UNIVERSITÀ di VERONA

Department of ECONOMICS

Working Paper Series
Department of Economics
University of Verona

Miscarriage of Justice in Judges' Mind: Theory and Experimental Evidence

Stefania Ottone, Ferruccio Ponzano, Margherita Saraceno, Luca Zarri

WP Number: 3 January 2022

ISSN: 2036-2919 (paper), 2036-4679 (online)

Miscarriage of Justice in Judges' Mind: Theory and Experimental Evidence

Stefania Ottone
University of Turin

Ferruccio Ponzano
University of Eastern Piedmont

Margherita Saraceno
University of Pavia

Luca Zarri[*]
University of Verona

Abstract

In this paper, we investigate – both theoretically and by means of a controlled lab experiment – judges' decisions when either "type-I" errors (i.e. convicting an innocent defendant) or "type-II" errors (i.e. acquitting a guilty defendant) can occur. Addressing this issue with field data is extremely challenging. Taken together, our findings indicate that participants are sensitive to both types of error, rather than to type-I avoidance only. Next, in both scenarios we interestingly detect "compensatory leniency" in judicial decision making, with participants seeming to balance the inherent trade-off between the errors by jointly managing the two key levers they are provided wiggle room on by our design: decision over (i) conviction/acquittal and (ii) severity of punishment. Finally, we show that participants are willing to pay to get further evidence and eliminate both type-I and type-II errors. We discuss implications of our core results for the design of behaviorally informed deterrence policies.

Keywords: Judicial Errors; Miscarriage of Justice; Economic Experiments; Law and Economics.
JEL classification: K42**,** K49, C91.

[*] Corresponding author. Economics Department, University of Verona, Via Cantarane 24, 37131 Verona, Italy.

1. Introduction

As noted by Sonnemans and van Dijk (2012), “Core business of judges is to transform uncertainty about the facts into the certainty of the verdict” (p. 687). However, despite substantial progress in applications of artificial intelligence and machine learning in many judicial systems across the world, human judging remains a highly “noisy” activity (Kahneman et al., 2021).[1] Crucially, the inherent uncertainty of human judging entails passing through frequently occurring errors, especially when available evidence is weak or contradictory. In this regard, we focus on *type-I* errors (false positives) vs. *type-II* errors (false negatives). This classical dichotomy is relevant in a wide array of real-life settings in which judges must decide under limited information:[2] in judicial terms, a false positive means “conviction of an innocent defendant”, whereas a false negative is provided by “acquittal of a guilty defendant”.

“Wrongful conviction” and “wrongful acquittal” are in an inherent tension, as “reducing one typically results in increasing the prevalence of the other. Thus, a sound criminal justice policy fundamentally entails striking an *acceptable tradeoff* between the two types of errors based on their respective costs” (Scurich, 2015; p. 23; italics added). As Cappelen et al. (2018) correctly point out, the question of how this trade-off should be handled is a fundamental challenge in policy design and implementation: the problem arises not only with regard to a variety of public policies – where a high risk of having undeserving claimants receiving benefits often occurs (from implementation of government immigration policies to granting unemployment subsidies or disability benefits to undeserving applicants) –, but is also a concern for private companies (e.g., as to the implementation of bonus policies).

As highlighted by Guthrie (2001), the quality of a judicial system depends on the quality of decisions that judges make. A large research area has shown that judge political affiliation is a relevant source of bias in various settings (Cohen and Yang, 2019), including criminal sentencing (Schanzenbach and Tiller, 2007). Next, newspaper coverage (Lim et al., 2015) and external pressure by non-independent actors may exacerbate biased decision making (Ottone et

[1] While there is growing consensus around the idea that in the next future artificial intelligence will play a far more relevant role in assisting judges in sentencing decisions, judges and jurors still exercise considerable discretion in many countries (see e.g. Hudja et al., 2021). In the US in 2005 the Supreme Court held that the “Federal Sentencing Guidelines” – originally adopted as mandatory – would be “effectively advisory”: as a result, the degrees of discretion granted to judges greatly increased (Cohen and Yang, 2019).

[2] When we think of the term “judging”, our mind tends to naturally recall judges’ and jurors’ decision-making process in court cases. However, whereas this is the commonly-held meaning of the word we will mainly refer to in the present paper, it is also important to note that judging *per se* is in fact a far broader notion, encompassing determinations on disputes or contests by a series of third-parties (Ottone et al., 2015), including regulators, arbitrators, tax auditors, administrative agencies, private companies, religious bodies, fact-finding commissions, and referees in sports events (see on this also Markussen et al., 2016).

al., 2015), especially for highly discretionary decisions (Cohen et al., 2022).[3] However, despite this wide and rapidly-growing literature identifying relevant sources of bias in judges' decision making, we still lack empirical work directly targeted to judges' viewpoint on the issue and convincingly dealing with the following key questions: *are judges' decisions sensitive to the possible occurrence of type-I and type-II error? If this is the case, are they averse to both types of errors or mainly concerned with one specific type of error? Compared to a no-error scenario, do judges react to the possibility to decide incorrectly by adjusting the mix of probability and severity of punishment? If this is the case, what kind of adjustments do they make?*

To properly address these crucial but still largely open questions, it is important to preliminarily acknowledge that, as argued by Chen and Schonger (2020), a classic divide separates the economic approach to optimal policy based on cost-benefit calculations from "non-consequentialist" views in which subjective evaluations of "what is right" play a key role (see on this also Galbiati and Vertova, 2008). Although deepening this philosophical discussion is beyond the scope of this paper, we believe – in light of a fast-growing empirical line of inquiry in law and economics – that judges' sensitivity towards type-I and type-II errors may reflect a mixture of consequentialist and non-consequentialist concerns. On the one hand, based on the established economic model of optimal deterrence pioneered by Becker (1968), both error types are equally detrimental and should be avoided (Png, 1986; Polinsky and Shavell, 2007). On the other hand, we cannot rule out that judges subjectively prioritize type-I error avoidance, due to a variety of consequentialist and non-consequentialist arguments. Individuals may be averse to this kind of errors as type-II errors, unlike type-I errors, allow to preserve the so called "expressive function" of the law (Cooter, 1998; Rizzolli and Stanca, 2012). As emphasized by Scurich (2015), a false positive has been historically considered more costly as it violates the social contract between the state and the individual and erodes the legitimacy of the justice system.[4] It is well known that in countries such as the US the total burden of evidence needs to meet the "beyond reasonable doubt" criterion justifying the high standard of evidence that is usually required in criminal procedure (Hudja et al., 2021).[5] However, judges might be very sensitive to

[3] Other studies document the role of individual drivers unrelated to the merits of the case – such as "cognitive illusions" and emotional shocks – in (possibly subconsciously) affecting judging despite the fact that decision makers are highly educated people (judges and juries) and that sentencing has hugely relevant consequences for defendants and society as a whole (Guthrie et al., 2001; Eren and Mocan, 2018). Abrams et al. (2022) show that also local sentencing practices shape judges' decisions.

[4] In April 2021, the National Registry of Exonerations, reporting every known exoneration in the US since 1989, revealed that exonerated defendants spent in prison more than 25,000 years for crimes they did not commit. Relatedly, high profile cases of unquestionably wrongful convictions, also due to subsequent public uproar, may lead to declines in conviction rates, likely due to judges becoming even more sensitive to false positives (Sonnemans and van Dijk, 2012).

[5] Also the so called Blackstone's (1769) adage (see on this also Section 3.2) is in line with this clearly pro-defendant philosophy. Rizzolli and Saraceno's (2013) theoretical analysis identifies conditions under which type-I errors are

type-II errors: false negatives entail that factually-guilty criminals are free to act and possibly commit further crimes and this is likely to generate important concerns in the lay public.

We believe that shedding light empirically on how judges decide when false positives and false negatives are likely to arise is a critical area of inquiry, especially for its behavioral and institutional implications (Cappelen et al., 2018). In this regard, key contributions may come from research at the intersection of behavioral economics and law (DeAngelo and Charness, 2012). In this study, we address the aforementioned questions by conducting an incentivized, controlled laboratory experiment guided by the theoretical model illustrated in the next section. The main reason why we have recourse to the experimental methodology, despite the well-known problems concerning the generalizability of laboratory findings (Levitt and List, 2007), is that for a variety of reasons carrying out rigorous field research on judicial decision making is inherently problematic (Sonnemans and van Dijk, 2012). Due to serious measurement error concerns, it is extremely challenging to produce reliable measures of the frequency of type-I and type-II errors in a given context (Anderson and Stafford, 2003; Gross et al., 2014) and rigorously analyze judges' decision making with regard to both probability and severity of punishment when type-I and type-II errors can occur.

There are three defining features of our design – detailed in Section 3 – that are key to interpreting our core findings. The first concerns the possibility to provide, within a unified experimental setting, a direct, clean comparison between conditions in which players know that their decisions occur under certainty (i.e., without errors) and conditions in which the same participants are informed that either type-I or type-II error may arise. The second distinguishing characteristic of our controlled environment regards the specificities of judges' decision set: in all conditions, participants not only decide upon *punishment* (i.e., convicting vs. acquitting the defendant), that might generate one of the two error types, but also upon the *severity* of punishment (which can be low, medium or high). In our view, also the possibility for our subjects to make decisions on both dimensions is a critical feature of experimental settings that aim to closely mirror naturally occurring judging, in which either type-I or type-II error may emerge. Finally, actual judges often fear that the evidence provided by the prosecutors over the defendant's behavior is not enough and call for further inquiry: to capture these decision-making options, our experimental setting includes rounds in which players can obtain *additional evidence* at a cost in order to decide on sounder evidential basis.[6]

more socially costly than type-II errors. Rizzolli and Stanca's (2012) experimental findings suggest that the two errors have asymmetric effects on deterrence, with an increase in the probability of type-I errors having a larger negative impact. Markussen et al.'s (2016) analysis on judicial errors and cooperation documents that, in their experimental setting, individuals dislike type-I errors more than type-II errors.

[6] The effects of evidentiary uncertainty – though in a different setting – have been studied experimentally by

The remainder of the paper proceeds as follows. Section 2 illustrates a theoretical model that frames our experimental analysis. Section 3 outlines the experimental design and in Section 4 we describe our major results. Section 5 concludes.

2. Theory

To provide guidance to our experimental analysis, we have recourse to a simple model based on Feess et al. (2018).[7] Our model shows that judges' attitude toward punishment is affected by the presence of erroneous reporting that induce type-I and type-II errors, since the mix of probability and severity of punishment is adjusted by accounting for the possibility to decide incorrectly. Next, it also sheds light on indirect effects induced by the possibility to generate erroneous verdicts.

Assumptions

In line with the experimental design described in Section 3, we assume that judges face a dichotomous decision: to punish or acquit a defendant, given a certain investigation report of guilt or innocence. The investigation report is erroneous with a certain probability. Let us define:

- $\varepsilon_1 = Prob.(guilt\, reporting|innocent)$ as the exogenous probability that an innocent is reported as guilty to the judge
- $\varepsilon_2 = Prob.(innocence\, reporting|guilty)$ as the exogenous probability that a guilty person is reported as innocent to the judge.

Both probabilities are common knowledge. Note that these probabilities are not yet probabilities of judicial errors; the fact that reports are erroneous with some probability simply places judges in the position of judging under the possibility of making a mistake. $(1 - \varepsilon_1 - \varepsilon_2) \geq 0$ by assumption, this is consistent with the experimental setup outlined in Section 3 and means that a minimum accuracy level in investigation reporting is granted.

- $q = \text{Prob.(guilty)}$ as the share of the population who commit an infringement. We temporarily consider q as given, although it will be ultimately determined endogenously because people decide to stay innocent or commit an infringement depending on judges' attitude towards

Grechenig et al. (2010) and Ambrus and Greiner (2012).

[7] Feess et al. (2018) present a two-sided game under incomplete information involving a potential violator and a judge. Without observing the defendant's behavior, the judge decides whether or not to impose an exogenous fee on the defendant. Their model focuses on the interdependency of judges' and potential violators' decisions and shows that higher fines reduce the punishment frequency and higher legal uncertainty increases the violation frequency. However, the effect of the fine size on the violation frequency and the effect of legal uncertainty on the punishment frequency are theoretically ambiguous.

punishment (*deterrence effect*).

- *H*>0 as the harm resulting from the infringement; for the sake of simplicity it corresponds to the private benefit of committing the infringement.
- *F*>0 as the fine that the defendant receives when punished.

Both the fine and the harm are common knowledge.

As in Feess et al. (2018), we make the following assumptions concerning judges' behavior:

- Each judge i (players Cs in the experiment) is characterized by two parameters:
  - $\alpha_i \epsilon (0,1)$ that is *i*'s *aversion to type-I error*. The disutility deriving from punishing an innocent is measured as $\alpha_i F$.
  - $\beta_i \epsilon (0,1)$ that is *i*'s *aversion to type-II error*. The disutility deriving from acquitting a guilty person is measured as $\beta_i H$.

Judge's utility from correct decisions is set to zero.

In order to model behavior of potential tortfeasors, we add the following assumptions:

- Each potential tortfeasor j (players As in the experiment) is risk neutral and characterized by a nonnegative parameter s_j, capturing the individual sensitivity towards social damage. $s_j H$ measures the individual benefit resulting from refraining from misbehavior.

Risk neutrality allows us to keep the model easily manageable. However, this assumption will be relaxed in the experimental setting since risk-aversion-elicitation allows us to control for this individual characteristic. Finally, let us define:

- p_I as the probabilities of being punished in the case of innocence reporting.
- p_G as the probabilities of being punished in the case of guilt reporting.

We assume $(p_G - p_I) \geq 0$. This sounds natural since it is plausible that people consider more (or at least equally) likely to be punished given a guilt reporting than given an innocence reporting. As for q, these probabilities will be endogenously determined in equilibrium, depending on how judges eventually decide to behave.

Judges' behavior for a given q (share of people who commit infringements)

A judge who assumes a certain q (that will be endogenously derived later) decides to punish only if the disutility of punishing an innocent is smaller than the disutility of acquitting a guilty person. Therefore:

- in the case of guilt reporting, judge i punishes if $\varepsilon_1(1-q)\alpha_i F < \beta_i(1-\varepsilon_2)qH$; that is when:

$$\frac{\alpha_i}{\beta_i} < \frac{(1-\varepsilon_2)q}{\varepsilon_1(1-q)} \frac{H}{F} \equiv r_G \tag{1}$$

where r_G is the judge's threshold-type in the case of guilt reporting such that all the judges who are characterized by a Blackstone ratio $\frac{\alpha_i}{\beta_i} < r_G$ prefer to punish.

- in the case of innocence reporting, judge *i* punishes if:

$$\frac{\alpha_i}{\beta_i} \leq \frac{\varepsilon_2 q}{(1-\varepsilon_1)(1-q)} \frac{H}{F} \equiv r_I \tag{2}$$

where r_I is the threshold-type in the case of innocence reporting such that all the judges who are characterized by a Blackstone ratio $\frac{\alpha_i}{\beta_i} \leq r_I$ prefer to punish.

(1) and (2) characterize judges' optimal decisions based on their beliefs about defendants' behavior (q). Note the symmetry in (1) and (2): without reporting errors, in the case of guilt reporting only "extreme laxity-lovers" who are characterized by a Blackstone ratio approaching to infinity decide in favor of acquittal; conversely, in the case of innocence reporting only "extreme punishers" who are characterized by a Blackstone ratio approaching to zero decide in favor of punishment.

Potential tortfeasors' behavior for a given probability to be punished

Let us now focus on behavior of potential tortfeasors. Although p_I and p_G (the probabilities of being punished in the case of innocence/guilt reporting, respectively) will be endogenously determined, we provisionally assume them as given. Therefore, potential tortfeasor *j* decides to commit the infringement if the private benefit from the infringement net of the expected sanction is greater than the private benefit of abstaining net of the expected sanction of being punished despite being innocent; that is to say, when: $H - F(\varepsilon_2 p_I + (1-\varepsilon_2)p_G) > s_i H - F\big((1-\varepsilon_1)p_I + \varepsilon_1\, p_G\big)$.

It follows that *j* commits the infringement if:

$$s_j < 1 - \frac{F}{H}(p_G - p_I)(1-\varepsilon_1-\varepsilon_2) \equiv \tilde{s} \tag{3}$$

$\tilde{s}$, nonnegative,[8] is the threshold-type of potential tortfeasors such that all the potential tortfeasors who are characterized by $s_i < \tilde{s}$ opt for infringement. (3) characterizes potential tortfeasors' optimal decision based on their expectations over judges' behavior (which, in turn, depends on reporting).

[8] We focus on cases corresponding to a parameter restriction such that $F(p_G - p_I)(1-\varepsilon_1-\varepsilon_2) \leq H$. This is consistent with our experimental setup and parameters (see Table 4, first line); it simply means that the sanction system does not allow for "over-punishment", i.e. free from error sanctioning and with judges who are perfectly able to separate guilty people from innocent ones ($p_G - p_I = 1$) $F \leq H$; otherwise adjustments are possible, but with limitations to avoid disproportionate sanctions.

By inspecting (1), (2) and (3), we derive the following proposition and the related implications (Proofs in Appendix).

Proposition 1 *Suppose that judges assume* q *as the probability that people commit infringement and that potential tortfeasors assume* p_G *and* p_I *as the probabilities of punishment in the case of guilt and innocence investigation reporting, respectively. Then,*

(i) $\frac{dr_G}{d\varepsilon_1} < 0$ $\quad \frac{dr_G}{d\varepsilon_2} < 0$ $\quad \frac{dr_G}{dF} < 0$ $\quad \frac{dr_G}{dH} > 0$ $\quad \frac{dr_G}{dq} > 0$

(ii) $\frac{dr_I}{d\varepsilon_1} > 0$ $\quad \frac{dr_I}{d\varepsilon_2} > 0$ $\quad \frac{dr_I}{dF} < 0$ $\quad \frac{dr_I}{dH} > 0$ $\quad \frac{dr_I}{dq} > 0$

(iii) $\frac{d\tilde{s}}{d\varepsilon_1} = \frac{d\tilde{s}}{d\varepsilon_2} \geq 0$ $\quad \frac{d\tilde{s}}{dF} \leq 0$ $\quad \frac{d\tilde{s}}{dH} \geq 0$ $\quad \frac{d\tilde{s}}{d(p_G - p_I)} \leq 0$

Implications

Parts *(i)* and *(ii)* of Proposition 1 illustrate how judges' punishment behavior – conditional on reporting and given a certain probability that people commit infringements – changes depending on the parameters of the model. For a give share of people who commit infringements (q), reporting errors determine intuitive effects on judges' punitive attitude: when the probability that an innocent is reported as guilty increases, punishment conditions (1) and (2) are hardly satisfied; conversely, when the probability that a guilty person is reported as innocent increases, conditions are more easily satisfied. Not surprisingly,[9] both a higher harm (*H*) and higher probability that people commit infringement (q) favor punitive attitude. Conversely, punitive attitude is moderated by tougher sanctions (*F*). Sanctions and punitive attitude act therefore as substitutes (Becker, 1968; Feess et al., 2018).

Part *(iii)* illustrates preliminary implications in terms of deterrence (given certain probabilities of being punished). Not surprisingly, errors in reporting are equally detrimental in terms of deterrence (Png, 1986). Tougher sanctions S discourage infringements, while higher private benefits H favor infringement. Finally, when reporting is at least partially informative ($(\varepsilon_1 + \varepsilon_2) < 1$), deterrence increases when judges can discriminate their punitive attitude on the basis of reporting (measured by $(p_G - p_I)$).

These theoretical implications can help guiding the interpretation of the results of the experiment. For given q, p_G, and p_I, errors in reporting and the related possibility to decide

[9] According to Feess et al. (2018), higher fines reduce both the violation and the punishment frequencies, whereas when there is greater uncertainty we have an increase in the violation frequency but a decrease in the punishment frequency.

incorrectly modify the punitive attitude of judges through a *complex pathway*: judges' are likely to adjust their decisions accounting both for their original concerns about making errors and concerns about inducing underdeterrence.

Equilibrium behavior

The considerations developed so far are derived by assuming q, p_G and p_I as given. However, these probabilities are ultimately *endogenous*, since behaviors of judges and potential violators are interdependent (Feess et al. 2018). In Bayesian Nash Equilibrium, expectations (LHS in the equation below) and actual probabilities (RHS in the equation below) coincide, such that we have:

$$\begin{cases} p_G = W(r_G) \\ p_I = W(r_I) \\ q = B(\tilde{s}) \end{cases} \tag{4}$$

where W and B denote the cumulative distribution functions of judges' r_i and potential violators' s_i, respectively. The system of total differentials of (4) is:

$$\begin{bmatrix} 1 & 0 & -w_G \dfrac{1-\varepsilon_2}{\varepsilon_1(1-q)^2}\dfrac{H}{F} \\ 0 & 1 & -w_I \dfrac{\varepsilon_2}{(1-\varepsilon_1)(1-q)^2}\dfrac{H}{F} \\ b\dfrac{F}{H}(1-\varepsilon_1-\varepsilon_2) & \dfrac{F}{H}(1-\varepsilon_1-\varepsilon_2) & 1 \end{bmatrix} \begin{bmatrix} dp_G \\ dp_I \\ dq \end{bmatrix}$$

$$= \begin{bmatrix} -w_G \dfrac{(1-\varepsilon_2)q}{\varepsilon_1^2(1-q)}\dfrac{H}{F} & -w_G \dfrac{q}{\varepsilon_1(1-q)}\dfrac{H}{F} & -w_G \dfrac{(1-\varepsilon_2)q}{\varepsilon_1(1-q)}\dfrac{H}{F^2} \\ w_I \dfrac{\varepsilon_2 q}{(1-\varepsilon_1)^2(1-q)}\dfrac{H}{F} & w_I \dfrac{q}{(1-\varepsilon_1)(1-q)}\dfrac{H}{F} & -w_I \dfrac{\varepsilon_2 q}{(1-\varepsilon_1)(1-q)}\dfrac{H}{F^2} \\ b\dfrac{F}{H}(p_G-p_I) & b\dfrac{F}{H}(p_G-p_I) & -\dfrac{b}{H}(p_G-p_I)(1-\varepsilon_1-\varepsilon_2) \end{bmatrix}$$

This allows us to conclude with the following proposition (Proofs in Appendix).

Proposition 2 *Name p_G^*, p_I^* and, q^* the Bayesian Nash equilibrium probabilities solving the system of equations* (4)*. Then,*

(i) $\dfrac{dp_G^*}{d\varepsilon_1} \gtreqless 0$ $\qquad$ $\dfrac{\partial p_G^*}{d\varepsilon_2} \gtreqless 0$ $\qquad$ $\dfrac{dp_G^*}{dF} < 0$

(ii) $\dfrac{dp_I^*}{d\varepsilon_1} \geq 0$ $\qquad$ $\dfrac{dp_I^*}{d\varepsilon_2} \geq 0$ $\qquad$ $\dfrac{dp_I^*}{dF} \gtreqless 0$

(iii) $\dfrac{dq^*}{d\varepsilon_1} \gtreqless 0$ $\qquad$ $\dfrac{dq^*}{d\varepsilon_2} \gtreqless 0$ $\qquad$ $\dfrac{dq^*}{dF} \gtreqless 0$

Implications

Unlike in Proposition 1, relations described in Proposition 2 are not univocal. Actually, when potential tortfeasors' behavior is endogenous, the pathway described above implies that reporting errors determine various effects.

In this regard, Part *(i)* and *(ii)* of Proposition 2 provides analogous implications, albeit differently articulated.[10] On the one hand, Part *(i)* shows that when the behavior of potential tortfeasors is endogenous, reporting errors can still produce the expected effect, inducing more caution (resp., less caution) in punishment when the probability that an innocent (resp., a guilty person) is reported as guilty (resp., innocent) increases. However, reporting errors might lead judges to increase the probability of punishment to compensate underdeterrence induced by errors themselves. Here, the relation of substitutability between probability and severity of punishment is confirmed since the probability of punishment in the case of guilt reporting is decreasing in fine size. On the other hand, Part *(ii)* shows that, when the behavior of potential tortfeasors is endogenous, the equilibrium probability of punishment in the case of innocence reporting is nondecreasing in both types of reporting errors: judges react to potential error-induced underdeterrence by keeping constant/increasing the probability of punishment even when the probability that an innocent is reported as guilty increases. On the other hand, the impact of fine size is ambiguous since the severity of punishment may be used either as a complement or as a substitute to *finally adjust* the overall punitive attitude.

Finally, Part (*iii*) shows that both types of errors in reporting and fine ambiguously affect the equilibrium probability of committing infringements. Again, we have forces working in opposite directions: on the one hand, errors imply underdeterrence; on the other hand, potential tortfeasors may fear an increase in the probability of being punished due to judges' desire to compensate error-induced underdeterrence.

In light of these theoretical results, we conclude that the overall attitude of judges toward punishment is affected by the presence of reporting errors. More specifically, the possibility to decide incorrectly because of erroneous investigation reporting can lead judges to adjust the mix of probability and severity of punishment accounting for both the uncertain nature of the defendant and underdeterrence that incorrect decisions may induce (in turn, error-induced variations in deterrence depend on how potential tortfeasors expect that judges adjust their punitive behavior). How these adjustments are made ultimately remains an empirical question.

First, we are interested in investigating empirically whether the two errors induce different effects (i.e., whether judges are more or less sensitive to one type of errors with respect

[10] Differences are due to how we defined the relevant thresholds as well as to computational reasons.

to the other). Second, we seek to understand whether and how judges modify their punitive attitude (probability and severity mix) accounting for error-induced underdeterrence. Third, relaxing the theoretical assumption of risk neutrality, the experiment allows us to account for risk aversion (that is elicited in the lab). In this regard, Rizzolli and Stanca (2012) already showed that risk aversion makes the impact of type-I and type-II errors different on deterrence. Similarly, we could expect a different sensitivity of judges towards the two types of errors depending on their risk aversion. However, considering risk aversion does not jeopardize the main implication of the model: errors affect punitive attitude both directly and indirectly, ultimately leading to adjustments in the mix of probability and severity of punishment. Fourth, our experimental setup allowed us to shed light on judges' sensitivity towards type-I and type-II errors along two further directions: (i) the individual Blackstone ratio ($\frac{\alpha_i}{\beta_i}$) of participants is directly elicited and (ii) participants can obtain additional evidence at a cost to eliminate the risk of reaching a wrong verdict (see the next section for details on this).

3. The Laboratory Experiment

The key features of real-life scenarios that we aim to mirror in the laboratory can be shortly described as follows: an individual must decide whether to commit an infringement or not; a judge has to decide whether to consider him as guilty and, if this is the case, to choose the size of the fine. This process may lead to a verdict characterized by one out of two types of judicial errors: innocent defendants might be mistakenly judged guilty (*type-I error*) and guilty defendants might be erroneously judged innocent (*type-II error*). As we anticipated in the previous sections, our experiment is aimed at testing whether judges are sensitive to these two errors, by addressing the following questions: does the possibility to reach a wrong verdict impact judges' punishment decisions? If this is the case, does this mainly occur with regard to one specific type of error or not? Do judges adjust the mix of probability and severity of punishment, in response to the possibility to make erroneous decisions? Are they willing to *spend more to eliminate the error* and make a decision under certainty? In Section 3.1 we provide a detailed description of our experimental design, while Section 3.2 illustrates the experimental procedure and Section 3.3. describes the experimental sample.

3.1. Experimental Design: the "Theft Game"

Each session involves 24 players: 8 players As (with A being the Dictator – the potential "Thief"), 8 players Bs (with B being the Dummy player – the potential "Victim") and 8 players

Cs (with C being the "Judge"). In each session, players participate in five rounds, corresponding to the five experimental conditions illustrated here below. At the beginning of each round, players receive their initial endowment according to their (randomly assigned) role. Each player A and B receives 100 tokens while each player C receives 200 tokens. Each round consists of two stages. In the first stage, A and B are paired and participate in a "Theft Game" (that is, a reverse dictator game), where A has the opportunity to commit an infringement, i.e. "stealing" 50 tokens from B.[11] During this stage, players As' choices are recorded through the computer. In the second stage, each player C is assigned to a couple who has played the Theft Game in the first stage and is asked to choose whether to punish or not A (*decision over punishment)* and, in case player C decides to punish, whether she is willing to punish A by 40, 50 or 60 tokens (*decisions over the severity of punishment*, which can be *low, medium* or *high*) for each of the two possible scenarios – the computer communicates that player A has/has not stolen the 50 tokens from B.[12] The sanction is costless for C. The difference among the five conditions associated with the different rounds lies in the *accuracy of judges' information* over the infringement. More specifically:

Round 1 ("No error" condition). In round 1, the computer *properly records* player A's choices and players Cs receive correct information. This implies that player C has the opportunity to make her choice *under certainty*, without any fear of making judicial errors. Since we implement the strategy method at this stage, we ask subjects Cs to choose for both potential events that may be communicated by the computer (the theft occurs vs no theft occurs), through a simple graphical representation describing the two possible scenarios (see Figure 1). Each player C perfectly knows that, at the end of the experiment, the computer will implement her choice according to the decision made by player A and actually recorded by the computer: this procedure is common to all five rounds.

Round 2 ("Type-I error" condition). In round 2, the computer *makes type-I error with 50% of probability* (i.e., a theft is recorded and reported to player C even if player A decided not to steal money from B). In this case, when player C decides to punish player A when the computer reports that a theft occurred, it is possible that an innocent is sanctioned and type-I error affects

[11] It is important to note that, in line with prior experimental work at the intersection between law and behavioral economics (see e.g. Falk and Fischbacher, 2002, Schildberg-Hörisch and Strassmair, 2012, and Faillo et al., 2013), we purposely used neutral, rather than loaded, language in the experiment, as we believe that our deliberately abstract, context-free presentation of experimental tasks transmits the key features of judicial decision making we are interested in shedding light on. Therefore, we never used terms such as e.g. "steal", "victim" or "crime" in our experimental instructions (see on this also Abbink and Hennig-Schmidt, 2006).

[12] Technically, we implement the strategy method (Falk and Fischbacher, 2002; Brandts and Charness, 2011; Jordan et al., 2016).

not only information but also the final sentence. On the other hand, if player C decides not to punish player A when the computer reports the infringement, *type-II error may emerge* (see player C's decision screen in Figure 2).[13]

Round 3 ("Type-II error" condition). In round 3, the computer *makes type-II error with 50% of probability* (i.e., the computer fails to record a theft and to report it to player C even if player A decided to steal money from B). The situation is totally symmetrical with respect to the previous round. When player C decides not to punish player A when the computer reports that a theft did not occur, it is possible that a guilty person remains unpunished and type-II error affects the final sentence. However, if player C decides to punish player A when the computer reports the infringement, *type-I error may emerge* (see player C's decision screen in Figure 3).

Rounds 4 and 5 ("Type-I and Type-II error correction" conditions). In rounds 4 and 5, *mistakenly recorded choices occur*: in round 4 they are characterized by the presence of type-I error, like in round 2, whereas in round 5 they occur in the presence of type-II errors, like in round 3. The novelty of these last two rounds, compared to rounds 2 and 3, is that player C can choose to spend 20 tokens (per round) to *eliminate the error* and decide whether to sanction A or not under certainty with regard to A being guilty or innocent, like in round 1. This feature of our two last experimental conditions aims at mirroring real-life situations in which, say, in courts the judge believes that the amount of evidence provided by the prosecutor over the suspect's behavior is not enough and decides to call for further inquiry (e.g., having access to defendant's DNA information, in case of crimes). In both rounds, when player C pays for the correct information, the decision screen is the same as the one used in the "no error" condition (round 1). If this is not the case, the decision screen in rounds 4 and 5 is the same as in rounds 2 and 3, respectively.

To minimize endowment and learning effects, at the beginning of each round players are rematched according to a stranger protocol and no feedback concerning punishment decisions is provided until the end of the session.

3.2. Experimental Procedure

Overall, 168 students enrolled at the University of Verona participated in the experiment. No student took part in more than one session. We ran all of the 7 sessions at VELE, the

[13] Therefore, our experimental design makes clear the existence of the well-known inherent trade-off affecting judging decisions recalled in the Introduction (see also Round 3, on this).

experimental lab at the Economics Department of the University of Verona. Decisions and performance were recorded through the computer, and the experiment was programmed and conducted with z-Tree (Fischbacher, 2007). Participants entered the laboratory and took a seat in front of a computer. They were immediately asked to switch off their mobiles and to stop talking to their colleagues. Participants on their computer screen read instructions while an experimenter read them aloud.[14] A copy of the graphical descriptions of players Cs' decisions was handed out round by round to let people analyze them before making their decisions. To collect all the parameters we needed to test our model, we asked people to fill in some brief (incentivized) questionnaires and we elicited subjects' beliefs concerning the behavior of the other participants.[15] More specifically, at the end of each session, we collected participants' socio-demographic characteristics, attitudes towards risk, concerns about others, their Blackstone's ratio (and its inverse with regard to different criminal contexts), player Cs' beliefs concerning the expected number of thieves in the room, players As' beliefs concerning players Cs' choices (elicited at the end of each round) and players Bs' beliefs concerning both players As' and players Cs' choices. The average duration was 75 minutes, and the average payoff was around €17.[16] The experiment adopted a single-blind procedure and preserved anonymity among participants.

3.3. Experimental Sample

168 subjects participated in the experiment (51% of them for the first time). They were on average 21.7 years old and 46% of them were males. Ninety-four percent of participants were Italian while the remaining 6% were from Eastern Europe. In terms of their political orientation and their religious beliefs, we found that 60% of the sample consisted of believers (mainly

[14] Original instructions were written and read in Italian. A translation of experimental instructions is provided in the supplementary material.

[15] As to participants' socio-demographics, we asked information about gender (MALE in the regression); age (AGE); first participation in experiments (FIRST); religion (BELIEVER); political orientation (POLITICAL ORIENTATION). As pointed out by Cohen and Yang (2019), a large literature documented the effects of judges' characteristics (including their political preferences) on their decision-making process. Participants' attitude towards risk has been assessed through the "Bomb" Risk Elicitation Task (BRET – Crosetto and Filippin, 2013). Participants' concerns about others has been measured by means of the Social Value Orientation questionnaire (SVO – Murphy, 2011). We also assessed participants' Blackstone's ratio and its inverse with regard to different criminal contexts, i.e. GENERALLY, in case of ROBBERY in a supermarket and in case of MURDER. Specifically, we asked subjects to answer the following questions: a) Generally/In case of robbery in a supermarket/In case of murder, what is the worst outcome between incorrectly acquitting a guilty party or incorrectly sanctioning an innocent person? 1. Incorrectly acquitting a guilty party; 2. Incorrectly sanctioning an innocent person; 3. No difference. b) If you chose 1, please tell us how many incorrectly sanctioned innocent people you are ready to imprison to avoid that a SINGLE guilty party is incorrectly acquitted. c) If you chose 2, please tell us how many incorrectly acquitted guilty parties you are ready to free to avoid that a SINGLE innocent person is incorrectly sanctioned. While the existing literature typically includes the direct measure, we decided to introduce also the inverse of the Blackstone's ratio, to allow for the possibility that some people may be willing to put in prison more than one innocent people to avoid the acquittal of a guilty person.

[16] The translation of this questionnaire is available upon request from the authors.

Catholics), around 40% is center-left oriented, 20% prefer center-right parties, 8% opt for a pure centrist position while 21% has no clear political tendency. Concerning their attitude towards risk, elicited through the BRET, 85% can be labeled as risk averse, 5% as risk neutral and 10% as risk lovers. According to the SVO questionnaire, 45% of the sample consists of Prosocial subjects, while the remaining 55% is represented by Individualists.

4. Experimental Results

In this section, we report the major findings of our data analysis, based on a description of players Cs' choices during the 5 rounds. In the laboratory, 56 participants were assigned the C player role. In the first three rounds they faced three different experimental conditions where the information provided by the computer concerning player A's behavior was correct (round 1), affected by the possibility of type-I error (round 2) and affected by the possibility of type-II error (round 3), respectively. In the last two rounds, they had the opportunity to spend money in order to avoid receiving wrong information, like in round 1. Consequently, in the fourth round, each player C could face either a certainty scenario (without errors) or an uncertain scenario where type-I error may occur, depending on her decision to buy the correct information or not. Symmetrically, in the last round, each player C could face either a certainty scenario (without errors) or an uncertain scenario where type-II error may occur. Due to the complexity of the last two rounds, for ease of exposition in the remainder of this section we illustrate our findings concerning players Cs' behavior in two steps: first, we focus on the first three rounds (Section 4.1); then we separately refer to the last two rounds (Section 4.2).

4.1. Rounds 1-3

Both in the first and in the third round (where information may be affected by type-II error), when the computer reports that player A subtracted 50 tokens to B, player C receives a correct information. On the other hand, in the second round, it is possible that information is affected by type-I error and that an innocent player A is wrongly accused. Similarly, when the computer reports that player A did not subtract 50 tokens to player B, it communicates the right information in the first two rounds, while it is possible that in round 3 a type-II error occurs and a theft remains wrongly undetected. While subjects Cs' reaction is more obvious when information is not affected by any type of error, the scenario is more complicated when the computer may communicate incorrect messages. Thus, in rounds 2 and 3, when the computer reports that A stole/did not steal money, respectively, players Cs' decision to punish A may depend on her aversion for a specific error as well as her beliefs, i.e. the subjective probability that A is actually a thief. The point can be made clearer through the following example. If player C is strongly averse to type-I error and thinks that players As have a high probability of being honest, in round 2, when the computer reports that player A stole money, probably she will not implement any punishment since she believes that the computer is likely to report an incorrect message. However, if in the same scenario the same (type-I-error averse) player C believes that, say, all players As are thieves, she will punish, due to her belief that type-I error may not occur. The same happens when we consider people who are strongly averse to type-II error. If player C thinks that players A hardly steal money, she will probably decide not to punish in round 3 when the computer reports that A did not steal money. However, if she believes that every A stole money, she will punish due to her belief that type-II error occurs for sure. Consider that, in both scenarios, players Cs' beliefs on players As' behavior might also be affected by the possibility that errors induce underdeterrence expectations.

As we made clear in Section 3.2, we collected both information concerning players' attitude towards type-I and type-II errors in three different situations (i.e. GENERALLY, ROBBERY, MURDER) and participants' beliefs on the number of thieves in each round. We then perform the following procedure:

1) according to players Cs' claim about their aversion to one of the two errors in the three different scenarios, we identify both the maximum number of thieves player C is ready to bear in order not to punish when the computer reports that player A stole money and the minimum number of thieves that leads player C to punish when the computer reports that player A did not steal money;
2) we compare players Cs' beliefs on the number of thieves in the different scenarios with the thresholds computed at step 1;
3) for each round, we compute the hypothetical proportion of punishment.

Tables 1 and 2 report the distribution of players Cs' attitude towards judicial errors in the three specific situations we propose and the distribution of their beliefs concerning the number of thieves in each round. What emerges is that players Cs are *mostly indifferent* between the two errors, while the second most popular answer supports type-I aversion. Moreover, generally, players Cs believe that a high number of players As are thieves: in particular, we find that player C's expectation is that, on average, more than 5 players A are thieves in each scenario (see Table 2 and Figure 4).

Table 3 reports the result of the previously illustrated procedure while Table 4 reports both the hypothetical and the actual values of punishment in each round. In most cases, hypothetical data are not far from actual ones. Tables 4 and 5 reveal that judges are strongly affected by the possibility that type-I error occurs (in round 2) when they have to decide the amount of the sanction, but not when they have to decide whether to punish or not. More specifically, the possibility that type-I error occurs strongly increases the number of cases in which player C chooses a fee of 40 tokens (30% vs 9% and 13%) and strongly decreases fees of 60 tokens (27% vs 52% and 59%).

Also in round 3, when information may be affected by type-II error, players Cs react to the possibility that a judicial error occurs. In particular, they strongly increase the probability of punishing player A when the computer reports that player A did not steal money (61% vs 22% and 25% in rounds 1 and 2, respectively). However, also in this case, the number of lenient punishments increases (39% vs 11% and 14% in rounds 1 and 2, respectively).

The econometric analysis corroborates this evidence (see Tables 6 to 9). We perform a series of multinomial probit regressions clustered at the subject level. The dependent variable is represented by player C's choices (a variable with four categories: NO FEE, FEE=40, FEE=50, FEE=60). The regressors are the socio-economic variables and the personal characteristics presented in Section 3, player C's belief on the number of thieves in each round and dummy variables for the experimental condition (NO ERROR, TYPE_I_ERROR, TYPE_II_ERROR). As to rounds 1-3, we then summarize the key converging findings obtained through the analyses illustrated in this section as follows:

Result 1. *Both types of errors affect judges' behavior.*

As predicted in the theoretical section, the possibility to decide incorrectly because of erroneous investigation reporting makes judges adjust the probability and severity of punishment mix. In particular:

Result 1a. *When information is affected by type-I error, players Cs opt for a more lenient fee when the computer reports that a theft occurred.*

Result 1b. *When information is affected by type-II error, players Cs increase the percentage of cases in which they punish player A when the computer reports that player A did not steal money. However, also in this case, the lenient fee is the favored one.*

Is this reaction due to some extent to judges' expectation of underdeterrence? Though we are aware that answering this question empirically is a challenging task, we seek to gain some insight about this relevant aspect and check whether player C's beliefs concerning the number of thieves change when errors may occur, with respect to the scenario where information is correct. Table 10 reports the results of a random-effects probit regression where the dependent variable is players Cs' beliefs on players As' behavior and the regressors include dummy variables for the experimental condition (TYPE_I_ERROR, TYPE_II_ERROR, TYPE_I_ERROR_CORR, TYPE_II_ERROR_CORR). According to the results we obtain, the only experimental condition where expectation of underdeterrence seems to occur – that is, the expected number of thieves significantly increases – is in the TYPE_II_ERROR condition (p = 0.056).[17]

4.2. Rounds 4 and 5

When analyzing players Cs' choices in the last two rounds, what emerges first is that they buy more frequently the correct information when type-I error may occur (48% vs 39% in the last round when type-II error may occur). However, this difference is not statistically significant (chi2 test, p=0.341) and 34% of players Cs buy the correct information in both experimental conditions. If we consider subjects who buy the correct information (27 and 22 in the fourth and fifth round, respectively), we find that, when the computer reports that player A stole the 50 tokens from player B, their choices are in line with round 1, as most of them (96% and 91%) decide to punish players As. On the other hand, when the computer reports that player A did not steal money, the percentage of subjects Cs who punish players As decreases drastically with respect to the first round (8% in both cases; see Tables 11 and 12). An interesting point is that, with respect to round 1, players Cs increase the number of severe fees (60) and decrease the number of lenient fees (40).[18]

When subjects Cs decide not to buy the correct information (24 subjects in round 4 and 31 in round 5), judicial errors may occur and the scenarios that we face are the same as in the second and in the third round. When information may be affected by type-I error and the computer reports that player A has stolen money, players Cs do not decrease their probability of punishing players As, but they are more likely to choose the lenient fee of 40 tokens (as in round 2). At the same time, in round 5, when information may be affected by type-II error and the computer reports that player A has not stolen any money, the probability of punishing player A drastically increases (as in round 3; see Tables 11 and 12).

[17] This finding may be due to a large extent to the fact that, over the whole experiment, players Cs believe that most of players As will be thieves.

[18] A possible explanation is that player C thinks that, if player A steals during the last rounds when judges have the opportunity to buy perfect information, maybe they did the same when errors cannot be avoided. Thus, they decided to punish more, also to compensate previously (potentially) unpunished thefts.

Again, the econometric analysis provides support to our results. We split our analysis between cases in which the correct information has been bought and cases where the errors are not deleted. We perform four series of multinomial probit regressions clustered at the subject level: 1) on data from rounds 1, 2 and 4 when the information is bought; 2) on data from rounds 1, 2 and 4 when the information is not bought; 3) on data from rounds 1, 3 and 5 when the information is bought; 4) on data from rounds 1, 3 and 5 when the information is not bought. The dependent variable is represented, again, by players Cs' choices (a variable with four categories: NO FEE, FEE=40, FEE=50, FEE=60). Socio-economic variables and the personal characteristics used as regressors are the same as in the previous slot of analyses. Dummy variables for the experimental condition are included (NO ERROR, TYPE_I_ERROR, TYPE_II_ERROR, TYPE_I_ERROR_CORR, TYPE_II_ERROR_CORR; see Tables from 13 to 20 on this).

Result 2. *When judges have the opportunity to invest resources to avoid errors, they do it with regard to both types of errors.*

5. Discussion and concluding remarks

The core finding we obtained through our laboratory experiment is that – contrary to a widespread view – judges are sensitive not only to the possibility of type-I error occurrence but to *both types of error*. Several pieces of evidence clearly point to this direction: subjects' attitude towards the two error types in different situations and (especially) subjects' punishment behavior when type-I and type-II error may occur (Result 1) and subjects' willingness to pay to eliminate errors (Result 2). More specifically, our findings indicated that, as far as type-I error is concerned (Result 1a), when the computer reports that a theft occurred, judges reduce punishment (compared to our baseline "no error" condition), as expected: however, this interestingly occurs not via lower punishment frequency, but with players opting for *more lenient sanctions*. We view this evidence as broadly consistent with internationally known verdicts in which judges eventually convicted defendants, but, at the same time, opted for relatively mild sentences. We speculatively argue that in these real-life cases media coverage played a non-negligible role in (more or less subconsciously) influencing judges' attitudes towards sentencing decisions for which – possibly also due to the fear of type-II error occurrence – they sought a solution to "close the case". However, at the same time, it may well be the case that judges in their minds also seriously considered the possibility of wrongful conviction. In other words, the finding on punishment behavior we obtained may be closely related to judges' overall management of the inherent trade-off between type-I and type-II errors recalled in our introductory section (Cappelen et al., 2018). As to type-II errors (Result 1b), when the computer reports that the defendant did not steal money, punishment frequency increases (compared to the baseline "no error" condition), as expected; however, also in case of false negatives, subjects opt for higher leniency in sanctioning. Though also the latter finding deserves further investigation, we believe that a plausible interpretation is that, also in this case, subjects are aware of the trade-off between the two error types and that, therefore, higher punishment frequency lowers the probability of type-II occurrence *but at the cost of increasing the probability of type-I error*: in other words, opting for a more lenient sentence in round 3 may underlie a subtle form of type-I error avoidance. Therefore, while our experimental findings deserve further investigation to shed light on participants' motives, the "compensatory leniency" that we detect in judges' decision making both when type-I errors and when type-II errors may occur suggests that participants seek to balance the inherent trade-off between the two errors by jointly managing the two key levers they are provided wiggle room on by our design: (i) the decision over acquittal vs. conviction and (ii) the decision regarding the amount of punishment in case of conviction. Finally, our experimental evidence from rounds 4 and 5 (Result 3) revealed that participants are willing to pay to reduce both type-I and type-II error occurrence, further corroborating the broad finding that judges are sensitive to both false positives and false negatives. Taken together, these pieces of evidence fail to provide support to the asymmetric, "pro-defendant view" recalled in the introductory section, with regard to the large and influential scholarship in law and economics arguing for a strong prioritization of type-I error aversion.

Certainty and severity of punishments are unquestionably core concepts in the economic theories of deterrence (Chalfin and McCrary, 2017). However, a thriving law and economics literature suggests that drawing implications in terms of deterrence from the usage of punishment frequency and severity is not straightforward (Harel and Segal, 1999; Anderson and Stafford, 2003; Durlauf and Nagin, 2011; Friesen, 2012; Schildberg-Hörisch and Strassmair, 2012; Buechel et al., 2020), also due to the risk that people fail to correctly understand the notion of deterrence (DeAngelo and Charness, 2012). As to the deterrence policy implications of our findings, it is important to note that, insofar as judges react to the possibility of making errors by opting for more lenient sentences and potential offenders anticipate this, deterrence-based policies increasing the magnitude of punishment risk being far less effective than expected in preventing future crime. In this regard, we view our results as complementing a fastly-growing empirical line of inquiry in law and economics questioning the old view that increasing fine size and probability of apprehension lead to lower crime rates (Chalfin and McCrary, 2017) and shedding light on the subtle and often counterintuitive "specific deterrence" implications of prosecution and punishment.[19] Feess et al. (2018) highlight that judges' fairness considerations may induce them to be reluctant to convict a defendant, especially in the presence of high fines to be imposed. Our results help qualify this broad idea by making clear that judges' reluctance to convict when they are aware of the possibility of making a type-I error may lead them to punish but, at the same time, to opt for relatively milder sentences – provided that they have wiggle room on this (as it is the case in our experimental setup as well as in a variety of real-world settings, also beyond judicial decision making).

This study is subject to at least three limitations, that also naturally open up new avenues for future research. First, the well-known external validity issue, as we conducted a laboratory experiment with students, rather than a field study on actual judges being asked to decide over potential violators' behavior. In the future, it will be important to see whether similar findings arise when judges – in courts or other contexts where third-party decision making occurs (see on this footnote 2, in this paper) – face the two types of errors that we investigate through our experimental setting. Second, in our experiment we do not distinguish between different types of judicial decisions, whereas it is plausible that the nature of offence (e.g., serious vs. minor crimes), the domain in which it occurs (e.g., "green" vs. non-environmental; see on this Cochran et al., 2018) as well as the specific sources of judicial error (e.g., "errors of observation" vs. "errors of execution"; see on this Markussen et al., 2016) affect judges' decision making. Third, our experimental analysis does not specifically dig into the channels underlying our major findings. In this regard, we speculatively conjecture that both "cognitive limitations" (biases and heuristics) as well as ideological factors (e.g., judges' political preferences) may play an important role in shaping judges' sensitivity towards the two error types and the resulting punitive mix. Relatedly, also non-cognitive drivers (such as judges' personality traits) might play a non-negligible role. We leave these interesting questions as important avenues for future law and economics research on the theme.

[19] In principle, unexpected behavior of judges can be rationalized by referring to modern, dynamic versions of Becker's (1968) seminal study, assuming that individuals' time preferences are characterized by hyperbolic discounting. Agan et al.'s (2021) causal analysis indicates that not prosecuting marginal nonviolent misdemeanor defendants substantially reduces their subsequent criminal justice contact. They also note that available empirical evidence on the impact of length of incarcercation on future crime is inconclusive.

References

Abbink, K., Hennig-Schmidt, H., 2006, Neutral versus loaded instructions in a bribery experiment. *Experimental Economics*, 9, 2, 103-121.

Abrams, D., Galbiati, R., Henry, E., Philippe, A., 2022. When in Rome… on local norms and sentencing decisions. *Journal of the European Economic Association*, forthcoming.

Agan, A., Doleac, J.L., Harvey, A., 2021. Misdemeanor prosecution. IZA Discussion Paper N. 14234.

Ambrus, A., Greiner, B., 2012. Imperfect public monitoring with costly punishment: An experimental study. *American Economic Review*, 102, 7, 3317-3332.

Anderson, L., Stafford, S., 2003. Punishment in a regulatory setting: experimental evidence from the VCM. *Journal of Regulatory Economics*, 24, 1, 91-110.

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

Brandts, J., Charness, G., 2011. The strategy versus the direct-response method: a first survey of experimental comparisons, *Experimental Economics*, 14, 375-398.

Buechel, B., Feess, E., Muehlheusser, G., 2020. Optimal law enforcement with sophisticated and naïve offenders. *Journal of Economic Behavior and Organization*, 177, 835-857.

Cappelen, A.W., Cappelen, C., Tungodden, B., 2018. Second-best fairness under limited information: The trade-off between false positives and false negatives. NHH Discussion Paper.

Chalfin, A., McCrary, J., 2017. Criminal deterrence: a review of the literature. *Journal of Economic Literature*, 55, 1, 5-48.

Chen, D.L., Schonger, M., 2020. Social preferences or sacred values? Theory and evidence of deontological motivations. TSE Working Paper n. 16-714.

Cochran, J.C., Lynch, M.J., Toman, E.L., Shields, R.T., 2018. Court sentencing patterns for environmental crimes: Is there a "green" gap in punishment? *Journal of Quantitative Criminology*, 34, 37-66.

Cohen, A., Neeman, Z., Auferoth, F. 2022. Judging under public pressure. *Review of Economics and Statistics*, forthcoming.

Cohen, A., Yang, C.S., 2019. Judicial politics and sentencing decisions. *American Economic Journal: Economic Policy*, 11, 1, 160-191.

Cooter, R. 1998. Expressive law and economics. *Journal of Legal Studies*, 27, 585-608.

Crosetto, P., Filippin, A., 2013. The "bomb" risk elicitation task. *Journal of Risk and Uncertainty*, 47, 31-65.

DeAngelo, G., Charness, G., 2012. Deterrence, expected cost, uncertainty and voting: Experimental evidence. *Journal of Risk and Uncertainty*, 44, 73-100.

Durlauf, S.N., Nagin, D.S., 2011. Imprisonment and crime: can both be reduced? *Criminology and Public Policy*, 10, 1, 13-54.

Eren, O., Mocan, N., 2018. Emotional judges and unlucky juveniles. *American Economic Journal: Applied Economics,* 10, 3, 171-205.

Faillo, M., Grieco, D., Zarri, L., 2013. Legitimate Punishment, Feedback, and the Enforcement of Cooperation. *Games and Economic Behavior*, 77, 271-283.

Falk, A., Fischbacher, U. 2002. "Crime" in the lab-detecting social interaction. *European Economic Review*, 46, 4-5, 859-869.

Feess, E., Schildberg-Hörisch, H., Schramm, M., Wohlschlegel, A. 2018. The impact of fine size and uncertainty on punishment and deterrence: Theory and evidence from the laboratory. *Journal of Economic Behavior and Organization*, 149, 58-73.

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

Friesen, L., 2012. Certainty of punishment versus severity of punishment: An experimental investigation. *Southern Economic Journal*, 79, 399-421.

Galbiati, R., Vertova, P., 2008. Obligations and cooperative behaviour in public goods games. *Games and Economic Behavior*, 64, 1, 146-170.

Grechenig, K., Nicklisch, A., and C. Thöni, 2010. Punishment despite reasonable doubt – a public goods experiment with sanctions under uncertainty. *Journal of Empirical Legal* Studies, 7, 847-867.

Gross, S.R., O'Brien, B., Hu, C., Kennedy, E.H. 2014. Rate of false conviction of criminal defendants who are sentenced to death. *Proceedings of the National Academy of Sciences of the United States of America*, 111, 20, 7230-7235.

Guthrie, C., Rachlinski, J., Wistrich, A. 2001. Inside the judicial mind. *Cornell Law Review*, 86, 777.

Harel, A., Segal, U., 1999, Criminal law and behavioral law and economics: observations on the neglected role of uncertainty in deterring crime. *American Law and Economics Review*, 1, 276-312.

Hudja, S., Ralston, J., Wang, S., Aimone, J., Rentschler, L., North, C., 2021. The effect of gender on tolerance of type 1 and type 2 error in judicial decisions. SSRN working paper.

Kahneman, D., Sibony, O., Sunstein, C., 2021. Noise. A Flaw in Human Judgment, Harpey Collins Publishers.

Jordan, J., McAuliffe, K., Rand, D. 2016. The effects of endowment size and strategy method on third party punishment. *Experimental Economics*, 19, 4, 741-763.

Levitt, S.D., List, J.A., 2007. What do laboratory experiments measuring social preferences reveal about the real world? *Journal of Economic Perspectives*, 21, 2, 153-174.

Lim, C.S.H., Snyder, J.M., Stromberg, D., 2015. The judge, the politician, and the press: newspaper coverage and criminal sentencing across electoral systems. *American Economic Journal: Applied Economics*, 7, 4, 103-135.

Markussen, T. Putterman, L., Tyran, J.R., 2016. Judicial error and cooperation, *European Economic Review*, 89, 372-388.

Murphy, R.O., Ackermann, K.A., Handgraaf, M.J.J., 2011. Measuring social value orientation. *Judgment and Decision Making*, 6, 8, 771-781.

Ottone, S., Ponzano, F., Zarri, L., 2015, Power to the people? An experimental analysis of bottom-up accountability of third-party institutions, *Journal of Law, Economics, and Organization*, 31, 2, 347-382.

Png, I.P.L. 1986. Optimal subsidies and damages in the presence of judicial error, *International Review of Law and Economics*, 6, 1, 101-105.

Rizzolli, M. Saraceno, M., 2013. Better that ten guilty persons escape: Punishment costs explain the standard of evidence, *Public Choice*, 155, 3, 395-41.

Rizzolli, M., Stanca, L., 2012. Judicial errors and crime deterrence: Theory and experimental evidence, *Journal of Law and Economics*, 55, 2, 311-338.

Schanzenbach, M.M., Tiller, E.H., 2007. Strategic judging under the U.S. sentencing guidelines: Positive political theory and evidence, *Journal of Law, Economics, and Organization*, 23, 1, 24-56.

Polinsky, A., Shavell, S., 2007. The theory of public enforcement of law, in Polinsky M., Shavell S., eds., *Handbook of Law and Economics*, Elsevier.

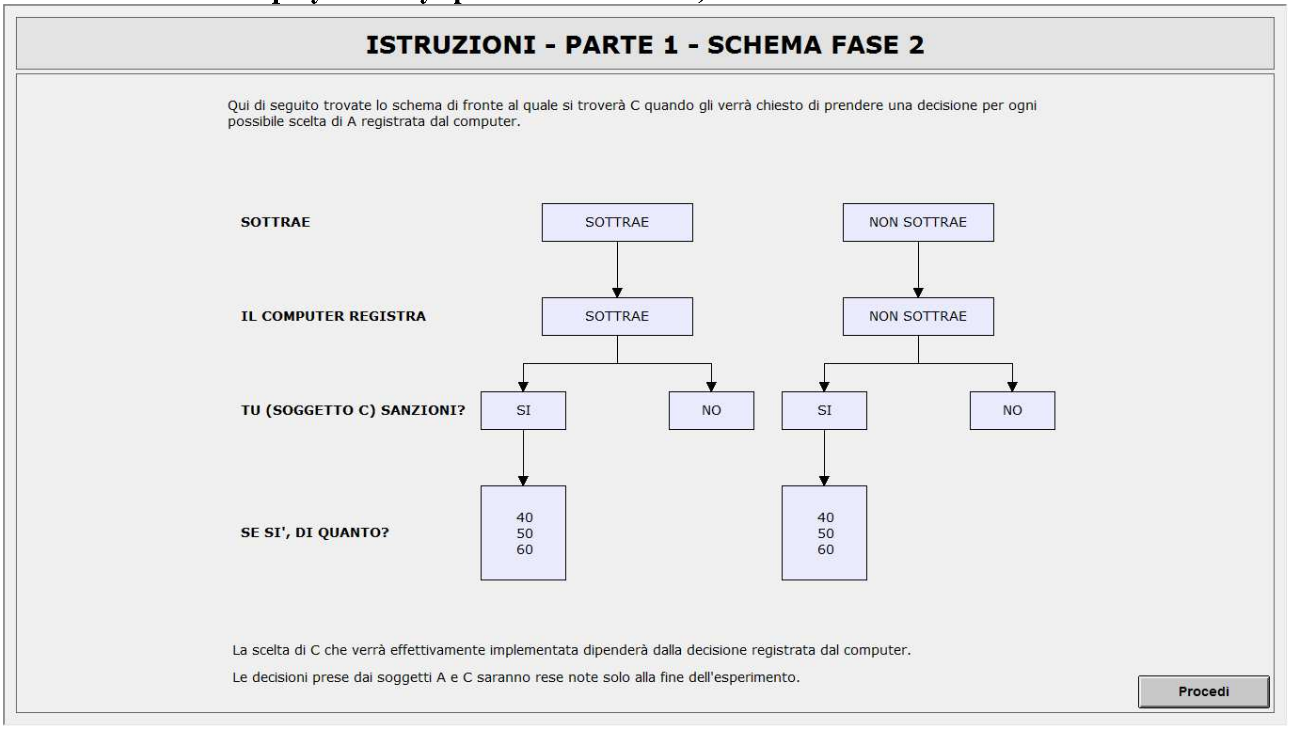
Schildberg-Hörisch, H., Strassmair, C., 2012. An experimental test of the deterrence hypothesis, *Journal of Law, Economics, and Organization*, 28, 447-459.

Scurich, N., 2015, Criminal justice policy preferences: Blackstone ratios and the veil of ignorance, *Stanford Law and Policy Review Online,* 26, 23-35.

Sonnemans, J., van Dijk, F. 2012. Errors in judicial decisions: Experimental results, *Journal of Law, Economics, and Organization*, 28, 4.

TABLES AND FIGURES

Figure 1. Decision Screen of participant C when information is affected by no error (round 1 and rounds 4 and 5 when player C buys perfect information)



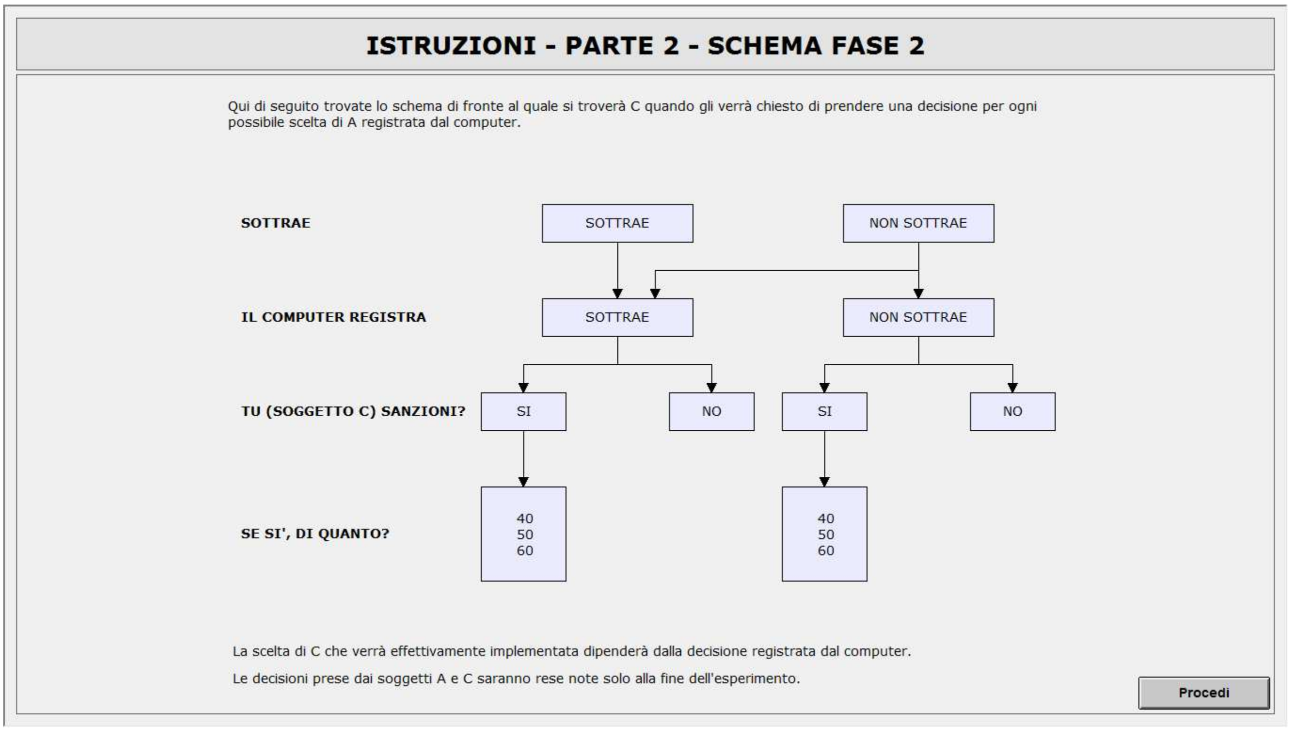
ISTRUZIONI - PARTE 1 - SCHEMA FASE 2

Qui di seguito trovate lo schema di fronte al quale si troverà C quando gli verrà chiesto di prendere una decisione per ogni possibile scelta di A registrata dal computer.

La scelta di C che verrà effettivamente implementata dipenderà dalla decisione registrata dal computer.

Le decisioni prese dai soggetti A e C saranno rese note solo alla fine dell'esperimento.

Figure 2. Decision Screen of participant C when information is affected by type-I error (round 2 and round 4 when player C does not buy perfect information)



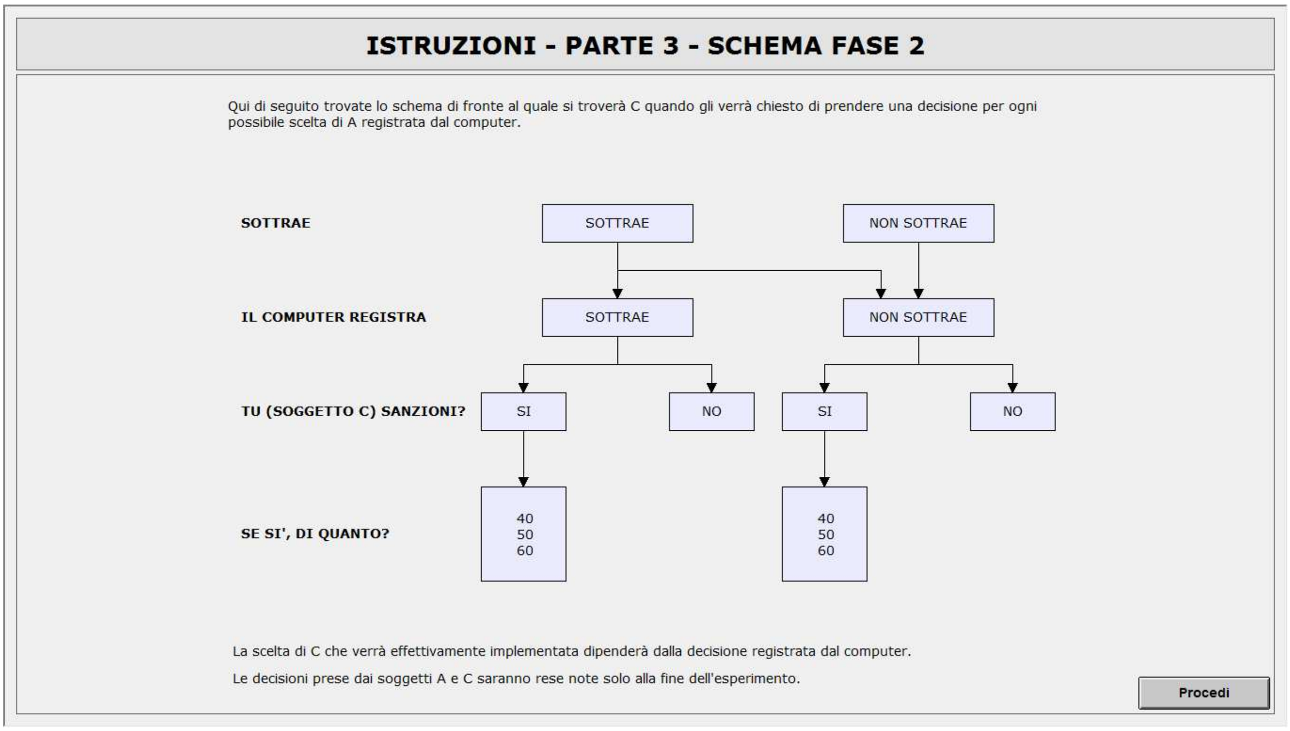
ISTRUZIONI - PARTE 2 - SCHEMA FASE 2

Qui di seguito trovate lo schema di fronte al quale si troverà C quando gli verrà chiesto di prendere una decisione per ogni possibile scelta di A registrata dal computer.

La scelta di C che verrà effettivamente implementata dipenderà dalla decisione registrata dal computer.

Le decisioni prese dai soggetti A e C saranno rese note solo alla fine dell'esperimento.

Figure 3. Decision Screen of participant C when information is affected by type-II error (round 3 and round 5 when player C does not buy perfect information)



ISTRUZIONI - PARTE 3 - SCHEMA FASE 2

Qui di seguito trovate lo schema di fronte al quale si troverà C quando gli verrà chiesto di prendere una decisione per ogni possibile scelta di A registrata dal computer.

La scelta di C che verrà effettivamente implementata dipenderà dalla decisione registrata dal computer.

Le decisioni prese dai soggetti A e C saranno rese note solo alla fine dell'esperimento.

Figure 4. Distribution of players Cs' belief on players As' number of thieves over the five experimental conditions

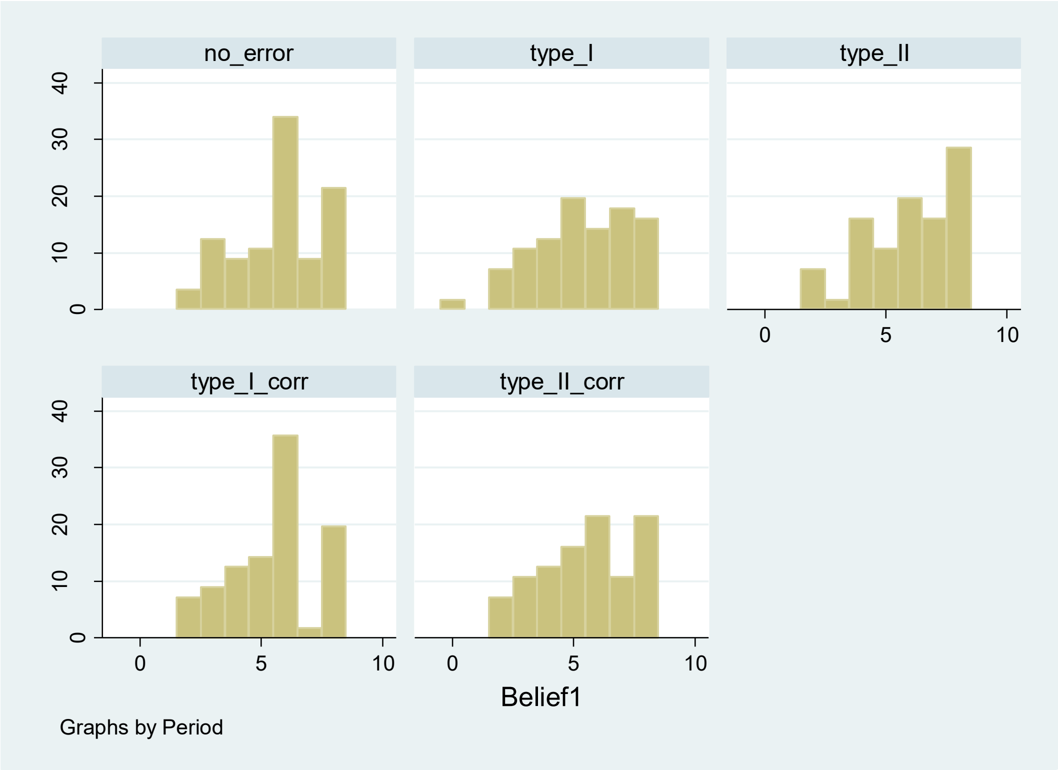


Table 1. Distribution of Player C's attitude towards judicial errors

	Type-I	Type-II	Equal
Generally	39%	12.50%	48%
Robbery	54%	14%	32%
Murder	27%	7%	66%

Table 2. Distribution of Player C's beliefs on the number of thieves

	Median	Mean
No error	6	5.7
Type-I	5	5.3
Type-II	6	5.9
Type-I_corr	6	5.5
Type-II_corr	6	5.5

Table 3. Player C's thresholds by their attitude towards judicial errors

When Blackstone Ratio is	Max # of thefts to avoid punishment (when the computer reports A stole)	When Executioner Ratio is	Min # of thefts to have punishment (when the computer reports A did not steal)
>7	7	>7	1
7	7	7	1
6	6	6	2
5	6	5	2
4	6	4	2
3	6	3	2
2	5	2	3
1	3	1	5
When C is averse to Type-II error	0	When C is averse to Type-II error	8

Table 4. Hypothetical and actual percentage of punishment in the first three rounds

	the computer reports that A stole money				the computer reports that A did not steal money	
	No error (N=56)	Type-I (N=56)	Type-II (N=56)	No error (N=56)	Type-I (N=56)	Type-II (N=56)
Actual	93%	84%	95%	22%	25%	61%
Hypothetical_generally	100%	71%	100%	0%	0%	59%
Hypothetical_robbery	100%	63%	100%	0%	0%	54%
Hypothetical_murder	100%	77%	100%	0%	0%	63%

Table 5. Level of punishment in the first three rounds

	C punishes A when the computer reports that A stole money		
	No error (N=56)	Type-I (N=56)	Type-II (N=56)
0	7%	16%	5%
40	9%	30%	13%
50	32%	27%	23%
60	52%	27%	59%
	100%	100%	100%

	C punishes A when the computer reports that A did not steal money		
	No error	Type-I	Type-II
0	78%	75%	39%
40	11%	14%	39%
50	7%	2%	9%
60	4%	9%	13%
	100%	100%	100%

Table 6. Multinomial Probit Regression – Marginal effects
Dependent variable: Players Cs' level of punishment in the first three rounds when the computer reports that player A stole money
Baseline condition: "no error" condition (round 1)
Cluster at the subject level

	No fee	*Fee = 40*	*Fee = 50*	*Fee = 60*
Type-I_error	0.019	.287***	-0.031	-0.275**
Type-II_error	-0.004	0.070	-0.127*	0.061
# hypothetical thieves	-0.000	-0.043***	0.040**	0.003
Socio-demographic controls	YES	YES	YES	YES
Personal characteristics controls	YES	YES	YES	YES
Obs	168	168	168	168
Subjects	56	56	56	56

*** 1% significance ** 5% significance * 10% significance

Table 7. Multinomial Probit Regression – Marginal effects
Dependent variable: Players Cs' level of punishment in the first three rounds when the computer reports that player A did not steal money
Baseline condition: "no error" condition (round 1)
Cluster at the subject level

	No fee	*Fee = 40*	*Fee = 50*	*Fee = 60*
Type-I_error	-0.080	0.062	-0.053*	0.071
Type-II_error	-0.428***	0.308***	0.009	0.111*
# hypothetical thieves	-0.040**	0.025	0.008	0.007
Socio-demographic controls	YES	YES	YES	YES
Personal characteristics controls	YES	YES	YES	YES
Obs	168	168	168	168
Subjects	56	56	56	56

*** 1% significance ** 5% significance * 10% significance

Table 8. Multinomial Probit Regression – Marginal effects
Dependent variable: Players Cs' level of punishment in the first three rounds when the computer reports that player A stole money
Baseline condition: "type-I error" condition (round 2)
Cluster at the subject level

	No fee	*Fee = 40*	*Fee = 50*	*Fee = 60*
No_error	-0.012	-.216***	-0.010	0.238**
Type-II_error	-0.015	-0.165***	-0.129	0.310***
# hypothetical thieves	-0.000	-0.043***	0.040**	0.003
Socio-demographic controls	YES	YES	YES	YES
Personal characteristics controls	YES	YES	YES	YES
Obs	168	168	168	168
Subjects	56	56	56	56

*** 1% significance ** 5% significance * 10% significance

Table 9. Multinomial Probit Regression – Marginal effects
Dependent variable: Players Cs' level of punishment in the first three rounds when the computer reports that player A did not steal money
Baseline condition: "type-II error" condition (round 3)
Cluster at the subject level

	No fee	*Fee = 40*	*Fee = 50*	*Fee = 60*
No_error	0.352***	-0.250***	-0.019	-0.083***
Type-I_error	-0.303***	-0.202***	-0.065**	-0.036
# hypothetical thieves	-0.040**	0.025	0.008	0.007
Socio-demographic controls	YES	YES	YES	YES
Personal characteristics controls	YES	YES	YES	YES
Obs	168	168	168	168
Subjects	56	56	56	56

*** 1% significance ** 5% significance * 10% significance

Table 10. Random-effects Probit Regression – Marginal effects
Dependent variable: Players Cs' belief on the number of player As are thieves
Baseline condition: "no error" condition (round 1)

	No fee
Type-I_error	0.135
Type-II_error	0.542*
Type-I_error_correct	-0.396
Type-II_error_correct	0.732
Socio-demographic controls	YES
Personal characteristics controls	YES
Obs	280
Subjects	56

*** 1% significance ** 5% significance * 10% significance

Table 11. Hypothetical and actual percentage of punishment in the last two rounds by player C's decision of buying perfect information

	the computer reports that A stole money		**the computer reports that A did not steal money**	
	Type-I_info (N=27)	**Type-II_info (N=22)**	**Type-I_info (N=27)**	**Type-II_info (N=22)**
Actual	96%	91%	8%	8%
Hypothetical_generally	100%	100%	0%	0%
Hypothetical_robbery	100%	100%	0%	0%
Hypothetical_murder	100%	100%	0%	0%

	the computer reports that A stole money		**the computer reports that A did not steal money**	
	Type-I_noinfo (N=29)	**Type-II_noinfo (N=34)**	**Type-I_noinfo (N=29)**	**Type-II_noinfo (N=34)**
Actual	83%	91%	38%	71%
Hypothetical_generally	79%	100%	0%	62%
Hypothetical_robbery	79%	100%	0%	62%
Hypothetical_murder	83%	100%	0%	68%

Table 12. Level of punishment in the last two rounds by player C's decision of buying perfect information

	C punishes A when the computer reports that A stole money			
	Type-I_info (N = 27)	**Type-II_info (N = 22)**	**Type-I_noinfo (N = 29)**	**Type-II_noinfo (N = 34)**
0	4%	9%	17%	9%
40	4%	5%	34%	9%
50	15%	9%	28%	23%
60	77%	77%	21%	59%
	100%	100%	100%	100%

	C punishes A when the computer reports that A did not steal money			
	Type-I_info	**Type-II_info**	**Type-I_noinfo**	**Type-II_noinfo**
0	92%	92%	62%	29%
40	4%	0%	18%	38%
50	0%	4%	10%	27%
60	4%	4%	10%	6%
	100%	100%	100%	100%

Table 13. Multinomial Probit Regression – Marginal effects
Dependent variable: Players Cs' level of punishment in the first, second and fourth round when the computer reports that player A stole money – subsample of subjects who buy the correct information
Baseline condition: "no error" condition (round 1)
Cluster at the subject level

	No fee	***Fee = 40***	***Fee = 50***	***Fee = 60***
Type-I_error	0.029	0.256***	-0.019	-0.266**
Type-I_correct	-0.009	-0.138***	-0.235***	0.382***
# hypothetical thieves	-0.002	-0.052***	0.031	0.023
Socio-demographic controls	YES	YES	YES	YES
Personal characteristics controls	YES	YES	YES	YES
Obs	139	139	139	139
Subjects	56	56	56	56

*** 1% significance ** 5% significance * 10% significance

Table 14. Multinomial Logit Regression – Marginal effects

Dependent variable: Players Cs' level of punishment in the first, third and fifth round when the computer reports that player A did not steal money – subsample of subjects who buy the correct information
Baseline condition: "no error" condition (round 1)

Cluster at the subject level

	No fee	*Fee = 40*	*Fee = 50*	*Fee = 60*
Type-II_error	-0.239**	0.045**	0.069	0.125*
Type-II_correct	0.094	-0.187***	0.038	0.055
# hypothetical thieves	-0.017	-0.006*	0.014	0.003
Socio-demographic controls	YES	YES	YES	YES
Personal characteristics controls	YES	YES	YES	YES
Obs	134	134	134	134
Subjects	56	56	56	56

*** 1% significance   ** 5% significance   * 10% significance

Table 15. Multinomial Probit Regression – Marginal effects

Dependent variable: Players Cs' level of punishment in the first, second and fourth round when the computer reports that player A stole money – subsample of subjects who buy the correct information
Baseline condition: "type-I error" condition (round 2)

Cluster at the subject level

	No fee	*Fee = 40*	*Fee = 50*	*Fee = 60*
No_error	-0.023	-0.211***	-0.007	0.241**
Type-I_correct	-0.020	-0.227***	-0.259***	0.506***
# hypothetical thieves	-0.002	-0.052***	0.031	0.023
Socio-demographic controls	YES	YES	YES	YES
Personal characteristics controls	YES	YES	YES	YES
Obs	139	139	139	139
Subjects	56	56	56	56

*** 1% significance   ** 5% significance   * 10% significance

Table 16. Multinomial Logit Regression – Marginal effects
Dependent variable: Players Cs' level of punishment in the first, third and fifth round when the computer reports that player A did not steal money – subsample of subjects who buy the correct information
Baseline condition: "type-II error" condition (round 3)
Cluster at the subject level

	No fee	*Fee = 40*	*Fee = 50*	*Fee = 60*
No_error	0.192**	0.035***	-0.064	-0.093*
Type-II_correct	0.293***	-0.241***	-0.022	-0.029
# hypothetical thieves	-0.017	-0.006*	0.014	0.003
Socio-demographic controls	YES	YES	YES	YES
Personal characteristics controls	YES	YES	YES	YES
Obs	134	134	134	134
Subjects	56	56	56	56

*** 1% significance ** 5% significance * 10% significance

Table 17. Multinomial Probit Regression – Marginal effects
Dependent variable: Players Cs' level of punishment in the first, second and fourth round when the computer reports that player A stole money – subsample of subjects who do not buy the correct information
Baseline condition: "no error" condition (round 1)
Cluster at the subject level

	No fee	*Fee = 40*	*Fee = 50*	*Fee = 60*
Type-I_error	0.071	0.289***	-0.080	-0.280***
Type-I_correct	0.062	0.402***	-0.083	0.381***
# hypothetical thieves	0.000	-0.025	0.007	0.018
Socio-demographic controls	YES	YES	YES	YES
Personal characteristics controls	YES	YES	YES	YES
Obs	141	141	141	141
Subjects	56	56	56	56

*** 1% significance ** 5% significance * 10% significance

Table 18. Multinomial Probit Regression – Marginal effects
Dependent variable: Players Cs' level of punishment in the first, third and fifth round when the computer reports that player A did not steal money – subsample of subjects who do not buy the correct information
Baseline condition: "no error" condition (round 1)
Cluster at the subject level

	No fee	*Fee = 40*	*Fee = 50*	*Fee = 60*
Type-II_error	-0.444***	0.350***	0.020	0.074
Type-II_correct	-0.493***	0.327**	0.163	0.003
# hypothetical thieves	-0.076***	-0.047**	0.024	0.005
Socio-demographic controls	YES	YES	YES	YES
Personal characteristics controls	YES	YES	YES	YES
Obs	146	146	146	146
Subjects	56	56	56	56

*** 1% significance ** 5% significance * 10% significance

Table 19. Multinomial Probit Regression – Marginal effects
Dependent variable: Players Cs' level of punishment in the first, second and fourth round when the computer reports that player A stole money – subsample of subjects who do not buy the correct information
Baseline condition: "type-I error" condition (round 2)
Cluster at the subject level

	No fee	*Fee = 40*	*Fee = 50*	*Fee = 60*
No_error	-0.061*	-0.261***	0.049	0.273***
Type-I_correct	-0.028	0.096	0.056	-0.149
# hypothetical thieves	-0.000	-0.025	0.007	0.018
Socio-demographic controls	YES	YES	YES	YES
Personal characteristics controls	YES	YES	YES	YES
Obs	141	141	141	141
Subjects	56	56	56	56

*** 1% significance ** 5% significance * 10% significance

Table 20. Multinomial Probit Regression – Marginal effects
Dependent variable: Players Cs' level of punishment in the first, third and fifth round when the computer reports that player A did not steal money – subsample of subjects who do not buy the correct information
Baseline condition: "type-II error" condition (round 3)
Cluster at the subject level

	No fee	*Fee = 40*	*Fee = 50*	*Fee = 60*
No_error	0.417***	-0.323***	-0.038	-0.056*
Type-II_correct	-0.091	-0.024	0.154	-0.039
# hypothetical thieves	-0.076***	-0.047**	0.024	0.005
Socio-demographic controls	YES	YES	YES	YES
Personal characteristics controls	YES	YES	YES	YES
Obs	146	146	146	146
Subjects	56	56	56	56

*** 1% significance ** 5% significance * 10% significance

SUPPLEMENTARY MATERIAL

Proofs

By applying Bayes' rule, we obtain the following probabilities of being innocent/guilty conditional upon innocence/guilt reporting:

$$Prob.(innocent|innocence\ reporting) = \frac{(1-\varepsilon_1)(1-q)}{q\varepsilon_2+(1-q)(1-\varepsilon_1)}$$

$$Prob.(guilty|guilt\ reporting) = \frac{(1-\varepsilon_2)q}{q(1-\varepsilon_2)+(1-q)\varepsilon_1}$$

$$Prob.(innocent|guilt\ reporting) = \frac{\varepsilon_1(1-q)}{q(1-\varepsilon_2)+(1-q)\varepsilon_1}$$

$$Prob.(guilty|innocence\ reporting) = \frac{\varepsilon_2 q}{q\varepsilon_2+(1-q)(1-\varepsilon_1)}$$

Proof of Proposition 1

(i)

$$\frac{dr_G}{d\varepsilon_1} = -\frac{(1-\varepsilon_2)q}{\varepsilon_1^2(1-q)}\frac{H}{F} < 0 \qquad \frac{dr_G}{d\varepsilon_2} = -\frac{q}{\varepsilon_1(1-q)}\frac{H}{F} < 0 \qquad \frac{dr_G}{dF} = -\frac{(1-\varepsilon_2)q}{\varepsilon_1(1-q)}\frac{H}{F^2} < 0$$

$$\frac{dr_G}{dH} = \frac{(1-\varepsilon_2)q}{\varepsilon_1(1-q)F} > 0 \qquad \frac{dr_G}{dq} = \frac{(1-\varepsilon_2)}{\varepsilon_1(1-q)^2}\frac{H}{F} > 0$$

(ii)

$$\frac{\partial r_I}{\partial\varepsilon_1} = \frac{\varepsilon_2 q}{(1-\varepsilon_1)^2(1-q)}\frac{H}{F} > 0 \qquad \frac{dr_I}{d\varepsilon_2} = \frac{q}{(1-\varepsilon_1)(1-q)}\frac{H}{F} > 0 \qquad \frac{dr_I}{dF} = -\frac{\varepsilon_2 q}{(1-\varepsilon_1)(1-q)}\frac{H}{F^2} < 0$$

$$\frac{dr_I}{dH} = \frac{\varepsilon_2 q}{(1-\varepsilon_1)(1-q)F} > 0 \qquad \frac{dr_I}{dq} = \frac{\varepsilon_2}{(1-\varepsilon_1)(1-q)^2}\frac{H}{F}$$

(iii)

$$\frac{d\tilde{s}}{d\varepsilon_1} = \frac{\partial\tilde{s}}{\partial\varepsilon_2} = \frac{F}{H}(p_G - p_I) \geq 0 \qquad \frac{d\tilde{s}}{dF} = \frac{(p_G - p_I)(1-\varepsilon_1-\varepsilon_2)}{H} \leq 0 \qquad \frac{d\tilde{s}}{dH} = \frac{F(p_G - p_I)(1-\varepsilon_1-\varepsilon_2)}{H^2} \geq 0$$

$$\frac{d\tilde{s}}{d(p_G - p_I)} = -\frac{F}{H}(1-\varepsilon_1-\varepsilon_2) \leq 0$$

Signs above are easy to be verified recalling that we assumed i. that people consider more (or at least equally) likely to be punished given a guilt reporting than given an innocence reporting and, ii. that a minimum accuracy in investigation reporting is required so that $(1-\varepsilon_1-\varepsilon_2) > 0$.

Proof of Proposition 2

In Bayesian Nash Equilibrium, expectations and actual probabilities coincide, so that we have:

$$\begin{cases} p_G = W(r_G) \\ p_I = W(r_I) \\ q = B(\tilde{s}) \end{cases} \rightarrow \begin{cases} p_G - W\left(\frac{(1-\varepsilon_2)q}{\varepsilon_1(1-q)}\frac{H}{F}\right) = 0 \\ p_I - W\left(\frac{\varepsilon_2 q}{(1-\varepsilon_1)(1-q)}\frac{H}{F}\right) = 0 \\ q - B\left(1 - \frac{F}{H}(p_G - p_I)(1-\varepsilon_1-\varepsilon_2)\right) \quad = 0 \end{cases}$$

The system of total differentials is:

$$\begin{bmatrix} 1 & 0 & -w_G\frac{1-\varepsilon_2}{\varepsilon_1(1-q)^2}\frac{H}{F} \\ 0 & 1 & -w_I\frac{\varepsilon_2}{(1-\varepsilon_1)(1-q)^2}\frac{H}{F} \\ b\frac{F}{H}(1-\varepsilon_1-\varepsilon_2) & \frac{F}{H}(1-\varepsilon_1-\varepsilon_2) & 1 \end{bmatrix}\begin{bmatrix} dp_G \\ dp_I \\ dq \end{bmatrix}$$

$$= \begin{bmatrix} -w_G\frac{(1-\varepsilon_2)q}{\varepsilon_1^2(1-q)}\frac{H}{F} & -w_G\frac{q}{\varepsilon_1(1-q)}\frac{H}{F} & -w_G\frac{(1-\varepsilon_2)q}{\varepsilon_1(1-q)}\frac{H}{F^2} \\ w_I\frac{\varepsilon_2 q}{(1-\varepsilon_1)^2(1-q)}\frac{H}{F} & w_I\frac{q}{(1-\varepsilon_1)(1-q)}\frac{H}{F} & -w_I\frac{\varepsilon_2 q}{(1-\varepsilon_1)(1-q)}\frac{H}{F^2} \\ b\frac{F}{H}(p_G-p_I) & b\frac{F}{H}(p_G-p_I) & -\frac{b}{H}(p_G-p_I)(1-\varepsilon_1-\varepsilon_2) \end{bmatrix}\begin{bmatrix} d\varepsilon_1 \\ d\varepsilon_2 \\ dF \end{bmatrix}$$

We compute the determinant of the first matrix: $\det = 1 + b\frac{(1-\varepsilon_1-\varepsilon_2)}{(1-q)^2}\left(w_G\frac{(1-\varepsilon_2)}{\varepsilon_1} + w_I\frac{\varepsilon_2}{(1-\varepsilon_1)}\right) > 0$

$$\begin{bmatrix} dp_G \\ dp_I \\ dq \end{bmatrix} = \frac{1}{\det}\begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}\begin{bmatrix} d\varepsilon_1 \\ d\varepsilon_2 \\ dF \end{bmatrix}$$

$$\frac{dp_G^*}{d\varepsilon_1} = \frac{1}{\det}a_{11} = \frac{1}{\det}w_G\frac{(1-\varepsilon_2)}{\varepsilon_1(1-q)}\left[-\frac{q}{\varepsilon_1}\frac{H}{F} - bw_I\frac{H}{F}\frac{(1-\varepsilon_1-\varepsilon_2)\varepsilon_2 q}{(1-\varepsilon_1)^2(1-q)^2\varepsilon_1} + b\frac{(p_G-p_I)}{(1-q)}\right] \gtreqless 0$$

$$\frac{dp_G^*}{d\varepsilon_2} = \frac{1}{\det}a_{12} = \frac{1}{\det}w_G\frac{1}{\varepsilon_1(1-q)}\left[-q\frac{H}{F} - bw_I\frac{H}{F}\frac{(1-\varepsilon_1-\varepsilon_2)q}{(1-\varepsilon_1)(1-q)^2} + b\frac{(1-\varepsilon_2)(p_G-p_I)}{(1-q)}\right] \gtreqless 0$$

$$\frac{dp_G^*}{dF} = \frac{1}{\det}a_{13} = -\frac{1}{\det}w_G\frac{(1-\varepsilon_2)}{\varepsilon_1(1-q)F}\left[q\frac{H}{F} + b\frac{(1-\varepsilon_1-\varepsilon_2)(p_G-p_I)}{(1-q)F}\right] < 0$$

$$\frac{dp_I^*}{d\varepsilon_1} = \frac{1}{\det}a_{21} = \frac{1}{\det}w_I\frac{\varepsilon_2}{(1-\varepsilon_1)(1-q)}\left[+\frac{q}{(1-\varepsilon_1)}\frac{H}{F} + bw_G\frac{H}{F}\frac{(1-\varepsilon_1-\varepsilon_2)(1-\varepsilon_2)q}{\varepsilon_1{}^2(1-q)^2}(1+\varepsilon_1) + b\frac{(p_G-p_I)}{(1-q)}\right] \geq 0$$

$$\frac{dp_I^*}{d\varepsilon_2} = \frac{1}{\det} a_{22} = \frac{1}{\det} w_I \frac{1}{(1-\varepsilon_1)(1-q)} \left[q\frac{H}{F} + bw_G \frac{H}{F} \frac{(1-\varepsilon_1-\varepsilon_2)q}{\varepsilon_1(1-q)^2} + b\frac{\varepsilon_2(p_G-p_I)}{(1-q)} \right] \geq 0$$

$$\frac{dp_I^*}{dF} = \frac{1}{\det} a_{23} = \frac{1}{\det} w_I \frac{\varepsilon_2}{(1-\varepsilon_1)(1-q)F} \left[-q\frac{H}{F} + b\frac{(1-\varepsilon_1-\varepsilon_2)(p_G-p_I)}{(1-q)F} \right] \gtreqless 0$$

$$\frac{dq^*}{d\varepsilon_1} = \frac{1}{\det} a_{31} = \frac{1}{\det} b \left[\frac{(1-\varepsilon_1-\varepsilon_2)q}{(1-q)} \left(\frac{w_G(1-\varepsilon_2)}{{\varepsilon_1}^2} - \frac{w_I\varepsilon_2}{(1-\varepsilon_1)^2} \right) + \frac{F}{H}(p_G-p_I) \right] \gtreqless 0$$

$$\frac{dq^*}{d\varepsilon_2} = \frac{1}{\det} a_{32} = \frac{1}{\det} b(1-\varepsilon_1-\varepsilon_2) \left[\frac{q}{(1-q)} \left(\frac{w_G}{\varepsilon_1} - \frac{w_I}{(1-\varepsilon_1)} \right) - \frac{(p_G-p_I)}{H} \right] \gtreqless 0$$

$$\frac{dq^*}{dF} = \frac{1}{\det} a_{33} = \frac{1}{\det} b(1-\varepsilon_1-\varepsilon_2) \left[\frac{q}{(1-q)F} \left(\frac{w_G(1-\varepsilon_2)}{\varepsilon_1} - \frac{w_I\varepsilon_2}{(1-\varepsilon_1)} \right) - \frac{(p_G-p_I)}{H} \right] \gtreqless 0$$

Experimental Instructions

Good morning. Thank you for accepting to participate in the experiment. There won't be specific difficulties, nor trick questions. You will have to carefully follow the instructions that will appear sequentially on your screen. The answers that you will provide will be completely anonymous: for those who will elaborate the data it won't be possible to trace it back to you.
During the experiment, you will earn experimental tokens. At the end of the experiment, the tokens you have earned will be converted in euros (considering that each token is worth **0.02 euros**) and you will receive your final earnings. Since the experiment consists of multiple rounds, your final earnings will be obtained by summing up the tokens earned in each round.

This experiment involves three types of participants (participant A, participant B, participant C).
At the beginning of the experiment, you will be randomly assigned to one of three roles: A, B or C. Your role will remain the same throughout the whole experiment.
The experiment consists of five rounds (rounds 1-5).
A common feature of all rounds in the experiment is that each of them consists of two stages (stages 1-2).
In the first stage (stage 1), each participant A has to decide whether to subtract or not some money from a participant B he/she will be matched with.
In the second stage (stage 2), each participant C has to decide whether to sanction or not the participant A he/she will be matched with. In particular, C will have to decide whether (and, if so, by how many tokens) to sanction A based on two possible choices made by A ("subtract" or "not subtract") recorded by the computer.
Each round (1-5) of the experiment has specific features that will be illustrated to you step by step.
In each round, you will be matched to other two participants so that each group will consist of one participant A, one participant B and one participant C. The participants you will be matched with will change from round to round. Neither participant is informed about the identity of the other participants he/she will be matched with, not even once the experiment will be over.

Now the **FIRST ROUND** of the experiment begins.
In the **FIRST STAGE**, in each group, participant A has an endowment of 100 tokens and has to decide whether to subtract or not 50 tokens from participant B. Participant B has an endowment of 100 tokens and does not have to take any decision.
In this stage the computer records the decision made by A.

In the **SECOND STAGE**, in each group, participant C has an endowment of 200 tokens.
In this stage, before the computer communicates the choice made by A, participant C has to decide whether to sanction A or not in the following two possible cases: the computer communicates that A has subtracted tokens from B / the computer communicates that A has not subtracted tokens from B.
If C decides to sanction A, he/she will also have to specify by how many tokens (opting for 40, 50 or 60 tokens).

Figure 1 illustrates the screen seen by participant C when he/she has to make a decision for each possible choice made by A recorded by the computer.

The choice made by C that will be implemented will depend on the decision recorded by the computer.
Information about the decisions taken by participants A and C will be provided only once the experiment will be over.

Let's make some examples to illustrate the experiment more clearly.
Three screens with examples follow.

Now the **SECOND ROUND** begins. In this round, each of you keeps his/her role but will be matched to other participants.
In the **FIRST STAGE**, in each group, participant A has an endowment of 100 tokens and has to decide whether to subtract or not 50 tokens from participant B. Participant B has an endowment of 100 tokens and cannot take any decision.
The computer records the decision made by A. However, in this stage, the computer may make errors in recording the decision made by A. In particular, the computer may communicate that 50 tokens have been subtracted even if participant A decided not to subtract tokens from B.
Therefore:

- If participant A subtracts 50 tokens, the computer correctly records his/her choice to subtract 50 tokens.

- If participant A does not subtract 50 tokens, there is a 50% probability that the computer fails to correctly record his/her choice and, therefore, communicates that tokens have been subtracted while this has not been the case. In this case, the computer communicates that there is a 50% probability that A has not subtracted 50 tokens and a 50% probability that A has subtracted 50 tokens.

This implies that:

- If the computer communicates that the 50 tokens have not been subtracted from B, then participant A certainly has not subtracted 50 tokens from B.

- If instead the computer communicates that the 50 tokens have been subtracted from B, then it is possible that A it is possible that A has not subtracted 50 tokens from B.

In the **SECOND STAGE**, in each group, participant C has an endowment of 200 tokens.
In this stage, before the computer communicates the choice made by A, participant C has to decide whether to sanction A or not in the following two possible cases: the computer communicates that A has subtracted tokens from B / the computer communicates that A has not subtracted tokens from B.
If C decides to sanction A, he/she will also have to specify by how many tokens (opting for 40, 50 or 60 tokens).
Please, consider that, if the computer can make errors in recording the decision made by A, various scenarios may emerge.
In particular:

- If the computer communicates that A has not subtracted 50 tokens from B and C decides to sanction A, participant C will sanction a participant A that CERTAINLY has not subtracted tokens from B.

- If the computer communicates that A has subtracted 50 tokens from B and C decides to sanction A, participant C will sanction a participant A that MIGHT have subtracted 50 tokens.

Figure 2 illustrates the screen seen by participant C when he/she has to make a decision for each possible choice made by A recorded by the computer.

The choice made by C that will be implemented will depend on the decision recorded by the computer.

Information about the decisions taken by participants A and C will be provided only once the experiment will be over.

Let's make some examples to illustrate the experiment more clearly.

Three screens with examples follow.

Now the **THIRD ROUND** begins. In this round, each of you keeps his/her role but will be matched to other participants.

In the **FIRST STAGE**, in each group, participant A has an endowment of 100 tokens and has to decide whether to subtract or not 50 tokens from participant B. Participant B has an endowment of 100 tokens and cannot take any decision.

The computer records the decision made by A. However, in this stage, the computer may make errors in recording the decision made by A. In particular, the computer may fail to communicate that 50 tokens have been subtracted even if participant A decided to subtract tokens from B.

Therefore:

- If participant A does not subtract 50 tokens, the computer correctly records his/her choice not to subtract 50 tokens.

- If participant A subtracts 50 tokens, there is a 50% probability that the computer fails to correctly record his/her choice and, therefore, that it fails to communicate that 50 tokens have been subtracted. In this case, the computer communicates that there is a 50% probability that A has not subtracted 50 tokens and a 50% probability that A has subtracted 50 tokens.

This implies that:

- If the computer communicates that the 50 tokens have been subtracted from B, then participant A certainly has subtracted 50 tokens from B.

- If instead the computer communicates that the 50 tokens have not been subtracted from B, then it is possible that A has subtracted 50 tokens from B.

In the **SECOND STAGE**, in each group, participant C has an endowment of 200 tokens.

In this stage, before the computer communicates the choice made by A, participant C has to decide whether to sanction A or not in the following two possible cases: the computer communicates that A has subtracted tokens from B / the computer communicates that A has not subtracted tokens from B.

If C decides to sanction A, he/she will also have to specify by how many tokens (opting for 40, 50 or 60 tokens).

Please, consider that, if the computer can make errors in recording the decision made by A, various scenarios may emerge.

In particular:

- If the computer communicates that A has subtracted 50 tokens from B and C decides to sanction A, participant C will sanction a participant A that CERTAINLY has subtracted tokens from B.

- If the computer communicates that A has not subtracted 50 tokens from B and C decides to sanction A, participant C will sanction a participant A that MIGHT have not subtracted 50 tokens.

Figure 3 illustrates the screen seen by participant C when he/she has to make a decision for each possible choice made by A recorded by the computer.

The choice made by C that will be implemented will depend on the decision recorded by the computer.

Information about the decisions taken by participants A and C will be provided only once the experiment will be over.

Let's make some examples to illustrate the experiment more clearly.

Three screens with examples follow.

Now the **FOURTH ROUND** begins. In this round, each of you keeps his/her role but will be matched to other participants.

In the **FIRST STAGE**, in each group, participant A has an endowment of 100 tokens and has to decide whether to subtract or not 50 tokens from participant B. Participant B has an endowment of 100 tokens and cannot take any decision.

The computer records the decision made by A. However, in this round, like in the second round, the computer may make errors in recording the decision made by A. In particular, the computer may communicate that 50 tokens have been subtracted even if participant A decided not to subtract tokens from B.

In the **SECOND STAGE**, in each group, participant C has an endowment of 200 tokens.

In this stage, before the computer communicates the choice made by A, participant C has to decide whether he/she is willing to eliminate possible errors in computer recording, by paying 20 tokens. Then, participant C has to decide whether to sanction A or not in the following two possible cases: the computer communicates that A has subtracted tokens from B / the computer communicates that A has not subtracted tokens from B.

If C decides to sanction A, he/she will also have to specify by how many tokens (opting for 40, 50 or 60 tokens).

If participant C decides to pay 20 tokens to eliminate any possible recording errors, the graphical description that he/she will see when he/she will have to decide with regard to each possible

choice made by A recorded by the computer will be the same as in the first round of the experiment (*see Figure 1*).
The choice made by C that will be implemented will depend on the decision recorded by the computer.
Information about the decisions taken by participants A and C will be provided only once the experiment will be over.

If participant C decides not to pay 20 tokens, various scenarios may emerge.
In particular:

- If the computer communicates that A has not subtracted 50 tokens from B and C decides to sanction A, participant C will sanction a participant A that CERTAINLY has not subtracted tokens from B.

- If the computer communicates that A has subtracted 50 tokens from B and C decides to sanction A, participant C will sanction a participant A that MIGHT have subtracted 50 tokens.

In this case, the graphical description that participant C will see when he/she will have to decide with regard to each possible choice made by A recorded by the computer will be the following (*see Figure 2*).
The choice made by C that will be implemented will depend on the decision recorded by the computer.
Information about the decisions taken by participants A and C will be provided only once the experiment will be over.

Now the **FIFTH ROUND** begins. In this round, each of you keeps his/her role but will be matched to other participants.
In the **FIRST STAGE**, in each group, participant A has an endowment of 100 tokens and has to decide whether to subtract or not 50 tokens from participant B. Participant B has an endowment of 100 tokens and cannot take any decision.
The computer records the decision made by A. However, in this round, like in the third round, the computer may make errors in recording the decision made by A. In particular, the computer may fail to record that 50 tokens have been subtracted even if participant A decided to subtract tokens from B.

In the **SECOND STAGE**, in each group, participant C has an endowment of 200 tokens.
In this stage, before the computer communicates the choice made by A, participant C has to decide whether he/she is willing to eliminate possible errors in computer recording, by paying 20 tokens. Then, participant C has to decide whether to sanction A or not in the following two possible cases: the computer communicates that A has subtracted tokens from B / the computer communicates that A has not subtracted tokens from B.
If C decides to sanction A, he/she will also have to specify by how many tokens (opting for 40, 50 or 60 tokens).

If participant C decides to pay 20 tokens to eliminate any possible recording errors, the graphical description that he/she will see when he/she will have to decide with regard to each possible

choice made by A recorded by the computer will be the same as in the first round of the experiment (*see Figure 1*).
The choice made by C that will be implemented will depend on the decision recorded by the computer.
Information about the decisions taken by participants A and C will be provided only once the experiment will be over.

If participant C decides not to pay 20 tokens, various scenarios may emerge.
In particular:

- If the computer communicates that A has subtracted 50 tokens from B and C decides to sanction A, participant C will sanction a participant A that CERTAINLY has subtracted tokens from B.

- If the computer communicates that A has not subtracted 50 tokens from B and C decides to sanction A, participant C will sanction a participant A that MIGHT have not subtracted 50 tokens.

In this case, the graphical description that participant C will see when he/she will have to decide with regard to each possible choice made by A recorded by the computer will be the following (*see Figure 3*).
The choice made by C that will be implemented will depend on the decision recorded by the computer.
Information about the decisions taken by participants A and C will be provided only once the experiment will be over.
