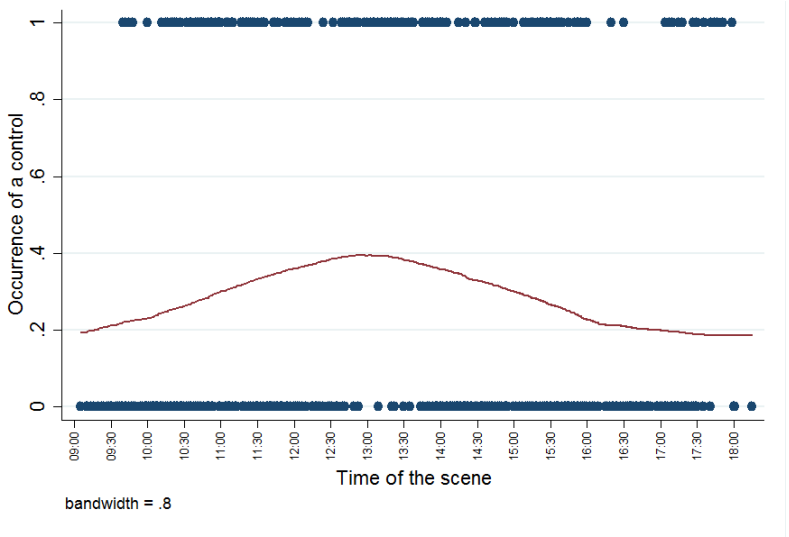

Figure A5. Relationship between frequency of inspections and time of the scene.

The figure depicts the smoothed running means of whether an inspection occurred or not as a function of the time at which the scene was played (which approximates the time of an inspection, when this occurred). The running means are computed using a band width of 0.8 (80% of the data) and are adjusted to equal the mean of the variable in the Y-axis.

The relationship takes the form of an inverted U, with more frequent ticket inspections between noon and 2:30pm.

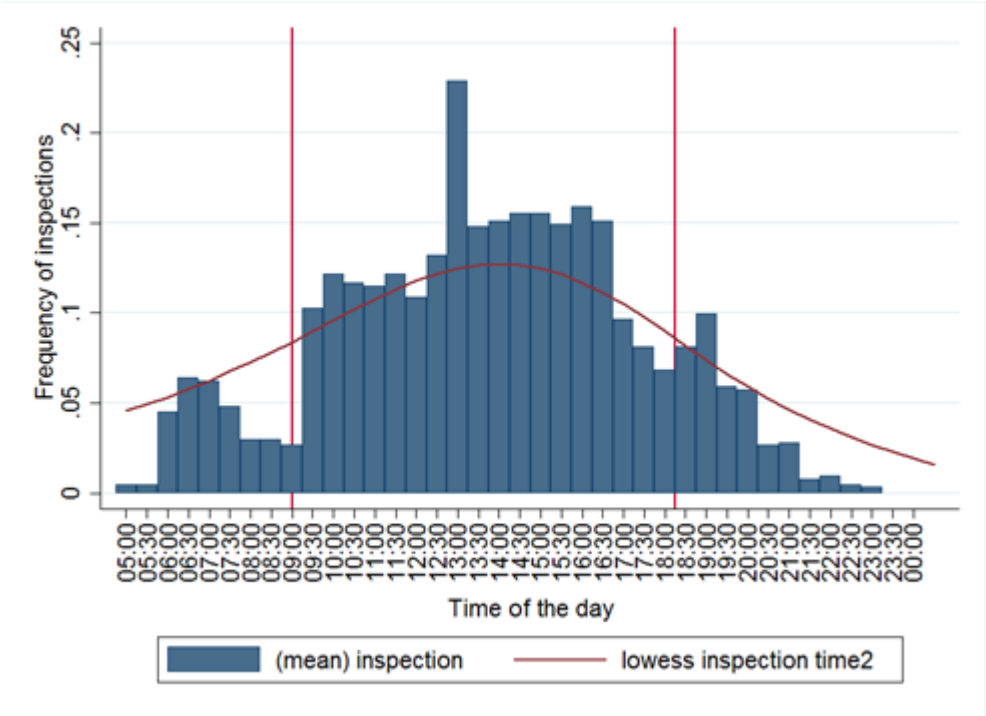

Figure A6. Relationship between frequency of inspections and the time of day based on the inspection plans of the public transport company

In order to check the consistency of our sampling strategy, we analyzed the data from the transport company regarding its monthly inspection plans. These data (available for three months: June, July and November 2017) contain information about the planned inspections for all tram lines (T1-T5). The figure depicts the distribution of the inspections over time. The inverted U-shape pattern is analogous to the one observed in our field data (see Figure A5). The figure depicts the frequency of inspections in the tram (pooling all lines together) from 5AM until midnight. The period is divided into half-hour slots. The vertical red lines identify the interval of time in which we conducted the quasi-experiment. The smoothed line corresponds to the running means of the data (computed using a band width of 0.8).

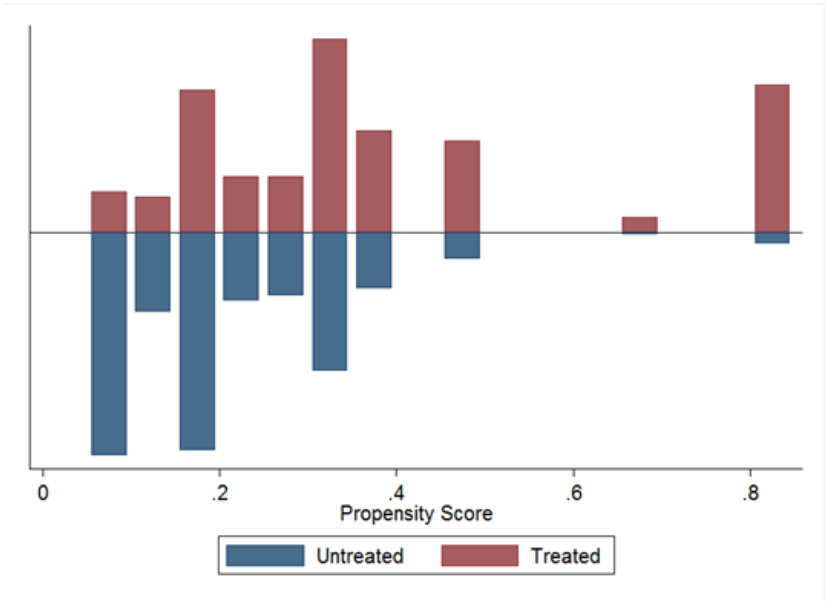

Figure A7: Histogram of the estimated propensity score

The figure shows the degree of overlap in the propensity scores between treated (bottom) and untreated units (top) in our data. Propensity scores have a similar and overlapping distribution in the treated and untreated groups.

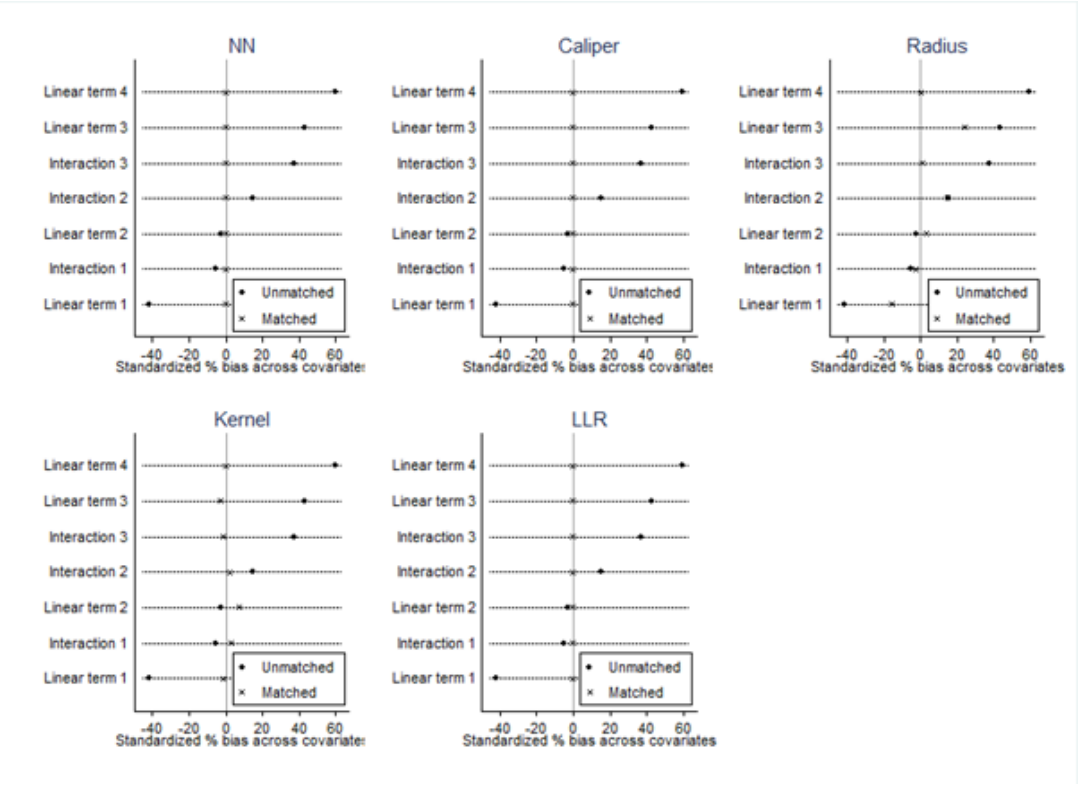

Figure A8: Covariate imbalance in terms of standardized percentage bias

The figure shows, for each matching technique, the extent of imbalance in terms of standardized percentage bias before and after matching using dot charts for each covariate. With NN, Caliper and LLM, all treated units are perfectly matched to their untreated counterparts, and all biases are removed. With Radius and Kernel, some small bias persists but it is drastically reduced compared to pre-matching.

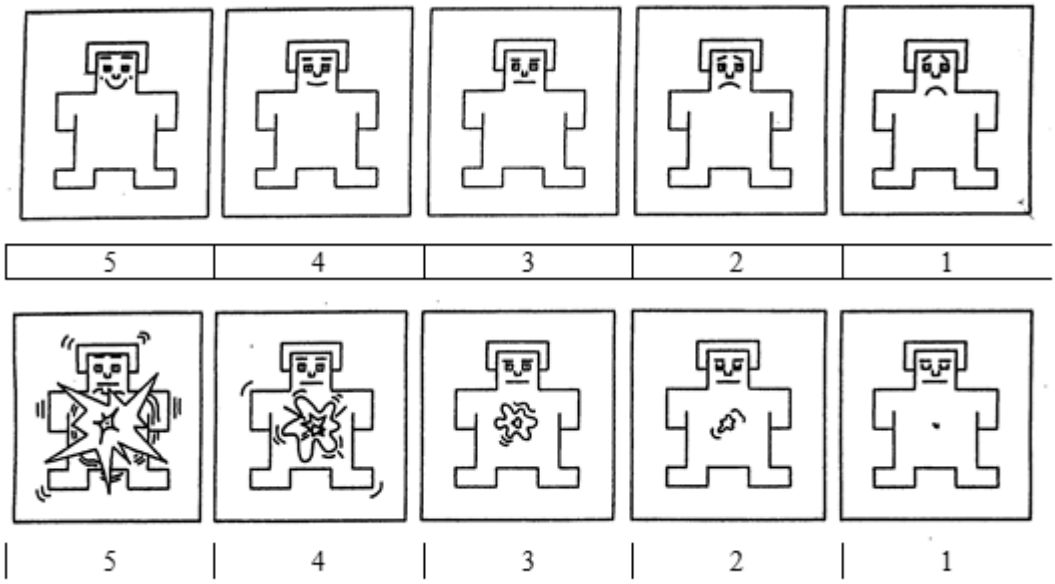

Figure A9. Self-Assessment Manikin (SAM) questions.

The figure depicts the pictures used in the survey conducted in public transport in Lyon to measure the emotional state (happiness on top and nervousness on bottom) of the participants on a scale from 1 to 5.

A research assistant identified non-fraudsters travelling on board of buses and trams and approached them when they were getting off the public vehicle, using the same identification procedure as in our quasi-experiment (N=160). The survey contained two questions using Self-Assessment Manikins (SAM) (Lang, 1980). The first question measured self-reported happiness on a scale from 1 (unhappy) to 5 (happy), while the second question measured emotional arousal on a scale from 1 (quiet, calm) to 5 (nervous). Each question was presented with five pictures associated with each possible answer, as shown in the figure.

APPENDIX 5: TABLES

Table A1. Descriptive statistics on targeted passengers' individual background variables in the NI, NI-A and I observations, in the quasi-experiment.

The table presents the share of observations collected in the NI, NI-A and I condition of the quasi-experiment, focusing on the mean individual characteristics of the targeted passengers. The last column present the *p*-values for the null hypothesis that the data for each variable are independent across conditions (two-sided χ^2 tests).
The sample is balanced with respect to gender, ethnicity, and when considering individuals who exhibit religious signs. While there are more poor individuals in the I condition, the proportion of fraudsters is not different depending on the estimated wealth when comparing the I and NI, NI-A conditions.

Characteristics	**NI & NI-A**	**I**	**χ^2 test (p-value)** **NI & NI-A vs. I**
Gender			
Female	48.02%	45.59%	0.558
Male	51.98%	54.41%	
Age			
18-24	27.18%	31.86%	
25-34	28.77%	25.49%	
35-44	16.07%	10.78%	0.24
45-59	18.25%	19.61%	
≥ 60	9.72%	12.25%	
Ethnicity			
Caucasian	61.43%	59.31%	
Arab	15.71%	20.1%	
African	15.71%	17.16%	0.517
Asian	4.64%	1.47%	
Other	2.5%	1.96%	
Wealth			
Poor	19.64%	30.39%	
Average	72.22%	60.29%	**0.005**
Rich	8.13%	9.31%	
Religious signs			
No	97.82%	97.06%	0.55
Yes	2.18%	2.94%	

Table A2. Descriptive statistics on targeted passengers' individual background variables in the quasi-experiment.

The table presents the mean individual characteristics of targeted passengers in the quasi-experiment by group and by condition. Columns (8) to (14) present the *p*-values for the null hypothesis that the data for each variable are independent across conditions (two-sided χ^2 tests). In column (8), the tests are conducted across all conditions. In columns (9)-(14), the tests are based on pairwise comparisons. N = 708.

Subjects differ across conditions in terms of estimated age, ethnicity and wealth (cf. column 9). This is mainly due to the difference between fraudsters and non-fraudsters (cf. columns 10-11). Fraudsters tend to be younger (χ^2 test, p = 0.006 in I) and with a lower apparent wealth (χ^2 test, p = 0.008 in I). This is in line with previous evidence on public transport users in Lyon (Dai *et al.*, 2018). We also checked whether the samples differ between I and NI, and between NI and NI-A (*cf.* columns 12-15). The only statistically significant difference is in apparent wealth between I and NI for fraudsters (p = 0.005). Inspected fraudsters are perceived as poorer than non-inspected fraudsters. In all the other comparisons we cannot reject the null hypothesis that the data are independent across conditions. We can thus conclude that, except for the anticipated difference between fraudsters and non-fraudsters, the sample is fairly balanced across conditions.

Characteristics	NI		I		NI-A		Total	χ^2 test (*P*-value)						
	Fraudsters	Non-fraudsters	Fraudsters	Non-fraudsters	Fraudsters	Non-fraudsters		All	NI fraud. *vs.* NI non-fraud.	I fraud. *vs.* I non-fraud.	NI fraud. *vs.* I fraud.	NI non-fraud. *vs.* I non-fraud.	NI fraud. *vs.* NI-A fraud.	NI non-fraud. *vs.* NI-A non-fraud.
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)
Gender														
Female	44%	51%	47%	44%	45%	52%	47%	0.758	0.281	0.691	0.677	0.316	0.854	0.869
Male	56%	49%	53%	56%	55%	48%	53%							
Age														
18-24	29%	23%	38%	26%	35%	24%	29%							
25-34	33%	29%	31%	20%	26%	26%	28%							
35-44	12%	17%	9%	13%	16%	19%	15%	**0.027**	0.527	**0.006**	0.400	0.320	0.606	0.529
45-59	16%	19%	17%	22%	15%	24%	19%							
≥ 60	10%	13%	5%	19%	8%	7%	10%							
Ethnicity														
Caucasian	54%	69%	55%	63%	52%	65%	60%							
Arab	19%	13%	24%	16%	21%	12%	17%	**0.061**	0.117	0.447	0.536	0.344	0.119	0.540
African	19%	12%	16%	18%	23%	19%	18%							

Asian	6%	4%	2%	1%	0%	3%	3%							
Other	3%	2%	3%	1%	5%	1%	2%							
Wealth														
Poor	21%	15%	40%	21%	25%	19%	23%							
Average	71%	74%	54%	66%	71%	73%	69%	**0.002**	0.325	**0.008**	**0.005**	0.410	0.482	0.475
Wealthy	8%	11%	6%	13%	5%	8%	8%							
Religious signs														
No	97%	98%	96%	98%	98%	98%	98%	0.886	0.702	0.380	0.627	0.905	0.594	0.825
Yes	3%	2%	4%	2%	2%	2%	2%							

Table A3. Descriptive statistics on the share of NI, NI-A and I observations for each public transport mode, each time of day, and each district.

The table compares the NI, NI-A and I observations for each public transport line, for each hour of day (from 9:00AM to 6:15PM), and for each district where we played the scene, respectively in Panel A, B and C. The line "Ticket" refers to the proportion of individuals with a validated ticket or pass in % (number) with respect to the overall number of observations collected in the NI, NI-A conditions or in the I condition. N = 708.

Regarding transport lines, between 17% and 34% of the total number of observations collected in each tram line refer to the I condition. We also travelled on a very wide and heterogeneous range of bus lines but this was done (i) less regularly, and (ii) typically when the transport company or other inspectors tipped us off about an inspection on a specific bus line, that is why our observations in the buses only refer to the I condition. However, both when considering the latter group of observations (*i.e.*, those collected in the buses or in the metro), and when considering the observations collected in the tram, our sample is well balanced when considering individuals that were holding a ticket (respectively 56.45% and 50% in the metro/bus and in the tram) or not holding a ticket (respectively 43.55% and 50% in the metro/bus and in the tram). Focusing on the tram lines we visited the most (tram lines T1, T2 and T4), inspections were more frequent in the tram line T4 (33.33%) compared to any other lines (20.36% in T1 and 20.18% in T2). Line T4 vertically crosses the metropole of Lyon, and stops at two main train stations and the University Campus. Regarding the time of day, inspections in our sample were more frequent between noon and 2:30pm. Regarding geolocation, the districts where we most frequently run our experiment are the second one, the third one, the seventh one and the suburbs, which are related to the itinerary of trams T1, T2 and T4.

Panel A

Line	NI & NI-A	I
67	0%	100%
Ticket	-	50%
	-	1/2
C12	0%	100%
Ticket	-	50%
	-	1/2
C14	0%	100%
Ticket	-	80%
	-	4/5
C17	0%	100%
Ticket	-	0%
	-	0/2

Line	NI & NI-A	I
C19	0%	100%
Ticket	-	0%
	-	0/1
C21	0%	100%
Ticket	-	100%
	-	1/1
C3	0%	100%
Ticket	-	33%
	-	7/21
C5	0%	100%
Ticket	-	100%
	-	1/1

Line	NI & NI-A	I
C8	0%	100%
Ticket	-	50%
	-	3/6
C9	0%	100%
Ticket	-	57%
	-	4/7
T1	80%	20%
Ticket	49%	42%
	172/352	38/90
T2	80%	20%
Ticket	57%	45%
	50/87	10/22

Line	NI & NI-A	I
T3	83%	17%
Ticket	67%	100%
	4/6	1/1
T4	67%	33%
Ticket	48%	69%
	28/58	20/29
T5	100%	0%
Ticket	0%	-
	0/1	-
D	0%	100%
Ticket	-	93%
	-	13/14

Panel B

Hour	NI & NI-A	I
9:00AM - 9:59AM	93%	7%
Ticket	38% 20/53	75% 3/4
10:00AM -10:59AM	72%	28%
Ticket	58% 47/81	48% 15/31
11:00AM -11:59AM	69%	31%
Ticket	58% 34/59	59% 16/27
12:00PM -12:59PM	68%	32%
Ticket	56% 22/39	50% 9/18
1:00PM -1:59PM	26%	74%
Ticket	47% 9/19	44% 24/55

Hour	NI & NI-A	I
2:00PM -2:59:PM	78%	22%
Ticket	47% 37/78	45% 10/22
3:00OM -3:59PM	68%	32%
Ticket	48% 29/61	48% 14/29
4:00PM -4:59PM	96%	4%
Ticket	57% 43/76	0% 0/3
5:00PM -5:59PM	70%	30%
Ticket	34% 12/35	87% 13/15
6:00PM – 6:15PM	100%	0
Ticket	33% 1/3	- -

Panel C

District	NI & NI-A	I
Suburbs	46%	54%
Ticket	53% 36/68	58% 46/79
1st	0%	100%
Ticket	- -	80% 4/5
2nd	65%	35%
Ticket	58% 67/116	61% 38/62
3hd	80%	21%
Ticket	40% 51/128	18% 6/33

District	NI & NI-A	I
5th	0%	100%
Ticket	- -	50% 1/2
6th	85%	15%
Ticket	47% 8/17	33% 1/3
7th	87%	13%
Ticket	55% 72/131	37% 7/19
8th	98%	2%
Ticket	45% 20/44	100% 1/1

Table A4. Determinants of the occurrence of an inspection in our quasi-experiment and according to the public transportation company's data.

In order to check the consistency of our sampling strategy, we analyzed the data from the transport company regarding its monthly inspection plans. These data (available for three months: June, July and November 2017) contain half-hourly information about the planned inspections for all tram lines (T1-T5). The Table reports the results of a linear probability model (Model (1)) and a logit model (Model (2)). The dependent variable is the planning of an inspection (= 1 if the public transport company planned an inspection and 0 otherwise). Independent variables are the time of day, its squared term and a dummy for each tram line. Note that several inspections were planned in overlapping segments of multiple lines; so, we can include in the regression all the lines without the need to interpret the coefficients with respect to a baseline line. The regressions are run on the data corresponding to the portion of day in which we conducted our quasi-experiment (from 9AM to approximately 6PM). Significance levels: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.

The results confirm that inspections are indeed more frequent around midday and in line T4 (consistently with our field data) and less common in T1 and T5.

	Regression 1		**Regression 2**	
	b	**se**	**b**	**se**
Time of the day	0.000***	0	0.000***	0
Time of the day × Time of the day	−0.000***	0	−0.000***	0
Line T1	−0.094***	0.011	−0.769***	0.1
Line T2	−0.016	0.015	0.053	0.137
Line T3	−0.01	0.015	0.104	0.137
Line T4	0.030***	0.01	0.379***	0.103
Line T5	−0.021*	0.011	−0.179*	0.098
Constant	−0.733***	0.08	−11.065***	0.84
Number of observations	12198		12198	
Adjusted or pseudo R-Square	0.026		0.037	

Table A5. Appropriateness scores across scenarios (N = 30) in Laboratory Experiment 1.

The table reports, for each scenario, the mean responses and the frequency of each possible response in the task eliciting the appropriateness of behavior in two scenarios: "Very socially inappropriate" (– –); "somewhat socially inappropriate" (–), "somewhat socially appropriate" (+), "very socially appropriate" (++). Modal responses are shaded in grey. Following Krupka and Weber (2013) to construct the mean score, we assigned a value of –1 to "very socially inappropriate", –1/3 to "somewhat socially inappropriate", 1/3 to "somewhat socially appropriate" and 1 to "very socially appropriate". The table also reports the *p*-values of Wilcoxon signed-rank tests comparing the distributions of responses in the two scenarios, with each subject taken as an independent observation.

Table A5 confirms that taking the banknote in Scenario 2 (where the banknote is found by person B) is collectively considered as socially inappropriate. The mean score is negative and statistically different from 0 ($p < 0.001$). The modal response ("very socially inappropriate") receives 63% of the responses. No one judged the decision of person A in Scenario 2 as socially appropriate.

Scenario	Mean	− −	−	+	+ +	Rank-sum test
B finds the banknote and asks A	−0.76	63.33%	36.67%	0%	0%	$p < 0.001$
A finds the banknote	0.29	0%	20%	66.67%	13.33%	

Table A6. Proportion of lab subjects who guessed "took" depending on the actual behavior of the field subjects (N = 45) in Laboratory Experiment 1.

This Table reports, for each condition and group of field subjects in the Lab Experiment 1, (*i*) the proportions of lab participants who guessed "took" when the field subjects took (first row) and did not take (second row) the banknote, respectively; (*ii*) the ability A to predict field behavior (third row); and (*iii*) the *p*-values of Wilcoxon signed-rank tests that A = 0, taking the participant's average A as the independent unit of observation (fourth row). Standard errors are reported in parentheses.

The table shows evidence that lab participants were not able to predict the behavior of the field subjects. In all conditions and groups of field subjects, the ability to predict is either not significantly different from zero or (weakly) significantly negative (in the I condition for fraudsters), meaning that lab subjects were, if anything, worse than chance in predicting field behavior.

Conditions and groups	All	NI fraudsters	NI non-fraudsters	I fraudsters	I non-fraudsters
Subject took, $F(T\|T)$	0.51	0.48	0.53	0.54	0.49
	(0.02)	(0.03)	(0.04)	(0.02)	(0.03)
Subject did not take, $F(T\|NT)$	0.51	0.49	0.5	0.6	0.51
	(0.02)	(0.03)	(0.03)	(0.04)	(0.03)
Ability to predict (A)	0.00	−0.01	0.03	−0.06	−0.02
Wilcoxon test, $A = 0$ (p-value)	0.969	0.852	0.955	0.053	0.663

Table A7. Effect of main treatments on guessing that a person took the banknote in Laboratory Experiment 1.

Even if our lab subjects were not able to predict behavior in the field (Table A6), they might still have perceived changes in the performance of the actors across conditions and groups of subjects, and assigned different probabilities of taking the banknote. To test this possibility, Table A7 reports the average marginal effects of a logit regression on the probability of guessing (correctly or not) that the person took the banknote. Independent variables are treatment dummies (using "NI fraudsters" as the baseline category) and fixed effects for the actors. Standard errors are clustered at the subject level to account for the fact that subjects made 48 guesses.

This table shows no evidence that subjects assigned a different probability of taking the banknote across conditions and groups of field subjects.

Dependent variable: 1 if subject guessed that the person took the banknote	dy/dx	Std. Err.	*p*>z
NI, non-fraudster	0.00	0.03	0.925
I, fraudster	0.05	0.03	0.129
I, non-fraudster	0.01	0.03	0.653
Actor fixed effects	Yes		
Number of observations	2160		
Pseudo R^2	0.006		
Prob > chi^2	0.088		

Table A8. Determinants of the decision to take the banknote (controlling for the presence of an experimenter).

The table reports the coefficients and robust standard errors of linear probability regressions. The dependent variable is the decision to take the banknote (= 1 if the banknote is taken and 0 otherwise). Significance levels: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$ (Wald tests).

Dependent variable: Decision to take the banknote	Model (3) Coeff.	se	Model (4) Coeff.	se
Inspection (baseline = No Inspection)	0.148**	0.067	0.144**	0.069
Audience (baseline = no audience)	0.050	0.061	0.028	0.061
Fraudster (baseline = no fraudster)	0.177***	0.057	0.173***	0.058
Inspection*Fraudster	-0.011	0.0897	-0.018	0.089
Audience*Fraudster	-0.096	0.087	-0.082	0.087
Constant	0.430***	0.123	0.538***	0.120
Actors/Actress (baseline = Higher−score actress)				
Lower-score actress	0.188***	0.048		
Higher-score actor	-0.016	0.071		
Lower-score actor	0.071	0.053		
Gender interaction (baseline = Female actress, Female passenger*)*				
Female actress, Male passenger			0.032	0.046
Male actor, Female passenger			-0.052	0.057
Male actor, Male passenger			-0.032	0.051
Male passenger (baseline = female passenger)	0.023	0.037		
Age	0.028*	0.014	0.245*	0.014
Time of day (baseline = 9:00AM - 11:59AM)				
12:00PM - 2:59PM	-0.071	0.044	-0.060	0.045
3:00PM - 6:15PM	-0.074	0.048	-0.057	0.048
Geolocation (baseline = Center Metropolitan Lyon)				
North-East Metropolitan Lyon	-0.030	0.047	-0.031	0.048
South-East Metropolitan Lyon	-0.070	0.051	-0.056	0.051
Line public Transport (baseline = other)				
T1	0.030	0.072	0.043	0.074
T2	-0.044	0.085	-0.090	0.085
T4	0.012	0.087	0.041	0.089
Experimenter	**0.007**	**0.048**	**0.035**	**0.048**
Ethnicity (baseline = Caucasian)				
Arab	0.062	0.055	0.036	0.054
African	0.063	0.051	0.053	0.051
Asian	0.079	0.117	0.079	0.124
Other	0.090	0.126	0.104	0.127
Social appearance (baseline = poor)				
Average	-0.205***	0.050	-0.199***	0.050
Rich	-0.254***	0.076	-0.254***	0.078
Religious signs (baseline = no religious signs)	-.0.226**	0.110	-0.215**	0.106

Crowded (baseline = No Crowded)	-0.048	0.038	-0.061	0.038
Weather (baseline = sunny)				
Cloudy	-0.019	0.055	-0.052	0.054
Rainy	-0.086	0.081	-0.151*	0.082
Someone could notice the scene (baseline = no one)	-0.041	0.039	-0.047	0.039
Obs	708		708	
R2	0.1483		0.1283	
Prob > F	0.0000		0.0000	

Table A9. Determinants of the decision to take the banknote (with interactions between the time of day and each transport line category (Model 1), and between the geolocation and the time of day (Model 2)).

The table reports the coefficients and the robust standard errors of linear probability regressions. We use Model (3) from the main text as a benchmark (see Table 1 in the main text). The results are qualitatively identical if we consider a different model. Significance levels: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$ (Wald tests).

Dependent variable: Decision to take the banknote	**Model (1)**		**Model (2)**	
	Coeff.	se	Coeff.	se
Inspection (baseline = No Inspection)	0.120*	0.066	0.136**	0.066
Audience (baseline = no audience)	0.042	0.061	0.043	0.061
Fraudster (baseline = no fraudster)	0.176***	0.058	0.174***	0.058
Inspection*Fraudster	0.016	0.087	0.005	0.088
Audience*Fraudster	-0.091	0.087	-0.085	0.086
Actors/Actress (baseline = Higher−score actress)				
Lower-score actress	0.177***	0.05	0.191***	0.048
Higher-score actor	-0.02	0.072	-0.019	0.072
Lower-score actor	0.054	0.052	0.075	0.051
Male passenger (baseline = female passenger)	0.027	0.037	0.022	0.037
Age	0.03**	0.014	0.029**	0.014
Time of day (baseline = 9:00AM - 11:59AM)				
12:00PM - 2:59PM	-0.057	0.210	-0.177**	0.088
3:00PM - 6:15PM	0.191	0.205	-0.11	0.091
Geolocation (baseline = Center Metropolitan Lyon)				
North-East Metropolitan Lyon	-0.017	0.048	-0.064	0.086
South-East Metropolitan Lyon	-0.075	0.051	-0.163*	0.084
Line public Transport (baseline = other)				
T1	0.164	0.195	0.03	0.072
T2	0.108	0.202	-0.036	0.085
T4	0.083	0.202	0.01	0.010
*Line Public Transport*Time of day*				
T1*12:00PM - 2:59PM	-0.026	0.220		
T1*3:00PM - 6:15PM	-0.293	0.214		
T2*12:00AM - 2:59PM	0.007	0.241		
T2*3:00PM - 6:15PM	-0.351	0.229		
T4*12:00AM - 2:59PM	0.117	0.237		
T4*3:00PM - 6:15PM	-0.226	0.285		
*Geolocation*Time of day*				
North-East*12:00PM - 2:59PM			0.119	0.115
North-East*3:00PM - 6:15PM			-0.042	0.117
South-East*12:00AM - 2:59PM			0.157	0.12
South-East *3:00PM - 6:15PM			0.125	0.115
Ethnicity (baseline = Caucasian)				
Arab	0.056	0.055	0.064	0.054
African	0.059	0.051	0.076	0.05
Asian	0.079	0.121	0.086	0.112
Other	0.086	0.128	0.094	0.119
Social appearance (baseline = poor)				
Average	-0.0203***	0.050	-0.199***	0.050
Rich	-0.248***	0.077	-0.244***	0.077

Religious signs (baseline = no religious signs)	-0.216*	0.110	-0.244**	0.115
Crowded (baseline = No Crowded)	-0.038	0.039	-0.052	0.0381
Weather (baseline = sunny)				
Cloudy	-0.014	0.052	-0.001	0.053
Rainy	-0.087	0.080	-0.089	0.091
Someone could notice the scene (baseline = no one)	-0.04	0.039	-0.034	0.039
Constant	0.306	0.216	0.47***	0.136
Obs	708		708	
R2	0.1571		0.1542	
Prob > F	0.0000		0.0000	

Table A10. Determinants of the decision to take the banknote (including hourly dummies for the time of day, a dummy for each minor line, and local district fixed effects for geolocation).
The table reports the coefficients and the robust standard errors of linear probability regressions. We use Model (3) from the main text as a benchmark (see Table 1 in the main text). Significance levels: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$ (Wald tests).

Dependent variable: Decision to take the banknote	**Model (1)**	
	Coeff.	se
Inspection (baseline = No Inspection)	0.138**	0.069
Audience (baseline = no audience)	0.045	0.063
Fraudster (baseline = no fraudster)	0.184***	0.059
Inspection*Fraudster	0.012	0.089
Audience*Fraudster	-0.089	0.088
Actors/Actress (baseline = Higher−score actress)		
Lower-score actress	0.186***	0.051
Higher-score actor	-0.018	0.073
Lower-score actor	0.052	0.053
Male passenger (baseline = female passenger)	0.018	0.037
Age	0.026*	0.014
Ethnicity (baseline = Caucasian)		
Arab	0.058	0.056
African	0.082	0.052
Asian	0.101	0.126
Other	0.144	0.127
Social appearance (baseline = poor)		
Average	-0.197***	0.051
Rich	-0.209**	0.082
Religious signs (baseline = no religious signs)	-0.167	0.124
Crowded (baseline = No Crowded)	-0.052	0.041
Weather (baseline = sunny)		
Cloudy	-0.015	0.052
Rainy	-0.106	0.079
Someone could notice the scene (baseline = no one)	-0.024	0.041
Dummies for each hour	YES	
Dummy for each minor line	YES	
Dummies for each local district	YES	
Constant	0.388***	0.132
Obs	708	
R2	0.1809	
Prob > F	0.0000	

Table A11. Measures of overall imbalance.

The table reports different measures of overall imbalance between treated and untreated units before (first row) and after (rows 2-6) matching. The second column (Pseudo R^2) displays the pseudo-R2 from the propensity score estimation using the unmatched (row 1) or matched data (rows 2-6), while the third column (*p*>chi2) reports the p-value of a likelihood ratio test on the joint significance of all coefficients in the model. A matching is successful in balancing the observations if the pseudo-R2 is low and the likelihood ratio test rejects the null hypothesis. The fourth (Mean Bias) and fifth (Median Bias) columns display the overall mean and median bias (from the distribution of the individual biases), respectively. The sixth (Rubin's B) and seventh (Rubin's R) columns display Rubin's B (the standardized difference in the means of the propensity scores in the treated and untreated units) and the Rubin's R (the ratio of the variances of the propensity score in treated and untreated units), respectively. A matching achieves a sufficiently balanced sample if $B < 25$ and $0.5 < R < 2$.

	Pseudo R^2	**p>chi2**	**Mean Bias**	**Median Bias**	**Rubin's B**	**Rubin's R**
Unmatched	0.153	<0.001	29.26	37.31	95.91	1.83
NN	0.000	>0.999	0	0	0	1
Caliper	0.000	>0.999	0	0	0	1
Radius	0.028	0.064	8.73	2.87	39.54	1.07
Kernel	0.002	0.997	2.60	1.95	9.96	1.09
LLR	0.000	>0.999	0	0	0	1

p>χ^2 = Likelihood ratio test on the joint significance of all coefficients in the model.

Table A12: Determinants of taking the banknote (robustness to matching techniques).

The table reports the estimates from linear probability regressions with robust standard errors in Column (1). Weighted least squares regressions with bootstrapped standard errors are reported in Columns (2)-(6). We use Model (3) from the main text as a benchmark (see Table 1 in the main text). This model is reported in Column (1) for comparison. The results are qualitatively identical if we consider a different model. Columns (2)-(6) replicate Model (3) but using matching techniques. The standard errors in Columns (2)-(6) are computed using a bootstrap routine with 500 replications that simultaneously estimates the propensity score and the weighted regressions (see *e.g.*, Whittaker et al., 2016). Significance levels: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.

	base		NN		Caliper		Radius		Kernel		LLR	
	b	se	b	se	B	se	b	se	b	se	b	se
Ticket inspection	0.15**	0.06	0.15**	0.08	0.15**	0.08	0.18**	0.07	0.16**	0.07	0.15**	0.08
Fraudster	0.18***	0.06	0.19**	0.08	0.19**	0.08	0.21***	0.07	0.18**	0.08	0.19**	0.08
Ticket inspection × Fraudster	-0.01	0.09	0.02	0.11	0.02	0.11	-0.01	0.1	0.02	0.1	0.02	0.11
Additional controls for actors, passengers, and environment	Yes		Yes		Yes		Yes		Yes		Yes	
Obs	708		670		670		677		677		670	
Pseudo R2	0.113		0.198		0.198		0.19		0.198		0.198	
Prob > chi2	0		0		0		0		0		0	

Table A13. Appropriateness scores across scenarios and conditions in Laboratory Experiment 2.

The table reports, for each scenario, the mean responses and the frequency of each possible response in the task eliciting the appropriateness of behavior in two scenarios in our second laboratory experiment: "Very socially inappropriate" (– –); "somewhat socially inappropriate" (–), "somewhat socially appropriate" (+), "very socially appropriate" (++). For a clean test of whether subjects who have been exposed to a ticket inspection changed their perception of the norm, we only focus on non-fraudsters (N = 56). Modal responses are shaded in grey. Following Krupka and Weber (2013) to construct the mean score, we assigned a value of –1 to "very socially inappropriate", –1/3 to "somewhat socially inappropriate", 1/3 to "somewhat socially appropriate" and 1 to "very socially appropriate". The table also reports the *p*-value of Wilcoxon rank-signed tests comparing the distributions of responses in the two scenarios, with each subject taken as an independent observation.

The table confirms the results from Laboratory Experiment 1 that taking the banknote in Scenario 2 (B finds the banknote and asks A) is collectively considered as socially inappropriate while taking the banknote in Scenario 1 (A finds the banknote) is not. If we compare the behavior of inspected and non-inspected non-fraudsters, we do not find statistically significant differences. The mean score of appropriateness is similar between the two groups of subjects in both Scenario 1 (Mann-Whitney tests, $p = 0.823$) and Scenario 2 ($p = 0.816$). The distribution of responses is also not statistically different (Fisher's exact test, $p = 0.784$ and 0.744 for Scenarios 1 and 2 respectively). This shows no evidence of a revision of their perception of the norm.

Situation and scenario	Mean	− −	−	+	+ +	Rank-sum test
Non-fraudsters no inspection (31)						
B finds the banknote and asks A	−0.57	41.94%	54.84%	3%	0%	$p < 0.001$
A finds the banknote	0.53	0%	6.45%	58.06%	35.48%	
Non-fraudsters inspection (25)						
B finds the banknote and asks A	−0.57	48%	44%	4%	4%	$p < 0.001$
A finds the banknote	0.49	0%	12%	52%	36%	

References of the Online Appendix

Belot, M., and van de Ven, J. (2017). How private is private information? The ability to spot deception in an economic game. *Experimental Economics*, 20: 19-43.

Bock, O., Baetge, I., and Nicklisch, A. (2014). hroot: Hamburg Registration and Organization Online Tool. *European Economic Review*, 71: 117-120.

Caliendo, M., Kopeinig, S. (2008). Some practical guidance for the implementation of propensity score matching. *Journal of Economic Surveys*, 22: 31-72.

Cochran, W. G., Rubin, D. B. (1973). Controlling bias in observational studies: A review. *Sankhyā: The Indian Journal of Statistics, Series A*, 417-446.

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

Heckman, J. J., Ichimura, H., Todd, P. E. (1997). Matching as an econometric evaluation estimator: Evidence from evaluating a job training programme. *The review of economic studies*, 64(4): 605-654.

Imbens, G., Rubin, D. B. **(**2015). *Causal Inference in Statistics, Social, and Biomedical Sciences.* Cambridge: Cambridge University Press.

Krupka, EL., Weber R.A. /2013). Identifying social norms using coordination games: Why does dictator game sharing vary? *Journal of the European Economic Association* 11(3): 495-524.

Leuven, E., Sianesi, B. (2003) PSMATCH2: Stata module to perform full Mahalanobis and propensity score matching, common support graphing, and covariate imbalance testing. Software, http://ideas.repec.org/c/boc/bocode/s432001.html.

Rosenbaum, P. R., Rubin, D. B. (1985). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician*, *39*(1), 33-38.

Rubin, D. B. (2001). Using propensity scores to help design observational studies: application to the tobacco litigation. *Health Services and Outcomes Research Methodology*, *2*(3-4), 169-188.

Whittaker, W., Anselmi, L., Kristensen, S. R., Lau, Y. S., Bailey, S., Bower, P., ... & Hodgson, D. (2016). Associations between extending access to primary care and emergency department visits: a difference-in-differences analysis. *PLoS Medicine*, *13*(9), e1002113.