Shaming as an incentive mechanism against stealing: behavioral and physiological evidence [*]

Isabelle Brocas
University of Southern California and CEPR

Juan D. Carrillo
University of Southern California and CEPR

Mallory Montgomery
Amazon.com, Inc.

October 2020

Abstract

We study experimentally the decision of an individual to steal or pay for an object that is produced at a cost by another individual. We consider two conditions. In the first condition, subjects caught stealing are sanctioned with a nominal fee. In the second condition, the sanction is increased by making the identity of the individual public (shaming). We also collect skin conductance responses to better understand the emotional arousal conducive to choices. Behaviorally, we show that stealing decreases significantly when shaming is introduced. More importantly, the emotional response at the time of decision strongly correlates with behavior. In particular, subjects who are more aroused are more likely to steal in the first condition, and also more likely to stop stealing in the second condition. Based on this physiological evidence, we develop a theoretical model where current decisions contribute to a "moral stock", which in turns affects the future cost of stealing and therefore the future decisions. The structural estimation of this model provides a good fit while capturing the heterogeneity across individuals.

Keywords: laboratory experiments, skin conductance, morality, non-choice data.

JEL Classification: C92, D91.

[*] Research was conducted while Mallory Montgomery was a graduate student at USC. We are grateful to Gary Charness and the members of the Los Angeles Behavioral Economics Laboratory (LABEL, `http://label-laboratory.org`) for their insights and comments in the various phases of the project. All remaining errors are ours. The study was conducted with the University of Southern California IRB approval UP-13-00248. We acknowledge the financial support of the Microsoft Corporation and the National Science Foundation grant SES-1851915. Address for correspondence: Juan D. Carrillo, Department of Economics, University of Southern California, 3620 S. Vermont Ave., Los Angeles, CA 90089, USA, `<juandc@usc.edu>`. Declarations of interest: none.

1 Introduction

Norm and rule violations cannot always be discouraged by well-crafted incentives. Crimes are often hard to detect and punishments are often low to act as deterrents. It has been emphasized that emotions, such as guilt and shame, are powerful drivers of social behavior (Elster, 1998). However, linking *explicitly* emotional states to decision-making can be challenging. Indeed, affective reactions are rarely measured. Instead, they are often implicitly deduced from choices. In this study, we join a small literature that exploits physical reactions–in our case electrodermal activity–to determine emotional arousal and then relates biological markers to decisions.

We design a very simple experiment, equivalent to a classic dictator game, but with loaded language and sanctions. One subject spends physical effort in developing a good of value \$20. The other subject (the dictator) can buy the good (pay \$7) or steal it (pay \$0). In Part 1, the potential sanction for stealing is a \$1 nominal fee. In Part 2, the sanction is increased via shaming, by showing the subject's picture to other participants and be labeled as a cheater. In either case, the dictator always keeps the good, so that stealing is the economically profitable strategy. Sanctions for stealing are random and subjects play multiple times with changing partners. In Elster (2011)'s terminology, the choice in Part 1 is affected only by moral norms (an internal feeling) whereas the choice in Part 2 is also affected by social norms (an expected behavior observed by others and enforced in some social way). During the experiment, we also measure each participant's skin conductance response, which is a proxy for emotional arousal, both when they decide whether to steal (decision time) and when they learn if they are caught (feedback time).

We first establish the positive role of social norms on behavior (section 4). Aggregate stealing decreases significantly when shaming is introduced. At the same time, there is large heterogeneity in behavior across individuals. Overall, and in accordance with most of the existing literature, norm compliance can be partly enforced through shaming when not through punishment.

We then study the ability of our physiological measures to predict behavior (section 5). We show that arousal at decision time is substantial and varies considerably across individuals. Subjects with the highest levels of emotional arousal when there is only a nominal fee are most likely to steal the good in that condition and also most likely to refrain from stealing when the social sanction is added. These individuals also exhibit a decrease in arousal between Parts 1 and 2, in part related to their change in behavior. Subjects who do not change their behavior when the social sanction is introduced exhibit a slightly higher arousal when they steal in the shaming condition. Also, and to our surprise, participants are not aroused at feedback time, which suggests that the consequences of stealing are fully integrated at the time of decision.

Last, we investigate the mechanisms through which emotions affect the choice of stealing (section 6). We perform a dynamic panel regression and find a very strong dynamic relationship between affective reactions and decisions: arousal at decision in round $t-1$ and choice in round t predict extremely well arousal at decision in round t. Based on that information, we build a dynamic theoretical model where stealing has a current monetary gain but a future (accumulated) immorality cost. We structurally estimate a reduced-form version of this model, where each subject makes decisions to keep the immorality stock as close as possible to an *immorality target.* Precisely, the individual prefers to steal (pay) when the current immorality stock is below (above) the target because the cost of choosing the reprehensible action is low (high). This raises (lowers) the immorality stock for the next period, which affects future behavior. Persistence in immorality accumulation is determined by a second parameter of the model, namely the *immorality depreciation* factor. The model fits well the observed behavior of our participants. Also, according to our estimation, individuals with similar behavior in both parts of the experiment display faster depreciation and a lower target compared to individuals who exhibit a treatment effect. This results in higher alternation between paying and stealing.

Summing up, the study demonstrates that social image is an important factor of choice in social games and an efficient deterrent of selfish play. More importantly, it shows that the underlying emotions we need to process in order to represent psychological and monetary consequences are integrated at the time of the decision and reflected in biological measures. This result is important in the context of behavioral models of social games. Some existing models already incorporate features to capture psychological or biological components. Still, the assumptions are largely untested, which limits the reliability of the predictions. By contrast, we can directly measure the relationship between behavior and physiology and use that information to build our model. The fact that the model is capable of predicting the dynamics of choice with reasonable accuracy implies that the study of biological markers can and should guide the modeling strategy whenever emotions affect decisions.

The article is organized as follows. In section 2, we position our paper within the existing literature. Section 3 describes the experiment. Section 4 presents the main behavioral results. Section 5 describes the physiological responses at decision time and feedback time. In section 6, we develop and test a behavioral model of decision-making affected by emotional states. Section 7 addresses a few final remarks and potential directions for future research.

2 Background and related literature

The idea that economic decisions have an emotional component is intuitive and the basis of numerous behavioral theories, with only a small fraction being reviewed here. Emotions may prevent us from taking risks, experiencing regret or feeling guilt (Loomes and Sugden, 1982; Loewenstein et al., 2001). Studies have shown that some individuals adhere to social norms and engage in prosocial behavior only to avoid suffering negative emotions (Tangney, 1995; Eisenberg, 2014) and that pre-play communication leads to more cooperation as a means to avoid feeling guilty (Charness and Dufwenberg, 2006). Also, cues have been found to affect prosocial choices in dictator games (Camerer and Fehr, 2004). Publicly implemented punishments tend to increase cooperation (Xiao and Houser, 2011), although shaming is not always effective (Van de Ven and Villeval, 2015). Finally, Xiao and Houser (2005) show that behavior is not only affected by emotions (with anger triggering rejection of offers) but also by the ability to express them (with relief triggering acceptance). Taken together, this evidence strongly suggests that emotions act as modulators of our motivations, triggered to represent social norms, memories of past experience, or anticipation of unwanted consequences.

As mentioned in the introduction, one difficulty of this research program is to be able to directly tie actions to emotions. The recent literature in economics has explored some creative methods to use biological markers as an objective indication of emotional arousal in the context of decision-making. Thinking about committing a norm violation may engage specific brain processes, trigger hormonal release, make us sweat more, make our heart beat faster or our pupil dilate, change our facial expressions or even change the way we pause to talk. Recent work has employed promising techniques that measure emotion-driven biological changes or their correlates, including brain imaging (Li et al., 2009), pupil dilation (Wang et al., 2010) and facial recognition (Van Leeuwen et al., 2018). The present paper joins a small literature in experimental economics that uses electrodermal activity to measure emotional arousal and use it to predict choices.

Skin conductance response (SCR) measures emotional arousal, but it does not indicate what kind of excitement the individual is experiencing: negative (embarrassment), positive (elation), neutral (surprise), etc. SCR has been extensively used in psychology and neuroscience to understand judgment and decision-making.[1] By contrast, it has received only limited attention in experimental economics. Coricelli et al. (2010) propose a tax evasion game and show that individuals with higher arousal at decision commit more fraud and commit fraud more often. Also, audits raise emotional responses and reduce evasion in the following period. Joffily et al. (2014) record emotions at four different stages in a voluntary contribution mechanism where punishments are and are not allowed. Subjects are more aroused if they contribute less and if other players free ride more. Arousal is also predictive of higher punishments in the treatment with sanctions. Kang and Camerer (2018) study games where individuals receive private signals about the value of an asset and choose when to sell. They show that when subjects cannot precommit to a strategy, they do not hold to their assets long enough. Decisions to sell excessively fast are correlated with emotional arousal, which is interpreted as a measure of anxiety. All three papers demonstrate a relationship between SCR and behavior in strategic games.

Our experiment has two major differences compared to this literature. First, we consider a much simpler game, where decisions are trivial in the absence of non-monetary considerations. While our setting is not as rich, it allows us to isolate the emotional impact of moral norms and social norms on decision-making. It also reduces the physiologically relevant measurements to two instances: decision time and feedback time.

Second, and most importantly, the experiment allows us to develop a novel biology-based theory that captures behavior more naturally than existing models. Indeed, an extensive theoretical literature posits that individuals have an intrinsic morality (or "honesty" or "cost of misbehaving", etc.) level that guides their decisions (see e.g., Tirole (1996); Carrillo (2000)). Since a fixed and unknown morality cannot rationalize the commonly observed alternance between moral and immoral behavior, the subsequent work has brought additional features. Examples include an imperfect self-knowledge about

[1] See e.g., Nikula (1991); Bechara et al. (2002); Crone et al. (2004); Nagai et al. (2004); Naqvi and Bechara (2006); Van't Wout et al. (2006); Figner et al. (2009); Reid and González-Vallejo (2009) out of a long list, as well as the selective survey by Figner and Murphy (2011).

one's morality and learning through self-experience (Bénabou and Tirole, 2011; Dal Bó and Terviö, 2013) or the interplay between brain systems with different access to information about morality (Brocas and Carrillo, 2019). Interestingly, the evidence reported here suggests an intuitive biological channel that can be captured by a simple dynamic optimization model. This model underscores the importance of tracing modeling features (such as morality) to biological markers instead of restricting theory to be an "as if" formalization.

3 Experimental design and procedures

The experiment was conducted in the Los Angeles Behavioral Economics Laboratory (LABEL) at the University of Southern California. It was programmed with the software z-Tree. In the experiment, 208 subjects were recruited via ORSEE. In the recruitment email, they were notified: "In this experiment, we may also record physiological measures (such as heart beat and skin conductance). These are non-invasive procedures. Please note that you cannot participate in the experiment if you do not accept these procedures." Each session had exactly eight subjects, randomly split into two even groups, with four *Consumers* and four *Producers*.

3.1 Non-choice measures

Emotional arousal. During the experiment, we collected biological data from Consumers using the Biopac MP150 and TEL100C physiological systems. We measured SCR, which is a proxy for emotional arousal. SCR measures the electrical conductance of the skin, which changes with minimal differences in moisture, as brought on by physiological arousal. A spike in SCR indicates that the subject is emotionally excited. SCR is considered a reliable measure of emotional arousal (Figner and Murphy, 2011; Boucsein, 2012; Dawson et al., 2017). Details of the recording procedure and technical aspects of the analysis are presented in Appendix A1.

Physical effort. We used the same Biopac physiological systems for an effort task that the Producers had to perform. We designed an effort task, which was physical, time-effective, unambiguously effortful, and individual-specific. We privately calibrated the strength of each Producer (without explanation of its purpose for the experiment) by telling them to squeeze a highly accurate hand dynamometer (Biopac MP3X) as hard as they could for 5 seconds. Three seconds into the task, they were reminded, "squeeze as hard as you can" to ensure maximum effort. We calculated an individual "grip threshold" as 50% of the average grip during the 5 second period. We recorded the threshold for use in the experiment but did not report it to the subjects. Unlike some other common effort inducement methods,[2] the task difficulty could be calibrated individually, making it closer to even difficulty for all subjects. It was physically (rather than cognitively) effortful and made intuitive sense to participants, who were told that they were putting effort into physically "creating" a good. Consumers also squeezed a disconnected hand dynamometer to experience for themselves the task the Producers would have to complete, and to be convinced of the effort and physical fatigue of the task.[3]

Picture. We took the photo of each Consumer, head only, against a neutral backdrop with a neutral expression.

3.2 Procedures

Once Producers were calibrated and Consumers had their photos taken, all 8 subjects were brought into the computer room. We seated them in a way that subjects were unable to see the faces or screens of individuals in the other role. We read aloud the instructions with screenshots on a projector, explaining the basics of SCR measurement. We cleaned each Consumer's left index and middle fingers of their non-dominant hand and had electrodes applied and attached. We also applied a long strip of gentle tape across their forearms and reminded them not to move that hand during the experiment.

In Part 1 of the experiment, we told each Producer to squeeze their hand dynamometer as hard as they could. After 7 seconds above their calibrated personal grip threshold, we moved to a screen informing them that they had created a digital good worth $20 to an anonymous Consumer they had been randomly matched with. If they did not exceed the threshold for 7 seconds, the software remained on the same screen telling them to squeeze as hard as they could. Producers did not received feedback regarding their threshold. The task was tiring and took, on average, 45 seconds for the group.

Once all Producers had completed the task, Consumers had two choices. They could

[2] For example, math operations (Corgnet et al., 2015), puzzles (Gneezy et al., 2003) or sliders (Gill and Prowse, 2012).

[3] To our knowledge, only Imas (2014) has used a hand dynamometer to record physical effort in an economics experiment. Charness et al. (2018) provide a comparison of the relative advantages of different real effort tasks.

"pay" \$7 to the Producer they were matched with, leaving them a net profit of \$13. Alternatively, they could "steal" the good, that is, pay \$0. If they chose to steal, there was a 40% chance they would not be caught, in which case they would obtain a net payoff of \$20. There was a 60% chance they would be caught and get a nominal \$1 fine, in which case they would obtain a net payoff of \$19, since they would still keep the good. The \$1 fine was then returned to the experimenter, not the Producer. There was no other consequence to stealing. Monetary payoffs were chosen in a way that an individual with no other-regarding or image concerns would unambiguously prefer to steal the good for any possible risk attitude. In particular, the fine had mostly a symbolic value. At the same time, we used loaded language in the instructions ("stealing", "caught", "fine", etc.) to elicit an emotional response that the physiological system could record.

Once all Consumers had made their decision, subjects saw their result screens. Consumers saw their own photo (regardless of their choice), their choice in that round and, if they stole the good, whether they were caught and charged a fine or not. Producers saw whether the Consumer they were matched with had stolen or paid for the good, since that decision determined their payoff. They saw no picture of Consumers. In case of stealing, they did not observe whether the Consumer was caught. These procedures were designed to minimize the social norm effect in this part. After everyone viewed their results and payoffs for the round, we proceeded to the next round, where Producers had to squeeze again the hand dynamometer to create a new digital good. Part 1 lasted 10 rounds.

After those 10 rounds, subjects were instructed that they were starting a new section of the experiment. In Part 2, the rules and consequences were identical, except for one change: any Consumer caught stealing would additionally have their photo displayed to Producers, along with a text indicating that they had been caught stealing. This change was common knowledge. To avoid peer considerations, Consumers could not see which other Consumers had cheated and/or had been caught, only whether they had been caught themselves. To maximize the exposure of being caught, the photos of the Consumers caught stealing were displayed to all Producers in the session. Part 2 lasted 20 rounds. We hypothesized that showing a photo when caught would (weakly) increase the emotional load of stealing. This, in turn, could potentially affect behavior.

After Part 2, we removed the electrodes from the Consumers and asked subjects to report their gender, which was used as a variable in our analysis. Subjects were paid their earnings in one randomly selected round from either Part 1 or Part 2 plus a \$5 show-up fee. The experiment adhered to the standard techniques in experimental economics, with a comprehension quiz and a practice round before each part. We did not tell participants in advance the number of rounds in each part to avoid last period effects. Sessions lasted on average 75 minutes and the earnings of Consumers and Producers averaged \$20.3 and \$7.7, respectively. Every Producer created the good every round, although it often took multiple tries, and up to 4 minutes. One session ended prematurely because one Producer was upset and refused to continue. Subjects in that session were all paid the show-up fee and the data was removed from the analysis.[4] This left us with data from 200 subjects, 100 Consumers and 100 Producers, collected in 25 sessions with 8 subjects each. Experimental instructions are included in Appendix B.

The entire analysis in the paper is performed on the 100 subjects who acted as Consumers. The role of Producers in our experiment is limited to "creating" the good and serving as real person counterparts to Consumers who can "pay" or "steal" from them. For the statistical analyses that involve repeated measures from each participant, we compute clustered bootstrapped standard errors.

4 Behavior

4.1 Choice over time

We first analyze the propensity of Consumers to "steal" the good over the course of the experiment. Figure 1 (left) presents the time series of behavior over the 30 rounds (10 in Part 1 and 20 in Part 2). Figure 1 (right) presents the cumulative distribution function (c.d.f.) of the fraction of rounds where the individual steals, separately for Parts 1 and 2.

Adding the picture of subjects who are caught has a very significant effect on behavior. Indeed, the percentage of observations in which Consumers steal the good drops from 78% in Part 1 to 58% in Part 2 (Figure 1 (left)). Using a Pettitt test of structural break with unknown break date, we find the existence of a break in the first round of Part 2 (p < 0.001). We also conducted a Mann-Kendall trend test in each part and found that the slopes within each part are not significantly different from zero. Overall, there is strong

[4] In retrospect, we should have given a higher show-up fee to Producers to compensate for their lower expected earnings. Also, our calibration was excessively demanding, and a 4 second (instead of 7 second) grip duration to create the good would have been effortful enough.

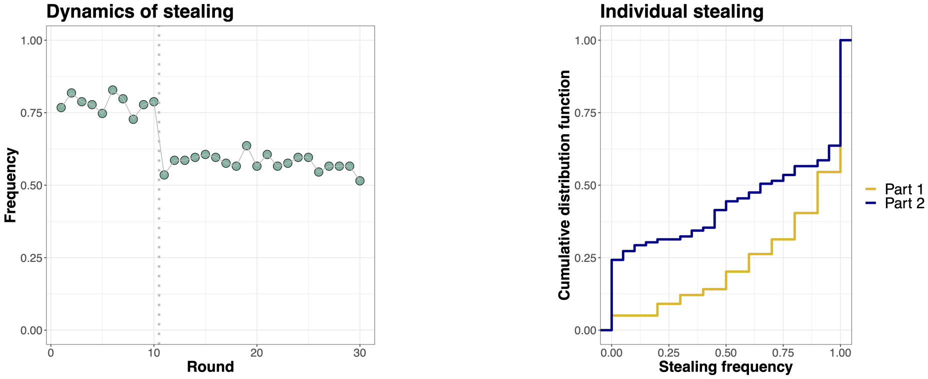

Figure 1: Stealing behavior without (Part 1) and with (Part 2) photo.

evidence of change in the aggregate rates of stealing between parts and no evidence of change within parts.[5]

A Wilcoxon rank sum test in Figure 1 (right) shows that the c.d.f. of the individual stealing frequency is significantly different between Parts 1 and 2 ($p < 0.001$). It reinforces the idea that stealing is substantially lower when we introduce the photo. The heterogeneity in behavior across individuals is also clear. For instance, in Part 2 there are 24% of individuals who never steal the good and 36% of individuals who always steal it.

4.2 Behavioral types

To further analyze heterogeneity across individuals, we group subject in each part into behavioral types. More precisely, we categorize an individual as “Always” [A] and “Never” [N] if he makes two or fewer deviations from always stealing and never stealing respectively, and we categorize him as “Sometimes” [S] otherwise.[6] Table 1 summarizes the number of subjects in each type, where the first and second letter in the brackets refer to their behavior in Parts 1 and 2, respectively. Unless otherwise stated, standard errors in all tables are presented in parenthesis.

[5] Ideally, one should randomize the order of Parts 1 and 2 across subjects to control for order effects. In our case, however, we preferred to keep the natural sequence of no social sanction followed by social sanction (that is, introducing one new element at the beginning of Part 2).

[6] We allowed 20% deviations in Part 1 and 10% in Part 2 because we figured that the first 1 or 2 decisions in the game might reflect some “testing the waters” or uncertainty about the consequences, and therefore not be representative of the subject’s overall intention. Small variations on the threshold do not affect the results significantly.

Type	Stealing fraction		number of
	Part 1	Part 2	subjects
[AA]	0.96 (0.01)	0.99 (0.01)	44
[SS]	0.50 (0.03)	0.50 (0.05)	11
[NN]	0.09 (0.04)	0.03 (0.02)	9
[AN]	0.95 (0.03)	0.00 (0.00)	10
[AS]	0.93 (0.02)	0.56 (0.04)	14
[SN]	0.57 (0.04)	0.02 (0.01)	11
[SA]	0.60	0.90	1
Total	0.78 (0.03)	0.58 (0.04)	100

Table 1: Behavioral types

We observe well-defined patterns of behavior in the population. Two-thirds of Consumers (64) behave similarly when the photo is not and is shown to Producers. Of these, the majority always steal [AA]. The rest are equally divided between those who never steal [NN] and those who steal sometimes [SS]. This latter group has similar stealing frequency in both parts of the experiment: none of these subjects changes behavior by more than 30% between Parts 1 and 2. The remaining one-third of subjects (36) exhibit a treatment effect. Of these, all but one decrease the frequency of stealing across parts: [AN], [AS] and [SN]. The changes between Parts 1 and 2 of these three types are statistically significant (t-tests of mean differences, p-value < 0.001). The remaining individual increases his stealing frequency in Part 2. We exclude this subject from the rest of the analysis, as (s)he is a clear outlier in our sample. Notice that the groups where subjects steal sometimes [S] have average stealing probabilities close to 0.5, with relatively low dispersion. This behavioral focal point is all the more interesting that it does not result in equal (or even similar) payoffs for Consumer and Producer.[7] In Appendix A2, we discuss in more detail the behavior of subjects who exhibit a treatment effect ([AN], [AS], [SN]), and show that the major difference is their significantly lower level of stealing in Part 2 (as opposed to a higher level of stealing in Part 1).

[7] Recall that the good is worth \$20, the price is \$7 and the fine is \$1 with probability 0.6. If the Consumer pays with probability 0.5, the expected monetary payoffs are 16.2 for the Consumer and 3.5 for the Producer. Even if the Consumer always pays, the payoffs would still be highly unequal (\$13 and \$7). We imposed this asymmetry to encourage paying for the good.

4.3 Summary

Overall, although the behavioral results are qualitatively in accordance with previous findings, the level of stealing in Part 1 is somewhat higher than we anticipated. Indeed, results in the dictator game suggest that two-thirds of individuals typically give a positive amount to recipients and the average amount shared is around 25% of the total (Engel, 2011). In our setting, with loaded language, effortful production and only two options (give 0% or 35%), we expected that participants would pay around half of the time in Part 1. Instead, they paid in only 22% of the observations, thereby sharing 7.7% of the value. On the other hand, showing the picture is more effective than we expected. This is reflected in the very significant decrease in stealing between Parts 1 and 2 (as reviewed in section 2 the existing literature finds mixed results on the effectiveness of public punishment). Finally, behavior is stable within each part of the experiment, with a majority of subjects stealing always, never, or half the time approximately. All in all, there is enough variation between subjects and conditions for an interesting analysis of the physiological data.

5 Physiology

5.1 Recording of electrodermal activity

There are two main events in our game, a *decision* event and a *feedback* event. Arousal at decision refers to the SCR while the subjects is deciding whether to steal the good or pay for it. It is a "flexible duration" event that goes from the moment where the choice appears on the screen until the (endogenously determined) time where the individual makes a selection. Arousal at feedback refers to SCR at the time the individual observes the result from his choice after stealing the good (caught or not caught). It is a "fixed" event with no specified duration, since the psychophysiological response is measured immediately after the feedback.

For flexible duration events, the software estimates three physiological measures: amplitude, latency and half-recovery time (Dawson et al., 2017). *Amplitude* refers to the phasic increase in conductance shortly following the stimulus onset, and it is measured in microsiemens (μS). *Latency* is the temporal interval between stimulus onset and SCR initiation, and it is measure in seconds (s). *Half-recovery time (HRT)* is the temporal interval between the SCR peak and the point of 50% recovery of SCR amplitude, and it is also measured in seconds (s). These measures help us capture how long it takes to respond to the stimulus (latency), how aroused the person is (amplitude), and how long the arousal lasts (HRT). For fixed events, the software only estimates amplitude and it assumes canonical parameters of latency and HRT. Figure 2 provides a generic graphical representation of this information.

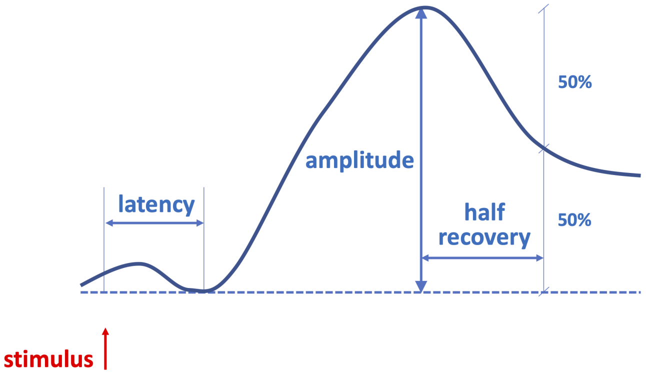

Figure 2: Representation of physiological measures.

For each individual, we also collected non-specific SCR data (as opposed to event-related SCRs) to compute a *Lability* index. This index is obtained by counting the number of SCRs during a fixed amount of time, in our case 2 minutes before the start of the experiment. The electrodermal response lability index is relatively consistent within individuals and it is associated with some personality traits. In particular, subjects with more SCRs ("labiles") tend to be emotionally undemonstrative, introverted, agreeable and calm. Subjects with fewer SCRs ("stabiles") tend to be emotionally expressive, extroverted, impulsive and antagonistic (Crider, 2008).[8] Research in psychology sometimes uses this classification to understand and predict behavior in social contexts (Raine et al., 1995; Lorber, 2004). To the best of our knowledge, the existing literature in economics has never looked at this measure.

5.2 Skin conductance response at decision and feedback

We first look at differences in SCR across treatments and events. Aggregating across events, we notice that maximum amplitude is significantly higher in Part 2 than in Part

[8] The inverse relationship between lability and the expression of emotions suggests that lability may reflect differences in the ability to control such expression.

1 (2.40 vs. 2.02, p = 0.005), which suggests a higher level of arousal when the photo is at stake. As we will see below, this is in large part because SCR is related to the subject's action, which itself evolves during the experiment. Also, the vast majority of subjects (79 in Part 1 and 78 in Part 2) are more aroused at decision than at feedback. This may be surprising since individuals receive new information only at feedback (and provided they steal the good). It means that the largest reaction is due to introspection. At the same time, the flexible duration (decision) event may have prompted more reaction than the fixed (feedback) event in our game simply because the anticipation of the physical act of selecting an alternative has a stronger impact on the emotional system.

We next look at correlation of the physiological reaction of individuals across parts and across measures using a standard Pearson correlation (PC) test. Reassuringly, there is a very strong positive correlation at the individual level between Parts 1 and 2 for all three measures of arousal at decision: amplitude (PC = 0.65, $p < 0.001$), latency (PC = 0.42, $p < 0.001$) and HRT (PC = 0.45, $p < 0.001$). There is also a very strong positive correlation of amplitude at feedback (PC = 0.83, $p < 0.001$). In Part 1, individuals with higher amplitudes at decision also respond faster (PC = -0.29, p = 0.004) and recover faster (PC = -0.44, $p < 0.001$). In Part 2, only the relationship between amplitude and recovery is significant (PC = -0.26, $p < 0.011$). Finally, there is also a positive correlation of amplitude at decision and amplitude at feedback, both in Part 1 (PC = 0.45, $p < 0.001$) and in Part 2 (PC = 0.38, $p < 0.001$). These initial findings indicate a substantial congruency between the different measures of arousal, but also interesting differences across parts and across events.

In section 4, we noted a large individual heterogeneity in behavior. We conjecture that this will map into large heterogeneity in arousal as well. To investigate this possibility, we grouped individuals by their behavioral type (see Table 1) and computed for each type and in each part the average amplitude (amp.), latency (lat.) and half-recovery time (HRT) at decision as well as the amplitude (amp.) at feedback (recall that for fixed events, the software only estimates amplitude). The results are presented in Table 2 (columns 2 to 5 refer to Part 1 and columns 6 to 9 refer to Part 2).

The amplitude at decision in Part 1 is higher for subjects who change behavior across parts ([AN], [AS], [SN]) than for those who do not ([AA], [SS], [NN]), with mean arousals of 0.90 and 0.46 respectively (p = 0.030). It suggests a relationship between emotional state and the decision to reduce or eliminate stealing in Part 2. Among subjects with a treatment

	decision (1)			feedback (1)	decision (2)			feedback (2)
	amp.	lat.	HRT	ampl.	amp.	lat.	HRT	ampl.
[AA]	0.45	3.14	1.57	0.27	0.51	2.24	1.10	0.27
	(0.06)	(0.26)	(0.11)	(0.05)	(0.07)	(0.25)	(0.09)	(0.05)
[SS]	0.44	2.98	1.36	0.23	0.66	2.37	0.93	0.33
	(0.07)	(0.50)	(0.15)	(0.04)	(0.18)	(0.45)	(0.07)	(0.06)
[NN]	0.50	4.05	1.72	0.32	0.35	2.11	1.28	0.26
	(0.17)	(1.38)	(0.36)	(0.09)	(0.11)	(0.41)	(0.19)	(0.08)
[AN]	1.30	2.39	1.31	0.38	0.91	2.19	0.98	0.22
	(0.51)	(0.45)	(0.33)	(0.10)	(0.20)	(0.25)	(0.18)	(0.07)
[AS]	0.92	2.31	0.97	0.29	0.54	1.84	0.86	0.27
	(0.24)	(0.40)	(0.12)	(0.06)	(0.10)	(0.24)	(0.06)	(0.06)
[SN]	0.51	3.04	1.73	0.26	0.60	2.18	1.08	0.20
	(0.24)	(0.08)	(0.37)	(0.06)	(0.20)	(0.31)	(0.17)	(0.05)
All	0.61	2.97	1.47	0.29	0.56	2.17	1.05	0.26
	(0.08)	(0.21)	(0.08)	(0.03)	(0.05)	(0.13)	(0.05)	(0.03)

Amplitude is measured in microsiemens (μS), latency and HRT are measured in seconds (s)

Table 2: Measures of arousal at decision and feedback

effect, we also found a decrease in amplitude between Parts 1 and 2, possibly affected by their change in behavior, although the difference is not significant at conventional levels ($p = 0.098$). As for the other measures (latency at decision, HRT at decision, and amplitude at feedback), we found no systematic differences across types and, in particular, there are no differences between participants who change behavior and those who do not. By contrast, we notice for all types significantly faster responses, faster recovery and lower amplitude at feedback in Part 2 than in Part 1 ($p < 0.05$). This may be due to SCR habituation, a phenomenon long established in the electrodermal literature (Dawson et al., 2017) and also present in previous research (Kang and Camerer, 2018). Given those findings, from now on we will focus our attention on amplitude, as the most relevant variable for the physiological analysis.

To further investigate the relationship between choice and arousal, we present in Figure 3 for each behavioral type the amplitude at decision in Parts 1 and 2 as a function of the choice (steal or pay) and the amplitude at feedback also in Parts 1 and 2 as a function of the outcome (caught or not caught).

The decrease in amplitude between Parts 1 and 2 emphasized in Table 2 is partly

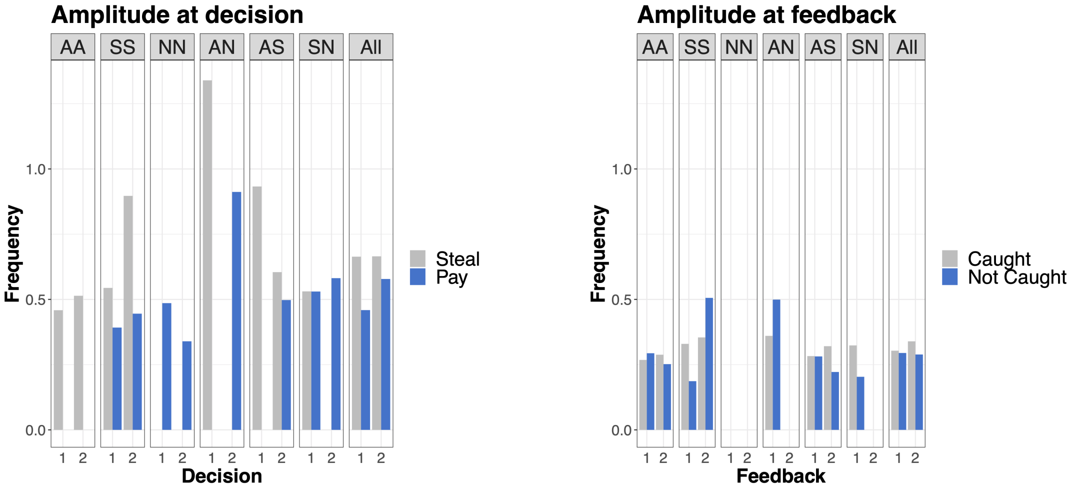

Figure 3: Amplitude at decision by behavioral type and choice (left) and amplitude at feedback by behavioral type and outcome (right) in Parts 1 and 2.

driven by the individuals with high arousal who steal in Part 1 and decide to steal less or never in Part 2 ([AS] and [AN]). Within the [SS] type, we also notice higher arousal when the subject steals and knows that the picture will be shown if he is caught. Overall, however, due to the reduced number of observations, the differences in arousal are not significant at conventional levels, neither in Part 1 ($p = 0.080$) nor in Part 2 ($p = 0.065$). In Appendix A3, we perform OLS regressions to study in more detail the effect of arousal and lability in the choice of individuals. The results confirm that the emotional response at decision in Part 1 is highly indicative of the stealing choice in Part 2.

Amplitudes at feedback are similar across all types in both parts of the experiment. Also, confirming previous results, arousal levels are significantly lower at feedback than at decision. There are no systematic differences in arousal when the individual is caught and not caught, even in Part 2 where stakes are higher. This is surprising, although it matches the previous behavioral result, which stated that the decision to steal does not depend on whether the individual was caught or not in the round before. There are three possible (non-competing) explanations for the absence of arousal differences. First, all arousal levels at feedback are small, hence relatively uninformative in our experiment. Second, there might be confounding effects: increased arousal when caught may reflect anxiety to the picture shown to Producers whereas increased arousal when not caught may reflect relief. Third, subjects make decision plans at the outset. The emotional consequences of stealing are then fully integrated at the time of decision.

We also study the round-by-round evolution of amplitude at decision in Parts 1 and 2. Figure 4 depicts this information separately for two groups of subjects: the individuals who change their behavior between Parts 1 and 2 ([AN], [AS], [SN]) and those who keep it constant ([AA], [SS], [NN]).

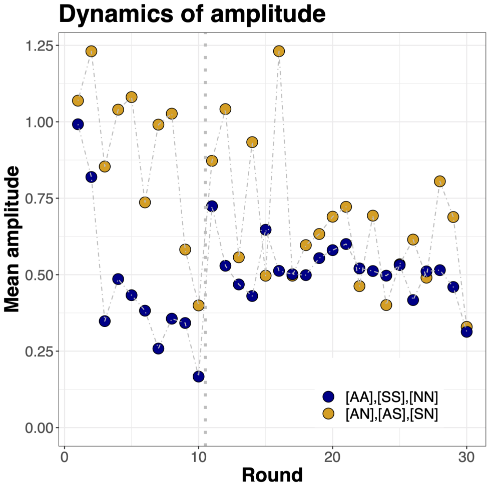

Figure 4: Evolution of amplitude at decision.

Figure 4 shows significant differences in arousal between individuals who reduce stealing when the picture is shown and those who do not. Both types of subjects start with high arousal in the first two rounds. However, subjects with constant behavior get habituated very fast, showing low and constant arousal from round 3 and until the end of the experiment. Using the Mann-Kendall test to determine if there is a time trend in amplitude, we find no trend between rounds 3 and 30 (p = 0.213) or even if we consider all rounds from 1 to 30 (p = 0.915). By contrast, subjects who steal less in Part 2 show a gradual decrease in amplitude during the entire experiment (p = 0.012 if we consider rounds 3 and 30 and p = 0.002 if we consider all rounds from 1 to 30). This is consistent with a habituation effect in arousal: with repeated exposure, the emotional consequence of choices gradually loses significance.[9] Also, arousal is higher on average among subjects who change their

[9] This, in turn, casts a warning flag regarding excessive repetition of decisions, and suggests that the first few measures of arousal within each part might be more indicative of the individual's feelings than the subsequent ones.

behavior, most notably in Part 1 as already discussed in Table 2 (0.90 vs. 0.46, p = 0.030) but also overall (0.78 vs. 0.49, p = 0.011). Interestingly, the volatility in amplitude at decision is also significantly higher among subjects who change their behavior than among subjects who do not (mean standard deviation 0.82 vs. 0.59, p = 0.018).

5.3 Electrodermal response lability

Last, we report the average lability score for each behavioral type. Following standard procedures, we classify individuals as labiles if their score is above the median, stabiles if their score is below the median and unclassified if their score is at the median. The information is reported in Table 3.

	Lability score	labiles	stabiles	unclassified
[AA]	12.5 (0.91)	20	21	3
[SS]	17.1 (2.31)	7	3	1
[NN]	12.8 (2.53)	6	3	0
[AN]	10.0 (1.84)	2	6	2
[AS]	13.5 (1.72)	7	7	0
[SN]	9.0 (1.74)	3	8	0
Total	12.5 (0.66)	45	48	6

Table 3: Lability scores and classification by type

Lability scores are similar across behavioral types. The only noticeable difference is that subjects who never steal when the picture is shown ([NN], [AN], [SN]) have typically lower lability scores than subjects who sometimes or always steal with the picture ([AA], [SS], [AS]), (p = 0.040). This fits with the personality traits described above: stabile individuals are emotionally demonstrative and impulsive, so they are more likely to refrain from stealing when the picture is shown than labile individuals. The effect, however, is small in magnitude.

5.4 Summary

The most interesting physiological responses occur at decision, and they are best captured by measures of amplitude. There are two distinct groups of subjects. For a large fraction of individuals, showing the picture has little to no impact on their behavior, although they feel the emotional load of stealing, especially if they do it sometimes and shaming is involved. For the remaining subjects, the threat of the picture substantially reduces the stealing rate. These subjects are more aroused in Part 1 than in Part 2, and also more aroused than the other subjects, suggesting that their high emotional reaction is linked to their decision to change behavior. Part of the sustained decrease in arousal over time can be linked to the change in behavior, from stealing (Part 1) to paying (Part 2). Finally, the physiological reaction to feedback is very modest. It suggests that choices result from stable preferences and that consequences are anticipated at the time of choice.

6 Unveiling the mechanisms of stealing

6.1 The dynamics of morality

In section 5, we have demonstrated that arousal at decision is typically higher when the subject steals than when the subject pays (Figure 3, left), suggesting that individuals feel the emotional load of engaging in the socially reprehensible action. We have also shown that subjects who are most aroused in Part 1 are also most likely to reduce stealing in Part 2 (Table 2). This implies that higher arousal levels are associated with the desire to decrease or stop stealing.

To further investigate the dynamic relationship between emotional arousal and choice, we perform a dynamic panel regression of amplitude at decision in round t as a function of amplitude at decision in $t-1$ and choice in t (steal = 1 and pay = 0). We include fixed effects to account for the fact that individual amplitude levels are likely to be correlated with unobserved individual characteristics. Also, we consider only individuals who exhibit variance in behavior in at least one part of the experiment ([AS], [SS], [SN]). Results are reported in Table 4.[10]

Amplitude levels at decision are strongly autoregressive, with a coefficient below 1 that indicates depreciation. The choice to steal in round t strongly correlates with amplitude in that round. Overall, Table 4 suggests a powerful dynamic relationship between choice and emotions.

In line with our previous argument regarding the literal interpretation of emotional arousal, we posit that amplitude is a biological marker of "immorality." We then consider

[10] The results of the regression are even stronger if we only consider individuals with variance in behavior in both parts ([SS]).

	Ampl.(t)
Ampl.(t-1)	0.194***
	(0.031)
Steal (t)	0.221**
	(0.070)
# Obs.	1080
R^2	0.02

** p = 0.01; *** p = 0.001

Table 4: Panel regression of amplitude in round t for types [AS], [SS] and [SN].

a dynamic theory of moral-based decision-making that captures the empirical relationship described above. Unfortunately, and as discussed in section 2, none of the existing models incorporates a dynamic link between choice and morality. We therefore develop a new framework where the psychological disutility when the individual engages in the reprehensible action follows a dynamic autoregressive process. In this model, the subject has a current stock of immorality that guides current choices, and those current choices affect the future stock of immorality. Formally, at each round t the individual is offered a good at a posted price x and decides whether to pay for it ($a_t = x$) or steal it ($a_t = 0$). The individual's stock of immorality at the end of round t, θ_t, follows a simple dynamic accumulation process:[11]

$$\theta_t = \alpha\theta_{t-1} + x - a_t \tag{1}$$

In words, the immorality stock depreciates at a per-round rate $\alpha \in (0, 1)$ as long as the individual pays for the good. Immorality is boosted each time the individual steals it. The increase is proportional to the difference between posted price and amount paid.

6.2 Predicting trial by trial behavior

Based on the morality accumulation process described above, we build and test a simple model of decision-making. Assume a per-round valuation v for the good, a discount factor δ, a cost of immorality $c(\theta)$ increasing and convex, and an initial stock of immorality

[11] This intertemporal link is perfectly anticipated by the individual. It is similar to models in the rational addiction literature (see e.g., Becker and Murphy (1988)), where current consumption affects the marginal cost of future consumption.

$\theta_0 = 0$. A rational, forward-looking, profit maximizing individual solves the following dynamic optimization problem:

$$\max_{a_t \in \{0,x\}} \quad \sum_{t=1}^{T} \delta^{t-1} \Big[v - a_t - c(\theta_{t-1}) \Big]$$
$$\text{s.t.} \quad \theta_t = \alpha\theta_{t-1} + x - a_t$$

As it is well-known, under certain parametric conditions, this class of problems has elegant solutions that can be analytically derived using recursive methods (Stokey, 1989).[12]

Given the restricted available data, we construct a behavioral reduced-form model that captures a similar relationship between choice and morality and can be fitted to the data. More precisely, we endow the subject with a simple strategy: in each round, he targets a fixed immorality stock m and decides to currently pay or steal in order to be as close as possible to the target. His behavior is therefore consistent with the following myopic optimization problem:

$$\max_{a_t \in \{0,x\}} \; v - \Big(\theta_t - m\Big)^2 \qquad \text{where} \;\; \theta_t = \alpha\theta_{t-1} + x - a_t$$

The optimal choice is then:

$$a_t^* = \begin{cases} 0 \;\; \text{if} \;\; -\Big(\alpha\theta_{t-1} + x - m\Big)^2 > -\Big(\alpha\theta_{t-1} - m\Big)^2 & \Rightarrow \;\; \theta_{t-1} < \theta^* \equiv \frac{2m-x}{2\alpha} \\ x \;\; \text{if} \;\; -\Big(\alpha\theta_{t-1} + x - m\Big)^2 < -\Big(\alpha\theta_{t-1} - m\Big)^2 & \Rightarrow \;\; \theta_{t-1} > \theta^* \equiv \frac{2m-x}{2\alpha} \end{cases}$$

This reduced-form model captures the main ingredients of the general theory. The individual steals the good ($a_t = 0$) when the current immorality stock is below a certain threshold θ^* (stealing has low moral cost), and pays for it otherwise. Stealing increases the stock for the next round, and therefore, the likelihood of surpassing the threshold. If it is surpassed, the next decision is to pay (stealing has a high moral cost), which then reduces the stock for the following round, and so on. The model has also some natural comparative statics predictions: $\partial\theta^*/\partial\alpha < 0$, $\partial\theta^*/\partial x < 0$ and $\partial\theta^*/\partial m > 0$. In words, the

[12] The problem is equivalent to choosing the action that solves the following recursive equation:

$$W_t(\theta_{t-1}) = \max \Big\{ v - c(\theta_{t-1}) + \delta W_{t+1}(\alpha\theta_{t-1} + x); v - x - c(\theta_{t-1}) + \delta W_{t+1}(\alpha\theta_{t-1}) \Big\}$$

where the first term of the maximization is the value function if the agent steal at t ($a_t = 0$) and the second term is the value function if the agent pays at t ($a_t = x$). With a convex cost $c(\theta)$, the solution typically involves a "target" stock θ_t^* at round t such that $a_t^* = x$ if $\theta_{t-1} > \theta_t^*$ and $a_t^* = 0$ if $\theta_{t-1} \leq \theta_t^*$. Under certain regularity assumptions, the target converges to a steady state.

threshold θ^* is lower (that is, the individual is more likely to pay in a given period) when immorality depreciates slowly (α high) since it means that the moral stigma of stealing persists longer. Paying is also more likely when stealing has a higher impact on the future moral stock (x high) and, by construction, when the target is lower (m low).

To test this theory, we consider a random utility model where each subject chooses between paying ($a_t = x$, which in the estimation is set to 7) and stealing ($a_t = 0$). The subject derives the following utility of action a_t (individual subscripts are omitted for compactness):

$$U_t(a_t) = v - \Big(\alpha\theta_{t-1} + x - a_t - m\Big)^2 + \varepsilon^t_{a_t}$$

We assume that the error terms are independent and identically distributed and follow an extreme value distribution with parameter γ (> 0).[13] For a given γ, the probability of choosing action $a_t = 0$ in round t, $\lambda(0, t)$, has the following logistic functional form (see Appendix A4 for details of the derivation):

$$\lambda(0,t) = \frac{1}{1 + e^{-\gamma\left[-\left(\alpha\theta_{t-1}+x-m\right)^2+\left(\alpha\theta_{t-1}-m\right)^2\right]}}$$

For a given α and the empirical series of choices $\{a_t\}_{t=1}^T$, there exists a unique series of immorality stocks $\{\theta_t\}_{t=1}^T$. The problem becomes standard and consists of structurally estimating by maximum likelihood γ and m for each individual using the series of moral stocks as the independent variables. The procedure is the following. We compute the overall likelihood of the model given a fixed α. Varying α generates a different (and unique) series of immorality stocks θ_t. Using these new stocks as independent variables, we compute the corresponding estimates of γ and m for each individual i, as well as the likelihood of the model. The estimates $(\hat{\alpha}_i, \hat{\gamma}_i, \hat{m}_i)$ are obtained by comparing the likelihoods obtained for all possible α.

Notice that many subjects never steal [N] or always steal [A] within a given part. For those individuals, the problem is not identified since there is a large range of parameter combinations that lead to such uniform behavior. We only estimate the parameters for the subjects with interior stealing probabilities [S]. Table 5 summarizes $(\hat{\gamma}, \hat{\alpha}, \hat{m})$, the average estimates of individuals within a behavioral type in Parts 1 and 2 of the experiment. It also includes the fraction of choices correctly predicted by the model as well as the average predicted payment to producers and its empirical counterpart.

[13] The cumulate distribution function of the error term is $F_i(\varepsilon^t_{a_t}) = \exp(-\exp(-\gamma\,\varepsilon^t_{a_t}))$ with $a_t \in \{0, x\}$.

	Estimates			correct	Average payment	
	$\hat{\gamma}$	$\hat{\alpha}$	$\hat{m}$	classifications	Predicted	Empirical
Part 1						
[SS]	0.38	0.34	5.86	0.80	3.96	3.25
	(0.22)	(0.09)	(0.86)	(0.04)	(0.45)	(0.27)
[SN]	0.99	0.41	8.05	0.87	2.55	3.04
	(0.35)	(0.08)	(1.09)	(0.03)	(0.67)	(0.41)
Part 2						
[SS]	0.34	0.23	4.73	0.70	3.48	3.45
	(0.22)	(0.06)	(0.57)	(0.03)	(0.78)	(0.36)
[AS]	0.33	0.38	7.43	0.74	3.24	3.05
	(0.18)	(0.06)	(1.44)	(0.03)	(0.68)	(0.29)

Table 5: Structural estimates and average payments

The model works well in Part 1, with 80% to 87% of choices correctly predicted. Subjects are driven by a moral stock that depreciates relatively fast ($\hat{\alpha}$ low), which results in frequent action switches. Individuals in [SN] have a higher target $\hat{m}$ than those in [SS], and therefore steal the good more frequently, as reflected in the lower average amounts paid to producers. Still, average stealing probabilities are not far from 50% in either case (which would correspond to a 3.5 average payment to producers). Notice that the model may be overfitted since we estimate three parameters given ten binary observations.

Results are also encouraging in Part 2: although the predictive accuracy is somewhat smaller (70% to 74%), it is based on twenty observations with the same three parameters, and the predicted payments to producers are within 6% of the empirical averages. We conclude that the model also performs reasonably well. Finally, the structural model predicts faster depreciation and lower targets–therefore more frequent switches between pay and steal–for individuals with interior behavior in both parts ([SS]) than for those whose behavior is constant in one part ([SN] and [AS]).

Some final comments are in order. We have built a model based on the empirical relationship between choices and affective states unveiled by our laboratory experiment. It is therefore not surprising that it performs well when fit to that same data, with stable and consistent estimates. Even though this is not a guarantee that it is the best model, and tests on new samples are required to conclude with confidence, the good fit is encouraging evidence that the channel emphasized in this paper is at work. Importantly, the model helps organize the data and better understand the relationship between norms, emotions and choices. At an exploratory level, the theoretical model proposes a causal relationship, where immoral decisions inherited from the past materialize in emotions that influence present decisions to comply with the norm. This decision in turn affects the stock of immoral decisions a future self inherits.

7 Conclusion

We have reported the results of an experiment that investigates physiological correlates of behavior under social and moral norms. We have found large heterogeneity in behavior but no (behavioral or emotional) effect of being caught stealing. Not surprisingly, stealing decreases significantly when social sanctions are introduced. Thus, norm compliance is enforced by shaming when not by punishment. The results are consistent with studies showing that knowledge by other players of one's intentions increases the incentives to comply with social norms whereas the ability to hide such intentions are associated with selfish play (Dana et al., 2006, 2007). Overall, our study shows in an extremely simple setting the importance of self-image in promoting prosocial behavior.

The physiological response of our participants is strongly indicative of their behavior. In particular, among subjects who always steal under no shaming, those who are more aroused when making their stealing decision are also more likely to stop stealing when shaming is introduced. By contrast, among subjects who sometimes steal in the first scenario, those who are intrinsically more impulsive are more likely to stop stealing in the second scenario. This suggests that not only *event-related emotions* (arousal at decision) but also *intrinsic emotions* (lability index) are indicative of choices in contexts where social norms are important. While the link between emotional predispositions and choice seems natural–almost trivial–we are not aware of any previous decision-making experiment that documents this connection. Further research on the relationship between emotions at rest and decision in moral contexts should be very valuable. More generally, the study provides additional support for the documented relationship between emotions and decisions. Yet, the mechanism that connects biology and social norms to individual traits and moral judgment is still imperfectly understood. In particular, is it the case that the biological expressions–our emotions–modulate our acquired social norms, or that social norms modulate our instinctive biologically-driven responses?

On this matter, the causal relationship between biology and choice is a fascinating but largely unresolved issue. Our exploratory model provides a modest step in that direction. Indeed, based on the autoregressive nature of arousal, we have built a theory where the current stock of immorality affects the current decision to comply with the norm, which itself influences the future stock of immorality. Our empirical test supports this dynamic relationship. However, using the data that inspires a model to test the predictions of that model is unsatisfactory. Future research should focus on collecting independent data to test this theory. Fortunately, the model is rich and can accommodate multiple variants. Most notably, new experiments could vary the price of the good (x) and allow interior payments, that is, the possibility of partial stealing ($a_t \in [0, x]$ instead of $a_t \in \{0, x\}$). Our model predicts a preference for partial stealing over alternance between no stealing and full stealing, and can be contrasted with the actual behavior of participants. It would be also interesting to extend the theory and parametrize the immorality cost function ($c(\theta)$ in the general model and $-(\theta - m)^2$ in the reduced-form version). Since it is reasonable to assume a higher parametric cost under public shaming, the cost differential could be estimated for individuals who participate in both treatments. Furthermore, one could determine the effects of cost variations both theoretically (predictions) and empirically. For instance, an increase would certainly result in lower rates of stealing. It may also result in more subtle effects such as a higher frequency of alternation between norm compliance and stealing. Overall, a structural model that links directly biology and choice can significantly improve our current understanding of norm-based behavior.

Finally, researchers often think of emotions either as states, such as anger or stress, that may alter otherwise non-related decisions (Garfinkel et al., 2016), or as by-products of feedback, such as relief or disappointment (Moretti et al., 2010). In our study, we show that their role as a byproduct of choice at the feedback stage is particularly limited. By contrast, their role as modulator of decisions is more significant. This is consistent with studies in neuroscience which show that regions of the brain (in particular, in the prefrontal cortex) integrate emotions into decision-making at the time of choice (Coricelli et al., 2007) and that damage in those regions produce a lack of emotional processing coupled with suboptimal choices (Bechara et al., 1999). It tells us that economic objects such as utility can be measured: far from being abstract ‘as if’ concepts, they have a biological foundation.

References

Dominik R Bach. A head-to-head comparison of scralyze and ledalab, two model-based methods for skin conductance analysis. *Biological psychology*, 103:63–68, 2014.

Dominik R Bach, Jean Daunizeau, Karl J Friston, and Raymond J Dolan. Dynamic causal modelling of anticipatory skin conductance responses. *Biological psychology*, 85 (1):163–170, 2010.

Antoine Bechara, Hanna Damasio, Antonio R Damasio, and Gregory P Lee. Different contributions of the human amygdala and ventromedial prefrontal cortex to decision-making. *Journal of neuroscience*, 19(13):5473–5481, 1999.

Antoine Bechara, Sara Dolan, and Andrea Hindes. Decision-making and addiction (part ii): myopia for the future or hypersensitivity to reward? *Neuropsychologia*, 40(10): 1690–1705, 2002.

Gary S Becker and Kevin M Murphy. A theory of rational addiction. *Journal of political Economy*, 96(4):675–700, 1988.

Roland Bénabou and Jean Tirole. Identity, morals, and taboos: Beliefs as assets. *The Quarterly Journal of Economics*, 126(2):805–855, 2011.

Wolfram Boucsein. *Electrodermal activity*. Springer Science & Business Media, 2012.

Isabelle Brocas and Juan D Carrillo. A neuroeconomic theory of (dis) honesty. *Journal of Economic Psychology*, 71:4–12, 2019.

Colin F Camerer and Ernst Fehr. Measuring social norms and preferences using experimental games: A guide for social scientists. *Foundations of human sociality: Economic experiments and ethnographic evidence from fifteen small-scale societies*, 97:55–95, 2004.

Juan D Carrillo. Corruption in hierarchies. *Annales d'Economie et de Statistique*, pages 37–61, 2000.

Tara M Chaplin and Amelia Aldao. Gender differences in emotion expression in children: A meta-analytic review. *Psychological bulletin*, 139(4):735, 2013.

Gary Charness and Martin Dufwenberg. Promises and partnership. *Econometrica*, 74(6): 1579–1601, 2006.

Gary Charness, Uri Gneezy, and Austin Henderson. Experimental methods: Measuring effort in economics experiments. *Journal of Economic Behavior & Organization*, 149: 74–87, 2018.

Brice Corgnet, Roberto Hernán-González, and Stephen Rassenti. Firing threats: Incentive effects and impression management. *Games and Economic Behavior*, 91:97–113, 2015.

Giorgio Coricelli, Raymond J Dolan, and Angela Sirigu. Brain, emotion and decision making: the paradigmatic example of regret. *Trends in cognitive sciences*, 11(6):258–265, 2007.

Giorgio Coricelli, Mateus Joffily, Claude Montmarquette, and Marie Claire Villeval. Cheating, emotions, and rationality: an experiment on tax evasion. *Experimental Economics*, 13(2):226–247, 2010.

Andrew Crider. Personality and electrodermal response lability: An interpretation. *Applied Psychophysiology and Biofeedback*, 33(3):141, 2008.

Eveline A Crone, Riek JM Somsen, Bert Van Beek, and Maurits W Van Der Molen. Heart rate and skin conductance analysis of antecendents and consequences of decision making. *Psychophysiology*, 41(4):531–540, 2004.

Ernesto Dal Bó and Marko Terviö. Self-esteem, moral capital, and wrongdoing. *Journal of the European Economic Association*, 11(3):599–633, 2013.

Jason Dana, Daylian M Cain, and Robyn M Dawes. What you don't know won't hurt me: Costly (but quiet) exit in dictator games. *Organizational Behavior and human decision Processes*, 100(2):193–201, 2006.

Jason Dana, Roberto A Weber, and Jason Xi Kuang. Exploiting moral wiggle room: experiments demonstrating an illusory preference for fairness. *Economic Theory*, 33(1): 67–80, 2007.

Michael E Dawson, Anne M Schell, and Diane L Filion. The electrodermal system. In John T. Cacioppo, Louis G. Tassinary, and Gary G. Berntson, editors, *Handbook of Psychophysiology*, chapter 7, pages 159–181. Cambridge University Press, 2017.

Nancy Eisenberg. *Altruistic emotion, cognition, and behavior (PLE: Emotion)*. Psychology Press, 2014.

Nicole M Else-Quest, Ashley Higgins, Carlie Allison, and Lindsay C Morton. Gender differences in self-conscious emotional experience: A meta-analysis. *Psychological bulletin*, 138(5):947, 2012.

Jon Elster. Emotions and economic theory. *Journal of economic literature*, 36(1):47–74, 1998.

Jon Elster. Norms. In Peter Bearman and Peter Hedstrom, editors, *The Oxford handbook of analytical sociology*, chapter 9. Oxford University Press, 2011.

Christoph Engel. Dictator games: A meta study. *Experimental Economics*, 14(4):583–610, 2011.

Bernd Figner and Ryan O Murphy. Using skin conductance in judgment and decision making research. *A handbook of process tracing methods for decision research*, pages 163–184, 2011.

Bernd Figner, Rachael J Mackinlay, Friedrich Wilkening, and Elke U Weber. Affective and deliberative processes in risky choice: age differences in risk taking in the columbia card task. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 35 (3):709, 2009.

Sarah N Garfinkel, Emma Zorab, Nakulan Navaratnam, Miriam Engels, Núria Mallorquí-Bagué, Ludovico Minati, Nicholas G Dowell, Jos F Brosschot, Julian F Thayer, and Hugo D Critchley. Anger in brain and body: The neural and physiological perturbation of decision-making by emotion. *Social cognitive and affective neuroscience*, 11(1):150–158, 2016.

David Gill and Victoria Prowse. A structural analysis of disappointment aversion in a real effort competition. *American Economic Review*, 102(1):469–503, 2012.

Uri Gneezy, Muriel Niederle, and Aldo Rustichini. Performance in competitive environments: Gender differences. *The quarterly journal of economics*, 118(3):1049–1074, 2003.

Alex Imas. Working for the “warm glow”: On the benefits and limits of prosocial incentives. *Journal of Public Economics*, 114:14–18, 2014.

Mateus Joffily, David Masclet, Charles N Noussair, and Marie Claire Villeval. Emotions, sanctions, and cooperation. *Southern Economic Journal*, 80(4):1002–1027, 2014.

Min Jeong Kang and Colin Camerer. Measured anxiety affects choices in experimental “clock” games. *Research in Economics*, 72(1):49–64, 2018.

Jian Li, Erte Xiao, Daniel Houser, and P Read Montague. Neural responses to sanction threats in two-party economic exchange. *Proceedings of the National Academy of Sciences*, 106(39):16835–16840, 2009.

George F Loewenstein, Elke U Weber, Christopher K Hsee, and Ned Welch. Risk as feelings. *Psychological bulletin*, 127(2):267, 2001.

Graham Loomes and Robert Sugden. Regret theory: An alternative theory of rational choice under uncertainty. *The economic journal*, 92(368):805–824, 1982.

Michael F Lorber. Psychophysiology of aggression, psychopathy, and conduct problems: a meta-analysis. *Psychological bulletin*, 130(4):531, 2004.

Laura Moretti, Irene Cristofori, and Angela Sirigu. Value functions incorporating disappointment and regret. *Wiley Encyclopedia of Operations Research and Management Science*, 2010.

Yoko Nagai, Hugo D Critchley, Eric Featherstone, Michael R Trimble, and Raymond J Dolan. Activity in ventromedial prefrontal cortex covaries with sympathetic skin conductance level: a physiological account of a “default mode” of brain function. *Neuroimage*, 22(1):243–251, 2004.

Nasir H Naqvi and Antoine Bechara. Skin conductance: A psychophysiological approach to the study of decision making. *Methods in mind*, pages 103–122, 2006.

Reiner Nikula. Psychological correlates of nonspecific skin conductance responses. *Psychophysiology*, 28(1):86–90, 1991.

Adrian Raine, Peter H Venables, and Mark Williams. High autonomic arousal and electrodermal orienting at age 15 years as protective factors against criminal behavior at age 29 years. *The American journal of psychiatry*, 1995.

Aaron A Reid and Claudia González-Vallejo. Emotion as a tradeable quantity. *Journal of Behavioral Decision Making*, 22(1):62–90, 2009.

Nancy L Stokey. *Recursive methods in economic dynamics*. Harvard University Press, 1989.

June Price Tangney. Recent advances in the empirical study of shame and guilt. *American Behavioral Scientist*, 38(8):1132–1145, 1995.

Jean Tirole. A theory of collective reputations (with applications to the persistence of corruption and to firm quality). *The Review of Economic Studies*, 63(1):1–22, 1996.

Jeroen Van de Ven and Marie Claire Villeval. Dishonesty under scrutiny. *Journal of the Economic Science Association*, 1(1):86–99, 2015.

Boris Van Leeuwen, Charles N Noussair, Theo Offerman, Sigrid Suetens, Matthijs Van Veelen, and Jeroen Van De Ven. Predictably angry—facial cues provide a credible signal of destructive behavior. *Management Science*, 64(7):3352–3364, 2018.

Mascha Van't Wout, René S Kahn, Alan G Sanfey, and André Aleman. Affective state and decision-making in the ultimatum game. *Experimental brain research*, 169(4):564–568, 2006.

Joseph Tao-yi Wang, Michael Spezio, and Colin F Camerer. Pinocchio's pupil: using eyetracking and pupil dilation to understand truth telling and deception in sender-receiver games. *American Economic Review*, 100(3):984–1007, 2010.

Erte Xiao and Daniel Houser. Emotion expression in human punishment behavior. *Proceedings of the National Academy of Sciences*, 102(20):7398–7401, 2005.

Erte Xiao and Daniel Houser. Punish in public. *Journal of Public Economics*, 95(7-8): 1006–1017, 2011.

Appendix A

Appendix A1. Technical details of SCR.

Recording. Data was collected using the AcqKnowledge software (version 4.3). Events were placed at least 7 seconds apart to allow skin conductance levels to return to their baselines (Figner and Murphy, 2011). The acquisition rate was set to 2000Hz while the channel sample rate was set to 1000Hz. Gain on the Biopac system was set to ×2000.

Data analysis. Skin conductance data was analyzed using SCRalyze (Bach et al., 2010; Bach, 2014). It was trimmed to include the entire period from initial instructions through end of the final round. Raw data was processed using a high-pass Butterworth filter with cutoff frequency 0.0159Hz and a low-pass filter with cutoff frequency 5Hz. SCRalyze uses model-based analysis to understand sudomodor nerve (SN) activity based on observed SCRs. Sympathetic nervous system arousal leads to SN activity, which in turn leads to SCRs. SCRalyze uses an inversion model, which observes SCRs and infers SN activity (the state of interest) from them, with an understanding of the processes which lead from sympathetic arousal to SN activity and from SN activity to SCRs. Specifically, it assumes that an event-related sympathetic arousal causes a Gaussian SN burst, after some delay. It assumes a canonical shape of the SCR function (based on over 1,000 recorded SCRs).

Appendix A2. Behavioral changes between Part 1 and Part 2.

To study in more detail the structural change between Parts 1 and 2, we report in Figure 5 the time series of choices among the subjects who change significantly their behavior between no shaming and shaming.

By construction, the difference across parts is exacerbated in this subsample, with stealing probabilities of 82% and 23% in Parts 1 and 2, respectively. In Part 1, participants who exhibit and do not exhibit a treatment effect behave similarly. By contrast, stealing in Part 2 is significantly lower for participants with a treatment effect. Formally, a two-sided t-test comparison of differences in the average stealing probabilities of subjects with and without a treatment effect reveals statistically significant differences in Part 2 ($p < 0.001$) but not in Part 1 ($p = 0.240$). Surprisingly, we find no systematic effect of feedback (caught vs. not caught) on the behavior of individuals in the round immediately after, either in Part 1 or in Part 2. Finally, there is an interesting gender effect: the percentage decrease in stealing between Parts 1 and 2 is higher in female than in male participants

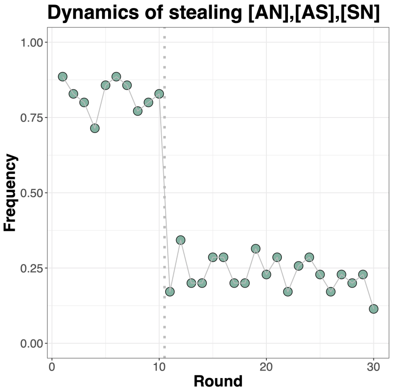

Figure 5: Stealing behavior of subjects with a treatment effect ([AN], [AS], [SN])

(26.5% vs. 11.2%, p = 0.018). The effect is mostly driven by the fact that 9 of the 10 individuals with the most severe reaction to the treatment ([AN]) are females. The result is consistent with the existing evidence of a systematic gender difference in the experience of shame (Else-Quest et al., 2012; Chaplin and Aldao, 2013).

Appendix A3. Using physiology to predict changes in behavior.

As indicated in section 5.2, participants who decrease their stealing rate when the picture is shown are more aroused in Part 1 than those who do not change their behavior. To further investigate this relationship, we present in Table 6 a number of regressions at the individual level. We first run OLS regressions on the entire population, where the dependent variable is the stealing probability in Part 2 (column (1), *steal2*) and the difference in stealing probabilities between Parts 1 and 2 (column (2), *steal-diff*). We then run Probit regressions where the dependent variable takes value 1 if the individual significantly decreases stealing between Part 1 and Part 2 (that is, if he belongs to types [AN], [AS] or [SN]) and 0 otherwise. The first Probit regression is performed in the entire population (column (3), [all]). Since we know from Table 2 that amplitude is not uniformly high among all subjects who reduce stealing, we perform the same Probit regression separately, only with subjects who always steal in Part 1 (column (4), [A -]) and only with subjects who sometimes steal in Part 1 (column (5), [S -]). In all cases, the independent physiological variables are the average amplitudes at decision and feedback in Part 1 (*decision-1* and *feedback-1*) and the non-event related lability score (*lability*). For column (1), we also include the stealing probability in Part 1 (*steal-1*).

	OLS		Probit		
	steal-2	*steal-diff*	[all]	[A -]	[S -]
steal-1	0.917***	–	–	–	–
	(0.111)				
decision-1	-0.143**	0.147**	0.756**	0.885*	0.536
	(0.048)	(0.047)	(0.293)	(0.366)	(0.936)
feedback-1	0.050	-0.051	-0.404	-0.471	0.088
	(0.127)	(0.127)	(0.587)	(0.714)	(3.099)
lability	0.009	-0.009	-0.044*	-0.013	-0.138*
	(0.005)	(0.005)	(0.022)	(0.028)	(0.057)
const.	-0.173	0.237**	-0.160	-0.651	1.455
	(0.114)	(0.075)	(0.309)	(0.394)	(0.868)
# obs.	99	99	99	68	22
adj. R^2	0.420	0.091	–	–	–
AIC	–	–	123.8	85.9	30.0

(significance levels: * = 5%, ** = 1%, *** = 0.1%.)

Table 6: Effect of arousal on stealing in behavior

From the OLS and Probit regressions in the full sample (columns (1), (2) and (3)), we notice that both stealing in Part 2 as well as the change in behavior between Parts 1 and 2 are related to arousal at decision but not at feedback. In other words, and reinforcing the findings in Table 2, emotional responses at the time of making the choices in Part 1 are very highly predictive of the reaction of individuals to the possibility of being shamed (remember that whether shaming actually occurs turns out not to be relevant). Naturally, the choice in Part 1 is also highly predictive of the choice in Part 2. From columns (4) and (5), we can see that the result is driven by subjects who always steal in Part 1: in the [A -] subsample, subjects with a low arousal amplitude are predicted to continue stealing in Part 2 while those with large arousals are predicted to decrease or stop stealing. By contrast, among the subjects who steal frequently but not always in Part 1 ([S -]), lability score is a better predictor of the treatment effect than amplitude at decision: individuals who are calm and emotionally undemonstrative (labiles) do not change their behavior when the picture is shown whereas individuals who are emotionally expressive and impulsive (stabiles) react significantly to it.

Appendix A4. Likelihood function.

Denote by $\Pr(a_t, t)$ the probability of choosing action a_t in round t, with $a_t = 0$ for stealing and $a_t = x$ for paying. We have:

$$\begin{aligned}\Pr(0,t) = \Pr[U_t(0) > U_t(x)] &= \Pr\Big[-\Big(\alpha\theta_{t-1} + x - m\Big)^2 + \varepsilon_0^t > -\Big(\alpha\theta_{t-1} - m\Big)^2 + \varepsilon_x^t\Big] \\ &= \Pr\Big[\varepsilon_x^t - \varepsilon_0^t < -\Big(\alpha\theta_{t-1} + x - m\Big)^2 + \Big(\alpha\theta_{t-1} - m\Big)^2\Big]\end{aligned}$$

Assume that the error terms are i.i.d. and follow an extreme value distribution, that is, the cumulative distribution function of the error term is $F_i(\varepsilon_{a_t}^t) = \exp(-\exp(-\gamma\,\varepsilon_{a_t}^t))$ with $\gamma > 0$ for all $a_t \in \{0, x\}$. Under these assumptions, for any γ, the probability of choosing option $a_t = 0$ in round t is the logistic function:

$$\begin{aligned}\lambda(0,t) &= \frac{e^{\gamma\left(U_t(0)\right)}}{e^{\gamma\left(U_t(0)\right)} + e^{\gamma\left(U_t(x)\right)}} = \frac{1}{1 + e^{-\gamma\left(U_t(0) - U_t(x)\right)}} \\ &= \frac{1}{1 + e^{-\gamma\left(-\left(\alpha\theta_{t-1}+x-m\right)^2 + \left(\alpha\theta_{t-1}-m\right)^2\right)}}\end{aligned}$$

which can be simplified into

$$\lambda(0,t) = \frac{1}{1 + e^{-\gamma\left(2m - 2\alpha\theta_{t-1} - x\right)x}}$$

From there, we can write a likelihood function and estimate both γ and m. Fixing α, the data consists in $a_t \in \{0, x\}$ and projected $\theta_t = \alpha\theta_{t-1} + x - a_t$.

The likelihood is:

$$\Pi_{t=0}^{T}\Big[\lambda(0,t)\mathbf{1}_{a_t=0} + \big(1 - \lambda(0,t)\big)\mathbf{1}_{a_t=x}\Big]$$

$$\theta_t = \alpha\theta_{t-1} + x\mathbf{1}_{a_t=0}$$

where $\mathbf{1}_{condition}$ is an indicator function that takes value 1 if the condition is satisfied.

Appendix B. Instructions

[add instructions here]