When Punishment Doesn't Pay: "Cold Glow" and Decisions to Punish

Aurélie Ouss[*], Alexander Peysakhovich[‡]

October 2012

Abstract

When will the aggregation of individual punishing behaviors lead to outcomes in line with those resulting from instrumental uses of sanctions? We present a model where individuals derive private utility from punishing norm-breakers ("cold glow"), and compare their choices to those made if penalties are only viewed as a means for social cooperation. Our theory predicts that cold glow punishers take into account their private share of the cost, care about their own contribution to overall punishment, and underweight the role of probability of capture. Instrumental punishers seeking optimal deterrence care about social costs and benefits of punishments, probability of apprehension, and total levels of punishment. This means that that different environments can predictably result in either over-punishment or under-punishment relative to the benchmark of optimal deterrence. We confirm this in a series of experiments.

[*]ouss@fas.harvard.edu

[†]apeysakh@fas.harvard.edu

[‡]We would like to thank Phillipe Aghion, Yochai Benkler, Tom Cunningham, Ed Glaeser, Roland Fryer, Oliver Hart, Drew Fudenberg, Louis Kaplow, Lawrence Katz, David Laibson, David Rand, Al Roth, Steven Shavell and the seminar participants at the Harvard Labor Lunch for helpful comments. Part of this research was funded by a grant from the Lab for Economic Applications and Policy at Harvard University.

"[The] Eighth Amendment is our insulation from our baser selves [when] a cry is heard that morality requires vengeance to evidence society's abhorrence of the act." - Thurgood Marshall

1 Introduction

Do individuals choosing punishments act in ways that are compatible with optimal levels of punishment to achieve social cooperation at minimal cost? If individuals derive private benefits from punishing norm-breakers, they will respond to different parameters than punishers only interested in maximizing material social welfare (optimal deterrence), and aggregate outcomes might differ radically. We build a theory of punishment decisions built on psychologically defensible assumptions, use it to focus on a set of parameters to consider, and test responses to variations in these parameters in a series of experiments.

There are many reasons for devoting resources towards sanctioning law breakers. For example, deterrence theory posits that higher potential punishments reduce law-breaking in a society, thus helping to maintain social cooperation. On the other hand, retributive theories see punishment as an end in itself. Motives such as deterrence could also be called 'public goods' motives of punishment, where the public good is increased social payoffs from increased cooperation; while motives such as retribution could also be called 'private goods' motives, since individuals receive personal utility from the punishment itself. Our main intuition is simple: if individuals view punishment more like a private good, then aggregations of individual decisions may not lead to outcomes in line with optimal punishments with a benchmark such as optimal deterrence in mind.[1]

We first formalize our intuition of punishment as a private good using a simple model. Our model builds strong reciprocity (see Gintis et al. (2005) for a survey) into a utility function. When an individual commits a socially praise-worthy act, that individual's payoff positively enters into the utility of others: they gain private benefits from increasing his welfare. Conversely, when an individual commits a norm violation, the utility of the norm violator negatively enters the utility of other individuals: individuals gain private

[1]This benchmark, formalized by Becker (1968), is the most frequent model used in the economics of crime literature. We later discuss other possible benchmarks motivated by the idea that psychic costs and benefits can be allowed to enter the cost-benefit calculation.

benefits when a norm-breaker's material welfare is reduced. We focus on the utility from punishment aspect of our model, which we term "cold glow."[2] We compare these decisions to those chosen by a Beckerian punisher interested in total social material payoffs. Compared to this benchmark, cold glow punishers respond to personal shares of cost of punishment and not to the total burden to the public, can be relatively insensitive to probability of apprehension, and punishment by others may not substitute for own punishments perfectly. These effects can lead to over or under punishment relative to the deterrence benchmark, depending on the structure of the environment.

Individual punishment decisions are important in many contexts, ranging from daily interactions to business organizations. We discuss one such domain of application: how individual decisions can shape aggregate outcomes in the criminal justice system. Our theory is most applicable to two such channels: voter behavior (the elections of judges and legislators) and juries (citizen juries set penalties in tort cases). We survey existing evidence on behavior of voters, judges and tort juries that is consistent with our theory of cold glow punishments. We then turn to another method for testing this theory: a series of laboratory experiments. Our experimental designs allow for transparent calculation of levels of punishment that would reach normative benchmarks.[3] This allows us to not only ask whether individual behavior responds to particular parameter changes but also whether aggregate behavior is, in some sense, 'optimal.'

We present three experiments, in which people can punish a norm violation: taking from a third party. We vary conditions of sanctions in order to test the role of different parameters in punishment choices, and how individual behaviors aggregate up.

Our first experiment looks at how punishment choices respond costs. The punishment available in this experiment is excluding norm breakers from the game: when this happens, they can neither make money nor take from other players. We show that environments where individuals can punish norm-breakers but do not personally bear the full cost of their decisions can lead

[2]In reference to warm glow theories of altruism, described in Andreoni (1990) and related works.

[3]Many papers consider the addition of punishment to public goods games (Ostrom et al. (1992)), dictator games (Fehr and Fischbacher (2004)). Others ask for individuals' impressions of 'fair punishments' in survey scenarios (Baron and Ritov (1993), Sunstein et al. (2000)). However in these games the calculations for material payoff maximizing punishments are not as transparent.

to socially inefficient over-punishment. Our setup is such that relatively small punishments can implement social goals consistent with motives of general deterrence, specific deterrence and incapacitation; yet when costs are not fully internalized, players over-punish. Results from this experiment allow us to conclusively rule out 'public goods' motivations as the sole drivers of high levels of punishment.

Our second experiment investigates the role of probability of apprehension in punishment choices. A player can take from a third party, and we experimentally vary the probability with which he is found (high or low). We compare ex-ante punishment choices and taking behavior across conditions. Consistent with our theory, choices of penalty do not react to changes in probability of apprehension, but taker behavior does. This leads to a different kind of inefficient punishment: levels too low to deter socially destructive behavior. We replicate these results in a third experiment where assigners give penalties as a reaction to decider behavior and not as an ex-ante deterrent.

Our final experiment looks at whether our 'cold glow' terminology is apt. The theory of 'warm glow' (Andreoni (1990)) posits that individuals gain private benefits from the *act* of contributing to a public good and not from the total share provided. In our final experiment, we ask whether individuals gain private benefit from overall levels of punishments imposed on norm-breakers, or whether these psychic benefits come from *their own* contributions to the punishment. In our study, two individuals make punishment decisions in sequence. We look at whether the second decision-maker's punishment decreases with the punishment of the first individual, and find that on average, no crowd-out occurs. We replicate these effects in an experiment where the first punisher's decision is made by a computer.

The rest of our paper is organized as follows: in the next section we present our model. In section 3, we review empirical evidence on field behaviors consistent with cold glow. Sections $4-6$ present our experiments, which examine respectively the impact of cost structures, the effects of probability of apprehension, and crowding out. Section 7 concludes.

2 A Model of Third-Party Negative Reciprocity

We first present a model punishing behaviors. We look at aggregate outcomes when punishers care about material social payoffs, and when they also derive

utility from punishment itself. We focus on third-party punishment, and ask what parameters affect decisions depending on the punisher's motivations.

2.1 General Setup

We begin with a simple game with three players: a taker (T), who can take or not take $\{t, nt\}$ from a victim, (V); and a punisher (P) who chooses s_P, how much to sanction the taker. If T chooses to take, V loses $s_T > 0$, and T gains αs_T. An individual who chose t is caught with probability p, in which case, they receive the sanction chosen by P, s_P. We treat V as a passive observer. The taker and the victim's utilities are given by their material payoffs:

$$\begin{aligned} U_T(s_T, s_P) &= \alpha s_T - s_P \\ U_V(s_T) &= -s_T \end{aligned}$$

We have $\alpha \in [0, 1]$, so that taking is a socially destructive action. Finally, we assume that sanction level s_P has a cost of $\beta > 0$ per unit paid for by the punisher P.

We will consider punishers with two types of social preferences: first, a punisher who cares about total material welfare; second, a punisher who gains personal utility from punishing socially destructive actions. We will apply these models to three types of sanctions: ex-post punishment, ex-ante punishment commitments, and punishment in the presence of several punishers.

2.2 Material Social Payoff Maximizing Punishers

We first consider the case of a punisher who cares about total material welfare: his goal is to minimize harm, subject to cost. While we allow for flexibility in assessment of harm and in the relative weight of pro-social and individual considerations, the punisher's problem is similar to that of the social planner in Becker (1968), so we will refer to him as a Beckerian punisher.

The Beckerian punisher's utility from the action pair (s_P, s_T) is given by

$$U_P(s_P, s_T) = -\beta s_P + \gamma(U_V(s_P, s_T) + \phi(s_T)U_T(s_P, s_T)).$$

The first term reflects P's material payoff, which can only be affected by his choice of sanction. The second term is P's social preference; γ measures how much weight P puts on maximizing social efficiency relative to his own

payoffs. So $\gamma = 0$ represents a standard self-interested actor, and $\gamma = \infty$ represents an individual who cares only about the total material payoff of the rest of society, ignoring his own payoff.[4] We let $\phi(s_T)$ represent the weight of the taker's utility in the punisher's maximization, which can depend on T's actions. When the taker does not take, $\phi = 1$ but if they choose to take, P may put a lower value on the taker's payoff than on the victim's.

The ex-post punishment case is trivial: in a one-shot interaction, P would choose a punishment level of 0, since there are only costs and no benefits to punishment. The ex-ante case, where P commits to a publicly known level of punishment s_P before T makes their decision, is more interesting. Recall that when he chooses to take, T is found with (exogenous) probability p, in which case the sanction applies. The taker is perfectly aware of this law when making her decisions, so she chooses to take if

$$U_T(s_T, s_P) \quad = \quad \alpha s_T - p s_P \quad > \quad 0$$

From this equation it follows that there is a level $s_P^{Deter}(p)$ above which T will not take, while below which T prefers to take and that this s_P^{Deter} increases in p. For simplicity, we assume that when indifferent, T chooses not to take. To avoid off equilibrium path dynamics, we assume that T trembles to an unintended action with probability ϵ which is small.

We can also look at P's utility in different cases:

$$U(s_P) = \begin{cases} 0 & \text{if } T \text{ didn't take} \\ -s_T + \phi(s_T)(\alpha s_T) & \text{if } T \text{ took and was not found} \\ -\beta s_P - s_T + \phi(s_T)(\alpha s_T - \beta s_P) & \text{if } T \text{ took, was found and } s_P \text{ was applied} \end{cases}$$

Assuming that the taker is rational as above, the punisher can maximize this utility with backward induction. We can show that the only levels of punishment that P ever chooses are 0 or $s_P^{Deter}(p)$. The punisher wishes to deter T from taking to maximize material social welfare, but if the potential costs (eg. in the case of a tremble) are too high, then this may not be worth it.

Note that this choice also depends on the level of γ, the weight that P puts on social material welfare relative to his personal costs. This gives

[4]The $\gamma = 1$ case, where an individual maximizes total material social payoff including his own in the welfare calculation, is most analogous in our context to the social planner Becker (1968) crime reduction function.

the following important implication: if P's chosen punishment is not paid for by himself, but by a fourth player (the public) then P's maximization problem becomes exactly that of a social planner maximizing total material welfare. Thus, moving punishment costs from being private to being borne by the public can only improve total material social welfare with a Beckerian punisher. If ϵ is small, under such a publicly funded punishment scheme Beckerian punishers will always set $s_P^{Deter}(p)$, and so punishments decisions will respond strongly to variations in probability of capture.

Note also that if we think about not a single Beckerian punisher, but two identical individuals P_1 and P_2 who each set a punishments, it is easy to show that in any equilibrium of the game we will have that

$$s_{P_1} + s_{P_2} \in \{0, s_P^{Deter}(p)\}.$$

Thus, Beckerian punishers will respond to variations in total social cost, probability of capture and care about total levels of punishment. We now introduce a model of a cold glow punishment choices and study how aggregate decisions differ.

2.3 Choices of Punishment with Negative Reciprocity

We now assume that individuals receive private benefits from negatively affecting the payoffs of those who have done socially inappropriate actions. We call these private benefits cold glow. Though we do not present it here, our model could be expanded to allow for individuals to get a private benefit, or warm glow (Andreoni (1990)), from positively affecting the payoffs of those who have done socially appropriate actions.

We argue that our assumptions about cold glow can be justified empirically. A large literature in behavioral economics points to the fact that individuals will often sacrifice personal payoffs to reduce the payoffs of individuals who behave selfishly in games such as the public goods game (Ostrom et al. (1992)) or the dictator game, even when the individual choosing to sanction is a third party and has no personal stake in the game itself (Fehr and Fischbacher (2004)). This occurs even when punishments can't be used to 'teach a lesson.[5][6]

[5]Fudenberg and Pathak (2010) show that individuals will pay to punish others who behave anti-socially in public goods games even when the effects of the punishment are not known until the end of the session.

[6]We also note that harmful acts appear to be punished more harshly when they are

Studies in social neuroscience give evidence that this behavior is driven by pleasure gained from the sanctions themselves: activity in the brain's reward areas during costless punishment can be used to predict punishment behavior in costly punishment situation (De Quervain et al. (2004)). Furthermore, individuals show reward activity (which correlates with subjective reports) when they watch another individual who cheated them in a trust game receive electric shocks, but not when the shocks are given to an individual who had been cooperative (Singer et al. (2006)).

Finally, research in moral psychology hints at the final part of our assumption, which is that this motive is very blunt: it is 'turned on' by harm itself and not by intention to harm. Cushman et al. (2009) ask individuals to play a modified dictator game, in which the dictator chooses between dice, with each different die yielding different probabilities of fair or selfish allocations. After the die is rolled, recipients are allowed to punish or reward the dictator. The authors find that outcomes predict punishment or reward behavior by the recipients, while intentions (choice of dice) have a smaller effect.

We first discuss these preferences in an ex-post decision. We then show how they affect ex-ante punishment decisions, that is, choices of punishment made when individuals can credibly commit to sanction a behavior in the future.

2.3.1 Ex-Post Behavior

First, we introduce a basic model of social preferences which depend on the *action* taken by another player.[7] We start with simple three player model, and then extend it to N players.

caused more directly. For example, Coffman (2011) shows that third parties punish a harmful act more when an individual himself commits it than when the same individual uses an intermediary to create the same outcome. To keep our discussion simpler, we omit such motivations from our model.

[7]Theories of social preferences in economics can be divided into several categories: theories such as inequity aversion (Fehr and Schmidt (1999)) take outcomes as the objects over which utility functions are defined, while fairness theories (eg. Rabin (1993)) take intentions as the important objects. By contrast, we take *actions* as well as payoffs as the primary focus of our theory, in this way we are similar to social norms theories such as Axelrod (1986)).

Three Players The utility functions of the Taker and the Victim are the same as in the previous sub-section. However, the Punisher's utility is now a function of both his material payoff (the first term), and his reaction to the Taker's action:

$$U_P(s_T, s_P) \quad = \quad -\beta s_P + \lambda(\Delta U_T(s_T, s_P), s_T)$$

The second term in P's utility, λ, captures the punisher's social preferences. The first argument of this function, $\Delta U_T(s_T, s_P)$ is the *total change* (relative to some baseline) to the taker's utility that occurs as a result of the punisher's action. Note that because T's utility is linear in s_P, the choice of baseline doesn't matter. Since s_T is fixed from the punisher's perspective when the punishment is carried out, we simplify the arguments to $\lambda(s_P, s_T)$.

The second argument, tells us how T's actions affect P's social preferences over T's payoff.

We make the following assumptions:

Assumption 1. *λ is smooth and concave in s_P.*

This is a standard assumption so that we can use our tools of maximization. Note that we do not very much constrain the shape of λ. In particular, for any s_T, λ can either be always increasing (bigger punishments are always better) or reach a global maximum for a certain value of s_P, which can be thought of as the 'perfectly fair' punishment, in line with just desserts theories.

Assumption 2. *We have that* $\dfrac{\partial^2 \lambda(\cdot, s_T)}{\partial s_P \partial s_T} > 0.$

Assumption 2 is the driving assumption of our model. It states that as the taker's action becomes more inappropriate, the punisher's attitude towards T's payoffs becomes increasingly negative (recall that higher s_T means a larger transfer from V when T chooses to take). Our final assumption is a normalization:

Assumption 3. *We have that* $\dfrac{\partial \lambda}{\partial s_P} = 0$ *if* $s_T = 0$.

This tells us that $s_T = 0$ is a 'neutral' action which causes P to not feel either positive or negative strong reciprocity towards T. Note that in a generalized model we could relax smoothness assumptions for λ and our

main results would hold. We maintain these assumptions to make exposition easier.

Note that because s_T is fixed for ex-post behavior, *levels* of λ in s_T are irrelevant for predicting P's static behavior. However, assumptions on this will make important statements about dynamic behavior or welfare. In particular, this allows for the both the situation where $\lambda(0,0) \geq \lambda(s_P, s_T \neq 0)$ implying that P would prefer to be in a situation where T takes a neutral action than where he has to exercise reciprocity or the opposite. We turn to this discussion later. However, without any assumptions on this we can still characterize behavior for a given s_T:

Proposition 1. *For any β, s_T there exists an optimal action for the punisher $s^*(\beta, s_T)$. Moreover*

1. *$s_P^*(\beta, s_T)$ decreases in β.*
2. *$s_P^*(\beta, s_T)$ increases in s_T.*

The comparative statics are easy to see: first, as the price of transfers (β) increases, P will provide less of it. Second, P's sanction of T is increasing in the inappropriateness of her behavior. There is already some evidence that punishment responds to both prices and inappropriateness in these ways: Anderson and Putterman (2006) find that punishment in public goods games responds to price effects as a normal good[8] while Peysakhovich and Rand (2012) find that reported inappropriateness ratings correlate with punishment decisions in a dictator game with third party punishment.

We note that we do not need to make these assumptions about how T's action affects V's payoff. Our model is perfectly consistent with a scenario in which T chooses an action s_T from a continuum, those s_T being linearly ordered by 'social inappropriateness', where higher s_T actions are considered more inappropriate by P.[9]

Four Players We now move to the case when several individuals observe the taker's behavior and can choose to affect his payoff. Suppose now there

[8]These are not exactly our scenarios as public goods game punishments are not third party.

[9]Because we take appropriateness as exogenously given, an important expansion of our project would be to consider how appropriateness of various actions can be endogenously generated. We leave this nuance for future work.

are four players: the taker T who can choose to take from the victim V, and two punishers, P and P_2, who move sequentially and can each choose how much to punish the taker, with P_2 knowing P's decision. T's utility now looks as follows:

$$U_T(s_T, s_P, s_{P_2}) \quad = \quad \alpha s_T - s_P - s_{P_2}$$

P's utility function is now as follows:

$$U_T(s_T, s_P, s_{P_2}) \quad = \quad -\beta s_P + \lambda(s_P, s_T, s_{P_2})$$

Keeping the old assumptions on the shape of the λ function, there exists a unique $s_P^*(\beta, s_T, s_{-b})$ describing the original punisher's optimal choice. What this setup gives us relative to the three-player setup is the possibility to discuss how P's punishment choices interact with that of all other punishing agents. Specifically our model allows for several types of behaviors:

Definition 1. *We say that if:*

1. $\dfrac{\partial s_P^*}{\partial s_{P_2}} < 0$ *(*$= -1$*), there is (perfect) crowding out*
2. $\dfrac{\partial s_P^*}{\partial s_{P_2}} > 0$ *(*$= 1$*), there is (perfect) crowding in*
3. $\dfrac{\partial s_P^*}{\partial s_{P_2}} = 0$*, P's punishment choice is independent of other punishers'*

Crowding out happens if a punisher considers that his and other players' punishment choices are substitutes to some degree. When crowding out is perfect (as was the case for the Beckerian punisher), a punisher only cares about is the overall level of punishment, similar to the maximum-deterrence punisher. When crowding out is imperfect, the punisher also cares about how *he* changes the taker's utility.[10]

Crowding in, on the contrary, implies that P's choice of punishment will be an increasing function of the other players' choices of punishment: the more other players punish, the more P punishes. Our model is mostly reduced form; one interpretation is that crowding in results from an imperfect

[10]This is the flip side of 'warm glow' as discussed in Andreoni (1993) or Cornes and Sandler (1994) for public goods contributions.

knowledge on P's part about of how wrong T's action was. With this uncertainty in place, other players' actions serve as a signal and so can lead to crowding in of punishments.[11]

Note, again, that we do not make any assumptions on the behavior of *levels* of λ in s_{-b}, as this variable is exogenous for P. For example, we make no assumptions about whether P prefers situations in which other individuals are also allowed to punish guilty individuals. Note that assumptions on this will also inform dynamic behavior.

Which behavior holds at individual level in an empirical question. The overall aggregate levels of punishment in society will depend on the relative proportions of decision-makers who display either behavior. We study this question in experiment 3.

2.3.2 Ex-Ante Punishments

So far, we have only looked at P's ex-post punishment decisions, taking T's action s_T as fixed. However, most punishment decisions are set ex-ante: laws and rules are set out and potential norm-breakers are presumed to know the laws. To better understand this situation, we now turn to incorporating ex-ante motives into our theory of punishment behavior. We do so in a highly reduced form way to get our main intuitions across.

First, we modify the order of the game and simplify the strategy space: P first sets out a sanction, s_P, to which he commits. T, having seen this sanction, makes a choice from the set $\{t, nt\}$ where t (taking) is some fixed $s_T > 0$ and NT is $s_T = 0$. If T chooses t, she gets a fixed benefit of k; she is caught and has P's sanction applied to her with probability p. Note here that P only pays for the sanction if it has to be implemented.

We assume that P has a map $\psi(s_P, p)$ which represents his probabilistic assessment that T will choose t given a sanction of size s_P. Further, we assume that the function is smooth, that $\psi(s_P, p)$ decreases in s_P, that ψ is bounded away from 0 to avoid off equilibrium dynamics and that the cross partial is negative. Intuitively, these assumptions correspond to P believing that higher sanctions decrease taking and that higher sanctions decrease taking more when probability of being caught is higher.

We leave open many possible choices of ψ,. For example, P could have rational expectations about T's behavior. One way to create a particular

[11] As in Bardsley and Sausgruber (2005) or Glazer and Konrad (1996).

choice for ψ is to assume a well behaved distribution of types k for T with $k \in [0, k_{max}]$ distributed according to pdf $f(\cdot)$. Each type gets utility k from choosing t and uses a simple cost benefit tradeoff between expected sanction and expected benefit to make decisions and trembles with probability ϵ. If P does not know T's type, but knows the distribution $f(\cdot)$, we will obtain a ψ function that satisfies our criteria. We also leave open the possibility that P may be partially strategically naive: ψ can be derived from a level-k thinking (Costa-Gomes et al. (2003)) or cognitive hierarchy (Camerer et al. (2004)) model.[12]

To make ex-ante decisions, P maximizes the following expected utility:

$$\psi(s_P, p)[p(\lambda(s_P, T) - \beta s_P) + (1-p)(\lambda(0, T))] + (1 - \psi(s_P, p))\lambda(0, 0).$$

Note that now the difference in *levels* $\delta(s_P) = \lambda(0,0) - \lambda(s_P, T)$ matters. If $\delta(s_P) > 0$ for all possible values of s_P, P prefers to be in the situations where T does not take and he does not punish than in a situation where T takes and P is forced to act. This means that P's ex-ante punishments incorporate a form of a deterrence motive.[13] Having an extra motive for punishments gives us the following result:

Proposition 2. *For generic choice of ψ there exists unique s_b^* that is the optimal ex-ante punishment. Moreover this ex-ante punishment is always weakly greater than what would be imposed for $s_T = t$ in the ex-post problem above.*

This proposition means that ex-ante and ex-post punishments are different in theory, but does not explain how large this difference is. What determines this difference is the relative shapes of λ and ψ. There are three interesting cases to consider. The easiest is where P is completely strategically naive and believes that T chooses T with a fixed probability no matter the sanction. This then reduces the ex-ante decision to the ex-post punishment case.

[12] We point out that understanding how accurate individuals are in their beliefs about how punishment levels affect decisions of potential criminals is an important topic at the intersection of law and psychology but we do not discuss it further here.

[13] This also means that our model nests a decision-maker who cares only about the deterrence aspects of punishments by setting $\dfrac{\partial \lambda}{\partial s_P}$ to be constantly 0. In this case $\delta(s_P)$ is exactly the weight that P puts on the social loss in payoffs that happens in T chooses t.

The second case is where most of the change in ψ happens at low levels of s_P. One such example sets $\psi(s_P, p) = 1$ if $s_P < \epsilon$ and $\psi(s_P, p) = q$ for $s_P \geq \epsilon$ with q and ϵ very small. Here cold glow motives push punishments above where they would be if P simply had a taste for deterrence (which would dictate that he simply set a punishment of ϵ).

However, there could be other possibilities. Consider a scenario where $\psi(s_P, p) = 1$ for $s_P < K$ where K is large and $\psi(s_P, p) = \epsilon$ for $s_P > K$. Thus, only very large punishments are deterring, but once the threshold is reached most taking behavior goes away (this could happen, for example, if T is rational, the benefits of T are modest and p is very low). Now, add to this a $\lambda(\cdot, T)$ which is single peaked in the first argument (that is, P has an optimal 'fair' punishment) and further suppose that this peak, $\overline{s}_b$, is much smaller than K. Set β very close to 0. Now an optimally deterring punishment would be one of size K, but P may choose a punishment much lower than this. Intuitively, this is because by setting a punishment K, P commits to choosing an action that is highly suboptimal, from his point of view, in the positive probability state of the world where T takes and is punished.

Thus, while cold glow gives P a deterrence motive for punishment, it also gives him other motives which he must trade off during his decision-making. We characterize the relative sizes of some of these motives in experiment 2.

2.4 Welfare Implications

We now compare parameters that matter for cold glow punishers relative to Beckerian punishers, and discuss other possible social benchmarks. Let's first discuss how cold glow P chooses a punishment whose cost is shared between P and P_2 (so P pays $\frac{\beta}{2}$ per unit of punishment).

We begin with the ex-post case where T has already chosen to take and has been caught. By our analysis above, P will make the choice that equates his marginal benefit from cold glow to its marginal cost (here $\frac{\beta}{N}$). This will lead to higher levels of punishment that those chosen by the Beckerian punisher, who factors in *total* costs. Furthermore, if cold glow is not included into aggregate welfare, or if it is a private good which only benefits the punisher, but not the rest of society, then sharing costs could lead to over-punishing. We will test this in experiment 1.

One could also assume that each member of society receives cold glow utility from punishment and has preferences identical to P, and that all choices are legitimate reflections of welfare. In this case, P acts as the rep-

resentative agent for society. However, even if we take cold glow to be a legitimate source of welfare, problems can arise. For example, we can consider a simple extension to our game where individuals can select into the role of punisher.[14] With sorting in place, individuals with 'the strongest' cold glow have incentives to sort into particular positions and it is unclear that individual maximization will lead to socially optimal outcomes *even if* cold glow enters into the calculation of social welfare.

We can also consider the opposite view. Behavioral economists (e.g. Kahneman et al. (1997)) often break utility down into two components: decision utility, the maximizer of which is P's choice, and experienced utility, which can be used for welfare comparisons. Taking such a point of view, cold glow reflects how individuals make decisions but doesn't tell us the whole story about how these decisions make them better or worse off. Finally, there is the important matter of how to weigh T's decrease in payoffs against the gains of other players. Moving to the ex-ante case (for example, setting laws or voting for politicians) adds even more complications to the discussion.

So far, we've given brief and by no means exhaustive list of possible ways to think about how cold glow motives should enter into aggregate welfare calculations. However, in each of these, one thing is clear: it is quite unlikely that the solution to the individual punisher's maximization problem, or to those of many such punishers, would in general aggregate up to produce socially optimal outcomes.

Our experiments test how parameters enter into individual level decisions, and they are set up in such a way that we can calculate what punishment would satisfy the Beckerian punisher's preferences. This lets us make statements both about what individuals seem to be doing and about whether their aggregate actions lead to socially optimal outcomes, and if not, how badly they miss the target.

3 Punishment Behavior in the Field: Criminal Justice

Before we turn to our experiments, we discuss how our investigation into the interaction between psychological motives and institutional structures fits

[14] In the criminal justice system, this could happen via matching mechanisms, for example if more punitive individuals choose to become criminal prosecutors.

into understanding important outcomes. So far, we have developed a stylized model of punishment behavior and how this behavior can lead to situations in which punishments set by individuals miss our normative benchmark. There are many important situations in which punishment decisions can affect aggregate outcomes and where we could apply such an analysis: organizations, work in groups, driving, and so on. Here, we limit our scope to a particular application: socially provisioned punishment via the criminal justice system.

We now review existing empirical work which is consistent with our model and discuss how cold glow motivations could affect aggregate outcomes through the behaviors and preferences of voters, juries and judges. We then turn to discussing how lab experiments can be integrated into an empirical strategy for understanding the aggregate effects of individual motivations.

Demand for punishment for private motives can affect aggregate outcomes through the behavior of elected officials. First, we note that if the punishment of criminals is indeed treated by voters as a private good which is provided at public cost, this would lead to demand for punishment even in the absence of clear effects on the crime reduction. There is qualitative discussion of this phenomenon: for example, legal sociologist David Garland argues that the most publicized measures (such as three strike laws, or Megan's law) have little effect on controlling crime but tend to become law due to "their immediate ability to enact public sentiment, to provide an instant response [or] to function as a retaliatory measure" (Garland (2001)).

In addition to descriptive evidence, causal links have been identified: Berdejo and Yuchtman (2009) analyze changes in sentencing behavior of judges during election cycles. They find that judge severity increases[15] when they are close to reelection and thus under political pressure from constituents, and sentences fall immediately afterwards. These results cannot be explained by differential work loads due to longer sentencing and variations in the month of nomination and election allow the authors to rule out seasonality or confounding political changes. This phenomenon of pre-election increase in sentences, immediately followed by a drop, is consistent with a model in which judges' preferences differ from individual voters' decisions, which are driven by the cold glow heuristic.

Cold glow could also affect outcomes in the criminal justice system through the behavior of judges themselves. We view that as a less likely place of in-

[15]Furthermore, the authors find that this variation is due to discretionary departure above sentencing guidelines, and not greater compliance to these guidelines.

fluence, since judges are specifically trained and make their decisions in a deliberate manner, perhaps mitigating the effects of cold glow. There has been a recent resurgence of interest in studying judicial behavior (Posner (2008), Danziger et al. (2011)) which has put forth at least some evidence that judges are subject to predictable biases, so perhaps it is not impossible that cold glow is partially at play during judicial decisions.

In addition, there is some evidence in law and economics pointing to the fact that individuals may not believe that it is "fair" to factor probability of capture into punishment decisions (see Polinsky and Shavell (2000) for a discussion and Sunstein et al. (2000) for two survey-based experiments). Insensitivity to probability of capture by punishers, an important input into optimal deterrence, is a behavior that cold glow punishers can display.

There has been no research directly assessing the effect of cost structures on demand for punishment, even though the question of costs of punishment has received attention from policy makers due to the budget crises in many states.[16] The only paper to investigates the effect of a change of costs on punishment decisions is Ater et al. (2012). They exploit a quasi-experimental change in costs of arrests in Israel: the responsibility of housing arrestees awaiting trial was transferred from local police to the prison authority. The authors find a sharp increase in arrests as a result of this policy, which is consistent with an imperfect factoring in of total costs of crime reduction when making arrest decisions.[17]

Additionally, whether individual punishment decisions are crowded out by already performed punishments could play a role in labor markets. There has been discussion on the role that having a criminal record plays employability of an individual (Bushway et al. (2007), Pager (2007)). One way this can occur is through a signaling channel (Rasmusen (1996)) where conviction is a signal of poor worker. However, if cold glow motives are not crowded out by already performed punishments, there may be a second channel for this effect: a lack of hiring can act as a sanction towards an individual who has committed an inappropriate act. The relative sizes of each of these effects matter quite a bit for choices of particular policies (for example, shrouding

[16] In particular, in California, one response has been to transfer housing of inmates from state prisons to county jails, with the argument that this would lower overall costs of criminal justice.

[17] We note there are many other possible explanations for these results: police officers' effort provision might respond to costs, police evaluations could depend on number of arrests, etc.

criminal records).

All in all, a lot of empirical facts can be explained by cold glow motivations playing a role in decisions which affect important aggregate outcomes. However, these decisions are a product of many factors: elections involve many non-judicial dimensions, juries are prompted to depart from emotions,[18] and exact magnitudes of costs or probabilities of apprehension are generally not known precisely by voters, juries or judges. In order to conclusively isolate the role and magnitude of cold glow in aggregate outcomes, we would ideally need data on voter, jury and judicial behaviors responding to (quasi) experimental variations in costs of judgments and probability of apprehension. Beyond the practical difficulties of implementing such a protocol, it would be difficult even in this scenario to isolate the exact mechanisms at play. To build our understanding of how cold glow interacts with institutions, we examine the behavior of individuals in a stylized setting using a series of laboratory experiments. These experimental methods allow us to study, in a controlled environment, punishment choices which are normally hard to observe in the field. For this reason, they are an important piece of a larger portfolio of methods that can help us to analyze and evaluate how cold glow motivations affect aggregate outcomes.

4 Experiment 1: Responses to Costs

In this first experiment we test an individual level hypothesis: when costs of punishment accrue to the group rather than to the individual, will individuals increase their punishment decisions? At the social level, the game is set up so that very low levels of punishment are sufficient to deter potential norm breakers. Simultaneously, our transfer of costs to society also increases the overall cost of the punishment beyond what would be consistent with using motives such as incapacitation as the social benchmark. We then ask: will individual punishment decisions meet or exceed our social benchmarks?

[18]For example, French jurors verbally pledge that they will "not listen to hatred or malice or fear or affection; [and decide] according to [their] conscience and [their] inner conviction, with the impartiality and rigor appropriate to an honest and free man."

4.1 Experimental Design

We run a series of experiments in which we vary the availability and cost structure of sanctions. In our game, participants gain Monetary Units (MU) throughout the experiment, which are converted into dollars at a rate of 50 MU per dollar. Players are randomly matched in groups of $n = 8$ to 12 players. Each group is given a public pot of $70 * n$ MU, which is equally split amongst all members of the group at the end of the game. Each player is also individually given 30 MU at the beginning of the game.

Participants play 20 rounds (one iteration) of the following game. They are asked to solve a simple math problem, for which they receive 4 MU upon completion. They are then given the possibility to "take." If a player chooses to take, she receives 2 MU, and another randomly selected player loses 3 MU. Taking, in this case, is a socially destructive behavior; yet, in the absence of sanctions, it is a dominant strategy. When a player chooses to take, she is found out in 50% of cases. Our conditions and treatments consist of varying what happens when a player is found out.

In the "No Punishment" condition, when a player is found out, she gets a message informing her that she has been found out, but nothing more happens. In both "Punishment" conditions, when a player is found out, another random player is chosen to be her "assigner." The assigner is able to punish found out players by excluding them from the game for up to 10 rounds. We elicit punishment using the strategy method: individuals choose a punishment after making their "take" decisions and seeing whether they were taken from, but before they are informed of whether they were found out, or if they were someone's assigner. They are asked at this point to enter an amount of penalty rounds that they would assign if they are chosen as an assigner for this round. Individuals can never be chosen as their own assigner, nor do they know which player they assign penalty rounds to. In particular, if they were taken from, there is no additional chance that they will assign a punishment to the player who took from them. In all conditions, only the assigner and the individual to whom penalty rounds are allocated learn about the punishment level chosen.

Each round of exclusion is costly, and we vary the cost structure. In the "Private Punishment" (hereafter Private) condition, if a player's punishment is chosen, they will pay 2MU from their private money for each round of punishment they have imposed. In the "Public Punishment" (hereafter Public) condition, if a player's punishment is chosen, each round costs 5 MU from the

public pot. This means that in the Public condition, the private share of the cost to a particular punisher is less than 2 MU per round. This experimental setup will allow us to investigate cost effects in demand for punishment, thus determining if demand for punishment looks like demand for a private good.

As a robustness check, we include one more condition. In the "One Round Take" condition, subjects play 1 round in which they can take and punish (with the public costs structure), followed by 10 rounds in which the take option is not available. In this case, since subjects cannot take for the following rounds of the interaction, future oriented motives (incapacitation or deterrence) cannot explain any choice of punishment. This is similar to the design employed by Fudenberg and Pathak (2010) who have individuals play multiple rounds of public goods games which include sanctions

In each experimental session, individuals are first put into a group to play one iteration of the No Punishment condition. After a random rematching into new groups, they play either one iteration of Public, one iteration of Private, or 3 iterations of One Round Take.[19] We implement this design for several reasons: it allows individuals to gain experience with the experiment in the first stage, and it allows us to look for correlations between individual behavior in No Punishment and their later behavior when punishment is available.

Our experimental design is different from other experimental designs assessing the role of non-altruistic motives for punishment. We vary the cost structure of punishment, which allows us both to discuss the institutional setup of financing sanctions, and to investigate the private benefits from punishment, using a basic economics framework. Second, the punishment in this game is not fines, as in prior experiments, but exclusion for a certain number of rounds. This is allows us to include an analysis of incapacitation, and therefore contribute to the discussion of different motives of incarceration motives in the economics of crime literature.

The experiment was conducted at the Harvard Decision Science Laboratory using the z-Tree software (Fischbacher (2007)), in June and July 2012.[20] The participants, recruited using the Decision Science Laboratory pool, were university students (mean age: 21.5 years old, 58% female) in the Boston area. We have a total of 91 participants: 39 in Public, 28 in Private and 24

[19]Participants are not informed about the full structure of the experiment, they are only given instructions for their current condition. However, participants are informed when the One Round Take condition is the final game in the experiment.

[20]Appendix 1 presents the experimental instructions

in One Round Take.

Participants were given a 10 dollar show-up fee, and their experimental earnings were converted at a rate of 50 MU per dollar. The experiment took between 40 and 50 minutes to complete. Participants earned between 17 and 23 dollars. They were informed of experimental earnings for each condition independently, and their final earnings were privately announced to them at the end of the experiment.

Our main outcome variable in this series of experiments is the choice of number of rounds of punishment for potential found takers. This is our measure of how much sanction players are willing to support when facing different cost structure.

4.2 Theories of Punishment

There are three major normative theories of punishment in the law and economics literature: incapacitation, general deterrence and specific deterrence. Our experimental setup allow us to discuss what kind of social benchmarks each of these motives sets. We briefly present these motives and how they form benchmarks in our experimental setup. Table 11 in the appendix summarizes predictions for these different motives.

4.2.1 Incapacitation

Incapacitation is the prevention of offending by removal of offenders. Shavell (1987) determines the optimal level of punishment to achieve cost-efficient incapacitation. He finds that incapacitation to be cost-efficient, the cost of incarceration (or, in our setup, of removing a player for N rounds) has to be lower than the expected harm that individual could do while incapacitated.

In our setup, even if we assume that an individual does not respond to deterrence incentives and always chooses to take, the maximal harm that individuals can do is to take 3 MU from one (random) player in each round. In the Public condition, the cost of removing this individual is 5 MU. Thus, from a perspective of maximizing social payoffs (even if we assume that 'bad' (taking) individuals' payoffs do not enter into this calculation), the cost of incapacitation outweighs its benefits.

In the Private condition, since the cost is 2 but the social benefit is 3 there may be pro-social incapacitation motives. However, from an individual payer's perspective, the expected harm per round of a rogue individual is $\frac{3}{n}$;

whereas the cost of removing the player is of $\frac{5}{n}$ per round. Thus there is no private incarceration motive either.[21]

Finally, in the One Round Take treatment, exclusion cannot be chosen for incapacitation motives, since the punishment applies to rounds in which it is impossible for the punished players to take.

4.2.2 Deterrence

General deterrence is the impact of the threat of future punishment on behaviors. In out setup, players cannot increase general deterrence by setting higher punishments. Only players who are found out learn about other players' punishment choices, and even then, only their assigner's choice of penalty rounds. General threats therefore cannot be emitted.

Specific deterrence, however, could be a consideration. In order to investigate this possibility, we consider several possible assumptions on takers' behaviors.

Assumption 1*: Takers are rational criminals.* In this case, average punishment should be ≤ 1 round. By being excluded for 1 round in 50% of cases, potential thieves lose in expectation 2 units,[22] which is exactly what they gain from taking. As long as participants taking are slightly risk averse, 1 round of punishment will be enough to deter them from taking. Excluding player for more than 1 round cannot be for only specific deterrence motives.[23]

Assumption 2 : *Takers can be "taught a lesson" if punishment is higher than a certain threshold.* We present a simple mathematical model of specific deterrence with reform in the appendix. The main results of our model is that though we can rationalize many different average levels of punishment,

[21] One may argue that risk averse players would prefer to pay a cost of $\frac{5}{n}$ for sure rather than lose $\frac{3}{n}$ with some probability, and thus that incapacitation can be seen as a form of private insurance against rogue group members. We note that this critique does not apply in the One Round Take condition.

[22] The math exercise – adding up 2 numbers – is easy: players get it right in 98,7% of cases. Furthermore, no participants systematically make mistakes: only 1 participant makes more than 2 mistakes, over the 40 additions participants are asked to do. We therefore assume that loss from exclusion for 1 round is equal to 4.

[23] One reason why individuals might choose punishments greater than 1 for specific deterrence motives is if they think that other players would punish less, because those players do not care about deterrence as a public good. There is however no reason for the average punishment in the public condition to be higher than average punishment in the private condition for this reason, unless individuals believe that others punish less in the latter compared to the former.

depending on individual beliefs, for fixed beliefs, the amount of punishment should decrease as the game gets closer to the end. This is because the value of reforming individuals decreases, since there are less rounds over which benefits from reform can be reaped, but the cost of punishment stays the same.

Assumption 3 : *Takers always take, and cannot be reformed.* This case reduces to the incapacitation case.

Finally, and regardless of our assumptions about takers' behaviors, in the One Round Take treatment, no positive exclusion can be rationalized by a specific deterrence motives, since taking is only possible in the first round of this treatment condition.

4.2.3 Cold Glow

Contrarily to pro-social motives, cold glow predicts that punishment in Public would be higher than in the Private. Private benefits from cold glow motives will be over consumed when costs are not fully internalized. Additionally, cold glow is the only motivation consistent with any non-zero punishment in the One Round Take condition.

4.3 Experiment 1 Results

This first section compares Public to the Private condition. We present graphs along with body text and regression analysis in the Appendix. We then present additional evidence from One Round Take as a robustness check.

4.3.1 Punishment Decisions

We first look at punisher's decisions. Figure 2 presents the number of rounds of punishment chosen in Public and Private conditions.[24]

The average begins at roughly the same level (approximately 3.5 rounds of exclusion). However, punishment decreases sharply in Private but not the Public conditions after the first 5 rounds. After this short learning period, average punishment settles to 1.3 rounds in Private and stays at 3.5 in Public.

The fact that punishment levels stay the same over rounds in the Public condition is a first indication that specific deterrence cannot be the only

[24] As a reminder: all players who are not currently excluded from the game can choose a punishment.

motivation at play: as participants get closer to the end of the game, the size of imposed punishment does not down. Furthermore, the average levels of punishment chosen in the Public condition far exceed than the levels in line with optimal deterrence or incapacitation.

Robustness check To conclusively rule out deterrence or incapacitation as the only motives for punishment, we also consider the One Round Take condition. Figure 3 shows the average punishment decisions made in rounds 6+ of the Private and Public conditions, and in all iterations of the One Round Take condition. Participants in One Round Take choose an average of 2.5 rounds of exclusion compared to 3.5 rounds in Public and 1.7 in Private. The fact that One Round Take punishments are positive, and higher than in the private condition shows that cold glow, as a private benefit to punishment, is a major motivating force of punishment decisions.

Table 2 presents regression results that confirm the intuitions presented in the graphs. We regress amount of punishment chosen on a dummy taking values 0 for Private and 1 for Public. Standard errors are clustered by participant.

Column 1 presents results for the full sample; column 3 presents decisions made from rounds 6 to 20. Participants in the private treatment choose smaller levels of punishment than in the public treatment. This holds when we control for round effects (column 2).

Column 1 (2) of table 3 shows the difference in number of rounds of exclusion chosen in One Round Take and Private (Public). In this specification, the One Round Take condition is significantly higher than Private and lower than Public. In column 3, we pool the data to tease apart the relative importance of public motives (deterrence and incapacitation) and cost structures in choices of punishment. We regress punishment choices on a dummy for costs being public (Public and One Round Take conditions) vs. Private; and a dummy for public good (deterrence or incapacitation) motives (Public and Private conditions) vs. One Round Take condition. The coefficients on these dummies represent the effects of cold glow vs. public goods motives in punishment decisions. The first dummy is significantly positive: people choose more rounds of exclusion when the costs are public. The second dummy is negative, smaller in magnitude but not significant implying that non-cold glow motives play a weak role in punishment behavior in our experiment.[25]

[25]Another possible explanation for the difference in behavior between One Round Take

Taken together, our regression analyses confirm that cold glow is a major motivation in punishment decisions. Other motives also exist, but cannot explain most of the variation in punishment. We now turn to see the effects of conditions on taking decisions.

4.3.2 Taking Decisions

Figure 1 shows taking decisions by availability of punishment, and table 1 presents our regression results. Taking behavior is significantly higher in Punishment and No Punishment conditions (column 1), which shows that general deterrence does matter: only 10% to 20% of participants who are able to take[26] choose to do so, even from round 1. However, there is no difference between the Public and Private conditions (column 3), and we find no session effects (column 2).

We find a slight learning effect in the No Punish condition. Approximately 70% of individuals take in the first round and by the 5^{th} round, 85% of participants choose to take. There is no significant difference between experimental sessions.

4.3.3 Individual Differences

So far, we have compared results across treatments. We now turn to individual variations within treatments. First, we ask what causes the learning effect that we find in the Private condition. We find that punishment levels decrease after a player's choice is implemented[27] in Private condition, but not in the Public condition. Table 4 shows the regression of punishment decisions on a dummy which takes value 1 in each round after an individual's punishment choice is implemented. On average(column 1), it appears that having paid for punishment does not influence choice of sentences (column 1). However, the effects are heterogeneous across treatment conditions (columns 2-4): in Private, subjects punish significantly less once their punishment has been chosen. We interpret this as a form of 'sticker shock.'

We also attempt to see whether behavior in No Punishment conditions predicts punishment behavior in later stages. We find an effect for individuals

and Public is that perhaps it is easier to ex-post rationalize punishment decisions in the former than in the latter.

[26] i.e. players who are not currently excluded from the game

[27] In other words, when she is randomly chosen to be a found individual's punisher.

who take less than 15 times in the original No Punishment rounds, whom we refer to as "low takers". They also give much smaller punishments on average (Column 5 of table 2).[28] This result would be interesting to investigate in future experiments, as it suggests negative correlation in warm and cold glow.

5 Experiment 2: Responses to Probability of Apprehension

Our second experiment asks whether punishers' decisions react when probability of apprehension (and thus optimally deterring punishments) change. We see how potential norm-breakers in turn react to punishers' behaviors. If punishers don't react to these changes, this can lead to an outcome where socially wasteful low levels of punishment can occur. In addition, we compare ex-ante and ex-post punishment decisions.

5.1 Experimental Setup

We use a game to test both how sentences are chosen, and how potential norm-breakers respond to expected punishments.[29] The basic setup is as follows: players are matched into groups of three to play a one shot game. They begin with a balance of 80 points.

Players are randomly assigned one of three roles: assigner, taker, or target. All rules of the game are known to all players before they begin the experiment. The game proceeds as follows: the assigner commits to a publicly known level of penalty units (between 0 and 10), each of these units corresponds to a 10 point sanction. *Knowing this level of sanction*, the taker decides to take or not from the target. If the taker choose to take, they gain 20 points, and the target loses 30 points. The taker is found out with probability p. If the taker is found out, they are imposed the sanction chosen by the assigner. The assigner is charged 1 point per 5 points of sanction they assign.

Our treatments vary in the probability that the taker will be found if he takes: in the "high probability" treatment, the taker is found with a

[28] Our results are robust to changes in the definition of "low takers". We chose this specification, as it took about 5 rounds for taking behavior to plateau at 90% in the No Punishment condition.

[29] Experimental instructions are presented in the appendix

probability 9/10; in the "low probability" treatment, with a probability 1/3.[30] All players are informed of all rules at the beginning of the game. Final payoffs depend on choices made by all of the players. Finally, the targets make no choice in our game, but we ask them to enter what they think would be a "fair" punishment for a taker who chooses to take.

We used the online labor market Amazon's Mechanical Turk (AMT) to recruit individuals to play the game for a show-up fee of .3 USD and an additional payment depending on points earned, using a conversion rate of 2 points per .01 USD at the end of the experiment.[31]

We recruited a total of 340 individuals (mean age: 28.8, 63% male) to play this game. Each individual played exactly one role in the interaction. To make sure that all participants understood the experiment they were first given a set of instructions followed by a three question comprehension quiz (see Appendix). If they failed to answer any of the quiz questions correctly, they were not allowed to play the game. Thus all of our results are from participants who answered all comprehension questions correctly. Dropping non-comprehenders, we are left with 243 individuals (a 71 % pass rate).

5.2 Experiment 2: Results

5.2.1 Punisher Behavior

We now consider the behavior of punishers across conditions. Part a of figure 4 presents assigners' average punishment levels for each of the probability conditions. Mean punishment levels are exactly the same in both treatments: probability of apprehension is not a parameter individuals respond to in

[30] Some studies in psychology have investigated the effects of probability of apprehension on punishment decisions. These studies directly ask participants to compare hypothetical punishments in different scenarios when probabilities of apprehension change (Baron and Ritov (2009)), or asked participants to assess the relative importance of deterrence or moral motives on punishment decisions (Carlsmith et al. (2002)). In these hypothetical contexts, players state that do not want to change behaviors based on probabilities of apprehension. Our experiment adds to this literature as a very strong test of whether punishers respond to probability and deterrence motives. In our games rules are perfectly transparent and deterring punishments are very easy to calculate.

[31] Several recent studies have been undertaken to examine the validity of experimental data collected using AMT at stakes of ~ 1 USD. They find that behavior on AMT matches well with standard laboratory results on economics games (Amir et al. (2012)) (Rand et al. (in Press)), and are based on samples that are more representative of the general population (Horton et al. (2011), Paolacci et al. (2010)).

punishment choices. The mean punishment level is 4.0 in the high probability condition and 4.1 in the low probability, and the difference non-significant (see table 5).

5.2.2 Decisions to Take

We find that takers' behaviors, however, *do* respond to probability of apprehension on the intensive margin. We use the strategy method to elicit choices of taking: takers are asked to enter their *maximum acceptable possible penalty* (MAPP). This is a number of penalty units such that if the assigner chooses a penalty below or equal to this level, the taker prefers to take. If the assigner chooses a larger penalty, the taker would prefer not to take. We perform analyses on choices of MAPP to understand takers' behaviors.

We first find that a relatively large amount of participants (approximately 30 %) who choose a MAPP of 0, indicating that they do not wish to take under any circumstances, in both conditions. Table 6 shows our regression results confirming there is no significant extensive margin response. However, focusing on the 70% of individuals who entered a MAPP > 0, we find that there is an effect on the intensive margin: as shown in part b of figure 4, individuals who choose to take at all choose different levels of MAPP between probability conditions (mean MAPP in low = 5.1, and mean MAPP in high = 3.8). Table 6 shows our regression results, confirming there is a significant intensive margin response.[32] Unlike punishers, takers respond to the probability of being caught,[33] and so the punishment levels chosen are too low to deter a lot of taker in the low probability condition.

5.3 Control Study: Ex-Post Punishments

A key part of our theory is that we allow for both an ex-ante (simulating a strategic motive such as deterrence) and an ex-post (or 'just desserts') component. To assess the size of these components, we ran a control experiment on AMT (n=194, age=28.9, 63 % male). The setup of the game in our con-

[32]We also find a gender effect. Women are less likely to take, and if they are willing to take, they enter lower maximum acceptable punishment levels. We note that this can be explained by higher risk aversion (Eckel and Grossman (2008)).

[33]This also allows us to control away a lack of attention or understanding by participants as the result of the null effect on punishment decisions as individuals are randomly assigned into roles.

trol study is identical, except that the order of moves is switched: takers first choose to take or not, and then assigners choose ex-post penalties to assign to takers who are caught. We use the same probability conditions in this study. This has the added benefit of acting as a robustness check on taker behavior from our original study where one possible confound is that takers could have found the strategy method confusing.

Figure 6 shows the results. We find that punishers again do not respond to probability of apprehension when choosing levels of ex-post punishment (mean punishment in low = 3.4, mean punishment in high = 3.2). Takers, however, do take probability into account: 25 % of individuals take in high probability condition and 43 % take in the low probability condition[34].

5.3.1 Comparisons

We now pool our data and compare the ex-ante and ex-post punishment conditions using regressions. Table 7 presents full sample results: as in the main sample, we find that neither choice to punish nor punishment level respond to probability of apprehension.

Furthermore, in the control condition, assigners still choose a positive level of punishment, even though this is a one-time interaction and punishments are privately costly. However, we confirm that levels of punishment are smaller when no deterrence motive is possible than when the assigner plays first: this indicates that some difference (approximately 20 percent) between ex-ante and ex-post punishments does seem to exist, however these differences are not significant. These results are consistent with the differences found in our first experiment between the One Round Take condition and the Public condition. We conclude that some form of deterrence motives *do* exist in the punishment choices, but ex-post 'just desserts' thinking seems to be the dominant motivator of punishment behavior in our samples.

5.4 Fairness Judgments

Finally, we look at judgments of 'fair punishments' for caught takers from the point of view of the target. Their answers do not appear to differ across

[34]This difference is significant, though only at the 10% level, due to sample size. The magnitude stays the same – 20 percentage points difference – and becomes significant at the 5% level when we control for gender

conditions (mean fair punishment in low, ex-ante = 4.3, high, ex-ante = 5, low, ex-post = 5.3, high, ex-post = 5.5).

Table 8 presents our regression analysis. Unsurprisingly, targets want higher punishments than assigners: this could be driven either by differences between second-party and third-party punishment (Fehr and Fischbacher (2004)), or because targets do not have to pay for chosen punishments. Interestingly, neither order of punishment assignment nor probability of being caught changes targets' beliefs about fairness: no extra retribution is demanded when probability of apprehension is lower: differences are not significant, and if anything the point estimates go in the wrong direction. All data taken together, neither punishers nor victims respond to probability of apprehension when choosing punishment levels, although this parameter seems to matter a lot in the decisions of potential norm-breakers.

6 Experiment 3: Crowding Out

Our final experiment asks an individual level question motivated by our theory: to what extent is punishment by one individual crowded out by known punishment choices of another individual? Our social level question asks whether a lack of crowding out can push aggregate punishment levels above particular benchmarks.

6.1 Main Experiment

In order to answer this question, we ran an experiment on AMT using a sample 476 individuals (mean age = 29.7, 56% male). Participants received a show-up fee of .5 USD and an additional payment depending on their earnings during the game, using a conversion rate of 1 points per .01 USD.[35]

We use a game similar to experiment 2 to explore crowding out behavior. Players are randomly assigned to groups of four and start the game with 100 points. Each individual is assigned one role: assigner 1, taker, target, or assigner 2.[36] All rules of the game are known to all players before they begin the experiment. Players act sequentially as follows: assigner 1 commits to a

[35]Given the average completion time of our experiment and average bonuses, total payoffs amounted to an hourly wage of approximately $\$8 - 10$ per hour.

[36]In experimental instructions taker and target are referred to as player 1 and player 2 respectively.

publicly known level of penalty units $(0-6)$, each penalty unit corresponds to a 10 point sanction. *Knowing this level of penalty*, the taker decides to take or not from the target. If the taker choose to take, they gain 30 points, and the target loses 40 points. The taker is found out in 3/4 cases. If the taker is found out, assigner 2 sees the punishment that assigner 1 chose, and is given a choice to assign an additional number of penalty units (up to 6). A found out taker is imposed the sum of the penalty units chosen by the assigner 1 and assigner 2 and both assigners are charged 1 point per 10 points of sanction they assign.

Again, although the target makes no choice in our game, we ask them to enter what they think would be a "fair" punishment for a taker who chooses to take. As in experiment 2, individuals see the instructions for the experiment and then take a quiz about the rules. Individuals who do not answer quiz questions correctly are not allowed to participate in the experiment. Overall, approximately 70% of participants answered the quiz questions correctly leaving us with 73 groups of four players.

Our main variable of interest is assigner 2's choice in level of punishment. As in the previous experiment, we use the strategy method to elicit this preference. Figure 7 presents the average punishment choice of assigner 2, for each possible assigner 1 choices. On average, there is no difference across assigner 1's choices, and thus no evidence of crowd-out behavior on aggregate.

We do find considerable heterogeneity in individual behavior. Because we use the strategy method we can look for different behavioral types in our population. Overall, we find that approximately 80% of assigner 2's can be classified into one of three types: individuals whose sanction choices decrease in assigner 1's choice (partial crowd-out types 35%), individuals whose sanction choices increases in assigner 1's choice (crowd-in types[37] 25%) and individuals whose sanctions do not change as a function of assigner 1's choice (constant types, 20%). Individual heterogeneity is not the main focus of this discussion, so we leave as an avenue for future work. However, we can use this analysis as a robustness check. If we restrict our analysis to the crowd-out types, we still see an imperfect crowding out of own punishment by the punishment of another and we can statistically reject the hypothesis of perfect crowding out even in this restricted subsample (table 9).

We can also look at the average behavior of the first assigner in this

[37]These individuals may be using assigner 1's decision as a signal of the inappropriateness of taking.

experiment and what the target deems to be a fair punishment. We find that the mean punishment assigned by the first assigner is 3.02 units (30 points). Combining this with the conditional punishments of assigner 2, we find that the average total punishment on a taking player is approximately 5 units of punishment, or 50 points. We note that this is 25% higher than the mean 'fair punishment' as viewed by the targets (mean fair punishment = 42 points).

6.2 Control Experiment

Experiment 3 uses a strategy method and a within subject design to look for the extent of crowd-out in punishment. We ran a second study as a robustness check using a between-subject design without the strategy method. We used AMT to recruit subjects, again dropping those who failed a comprehension quiz. We were left with 243 participants (mean age = 29, 57 % male) between two conditions.

In our control experiment, players are put into groups of three and assigned a role: taker, target or assigner. All rules of the game are known to all players before they begin the experiment. The game proceeds as follows: the taker decides to take or not from the target. If the taker chose to take, they gain 30 points, and the target loses 40 points. The taker is found out in 3/4 cases. If the taker is found out, they automatically lose c points, where c is varied to be 0 or 40 by condition. If the taker is found out, the assigner can assign up to 6 penalty units, each of which amounts to a 10 point sanction. The assigner is charged 2 point for every 1 penalty unity.

This control lets us look at crowd-out effects when punishment is assigned by an outside figure instead of another player in the game. Figure 8 shows the average chosen levels of punishments in the two conditions. Assigner punishment levels chosen are slightly lower when $c = 40$ than when $c = 0$, but this difference is not statistically significant, and it is in any case much smaller than a one-for-one crowding out: punishments are of on average 2 units in the $c = 0$ condition, and 1.7 in the $c = 40$ condition. Thus realized sanction are approximately 20 points in the $c = 0$ condition and 57 points in the $c = 4$ condition.

In the interest of space, we skip discussion of taker behavior and fairness evaluations by the target, as they only replicate the qualitative results of experiments 1 and 2.

This last set of experiments therefore indicates that punishment is not

crowded out one for one by pre-set levels of sanctions. On average, there is no effect of pre-set sanctions on average punishment. We note that there is considerable heterogeneity in this behavior, but never observe perfect crowding out.

7 Conclusion

Though many legal scholars and philosophers think of moral reasoning as driven by rational, calculating processes, the nascent field of moral psychology suggests that moral behaviors, including the punishment of those who break social norms, are mostly driven by emotional reactions which are then rationalized by conscious processing (Greene and Haidt (2002), Haidt (2001)). Using such a blunt psychological mechanism motivated by affective factors and not rational reasoning to make punishment decisions may sometimes collaterally result in social harmony, but in other domains can result in either highly inefficient over punishing or inefficient under punishing. We have presented a simple theory based on this observation which predicts that punishment decisions will be driven by personal cost, and not public cost, will not respond to probability of apprehension, as optimal deterrence might and may not necessarily crowd-out one-for-one as punishment might in a theory of 'just desserts.' We confirm these predictions in our experiments and find little evidence that standard rational motives (deterrence, incapacitation) are major drivers of individual punishment decisions.

We argue that understanding the role more emotional or automatic mechanisms at play in choosing levels of punishments could be important in our understanding of many types of social behaviors including aggregate outcomes in the criminal justice system. We have presented several possible channels through which we believe our theory of behavior can affect these aggregate outcomes. More empirical research is needed in understanding to what extent cold glow motives drive the behaviors of voters, judges and juries, as well as everyday punishment behaviors in social groups.

Simultaneously with field data, further lab experiments could be used to investigate the mechanisms at play in choosing levels of sanctions. In particular, does feedback on deterrence appear to have effects on choices of levels of punishment? Does drawing people's attention to the cost of sanctions modify their choices? Does professional training change the methods of decision-making employed by individuals?

Behavioral and social scientists have increasingly gone beyond studying how aggregate outcomes come about, and have taken a plunge into the practice of using their skills to to help design "rules of the game" that achieve normatively desired outcomes.[38] We note that, especially in the case of punishment institutions, it seems that effective rules of the game will depend on the psychological motivations of the players. This is particularly stark if we consider the difference in assumptions that individuals punish for public goods motives (theories of deterrence, incapacitation) or for private benefits (cold glow). In the former case, punishment will be under provided due to free-riding motivations and so mechanisms which subsidize the costs of punishment decisions will improve overall efficiency. However, if individuals are motivated by cold glow, the same subsidies may lead to highly inefficient outcomes. Economics as "rule design" is a growing and important part of modern social science and we hope that our results contribute to this important conversation.

[38]For a survey of recent work in the field of market design see Roth (2003).

References

Amir, O., D.G. Rand, and Y.K. Gal, "Economic Games on the Internet: The Effect of $1 Stakes," *PloS one*, 2012, *7* (2), e31461.

Anderson, C.M. and L. Putterman, "Do non-strategic sanctions obey the law of demand? The demand for punishment in the voluntary contribution mechanism," *Games and Economic Behavior*, 2006, *54* (1), 1–24.

Andreoni, J., "Impure altruism and donations to public goods: a theory of warm-glow giving," *The Economic Journal*, 1990, *100* (401), 464–477.

_ , "An experimental test of the public-goods crowding-out hypothesis," *The American Economic Review*, 1993, pp. 1317–1327.

Ater, I., Y. Givati, and O. Rigbi, "Organizational Structure, Police Activity and Crime: Evidence from an Organizational Reform in Jails," *Working Paper*, 2012.

Axelrod, R., "An evolutionary approach to norms," *The American Political Science Review*, 1986, pp. 1095–1111.

Bardsley, N. and R. Sausgruber, "Conformity and reciprocity in public good provision," *Journal of Economic Psychology*, 2005, *26* (5), 664–681.

Baron, J. and I. Ritov, "Intuitions about penalties and compensation in the context of tort law," *Journal of Risk and Uncertainty*, 1993, *7* (1), 17–33.

_ **and** _ , "The role of probability of detection in judgments of punishment," *Journal of Legal Analysis*, 2009, *1* (2), 553–590.

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

Berdejo, C. and N. Yuchtman, "Crime, Punishment and Politics: An Analysis of Political Cycles in Criminal Sentencing," *Unpublished manuscript, Harvard University*, 2009.

Bushway, S., M.A. Stoll, and D.F. Weiman, *Barriers to Reentry?: The Labor Market for Released Prisoners in Post-industrial America*, Russell Sage Foundation Publications, 2007.

Camerer, C.F., T.H. Ho, and J.K. Chong, "A cognitive hierarchy model of games," *The Quarterly Journal of Economics*, 2004, *119* (3), 861–898.

Carlsmith, K.M., J.M. Darley, and P.H. Robinson, "Why do we punish?: Deterrence and just deserts as motives for punishment.," *Journal of Personality and Social Psychology; Journal of Personality and Social Psychology*, 2002, *83* (2), 284.

Coffman, L.C., "Intermediation reduces punishment (and reward)," *American Economic Journal: Microeconomics*, 2011, *3* (4), 77–106.

Cornes, R. and T. Sandler, "The comparative static properties of the impure public good model," *Journal of Public Economics*, 1994, *54* (3), 403–421.

Costa-Gomes, M., V.P. Crawford, and B. Broseta, "Cognition and Behavior in Normal-Form Games: An Experimental Study," *Econometrica*, 2003, *69* (5), 1193–1235.

Cushman, F., A. Dreber, Y. Wang, and J. Costa, "Accidental outcomes guide punishment in a ?trembling hand? game," *PloS one*, 2009, *4* (8), e6699.

Danziger, S., J. Levav, and L. Avnaim-Pesso, "Extraneous factors in judicial decisions," *Proceedings of the National Academy of Sciences*, 2011, *108* (17), 6889–6892.

Eckel, C.C. and P.J. Grossman, "Men, women and risk aversion: Experimental evidence," *Handbook of experimental economics results*, 2008, *1*, 1061–1073.

Fehr, E. and K.M. Schmidt, "A theory of fairness, competition, and cooperation," *The Quarterly Journal of Economics*, 1999, *114* (3), 817–868.

_ **and U. Fischbacher**, "Third-party punishment and social norms," *Evolution and human behavior*, 2004, *25* (2), 63–87.

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

Fudenberg, D. and P.A. Pathak, "Unobserved punishment supports cooperation," *Journal of Public Economics*, 2010, *94* (1-2), 78–86.

Garland, D., *The culture of control: Crime and social order in contemporary society*, Oxford University Press US, 2001.

Gintis, H., S. Bowles, R.T. Boyd, and E. Fehr, *Moral sentiments and material interests: The foundations of cooperation in economic life*, Vol. 6, MIT press, 2005.

Glazer, A. and K.A. Konrad, "A signaling explanation for charity," *The American Economic Review*, 1996, *86* (4), 1019–1028.

Greene, J. and J. Haidt, "How (and where) does moral judgment work?," *Trends in cognitive sciences*, 2002, *6* (12), 517–523.

Haidt, J., "The emotional dog and its rational tail: a social intuitionist approach to moral judgment.," *Psychological Review; Psychological Review*, 2001, *108* (4), 814.

Horton, J.J., D.G. Rand, and R.J. Zeckhauser, "The online laboratory: Conducting experiments in a real labor market," *Experimental Economics*, 2011, *14* (3), 399–425.

Kahneman, D., P.P. Wakker, and R. Sarin, "Back to Bentham? Explorations of experienced utility," *The Quarterly Journal of Economics*, 1997, *112* (2), 375–406.

Ostrom, E., J. Walker, and R. Gardner, "Covenants with and without a sword: Self-governance is possible," *The American Political Science Review*, 1992, pp. 404–417.

Pager, D., *Marked: Race, crime, and finding work in an era of mass incarceration*, University of Chicago Press, 2007.

Paolacci, G., J. Chandler, and P.G. Ipeirotis, "Running experiments on amazon mechanical turk," *Judgment and Decision Making*, 2010, *5* (5), 411–419.

Peysakhovich, A. and D. Rand, "Building cooperative norms," *Working paper*, 2012.

Polinsky, A.M. and S. Shavell, "The fairness of sanctions: some implications for optimal enforcement policy," *American Law and Economics Review*, 2000, *2* (2), 223–237.

Posner, R.A., *How judges think*, Harvard University Press, 2008.

Quervain, D.J.F. De, U. Fischbacher, V. Treyer, M. Schellhammer, U. Schnyder, A. Buck, and E. Fehr, "The neural basis of altruistic punishment.," *Science; Science*, 2004.

Rabin, M., "Incorporating fairness into game theory and economics," *The American Economic Review*, 1993, pp. 1281–1302.

Rand, DG, J.D. Greene, and M.A Nowak, "Spontaneous giving and calculated greed," *Nature*, in Press.

Rasmusen, E.B., "Stigma and self-fulfilling expectations of criminality," *Journal of Law and Economics*, 1996, *39*, 519–544.

Roth, A.E., "The economist as engineer: Game theory, experimentation, and computation as tools for design economics," *Econometrica*, 2003, *70* (4), 1341–1378.

Shavell, S., "A model of optimal incapacitation," *The American Economic Review*, 1987, *77* (2), 107–110.

Singer, T., B. Seymour, J.P. O'Doherty, K.E. Stephan, R.J. Dolan, and C.D. Frith, "Empathic neural responses are modulated by the perceived fairness of others," *Nature*, 2006, *439* (7075), 466–469.

Sunstein, C.R., D. Schkade, and D. Kahneman, "Do people want optimal deterrence," *J. Legal Stud.*, 2000, *29*, 237.

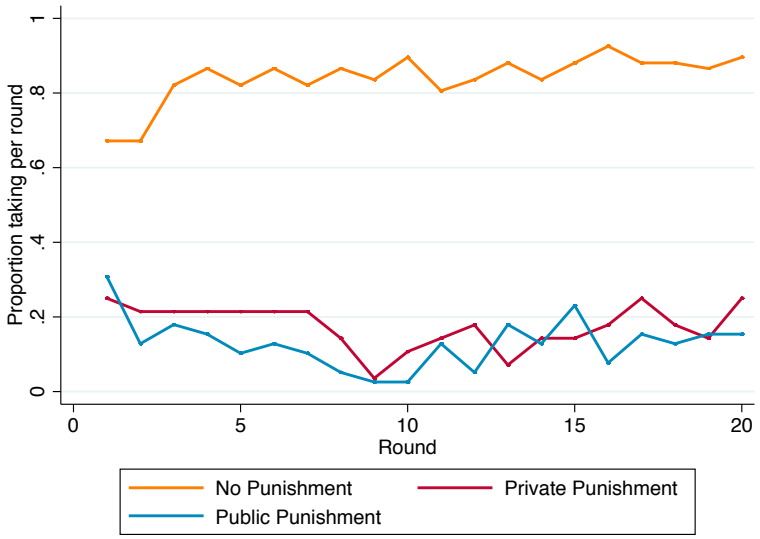

Figure 1: Percent choosing take

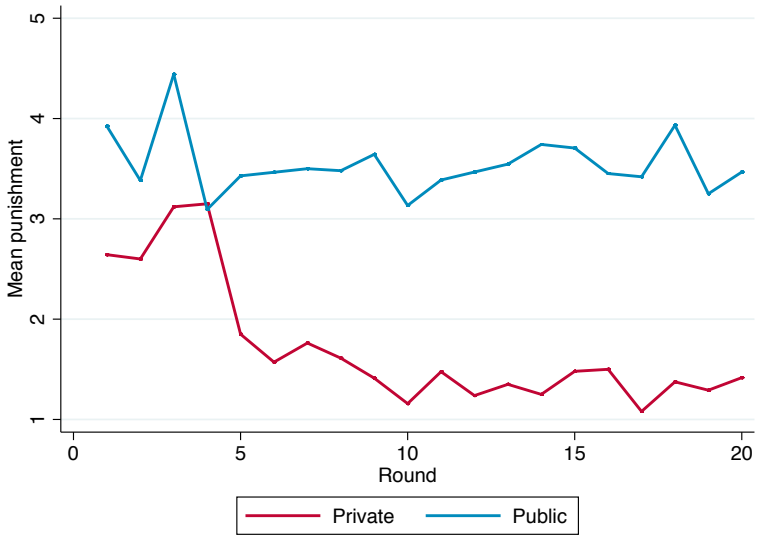

Figure 2: Mean punishment level chosen by round

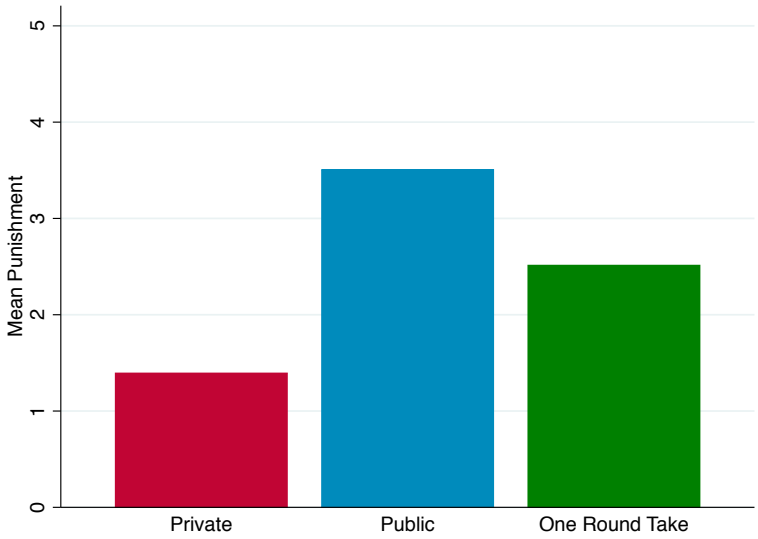

Figure 3: Mean punishment level chosen, round ≥ 5

Figure 4: Experiment 2: punishers' and takers' behaviors

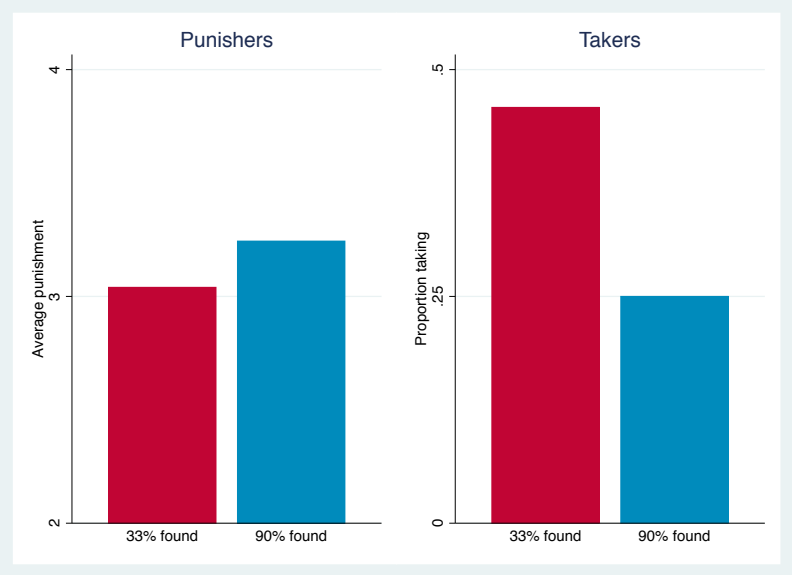

Figure 5: Study 2 Control Decisions

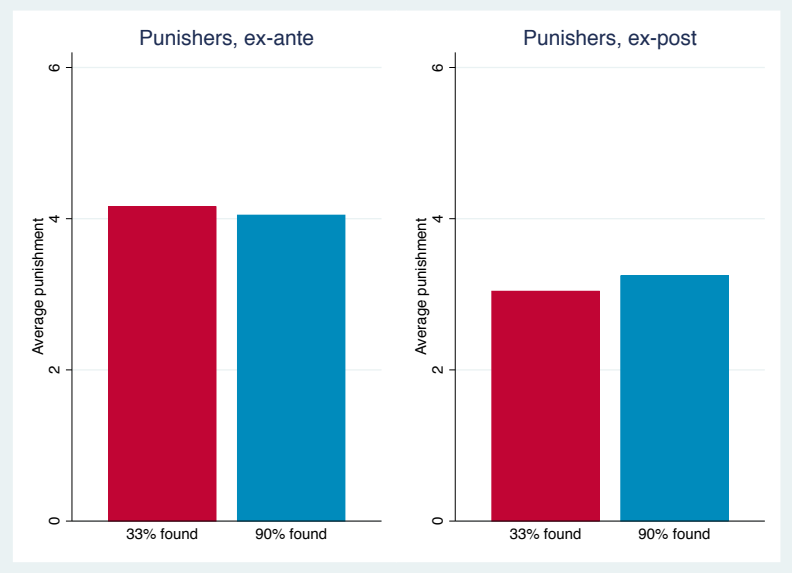

Figure 6: Ex Ante vs. Ex Post punishment decisions

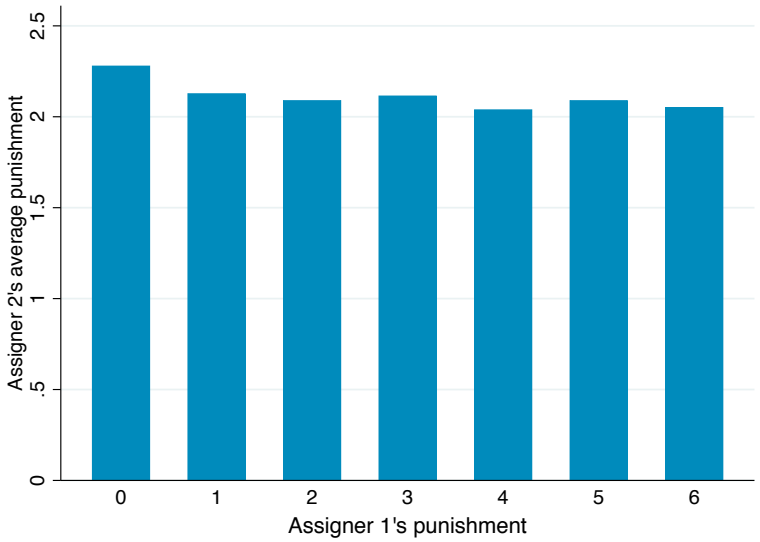

Figure 7: Experiment 3: Assigner 2's behavior

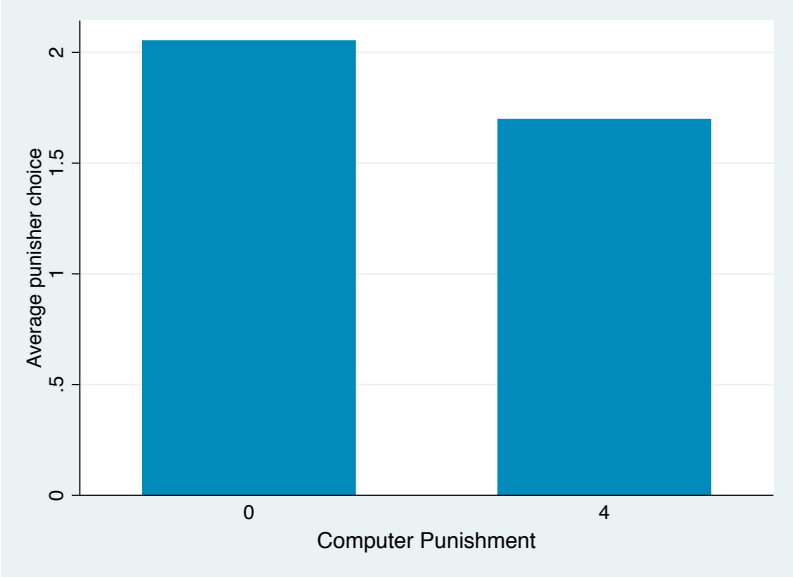

Figure 8: Study 2 Control Decisions

Table 1: Experiment 1 - Taking behavior, by condition

	(1) No vs. With Sanctions	(2) Punishment Cost Structure
1=With Sanctions	-0.655**	
	(0.0371)	
Public		-0.0556
		(0.0728)
Constant	0.841**	0.219**
	(0.0256)	(0.0625)
Observations	2407	1067

Results clustered at the subject level

$^{+}$ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$

Table 2: Experiment 1 - Length of punishment, by treatment, across rounds

	(1) All	(2) All	(3) Rounds 6-20	(4) Rounds 1-5
Public	1.818*	1.809*	2.113**	0.990
	(0.754)	(0.754)	(0.771)	(0.840)
Round		-0.0406*		
		(0.0186)		
Constant	1.734**	2.166**	1.394**	2.686**
	(0.455)	(0.530)	(0.457)	(0.606)
Observations	1067	1067	782	285

Results clustered at the subject level

$^{+}$ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$

Table 3: Experiment 1 - Length of punishment: public motives vs. cost structure

	(1) Public vs. Private	(2) Deterrence vs. None	(3) Both Effects
Public Costs	0.780*		1.818*
	(0.357)		(0.752)
No Deterrence		-1.039*	-1.039
		(0.459)	(0.767)
Constant	1.734**	3.553**	1.734**
	(0.133)	(0.148)	(0.454)
Observations	520	691	1139

Results clustered at the subject level

$^{+}$ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$

Table 4: Experiment 1 - Length of punishment: individual differences

	(1)	(2)	(3)	(4)	(5)
Public	1.887*			1.777*	2.317**
	(0.778)			(0.750)	(0.823)
Punishment Chosen	0.579	-1.214+	1.936*		
	(0.640)	(0.686)	(0.936)		
Stolen From				-0.369	
				(0.287)	
Low Taker					-2.305**
					(0.806)
Constant	1.437*	2.358**	2.789**	1.896**	1.992**
	(0.640)	(0.733)	(0.587)	(0.515)	(0.486)
Observations	1067	448	619	1067	1067

Results clustered at the subject level

Low Taker: took less than 15 times in the no punishment condition

$^{+}$ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$

Table 5: Experiment 2: Choice of punishment type, by treatment

	(1) Punish	(2) Level, Full sample	(3) Level, if Punish = 1
1 = High	-0.131	-0.154	0.584
	(0.0799)	(0.744)	(0.741)
1 = Female	-0.0296	-0.788	-0.749
	(0.0817)	(0.760)	(0.753)
Age	-0.0000964	0.0140	0.0156
	(0.00439)	(0.0409)	(0.0397)
Constant	0.938**	4.121**	4.396**
	(0.139)	(1.291)	(1.277)
Observations	81	81	69

Standard errors in parentheses

High: found with a 90% chance; Low: found with a 33% chance.

Punish=1 if assigner entered a positive level of punishment.

Level = amount of punishment chosen

$^{+}$ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$

Table 6: Experiment 2: MAPP, by treatment

	(1) Take	(2) Level, Full Sample	(3) Level, if Take = 1
1 = High	0.109	-0.593	-1.799*
	(0.0991)	(0.722)	(0.813)
1 = Female	-0.211+	-1.981*	-2.109*
	(0.107)	(0.780)	(0.921)
Age	-0.00490	-0.0428	-0.0645
	(0.00573)	(0.0417)	(0.0506)
Constant	0.862**	5.322**	7.775**
	(0.179)	(1.307)	(1.616)
Observations	82	82	58

Standard errors in parentheses

High: found with a 90% chance; Low: found with a 33% chance.

MAPP = Maximum Acceptable Possible Penalties

Take=1 if taker entered a positive level of acceptable punishment.

Level = amount of acceptable punishment

$^{+}$ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$

Table 7: Experiment 2: Punishment choice, all data pooled (control)

	(1) Extensive: Punish	(2) Intensive: Level
1 = High	-0.0728	-0.0692
	(0.0624)	(0.527)
1 = Assigner First	0.0290	0.939^{+}
	(0.0616)	(0.520)
1 = Female	0.0809	-0.256
	(0.0637)	(0.538)
Age	-0.00542^{+}	-0.0315
	(0.00316)	(0.0267)
Constant	0.986**	4.251**
	(0.112)	(0.944)
Observations	147	147

Standard errors in parentheses

High: found with a 90% chance; Low: found with a 33% chance

$^{+}$ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$

Table 8: Experiment 2: Target's opinion on fair punishment level, by treatment

	(1) With Deterrence	(2) No Deterrence	(3) Comparing Conditions
1 = High	0.620	0.180	0.438
	(0.715)	(0.846)	(0.539)
1 = Female	-0.186	-0.154	-0.193
	(0.817)	(0.866)	(0.582)
Age	-0.0899*	-0.0536	-0.0736**
	(0.0396)	(0.0416)	(0.0282)
1 = Assigner First			-0.758
			(0.535)
Constant	6.972**	6.947**	7.374**
	(1.190)	(1.387)	(0.954)
Observations	80	64	144

Standard errors in parentheses

$^{+}$ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$

Table 9: Experiment 3: 2nd punisher's choice, by 1st punisher choice

	(1) Full sample	(2) Crowding Out
Player 1 sanction	-0.0289	-0.569**
	(0.0620)	(0.0585)
Constant	2.199**	3.380**
	(0.237)	(0.363)
Observations	553	196

Standard errors in parentheses

$^{+}$ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$

A Summary of Experiments

Table 10: Summary of experiments

Experiment	Conditions	Hypotheses tested
Experiment 1	Private costs of punishment	Role of private costs
Cost Structures	Public costs of punishment	vs. public costs
	One Round Take	No social motives
Experiment 2	Ex-ante vs. ex-post	Effects of probability of capture
Probabilities	$p = .33$ vs. $p = .9$	Ex-ante vs. ex-post behavior
Experiment 3	2 assigners	Crowding out behavior
Multiple punishers	Computer + assigner	

B Summary Predictions, Experiment 1

C A Mathematical Model of Specific Deterrence

The theory of specific deterrence which we will model here is as follows: individuals start with a propensity to choose take in every round, each individual has a type $\theta \in [0, \theta^{max}]$ and a threshold level of punishment that depends on his type. The probability distribution over types is given by $p \in \Delta([0, \theta^{\max}])$ and is smooth and well behaved with density f that has strictly negative first derivative (that is, higher types are rarer).

If an individual of type θ receives a punishment of size at least θ, he 'learns his lesson' and never takes again. If he receives a punishment of size less than θ he continues to take in all rounds after.

To formalize our theory we consider a group of N honest individuals with one individual i who has been found out for taking and follows the behavioral rule outlined above, there are k rounds left in the game. We consider a benevolent social planner who does not know the type θ of the taking individual. The social planner wants to maximize the monetary rewards that will accrue to honest individuals. We now ask, given such assumptions, what can

Table 11: Experiment 1: Behaviors Predicted by Various Punishment Theories

Punishment Motive	Result
Incapacitation	Public Punishment = 0 One Round Take = 0
General Deterrence	Public Punishment = 0 Private Punishment = 0 One Round Take =0
Specific Deterrence	Punishment decreases in rounds One Round Take =0
Cold Glow	Public Punishment > Private Punishment One Round Take > 0

we say about the optimal punishment strategy? For simplicity, we suppose that punishments can be delivered in continuous amounts $c \in [0, \infty)$ and has a social cost of v per unit to make the math easier.

Proposition 3. *There exists a unique optimal punishment level c^* which is given by the first order condition:*

$$3f(c^*)k = v.$$

c^* is decreasing in both number of rounds left and public cost of punishment.

The intuition for the first-order condition is as follows: by marginally increasing c the social planner increases the probability that the individual in question learns their lesson from the punishment. The marginal benefit of this is exactly 3 units times the number of rounds left. The marginal cost is exactly v. When there are less rounds left, the marginal benefit is lower so optimal punishments are lower. The exact solutions, however, depend on assumptions about the distribution of types.

D Proofs of Propositions

Proof of Proposition 1. The utility function P maximizes is

$$-\beta s_P + \lambda(s_P, s_T)$$

the first order conditions of the maximization are simply

$$\beta = \frac{\lambda(s *_b (\beta, s_T), s_T)}{\partial s_P}$$

which by assumption are unique (λ is concave in s_P) and give us the comparative statics directly. □

Proof of Proposition 2. Recall that we can write P's maximization problem as:

$$\psi(s_P, q)[q(\lambda(s_P, T) - \beta s_P) + (1-q)(\lambda(0, T))] + (1 - \psi(s_P, q))\lambda(0, 0).$$

We can set $\lambda(0, 0)$ to be 0 and drop the dependence of λ on the second argument to save notation. Our maximization becomes

$$\psi(s_P, q)[q(\lambda(s_P) - \beta s_P) + (1-q)(\lambda(0))].$$

Note that if we take the derivative we get

$$\psi'(s_P, q)[q(\lambda(s_P) - \beta s_P) - (1-q)(\lambda(0)] + \psi(s_P, q)[q(\frac{\partial \lambda}{\partial s_P} - \beta)].$$

The first term is positive because ψ' is negative and the quantity in parentheses which it multiplies is negative from the assumption that $\delta(s_P) > 0$.

Now, consider the ex-post problem with the same λ. The answer to this problem is given $\overline{s}$ that sets

$$\beta = \frac{\partial \lambda}{\partial s_P}$$

this means that for $s < \overline{s}$ we have that the second term must be also positive and hence the overall utility only increases for $s \in [0, \overline{s}]$ so any maximizer of the ex-ante problem must be above the maximizer of the ex-post problem. Additionally, we may have that the original maximization problem has several local maxima (and hence we cannot, without more conditions, describe the maximum using derivatives), however it is a continuous function on a convex set so it will generically have one global maximum which is, by the argument above, guaranteed to lie about $\overline{s}$.

□