Deterrence of Unwanted Behavior: a Theoretical and Experimental Investigation[1]

Penélope Hernández[2] Zvika Neeman[3] Ro'i Zultan[4]

September 1st, 2023

[1] Acknowledgments to be added.

[2] ERI-CES Universidad de Valencia

[3] Economics, Tel Aviv University

[4] Economics, Ben-Gurion University of the Negev

Abstract

Suppose that spreading enforcement resources uniformly across time and space allows sanctioning anyone who engages in an unwanted activity with probability p. However, by concentrating enforcement resources, it is possible to split the probability p into a higher probability of sanction $p_H > p$ in some targeted areas or times, at the expense of a lower probability of sanction $p_L < p$ elsewhere. If the objective is to minimize the overall level of the socially unwanted activity, irrespective of its specific location or time, does splitting the probability of sanction p help achieve this goal?

We present a theoretical model of this situation, and undertake an experiment that allows us to answer this question empirically. Since the idea of beneficial splitting of prior beliefs is central to Bayesian persuasion literature, our investigation presents an experimental investigation into whether Bayesian persuasion can indeed yield practical benefits in a realistic parametrized setting.

1 Introduction

Suppose that the chief of police in a certain town aims to deter crime, or some other socially unwanted activity. Suppose that spreading enforcement resources uniformly across time and space allows sanctioning anyone who engages in the unwanted activity with probability p. However, by concentrating enforcement resources, it is possible to split the probability p into a higher probability of sanction $p_H > p$ in some targeted areas or times, at the expense of a lower probability of sanction $p_L < p$ elsewhere.[1] If the objective is to minimize the overall level of the socially unwanted activity, irrespective of its specific location or time, does splitting the probability of sanction p help achieve this goal?

We present a theoretical model that describes this situation, and undertake an experiment that allows us to answer this question empirically. Since the idea of beneficial splitting of prior beliefs is central to Bayesian persuasion literature, our investigation presents an experimental investigation into whether Bayesian persuasion can indeed yield practical benefits in a realistic parametrized setting.

Specifically, we consider a model with a large number of individuals. Each individual faces a choice between a benign and a socially unwanted action. For example, individuals may choose between parking legally and illegally. We assume that the benign action generates a certain payoff for the individual. By contrast, the socially unwanted action induces a risky binary lottery, whose outcome depends on whether the individual is sanctioned or not. Each individual is characterized by the threshold probability of sanction above which she prefers the benign action over the risky lottery, which for simplicity we assume to be independent of time and place.

As mentioned above, we conduct an experiment to assess our theoretical model. Since experimental results tend to be subject to noise, we enhance the realism of our model by assuming that each individual's threshold probability is normally distributed. As a result, individuals' choices between the benign and unwanted actions are also noisy in the theoretical model. Notably, each individual's violation function, which relates the probability that the individual chooses the socially unwanted action to the probability of sanction, is decreasing and S-shaped in the probability of sanction. This decreasing S-shape form is consistent with the intuition that, on the one hand, very small probabilities of sanction should hardly affect individuals' propensities to engage in the unwanted activity, and on the other hand, sufficiently large probabilities of sanction should deter almost everyone from engaging in the un-

[1] Of course, for splitting the probability of sanction to have any effect at all, it must be observable. Namely, it must be known that in certain locations and times, enforcement is stricter.

wanted activity. Indeed, a famous experiment that was conducted in Kansas City in 1974 (Kelling et al., 1974) found that a doubling of police patrols had virtually no statistically significant effect on street crime.[2]

Summing up the individuals' violation functions produces an aggregate violation function that relates the probability of sanction to the share of the population that engages in the socially unwanted action, which is also decreasing and S-shaped.

If a violation function is decreasing and *convex* throughout its range, then splitting the sanction probability p would increase the overall likelihood that the individual would engage in the unwanted activity. However, if the violation function is *concave* throughout, then splitting the sanction probability p would decrease the overall likelihood that the individual would engage in the unwanted activity. The fact that individuals', as well as the aggregate, violation functions are decreasing and S-shaped implies that they are first concave, and then convex. This suggests that small values of the sanction probability p may be split in a way that promotes overall deterrence, but large values cannot.

Moreover, for any fixed probability of sanction, decreasing the magnitude of the sanction, or increasing the reward from choosing the socially unwanted action without being sanctioned, increases the relative attractiveness of the socially unwanted action, and so shifts each individual's as well as the aggregate violation function to the right. Accordingly, we say that decreasing the magnitude of the sanction, or increasing the reward from choosing the socially unwanted action, increases the *temptation* to choose the socially unwanted action. Intuitively, when temptation is very low, the violation function is shifted so much to the left that it becomes convex on the entire relevant range of sanction probabilities, which in turn implies that splitting increases the violation rate (hurts deterrence). By contrast, when temptation is very high, the violation function is shifted so much to the right that it becomes concave on the entire relevant range, which implies that splitting decreases the rate of violation (improves deterrence).

The main theoretical result of this paper formalizes this intuition. We show that

[2] This finding had a big effect on the thinking on deterrence. It convinced both academics and the police itself that "police presence does not deter" (Sherman and Weisburd, 1995). Sherman and Weisburd (1995) famously criticized the Kansas City experiment by claiming that Kansas City is too large a unit of analysis for a doubling of patrols to produce an effect, or for a true reduction in crime to be statistically significant. Sherman and Weisbrud repeated the Kansas City experiment in Minneapolis two decades after the Kansas City experiment, but restricted it to crime "hot spots," which can be as small as a street corner or a city block. They found that a doubling of police patrols in crime hot spots produced reductions in total crime that ranged from 6 percent to 13 percent (however, "observed disorder" decreased by one-half). Their findings are consistent with the prevailing view that "large increases in dosage may be essential if any effect on crime is to be observed" (Sherman and Weisburd, 1995).

for any fixed probability of sanction p, if it is possible to improve individual or aggregate deterrence by splitting p, then it is also possible to improve individual or aggregate deterrence, respectively, by splitting p under higher temptation. And conversely, for any fixed probability of sanction p, if it is impossible to improve individual or aggregate deterrence by splitting p, then it is also impossible to improve individual or aggregate deterrence, respectively, by splitting p under lower temptation.

We confirm these theoretical predictions in a laboratory experiment, in which subjects learn the probability of sanction *from experience*. In each round of the experiment, subjects can choose between a safe action, which pays 5 Experimental Currency Units (ECU), and a binary lottery, which pays either a positive or a negative amount.[3] A subject who chooses the safe action is said to be deterred. We implement splitting by telling subjects to pay attention to a color that is flashed in front of them, because it is related to the probability of receiving the positive payment in the risky binary lottery. The fact that the subjects in the experiment learn the sanction probability from experience, rather than being told what it is, supports the view that, for those parameter values when it is successful, splitting can also be useful in practice.

The importance of the experiment lies in that it allows us to quantify exactly how high temptation needs to be in order for splitting to be effective in promoting overall deterrence. In our experiment, the sanction probability $p = 0.6$ can be split in a way that promotes overall deterrence if the binary lottery pays -10 and 50 ECU, when the individual is and is not sanctioned, respectively. In such an environment, not splitting the sanction probability $p = 0.3$ implies that 59% of participants' choices are for the socially unwanted action.[4] If the binary lottery pays -30 and $+30$ ECU upon sanction and no sanction, respectively, then splitting has no effect on overall deterrence. In such an environment, not splitting the sanction probability $p = 0.6$ implies that 33% of participants' choices are for the socially unwanted action. Finally, if the binary lottery pays -50 and $+10$ ECU upon sanction and no sanction, respectively, then splitting hurts overall deterrence. In such an environment, not splitting the sanction probability $p = 0.6$ implies that 12% of participants' choices are for the socially unwanted action.

The experiment thus both confirms and quantifies the observation that splitting can be effective in promoting deterrence in environments in which the temptation to

[3] At the conclusion of the experiment, the ECU is converted to euros so that the average payment to the participants is held constant through the different sessions.

[4] In choosing the parameters for our experiment, we have relied on the estimates produced by Erev et al. (2017)'s *Best Estimate and Sampling Tools* (BEAST) model of choice under uncertainty. Accordingly, our findings both rely on and validate the theoretical predictions produced by the BEAST model.

commit the socially unwanted action is strong and the choice of this action is relatively common, but not in environments in which the temptation to commit the socially unwanted action is weak and the choice of this action is relatively uncommon.

Estimates of the extent of illegal behavior are generally hard to get. But to put the numbers above in perspective, it is noteworthy that between 25%-35% of the bus passengers in Santiago, Chile, reportedly evaded payment of the required travel fare between 2015 and 2019 (Cantillo, Raveau and Muñoz, 2022). According to the Internal Revenue Service (IRS), roughly one out of every six dollars owed in federal taxes between 2008 and 2010 went unpaid.[5] Recent research conducted in Barcelona and New York City's Murray Hill, Midtown Manhattan, revealed an average of 1.32 and 0.28 illegally parked vehicles per 100 meters of road, respectively, suggesting that approximately 6% and 1.25% of vehicles in these areas were parked illegally (Morillo and Campos, 2014). Lastly, the National Coalition Against Domestic Violence reports that over 10 million adults in the United States experience domestic violence annually, indicating that that approximately 3% of Americans are involved in perpetrating domestic violence.[6]

Related Literature

The idea that, in a game with incomplete information, it may be possible to profitably manipulate players' choices through the splitting of prior probabilities dates back at least to work of Aumann and Maschler (1995), and is a key observation of the literature on Bayesian persuasion, which originated in Kamenica and Gentzkow (2011). For a recent review of this literature, see Kamenica, Kim and Zapechelnyuk (2021).

We are aware of only three experimental studies of Bayesian persuasion. All three papers have a very different focus from ours. Fréchette, Lizzeri and Perego (2022) study the role of commitment in communication and show that a form of commitment blindness leads some senders to overcommunicate when information is verifiable and undercommunicate when it is not. Au and Li (2018) perform an experimental study of the relationship between Bayesian persuasion and reciprocity, and Nguyen (2017) studies experimentally whether subjects design their signals in a way that maximizes their expected payoff.

The hotspots literature in criminology (see, e.g., Braga, Papachristos and Hureau,

[5] See the IRS publication titled Federal Tax Compliance Research: Tax Gap Estimates for Tax Years 20082010 Publication 1415 (5-2016), https://www.irs.gov/pub/irs-soi/p1415.pdf.

[6] See the national intimate partner and sexual violence survey: 2010 summary report (http://www.cdc.gov/violenceprevention/pdf/nisvs_report2010-a.pdf).

2014; and Braga et al., 2019) studies how to focus enforcement resources where they make the most difference. For example, in a town with two neighborhoods and two police cruisers, is it better to deploy these cruisers in the first or second neighborhood, or to split them between the two neighborhoods? By contrast, we focus mainly on the question of whether it is possible to improve deterrence through resource allocation, under the constraint that amount of resources is fixed. Namely, in a town with two neighborhoods as above, is it better to have the two cruisers patrol together, or separately?

Lando and Shavell (2004) and Eeckhout, Persico and Todd (2010) both considered the question of how to allocate enforcement resources, and argued that it may be beneficial to concentrate enforcement on a subset of the population. Eeckhout, Persico and Todd also demonstrated this idea empirically using traffic data gathered by the Belgian Police Department. More recently, Hernández and Neeman (2022) have generalized their theoretical results by considering any number of locations, adding uncertainty, considering the question of how to further improve deterrence through Bayesian persuasion, or communication.

The rest of the paper proceeds as follows ... All proofs are relegated the the Appendix.

2 Model

We consider the following choice problem. An individual faces a choice between two actions. One is benign, and the other is socially unwanted but beneficial for the individual.

We assume that choice of the benign action generates a certain payment to the individual, which we normalize to zero. We refer to the choice of this action as "compliance." Because choice of the socially unwanted action may by subject to sanction, choice of this action induces a risky binary lottery $L(p)$, which is parameterized by the probability of sanction $p \in [0,1]$. With probability p the individual is sanctioned, and the lottery generates a loss $L < 0$ to the individual, and with probability $1-p$ the individual is not sanctioned, and the lottery generates a reward $R > 0$. We refer to the choice of this action, which generates an expected payment of $\mathbb{E}[L(p)] = p \cdot L + (1-p) \cdot R$ to the individual, as "committing a violation." Accordingly, individuals who are induced to comply are said to be deterred from committing a violation.

The choice environment we consider is thus characterized by two parameters:

a reward $R > 0$, and a loss $L < 0$. Obviously, for any fixed probability of sanction $p \in [0,1]$, increasing the reward R, or decreasing the (absolute value of the) loss $|L|$ strengthens the individual's temptation to commit a violation. In other words, temptation introduces a binary relation over choice environments, which is defined as follows.

Definition 1 *A choice environment $\langle R, L \rangle$ induces a stronger temptation to commit a violation than the choice environment $\langle R', L' \rangle$ if either $R \geq R'$ or $L \geq L'$ and at least one of these inequalities is strict.*

If a choice environment $\langle R', L' \rangle$ induces a stronger temptation to commit a violation than the choice environment $\langle R, L \rangle$ then we say that "temptation increases" from $\langle R, L \rangle$ to $\langle R', L' \rangle$.

Denote by $\tilde{p}_{R,L}$ the threshold probability of sanction above which the individual prefers to comply in choice environment $\langle R, L \rangle$. That is, when faced with choice problem $\langle R, L, p \rangle$, where R and L denote the Reward and Loss, respectively, and p denotes the probability of sanction, if $p > \tilde{p}_{R,L}$ then the individual would comply; if $p < \tilde{p}_{R,L}$ the individual would commit a violation; and if $p = \tilde{p}_{R,L}$ the individual would be indifferent between compliance and the commitment of a violation. To simplify notation, as long as it does not cause confusion, we drop the subscript from the threshold sanction and denote it as $\tilde{p}$.

As explained in the introduction, we test our theoretical model experimentally. Because experimental results tend to be subject to noise, we enhance the realism of our model by using a random preference model. Specifically, we assume that when faced with a choice environment $\langle R, L \rangle$, the individual's threshold sanction $\tilde{p}$ is distributed according to a normal distribution with mean $\mu_{R,L}$ and standard deviation $\sigma_{R,L}$.

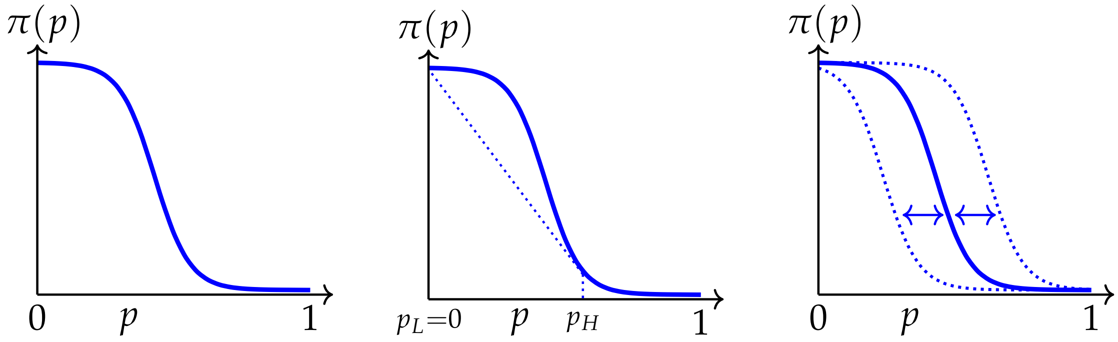
Thus, when faced with a choice problem $\langle R, L, p \rangle$, where $p \in [0,1]$, the probability that the individual would commit a violation is given by:

$$\begin{aligned} \pi_{R,L}(p) &\equiv \Pr\left(p \leq \tilde{p}_{R,L}\right) \\ &= 1 - \Phi_{R,L}(p), \end{aligned}$$

where $\Phi_{R,L}$ denotes the cumulative distribution function of a Normal distribution with mean $\mu_{R,L}$ and standard deviation $\sigma_{R,L}$. We refer to the function $\pi_{R,L}(\cdot)$ as the "violation curve" for environment $\langle R, L \rangle$. As before, to simplify notation, we drop the subscript from the violation curve and denote it as $\pi(\cdot)$.

The function $\Phi_{R,L}$ is an increasing S-shaped function, or a sigmoid function. That is, $\Phi_{R,L}$ is convex and then concave in its argument. Thus, the violation curve $\pi(p)$

is a decreasing S-shaped function of the sanction probability p, which is concave and then convex in p, as depicted in Figure 1a below.



(a) S-shaped violation (b) Optimal splitting (c) Changing fine/reward

Figure 1: The Violation Curve

3 Splitting

Splitting the probability of a sanction p into two probabilities $p_L < p < p_H$ may facilitate compliance while maintaining the same amount of enforcement resources. The basic idea is the following. Instead of being faced with the choice problem $\langle R, L, p \rangle$, the individual would be faced with one of two choice problems. With probability λ, the individual would be faced with the choice problem $\langle R, L, p_L \rangle$; and with probability $1-\lambda$, the individual would be faced with the choice problem $\langle R, L, p_H \rangle$. The numbers p_L, p_H and λ are chosen such $0 \leq p_L < p < p_H \leq 1$, and $\lambda p_L + (1-\lambda)p_H = p$. This ensures that the mean probability of a sanction across the two choice problems $\langle R, L, p_L \rangle$ and $\langle R, L, p_H \rangle$, $\lambda p_L + (1-\lambda)p_H$, remains fixed at p. Being faced with one of two choice problems instead of just with a single choice problem is called splitting because the probability p used in the single choice problem $\langle R, L, p \rangle$ is split into the probabilities p_L and p_H used in the two choice problems $\langle R, L, p_L \rangle$ and $\langle R, L, p_H \rangle$ in a way that preserves the overall probability that an individual who chooses to commit a violation is sanctioned.

An individual who is faced with the choice problem $\langle R, L, p_L \rangle$ chooses the lottery with probability $\pi(p_L)$; and an individual who is faced with the choice problem $\langle R, L, p_H \rangle$ chooses the lottery with probability $\pi(p_H)$. It follows that the expected probability that an individual who is faced with one of the two choice problems

$\langle R, L, p_L \rangle$ and $\langle R, L, p_H \rangle$ as described above would commit a violation is equal to:

$$\lambda\pi(p_L) + (1 - \lambda)\pi(p_H).$$

The next definition formalizes the sense in which a probability of a sanction p can be split in a way that promotes social welfare.

Definition 2 *Fix an environment $\langle R, L \rangle$. A violation curve $\pi(\cdot)$ is said to be profitably convexifiable at $p \in (0, 1)$ if there exist two probabilities $p_L < p < p_H$ such that*

$$\lambda\pi(p_L) + (1 - \lambda)\pi(p_H) < \pi(p)$$

for a probability $\lambda \in (0, 1)$ that satisfies the equation $\lambda p_L + (1 - \lambda)p_H = p$.

If a violation curve $\pi(\cdot)$ is profitably convexifiable at p, then there exist two probabilities $p_L < p < p_H$ such that the straight line that connects the points $(p_L, \pi(p_L))$ and $(p_H, \pi(p_H))$ lies below $\pi(\cdot)$ on the interval (p_L, p_H). This is depicted in Figure 1b, where the value of p_L is taken to be equal to zero.

Notably, while a function that is concave on an open interval is profitably convexifiable at any point in this interval, a function can also be profitably convexifiable at points in which it is convex. Figure 1b depicts a violation curve that is both locally convex and profitably convexifiable at points that are sufficiently close to p_H from below.

If it is possible to split the probability of a sanction in a way that reduces the probability of committing a violation, or that increases compliance, then we say that splitting is socially beneficial. In principle, there could be many pairs of sanction probabilities $p_L < p < p_H$, which make splitting socially beneficial. The pair p_L and p_H that is depicted in Figure 1b is the optimal pair, which maximizes the probability of compliance.

Lemma 1 *Fix an environment $\langle R, L \rangle$. A probability of a sanction p is profitably convexifiable if and only if $p < p^*_{R,L}$ where $p^*_{R,L}$ is given by the unique solution of the problem*

$$\min_{p \in [0,1]} \frac{\pi(0) - \pi(p)}{p}, \tag{1}$$

provided that $p^*_{R,L} \leq 1$. If $p^*_{R,L} > 1$, then any sanction probability $p < 1$ is profitably convexifiable.

Increasing temptation makes non-compliance relatively more attractive for the individual for every sanction probability. It is therefore natural to assume that increasing temptation increases the rate of violation $\pi(\cdot)$ for every sanction probability $p \in (0,1)$.

Lemma 2 *Suppose that increasing temptation increases the rate of violation $\pi(\cdot)$ for every sanction probability $p \in (0,1)$. Then if temptation increases from choice environment $\langle R, L\rangle$ to $\langle R', L'\rangle$, then $\mu_{R',L'} > \mu_{R,L}$.*

Intuitively, because a larger reward R and a smaller (absolute value of) loss $|L|$ make any lottery $L(p)$ more attractive, the mean $\mu_{R,L}$ is weakly increasing in both R and L. For simplicity, we assume that the standard deviation $\sigma_{R,L}$ is constant in R and L.[7] Intuitively, we expect that both when temptation is very high and when it is very low, the variance of the individual's choice will be small (because the individual will almost always violate and comply, respectively). Note that our model captures this intuition despite $\sigma_{R,L}$ being constant. This is because when temptation is very high or very low $\tilde{p}$ is almost always either larger or smaller than any $p \in [0,1]$, respectively.

The next proposition describes the main theoretical result of the paper.

Proposition 1 *Increasing temptation shifts the violation curve to the right. Specifically, suppose that the choice environment $\langle R, L\rangle$ induces a stronger temptation to commit a violation than choice environment $\langle R', L'\rangle$. Then, if the violation curve $\pi(\cdot)$ is profitably convexifiable at sanction probability p in choice environment $\langle R', L'\rangle$, then it is also profitably convexifiable at sanction probability p in choice environment $\langle R, L\rangle$.*

Because the violation curve is decreasing and S-shaped, Proposition 1 implies that increasing temptation shifts the violation curve to the right. As shown in Figure 1c, it follows that the range in which the violation curve is concave expands as temptation increases, and the range in which it is convex expands as temptation decreases. By moving the violation curve sufficiently to the right, it can be made concave over the entire range of sanction probabilities, and by moving the violation curve sufficiently to the left, it can be made convex over the entire range of sanction probabilities.

The following immediate corollary of Proposition 1 provides a testable empirical prediction of our main result.

Corollary 1 *Suppose that the choice environment $\langle R, L\rangle$ induces a stronger temptation to commit a violation than choice environment $\langle R', L'\rangle$. Then, if a sanction probability p is profitably convexifiable through the splitting of p into p_L and p_H in environment $\langle R, L\rangle$, then p*

[7] Our results continue to hold as long as the standard deviation $\sigma_{R,L}$ does not decrease too fast in R and L.

is also profitably convexifiable through splitting it into p_L and p_H in choice environment $\langle R', L' \rangle$.

Proposition 1 is formulated for the case of a single individual. By aggregating individuals' violation functions, it is possible to obtain an aggregate analog of Proposition 1 as follows.

Suppose that there are n different individuals. Let $\tilde{p}^i_{R,L}$ denote the threshold probability of sanction above which individual i prefers to comply in choice environment $\langle R, L \rangle$. Suppose that individuals compliance decisions are stochastically independent. That is, when faced with choice problem $\langle R, L, p \rangle$, if $p > \tilde{p}^i_{R,L}$ then individual i complies; if $p < \tilde{p}^i_{R,L}$ then individual i commits a violation; and if $p = \tilde{p}^i_{R,L}$ then individual i is indifferent between compliance and the commitment of a violation, independently of whether other individuals' comply or not. As before, to simplify notation, we drop the subscript from the threshold sanction and denote it as $\tilde{p}^i$.

Each individual i's threshold sanction $\tilde{p}$ is normally distributed, with mean $\mu^i_{R,L}$ and standard deviation $\sigma^i_{R,L}$. We denote individual i's distribution by $\Phi^i_{R,L}(p)$. Thus, when faced with a choice problem $\langle R, L, p \rangle$, the mean fraction of individuals who would commit a violation is given by:

$$\begin{aligned}
\pi^{\Sigma}_{R,L}(p) &\equiv \frac{1}{n}\sum_{i=1}^{n} \Pr\left(p \le \tilde{p}^i_{R,L}\right) \\
&= \frac{1}{n}\left(\sum_{i=1}^{n}(1 - \Phi^i_{R,L}(p)\right) \\
&= 1 - \Phi^{\Sigma}_{R,L}(p),
\end{aligned}$$

where $\Phi^{\Sigma}_{R,L}$ denotes the cumulative distribution function of a Normal distribution with mean $\mu^{\Sigma}_{R,L} = \frac{1}{n}\sum_{i=1}^{n}\mu^i_{R,L}$ and standard deviation $\sigma^{\Sigma}_{R,L} = \sqrt{\frac{1}{n^2}\sum_{i=1}^{n}(\sigma^i_{R,L})^2}$. We refer to the function $\pi^{\Sigma}_{R,L}(\cdot)$ as the "aggregate violation curve" for environment $\langle R, L \rangle$. As before, to simplify notation, we drop the subscript from the violation curve and denote it as $\pi^{\Sigma}(\cdot)$. Because it is equal to a sum of decreasing S-shaped functions, the aggregate violation curve $\pi^{\Sigma}(p)$ is a also a decreasing S-shaped function of the sanction probability p, which is concave and then convex in p, as depicted in Figure 1a above.

We thus have the following proposition, which describes the aggregate analog to Proposition 1.

Proposition 2 *Suppose that the choice environment $\langle R, L \rangle$ induces a stronger temptation to commit a violation than choice environment $\langle R', L' \rangle$. Then, if the aggregate violation* *curve $\pi^{\Sigma}(\cdot)$ is profitably convexifiable at sanction probability p in choice environment $\langle R', L' \rangle$, then it is also profitably convexifiable at sanction probability p in choice environment $\langle R, L \rangle$.*

The following corollary of Proposition 2 provides a testable empirical prediction of the aggregate behavior.

Corollary 2 *Suppose that the choice environment $\langle R, L \rangle$ induces a stronger temptation to commit a violation than choice environment $\langle R', L' \rangle$. Then, the number of individuals whose violation curve $\pi^{i}(\cdot)$ is profitably convexifiable at sanction probability p through splitting into probabilities p_L and p_H in choice environment $\langle R, L \rangle$ is larger than or equal to the number of individuals whose violation curve is profitably convexifiable at p through splitting into probabilities p_L and p_H in choice environment $\langle R', L' \rangle$.*

4 Experiment

4.1 Design and Procedure

We ran an experiment in which we manipulated splitting within-subjects and temptation between-subjects to test the benefit from splitting under different levels of temptation. The experiment consisted of 200 trials in two blocks, which were divided into one block of 100 splitting trials and another block of 100 pooling trials. The order of the blocks was randomized at the participant level. In each trial, each participant observed a signal in the shape of a colored circle and chose between a safe option (comply) and a risky option (violate). The safe option always yielded a payoff of 5 ECU (Experimental Currency Units). The possible payoffs obtained from the risky option varied depending on the treatment and color of the circle. Each treatment was associated with two possible payments, one positive and one negative. The positive payment captured the benefit from committing a violation without being sanctioned, and the negative payment captures the loss from committing the violation and being sanctioned.

In the pooling trials, the circle was always yellow and the probability of sanction upon choosing to commit a violation was fixed at $\frac{3}{5}$. In the splitting trials, the circle was either red with a probability of .56 or blue with a probability of .44. The color red corresponded to a high rate of enforcement, or probability of sanction, and the blue color corresponded to a low rate of enforcement, or probability of sanction. Accordingly, a red circle indicated that the probability of a sanction was 1, that is, when the circle was red, a subject who chose to commit a violation was sanctioned with probability 1. The blue circle indicated a small probability of sanction. When the circle was

Table 1: Experimental Treatments

Temptation	Safe	Reward	Loss
Weak	5	10	-50
Medium-weak	5	10	-30
Medium	5	30	-30
Medium-strong	5	30	-10
Strong	5	50	-10

blue, a subject who chose to commit a violation was sanctioned with probability $\frac{1}{12}$. The mean probability of a sanction was thus $.56 \cdot 1 + .44 \cdot \frac{1}{12} = .597$, slightly less than in the pooling trials.

Table 1 presents the between-subjects treatments. The treatments manipulated temptation by gradually varying the negative or loss payments and the positive or reward payments. Starting from weak temptation, with a low reward of 10 ECU and a high absolute loss of -50 ECU, temptation increased by first decreasing the loss payment to -30 in the Medium-weak temptation treatment, and then also increasing the reward payment to 30 in the Medium temptation treatment. This was followed by further decreasing the loss payment to -10 in the Medium-strong temptation treatment and finally by further increasing the reward payment to 50 in the Strong temptation treatment.

For each treatment, we ran one session with 50 participants. The total number of participants in the experiment was thus 250. The experiment was carried out at the Laboratory for Research in Experimental Economics (LINEEX) in the University of Valencia in December, 2022, and March, 2023. The instructions, available in the appendix, were read aloud at the beginning of each session/treatment. Each session lasted approximately 50 minutes. The average payment to participants was 13.8 Euros.

4.2 Hypotheses

Because participants in the experiment faced three different probabilities of sanction: $\frac{1}{12}$, $\frac{3}{5}$, and 1, the experiment generated three points on the participants' violation curve. We refer to the piecewise linear curve resulting from connecting these three points as the *experimental violation curve*. Whether splitting is beneficial is directly tied to the curvature of this curve. If the line that connects the points with $p = \frac{1}{12}$ and $p = 1$ lies below the point with $p = \frac{3}{5}$, then the sanction probability $p = \frac{3}{5}$ is beneficially convexifiable through splitting to $p_L = \frac{1}{12}$ and $p_H = 1$.

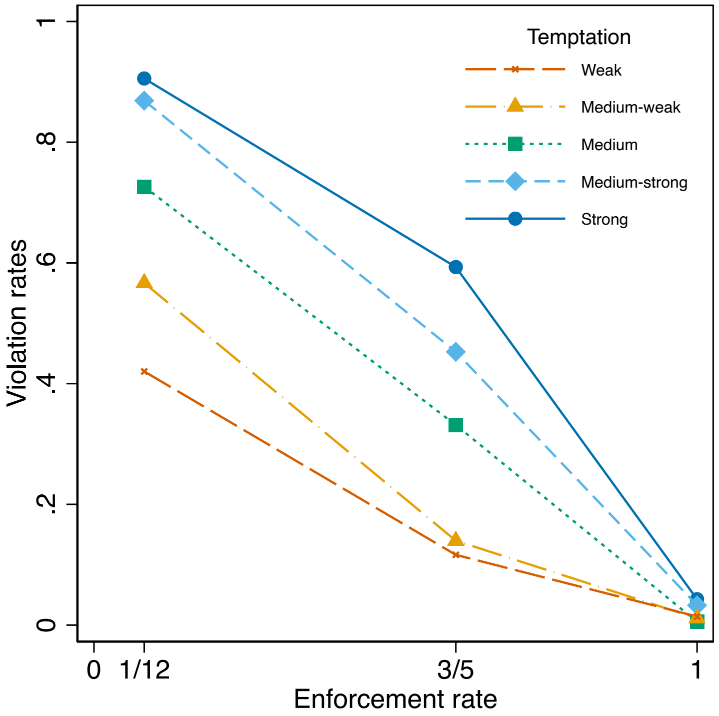


Figure 2: Violation rates across treatments.

Our first hypothesis tests Corollary 1 as reflected in the concavity of the experimental violation curve.

Hypothesis 1 *If the experimental violation curve is concave for any treatment, it is also concave to all treatments with higher temptation.*

Our second hypothesis tests Corollary 2.

Hypothesis 2 *As temptation increases, the violation curves of more individuals become concave.*

4.3 Results

We first analyze the effects of temptation on the concavity of the experimental violation curve and the effectiveness of splitting at the aggregate level. We proceed with analyses and tests at the individual level.

Table 2: Curvature of the violation curves.

	Term	Convexity	t-value	p-value
Weak	β_2	0.357	2.98	0.003
Medium-weak	$\beta_2 + \beta_5$	0.544	6.39	0.000
Medium	$\beta_2 + \beta_8$	-0.051	-0.36	0.722
Medium-strong	$\beta_2 + \beta_{11}$	-0.269	-1.65	0.099
Strong	$\beta_2 + \beta_{14}$	-0.843	-4.43	0.000

Notes: Marginal self interactions of enforcement rate based on an OLS regression of subject-level mean violation rate by enforcement rate with robust standard errors clustered on participants.

4.3.1 Aggregate Violation

Figure 2 depicts the aggregate violation rates by treatment and probability of sanction. The experimental violation curve is approximately linear for medium temptation, concave for higher temptations and convex for lower temptation, confirming Hypothesis 1.

Let VR_{ip} denote the violation rate of participant i across the periods in which the enforcement rate was p. To formally measure the curvature of the experimental violation curves, we estimated the following regression with robust standard errors clustered on individuals:

where M_i, MW_i, MS_i, and S_i are dummy variables for the temptation treatments.[8] Table 2 presents terms based on the regression coefficients that capture the convexity (if positive) and concavity (if negative) of the experimental violation curve across the five treatments. The results support the observation based on Figure 2. The experimental violation curve is significantly convex under Weak and Medium-weak temptation. It is indistinguishable from linear under Medium temptation, and significantly concave under Medium-strong and Strong temptation (at the 10% for the former).[9]

Table 3 presents the results of a logistic regression of the violation rate on treatment interacted with splitting with robust standard errors clustered on participants. Figure 3 plots the results from the regression, with the violation rate presented in the

[8] Each of the 250 participants was faced with three different enforcement rates, resulting in five treatments and 750 observations overall.

[9] Note that concavity does not increase monotonically with temptation as the experimental violation curve is more concave under Weak compared to Medium-weak temptation. Although at first sight this appears to contradict the intuition behind our prediction, it is consistent with our model. It is only convex*ifiabil*ity and not concavity that increases monotonically with temptation.

Table 3: Regression on violation rates.

	Coefficient	Robust std. error	z-statistic	p-value
Medium-weak	0.210	0.213	0.99	.324
Medium	1.325	0.216	6.15	.000
Medium-strong	1.837	0.220	8.34	.000
Strong	2.404	0.239	10.04	.000
Split	0.624	0.1997	3.16	.002
Medium-weak × Split	0.139	0.227	0.61	.541
Medium × Split	-0.652	0.240	-2.72	.007
Medium-strong × Split	-0.869	0.239	-3.64	.000
Strong × Split	-1.297	0.254	-5.11	.000
Constant	-2.027	0.168	-12.10	.000

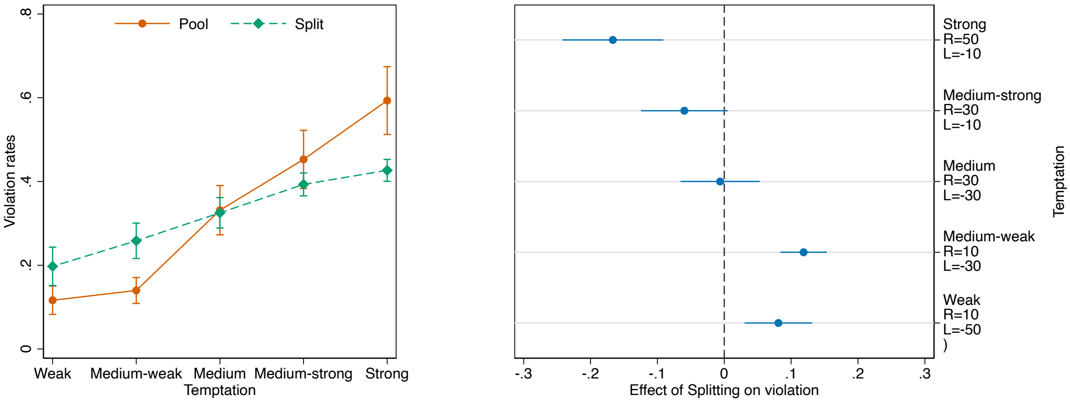
Notes: Logistic regression of violation rates on treatment interacted with splitting with robust standard errors clustered on participants.

left panel and the estimated marginal effect of splitting presented in the right panel. The regression confirms the results from the non-parametric tests. Splitting of the sanction probability $\frac{3}{5}$ into the sanction probabilities $\frac{1}{12}$ and 1 reduces violations significantly under Strong temptation ($z = 4.32, p < .001$) and weakly significantly under Medium-strong temptation ($z = 1.80, p = .072$), significantly increases violations under Medium-weak ($z = 6.71, p < .001$) and Weak temptation ($z = 3.14, p = .002$), and has no significant effect under medium temptation ($z = 0.21, p = .837$).[10]

A pairwise comparison of the effect of splitting on the violation rate between adjacent treatments reveals that the treatment effects arise from changes in the reward R rather than in the loss L. Splitting is less detrimental (more beneficial) when increasing R from Medium-weak to Medium and from Medium-strong to Strong temptation ($z = 3.57, p < .001$ and $z = 2.10, p = .035$, respectively). In contrast, increasing L from Weak to Medium-weak or from Medium to Medium-strong temptation has no significant effect ($z = -1.20, p = .229$ and $z = 1.19, p = .233$, respectively).

Result 1 *the results strongly support Hypothesis 1. Splitting is socially beneficial for high temptation and socially detrimental for low temptation. The benefit from splitting is sensitive to the reward R but not to the loss L.*

[10] Wilcoxon signed-rank tests comparing violation rates between splitting and pooling in each treatment yield identical results.



(a) Violation rates.

(b) Marginal effects.

Figure 3: Effects of splitting enforcement on violation rates.

4.3.2 Individual Heterogeneity

To test Hypothesis 2, we calculated for each participant the absolute difference in violation rate between the splitting and pooling trials. Figure 4 presents the share of individuals for whom splitting is beneficial or detrimental (i.e., reduces or increases violations, respectively) by treatment. Significance is based on t-tests at the individual level. The share of individuals for whom splitting is beneficial increases with temptation, while the share of individuals for whom splitting is detrimental mostly decreases with temptation. The exception is that splitting is detrimental for more individuals under Medium-weak compared to Weak temptation. This deviation from monotonicity is, perhaps, not surprising considering that our model predicts that the violation curve flattens if shifted enough to the left or to the right. The fact compliance is the modal choice even with $p = \frac{1}{12}$ under Weak temptation suggests that this is the case. Furthermore, this non-monotonicity disappears when excluding nine participants who never violated in the Weak and Medium-weak treatments.

An ordered logistic regression of the five types presented in Figure 4 on treatment reveals significant effects in the predicted direction. Mirroring out results for beneficial splitting, the shift in the type distribution is significant when increasing R from Medium-weak to Medium temptation and from Medium-strong to Strong temptation ($z = 3.07, p = .002$ and $z = 2.70, p = .007$, respectively) but not when increasing L from Weak to Medium-weak or from Medium to Medium-strong temptation ($z = -0.98, p = .329$ and $z = 0.65, p = .513$, respectively).

Result 2 *The individual-level analysis supports Hypothesis 2. The share of individuals for whom splitting is socially beneficial weakly increases with temptation. This share is sensitive*

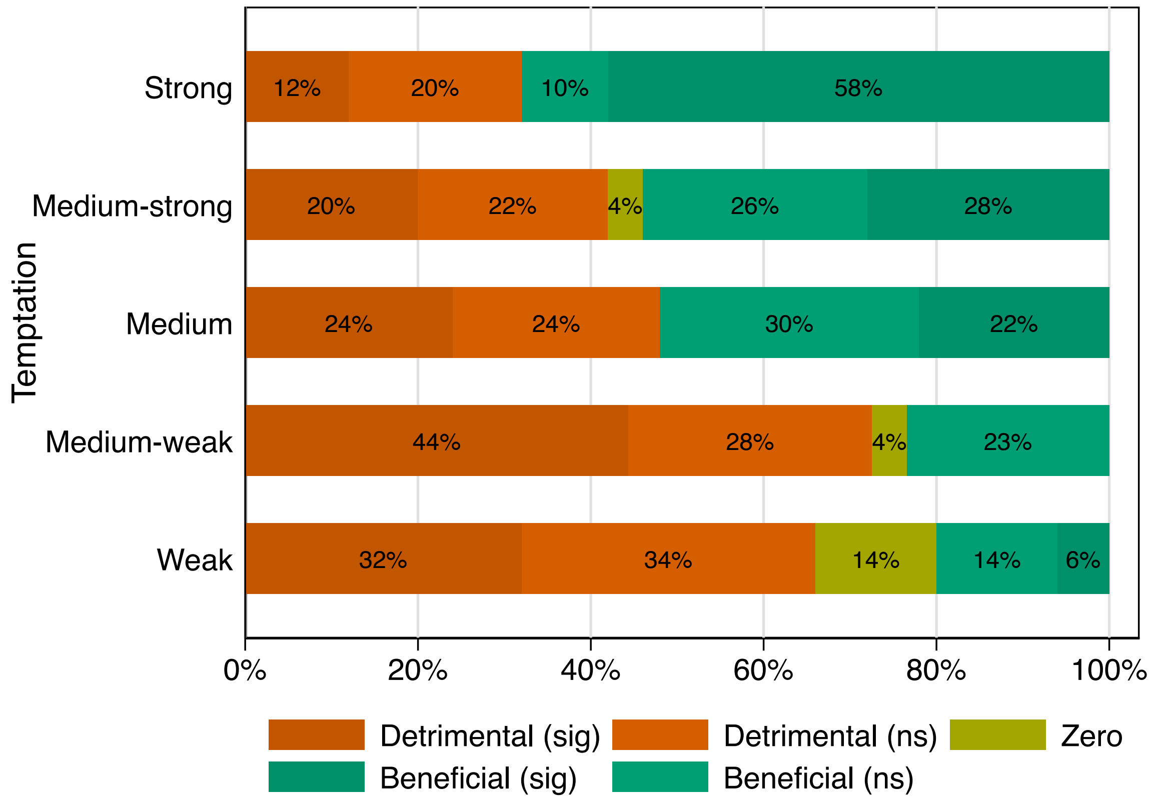


Figure 4: Individual heterogeneity.

to the reward R but not to the loss L.

5 Conclusion

References

Aumann, Robert J, and Michael Maschler. 1995. *Repeated Games with Incomplete Information.* MIT Press.

Au, Pak Hung, and King King Li. 2018. "Bayesian persuasion and reciprocity: theory and experiment." *Available at SSRN 3191203*.

Braga, Anthony A, Andrew V Papachristos, and David M Hureau. 2014. "The effects of hot spots policing on crime: An updated systematic review and meta-analysis." *Justice quarterly*, 31(4): 633–663.

Braga, Anthony A, Brandon S Turchan, Andrew V Papachristos, and David M Hureau. 2019. "Hot spots policing and crime reduction: An update of an ongoing systematic review and meta-analysis." *Journal of experimental criminology*, 15: 289–311.

Cantillo, Angel, Sebastián Raveau, and Juan Carlos Muñoz. 2022. "Fare evasion on public transport: Who, when, where and how?" *Transportation Research Part A: Policy and Practice*, 156: 285–295.

Eeckhout, Jan, Nicola Persico, and Petra E Todd. 2010. "A Theory of Optimal Random Crackdowns." *American Economic Review*, 100: 1104–1135.

Erev, Ido, Eyal Ert, Orly Plonsky, Dana Cohen, and Or Cohen. 2017. "From anomalies to forecasts: Toward a descriptive model of decisions under risk, under ambiguity, and from experience." *Psychological Review*, 124(4): 369–409.

Fréchette, Guillaume R, Alessandro Lizzeri, and Jacopo Perego. 2022. "Rules and commitment in communication: An experimental analysis." *Econometrica*, 90(5): 2283–2318.

Hernández, Penélope, and Zvika Neeman. 2022. "How Bayesian Persuasion Can Help Reduce Illegal Parking and Other Socially Undesirable Behavior." *American Economic Journal: Microeconomics*, 14(1): 186–215.

Kamenica, Emir, and Matthew Gentzkow. 2011. "Bayesian Persuasion." *American Economic Review*, 101: 2590–2615.

Kamenica, Emir, Kyungmin Kim, and Andriy Zapechelnyuk. 2021. "Bayesian persuasion and information design: Perspectives and open issues." *Economic Theory*, 72(3): 701–704.

Kelling, George, Anne M Pate, Dennis Dieckman, and Christine Brown. 1974. “The Kansas City preventive patrol experiment: Technical report.” Washington DC: Police Foundation.

Lando, Henrik, and Steven Shavell. 2004. “The advantage of focusing law enforcement effort.” *International Review of Law and Economics*, 24: 209–218.

Morillo, Carlos, and José Magín Campos. 2014. “On-street Illegal Parking Costs in Urban Areas.” *Procedia - Social and Behavioral Sciences*, 160: 342–351. XI Congreso de Ingenieria del Transporte (CIT 2014).

Nguyen, Quyen. 2017. “Bayesian persuasion: evidence from the laboratory.” *Work. Pap., Utah State Univ., Logan*.

Sherman, Lawrence W, and David Weisburd. 1995. “General deterrent effect of police patrol in crime “hot spots”: A randomized controlled trial.” *Justice Quarterly*, 12(4): 625–648.

Appendix

Proof of Lemma 1

Proof. The first order condition of the optimization problem yields

$$\pi(p^*) = \pi(0) + \pi'(p^*)p^*. \tag{2}$$

The solution of Equation (2) describes the unique probability of a sanction p^* that has the property that the tangent of the violation curve $\pi(p)$ at p^* is such that: (i) the tangent lies below $\pi(p)$ on the interval $(0, p^*)$; and (ii) the tangent intersects the violation curve at the point $(0, \pi(0))$. It follows that all the probabilities $p < p^*$ are profitably convexifiable, and no probability $p > p^*$ is profitably convexifiable. ■

Proof of Lemma 2

Proof.

Notice that if temptation increases from choice environment $\langle R, L\rangle$ to $\langle R', L'\rangle$, then $\pi_{R',L'}(p) \geq \pi_{R,L}(p)$ for all $p \in [0, 1]$.

We express the function $\pi(.)$ by using the error function $erf(\cdot)$ as follows:

$$\pi_{R,L}(p) = \frac{1}{2} - \frac{1}{\sqrt{\pi}} \int_0^{\frac{p-\mu}{\sigma\sqrt{2}}} e^{\frac{-t^2}{2}} dt$$

Writing the above inequality for $p = \mu$, we get:

$$\frac{1}{2} - \frac{1}{\sqrt{\pi}} \int_0^{\frac{\mu-\mu'}{\sigma'\sqrt{2}}} e^{\frac{-t^2}{2}} dt \geq \frac{1}{2} - \frac{1}{\sqrt{\pi}} \int_0^{\frac{\mu-\mu}{\sigma\sqrt{2}}} e^{\frac{-t^2}{2}} dt$$

or

$$-\frac{1}{\sqrt{\pi}} \int_0^{\frac{\mu-\mu'}{\sigma'\sqrt{2}}} e^{\frac{-t^2}{2}} dt \geq 0$$

Consequently, $\mu' \geq \mu$ and the condition holds. ■

Proof of Proposition 1

Proof. As the choice environment $\langle R, L\rangle$ induces a stronger temptation to commit a violation than the choice environment $\langle R', L'\rangle$ then $\mu \geq \mu'$ and $\sigma = \sigma'$. Notice that the violation curve $\pi_{R,L}$ under environment $\langle R, L\rangle$ is the violation curve $\pi_{R',L'}$ under environment $\langle R', L'\rangle$ shifts to the left in a parallel way. Consequently,

$$\pi_{R',L'}(p) = \pi_{R,L}(p + (\mu - \mu')) \tag{3}$$

for all $p \in \mathbf{R}$.

Let $p^*_{R',L'}$ be the solution of 2 and $\tilde{p} < 0$ such that $\pi_{R,L}(\tilde{p}) + (\mu - \mu') = \pi_{R',L'}(0)$. As the line that connect $(0, \pi_{R',L'}(0))$ and $(p^*_{R',L'}, \pi_{R',L'}(p^*_{R',L'}))$ is tangent at $p^*_{R',L'}$, then the line that connect $(\tilde{p}, \pi_{R',L'}(\tilde{p}))$ and $(p^*_{R',L'}, \pi_{R',L'}(p^*_{R',L'}))$ is also bellow than $\pi_{R',L'}$.

Let $p \in [0, p^*_{R',L'}]$ profitably convexifiable, then we can write $p = \lambda_{R',L'} p^*_{R',L'}$ for $\lambda_{R',L'} \in [0, 1]$ and

$$(1 - \lambda_{R',L'})\pi_{R',L'}(0) + \lambda_{R',L'}\pi_{R',L'}(p^*_{R',L'}) \leq \pi_{R',L'}(\lambda_{R',L'} p^*_{R',L'}) = \pi_{R',L'}(p)$$

Let us check that p is also convexifiable under the environment $\langle R, L\rangle$ by splitting at 0 and $p^*_{R',L'} + (\mu - \mu')$. Let $\tilde{\lambda}$ be such that $p = (1 - \tilde{\lambda})0 + \tilde{\lambda}(p^*_{R',L'} + (\mu - \mu'))$.

$$\begin{aligned}
(1-\tilde{\lambda})\pi_{R,L}(0) + \tilde{\lambda}\pi_{R,L}(p^*_{R',L'} + (\mu - \mu')) &= (1-\tilde{\lambda})\pi_{R,L}(0) + \tilde{\lambda}\pi_{R',L'}(p^*_{R',L'}) \\
&= (1-\tilde{\lambda})\pi_{R',L'}(-(\mu - \mu')) + \tilde{\lambda}\pi_{R',L'}(p^*_{R',L'}) \\
&< \pi_{R',L'}((1-\tilde{\lambda})(-(\mu - \mu')) + \tilde{\lambda}(p^*_{R',L'})) \\
&= \pi_{R',L'}(\tilde{\lambda}(p^*_{R',L'} + (\mu - \mu')) - (\mu - \mu'))) \\
&= \pi_{R,L}(\tilde{\lambda}(p^*_{R',L'}) + (\mu - \mu')) \\
&= \pi_{R,L}(p)
\end{aligned}$$

where the first, second, and fifth equality come from the equation 3 and the third inequality holds because the line that connect $(-(\mu - \mu')), \pi'(-(\mu - \mu'))$ with $(p^*_{R',L'}, \pi'(p^*_{R',L'}))$ is below to the tangent at $p^*_{R',L'}$. ∎

Proof of Proposition 2

The proof is similar to the proof of Proposition 1.