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

Aurélie Ouss†, Alexander Peysakhovich‡

February 2015

FORTHCOMING, *Journal of Law and Economics*

Abstract

Economic theories of punishment focus on determining the levels that provide maximal social material payoffs. In other words, these theories treat punishment as a public good. Several parameters are key to calculating optimal levels of punishment: total social costs, total social benefits and the probability that offenders are apprehended. However, levels of punishment are often determined by aggregating individual decisions. Research in behavioral economics, psychology and neuroscience shows that individuals appear to treat punishment as a private good ("cold glow"). This means that individual choices may not respond "appropriately" to the social parameters above. We present a simple theory and show in a series of experiments that individually chosen punishment levels can be predictably too high or too low relative to those that maximize social material welfare. Our findings highlight the importance of the psychology of punishment for understanding social outcomes and for designing social institutions.

*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, Steve Raphael, Al Roth, Steven Shavell, Bruce Western 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.

†Department of Economics, Harvard University, 1805 Cambridge St., Cambridge, MA 02138. ouss@fas.harvard.edu. Ouss is a Terence M. Considine Fellow in Law and Economics at Harvard Law School and acknowledges support from the Considine Family Foundation and the School's John M. Olin Center for Law, Economics, and Business

‡Program for Evolutionary Dynamics, Harvard University; Department of Psychology, Yale University; apeysakh@fas.harvard.edu. Peysakhovich thanks the John Templeton Foundation for financial support.

1 Introduction

The criminal justice system is an expensive part of modern society, but it has an important instrumental role: it helps ensure cooperation and social order. Since Becker (1968) there has been a large interest in the economics of crime and punishment (see Levitt and Miles (2007) for a empirially-minded review). The Beckerian framework focuses on levels of punishment which yield optimal outcomes, where marginal (material) costs of punishment equal the marginal (material) benefits of decreased crime.[1] Thus, the Beckerian approach can be used as a normative theory of how to set punishment levels.

In many cases, however, levels of punishment in society are determined by aggregating individual decisions. For example, voters change laws (directly or via representatives), juries of civilians deliver verdicts and sometimes groups or individuals mete out social punishments themselves. In this paper we use experimental methods to ask two questions: first, do individual decisions about punishment respond to parameters that are important for setting Beckerian optimal punishments? Second, when levels of punishment are chosen by individuals, are socially optimal outcomes reached?

Researchers in the behavioral sciences have recently become interested in understanding human punishment behavior. In lab settings, where punishment is formally defined as the willingness to take actions that reduce the payoffs of others, a large portion of individuals are willing to pay costs to punish those who act in inappropriate ways (Ostrom, Walker, and Gardner (1992), Fehr and Gachter (2000), Peysakhovich, Nowak, and Rand (2014)) even when they have no personal stake (Fehr and Fischbacher (2004)) or no possible strategic motive (Fudenberg and Pathak (2010)). These findings suggest that individuals punish because they directly derive utility from reducing the payoffs of those who violate norms of cooperation.[2] Studies in social neuroscience support this theory: 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)), and reward activity is not visible when cooperative players are punished (Singer et al. (2006)). We refer to this broadly defined set of individual motivations as "cold glow," in reference to warm glow theories of altruism in which individuals receive utility from the act of being cooperative itself, beyond the

[1] There exist multiple channels by which punishment can increase cooperation: deterrence, specific deterrence and incapacitation (Shavell (1987)) are often mentioned. These economically-motivated analyses share a common thread: they each view punishment as a means to ensure social cooperation.

[2] Though there is considerable evidence that what actions constitute a norm violation of cooperation appear to vary by society (Henrich et al. (2010), Herrmann, Thöni, and Gächter (2008)) and can be manipulated in the lab (Peysakhovich and Rand (2015)).

final social consequences of cooperation (Andreoni (1990)). Additional evidence suggests that the proximal mechanism that drives cold glow involves affective considerations: strong negative emotions are engaged when other individuals break social norms (Xiao and Houser (2005), Fehr and Gächter (2002)).[3] Finally, research in moral psychology hints that this motive is very blunt.[4] Taken together, this broad array of evidence raises the question of whether individual punishment decisions will be reactive to changes in more abstract parameters, such as probability of apprehension or total social costs and benefits. If not, aggregates of individual punishment behaviors might not result in optimally deterring levels of punishment and may indeed lead to inefficient social outcomes.

We note that our focus is not on pinning down the exact mechanisms driving punishment behavior. Rather, we treat cold glow motivations as a first-order approximation and focus on asking how these well-documented facets of *individual* psychology can interact with institutions, modeled as available punishment technologies,[5] creating inefficiencies at *aggregate* level. More specifically, we ask whether aggregates of individual decisions hit the target of Beckerian optimal deterrence.

We hypothesize that individuals may respond to private costs much more than to social costs and thus when these costs are externalized, for example via group-funded punishment, individuals may demand a much higher level of punishment than is consistent with optimal deterrence. Additionally, if individuals are driven by more blunt 'just desserts' motivations, they may ignore the role of probability of apprehension. In this case, environments where individuals are rarely caught may exhibit aggregate punishment levels too low to deter expected utility maximizing offenders.

[3]Recent research in neuroscience (Sanfey et al. (2003), Knoch et al. (2006)) suggests that both affective and controlled processes are important for punishment behavior, and Buckholtz et al. (2008) find that controlled processes might matter more in determining criminal responsibility, while affective processes are more engaged when choosing magnitude of sanctions.

[4]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. In a similar vein Coffman (2011) finds that when defection is done via an intermediary punishment behaviors are reduced.

[5]There is a substantial literature that looks at the differential effectiveness of different punishment mechanisms in various games (Ostrom, Walker, and Gardner (1992), Xiao and Houser (2011), Houser et al. (2008), Andreoni and Gee (2012), Sutter, Haigner, and Kocher (2010), Nikiforakis (2008), Casari and Luini (2009), Balafoutas, Grechenig, and Nikiforakis (2014)). In our analysis, however, we set a mechanism and vary parameters of the environment as opposed to setting an environment and varying the mechanism. Integrating findings from psychology into designing punishment mechanisms robust to changes in parameters is an important topic for future research.

Finally, optimally deterring punishments take into account total levels of sanction but cold glow punishers' decisions may not be crowded out by other punishers' choices.[6] As our main contribution, we explore cold glow punishment in a series of lab experiments, which are designed not only to look for individual motives but also importantly to relate individual decisions to aggregate outcomes. To look at the effects of cost sharing, probability of apprehension and crowding out, we present three experiments in which people can punish a particular norm violation: taking from a third party. Our experimental designs allow for transparent calculation of levels of punishment that would reach the optimal deterrence benchmark: not only can we ask whether individual behavior responds to particular parameters, but also explore whether aggregate outcomes are 'optimal' in some sense.

Our first experiment looks at how punishment choices respond to cost structures: we vary whether the costs of implementing punishment are borne by the individuals making this choice, or by the group. 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. Our setup is such that relatively small punishments can implement social goals consistent with maximizing overall cooperation; yet when costs are not fully internalized, players over-punish. 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 caught and punished. We compare ex-ante punishment choices and taking behavior across conditions. 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 are too low to deter socially destructive behavior.

Our final experiment looks at whether our 'cold glow' terminology is apt. The theory of warm glow posits that individuals gain private benefits from the *act* of contributing to a public good and not from the overall amount. 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 note that some of these effects have been demonstrated in *second party* contexts, when the punisher's material welfare had been affected by

[6]We choose these three facets of punishment as they have particular relevance for important field behaviors. Because we view experiments as part of a larger empirical portfolio, we survey existing field evidence for the important of cold glow motives in section 5.

the offense (eg. Anderson and Putterman (2006), Nikiforakis and Normann (2008) demonstrate a 'demand curve' for punishment, while Casari and Luini (2012), Duersch and Müller (2010) discuss imperfect crowding out). In these experiments, unlike in our setups, a personal revenge motive always exists when punishing a defector. Our results complement the literature on peer punishment by showing that many results continue to hold even in third-party punishment situations. Thus, our experiments also indirectly shed some light on the question of whether second and third party punishment are instantiated via similar psychological mechanisms. Our setup is also more representative of settings of interest to legal scholars, as conviction and sentencing are more akin to third-party than second-party punishment.

2 Beckerian vs. Cold Glow Punishers

2.1 A simple reduced form punishment model

Traditional economic theories of crime take the case of rational criminals. For the main text, we derive our predictions in a simple reduced form model and in the online appendix we present a more detailed game theoretic derivation of these results for the case of a single actor.

Consider a single-shot scenario where a continuum of individuals can choose to engage in an action that is personally beneficial (they gain benefit b) and also socially costly. If individuals choose to take this action, they are 'caught' with probability p and receive a punishment of size l. Suppose that b varies across individuals for various exogenous reasons and, though it is an important subject, we do not discuss the responses of punishers to changes in the distribution of b in the population. This means that for a given probability of getting caught and punishment level we can write a demand curve $D(p, l)$, which is the amount of socially inefficient action that occurs, and is decreasing in both p and l. We assume that this demand is smooth and downward sloping. For simplicity we assume that even at $l = \infty$, there is some wasteful action taken either due to trembles or random utility models.[7]

First, let's suppose that a money-maximizing planner chooses the level of punishment l with all other parameters held exogenously fixed. He considers a social loss function $V(D(p, l))$ (which we assume is convex) and a cost function which we take as linear for simplicity, cl. This means for a given set of parameters there is an optimal punishment level $l^*_{Becker}(p, D, c)$

[7]This could be achieved, for example, if all individuals had logistic random utility as in a quantal response equilibrium model (McKelvey and Palfrey (1995)) or if they always trembled to an unintended action with small probability.

which minimizes the total social costs

$$V(D(p,l)) + cl.$$

We refer to this as the Beckerian optimum.

Note that the Beckerian optimum has two important comparative statics: first,

$$\frac{\partial l^*_{Becker}}{\partial c} < 0$$

that is, the optimal Beckerian punishment decreases in social cost per unit of punishment. This is because the optimum equates the marginal benefit of another unit of punishment (ie. decreases in defection) with the marginal cost. Second, we have that

$$\frac{\partial l^*_{Becker}}{\partial p} < 0.$$

As the probability of being caught decreases, the marginal benefit of a unit of punishment decreases (by the convexity assumption above) thus levels of punishment decrease.

Now, suppose that the planner is actually an individual from the society who can set l but bears a fraction γ of society's cost (for example, through taxes). The individual puts weight θ on total social welfare (ie. has social preferences). They also might derive some direct utility $G(l)$ from the act of punishment itself.[8]

The individual's maximization problem becomes to choose l to maximize

$$-\gamma cl + G(l) - \theta(-V(D(p,l)) - (1-\gamma)cl).$$

We refer to the case of $G(\cdot) = 0$ as a Beckerian punisher.

This simple model gives us some immediate insights. First, if punishers are Beckerian for any non-zero γ we get punishment levels that are different from the socially optimal level (essentially because punishment here is a public good). On the other hand, if punishers display a demand for punishment as a private then shifting costs to society may actually increase punishment above the Beckerian optimum. In addition, we also see that if θ is small relative to $G(\cdot)$ then individual decisions will not respond to changes in probability of apprehension as much as they should.

Finally, suppose that punishment is split into two parts, so that individuals who are caught first receive a punishment of l_1 and then a punishment of l_2. Note that the Beckerian optimum is applicable to the joint punishment $l = l_1 + l_2$, so optimal choices of l_2 are decreasing in l_1. This means that Beckerian punishers should show 'crowding out' effects on their own

[8]Note that $G(l)$ need not necessarily be strictly increasing in sanction. For example, a person might want a "fair punishment" which fits the offense.

punishment by others' punishments. On the other hand, if punishers care about their own contribution rather than total levels[9] then crowding out may not occur and can lead to higher levels of punishment than the Beckerian optimum.

These very different responses to parameters lead to very different predictions of the effects of various institutions. Making punishment easier by reducing or shifting the costs, or having multiple punishers, leads to good outcomes in a world where individuals punish for Beckerian reasons but can lead to inefficiently high punishments in a world where individuals choosing punishments are also motivated by cold glow. Changing probability of apprehension to be very low may lead to punishment levels that are too low and finally allowing for multiple punishers can allow for total amounts of punishment that could be seen as excessive.

2.2 Experimental Design

We now turn to evaluating our discussion above empirically. Our experiments test two types of questions. The first are comparative statics: do punishment levels respond to social or private costs? Is punishment crowded out? And does an other important parameter in the Beckerian model – probability of apprehension – change punishment decisions made by individual punishers?[10] We note that these individual-level questions ask about comparative statics rather than levels and so do not require strong assumptions on the form of utility functions.

The second set of questions we seek to address are mechanism design questions: are punishment too low or too high relative to Beckerian benchmarks? To make these statements, we require some assumptions on utility functions and for simplicity we choose risk neutrality and the assumptions that individuals tremble to unintended actions with probability ϵ. Without this assumption punishment levels of infinity lead to infinite deterrence and are weakly preferred to any other punishment level. While these assumptions are necessarily restrictive they allow us to make statements about whether aggregate actions lead to socially optimal outcomes, and if not, how badly they miss the target. Table 1 summarizes the experimental designs.

[9]This exact logic leads to predictions of a lack of crowding in cooperation under the 'warm glow' theory of public good provision (Andreoni (1993), Cornes and Sandler (1994)).

[10]In the full model, available in the online appendix, we show that even for ex-ante punishments, if decision-makers have preferences for punishments that 'fit the crime', when probability of apprehension is low, punishers might still shy away from the (high) levels of punishment that would sustain low levels of offending.

3 Experiment 1: Responses to Costs

In this first experiment, we ask an individual level and a group level question. At the individual level, we test whether costs of punishment accruing to the group rather than to the individual lead to higher demand for punishment. At the social level, the game is set up so that very low levels of punishment are sufficient to deter potential norm breakers. We then ask: will aggregate outcomes be in line with the Beckerian benchmark of optimal deterrence?[11]

3.1 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 into 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 is a socially destructive behavior in this case; 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

[11] There are other potential 'public good' motivations at play here beyond deterrence such as incapacitation or specific deterrence. We discuss them in more detail later as well as in the online appendix.

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 2 MU 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 public good, as stipulated in most economic models of law enforcement.

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 but these total levels of sanctions chosen are only revealed at the end of the session.

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.[12,13] 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 punish-

[12]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.

[13]Given lab size constraints, several sessions were run for each treatment, but subjects could only participate once in the experiment. They were not informed that later sessions of the same experiment would take place, making implausible that players had ulterior deterrence in this experiment in mind.

ment 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.[14] 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 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 of interest 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.

3.2 Theories of Punishment

There are three major 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 predictions in our experimental setup of these different theories for choices of punishment; a full discussion is developed in the online appendix.

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 for 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 the Public condition the cost of incapacitation outweighs its benefits, making it an insufficient motive to explain positive punishment levels.

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 they only know their assigner's choice of penalty rounds. General threats cannot be emitted.

[14] Experimental instructions are available in the online appendix

Specific deterrence would be the impact of received sanctions on punished offenders' future behaviors: giving larger sanction could make caught offenders less likely to take in future rounds. This motive could be a consideration, which we explore formally in the online appendix, making different assumptions on takers' behaviors. The main result is that sanctions should be decreasing as the game nears its end; and, regardless of our assumptions about takers' behaviors, in the One Round Take treatment, no sanction can be rationalized by a specific deterrence motives, since taking is only possible in the first round of this treatment condition.

Unlike with pro-social motives, cold glow predicts that punishment in Public would be higher than in the Private condition. 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.

3.3 Experiment 1 Results

This first section compares Public to the Private condition. We present graphs along with body text and regression analyses in tables 2 to 4. We then present additional evidence from One Round Take as a robustness check.

3.3.1 Punishment Decisions

We first look at punishers' decisions.[15] Figure 1 presents the number of rounds of punishment chosen in Public and Private conditions.[16]

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 motivation at play: as participants get closer to the end of the game, the size of imposed punishment does not decline. The idea behind specific deterrence is that punished players will have some period of time during which to apply the "lessons" that they have learned and no longer take;

[15]We note for completeness that in two of the experimental sessions, a bug in the software caused group accounts to unintentionally gain an extra 20-30 MU in the middle of the session. No participants reported noticing anything odd happening, participant behavior appears not to have been affected by the event and all our results are robust to restricting our analyses to rounds before this occurrence.

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

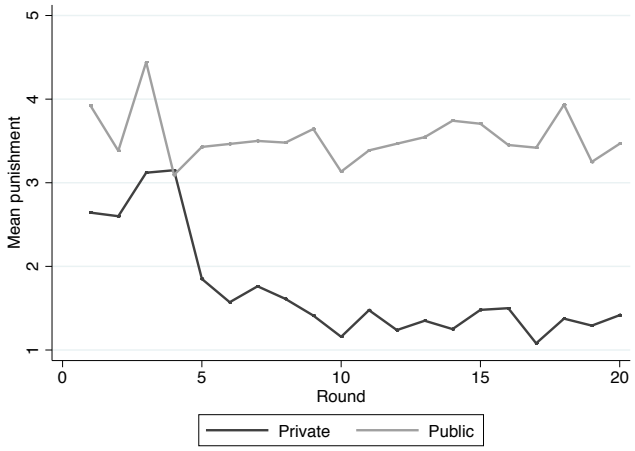

Figure 1: Mean punishment level chosen by round. There is a learning effect, but after a short time punishment levels in Private drop substantially below those in the Public condition.

less rounds would be necessary for that as the end of the game gets closer.[17] Furthermore, the average levels of punishment chosen in the Public condition far exceed the levels which would be 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 2 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.

Columns 1 - 4 of table 2 present 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

[17]Players are told that if the interaction ends before the Penalty Rounds are up, they will not be charged for the extra rounds. In those analyses, we look at the number of rounds that players chose, and not the number of rounds that they ended up doling out, as these would mechanically decline as the end of the game gets closer.

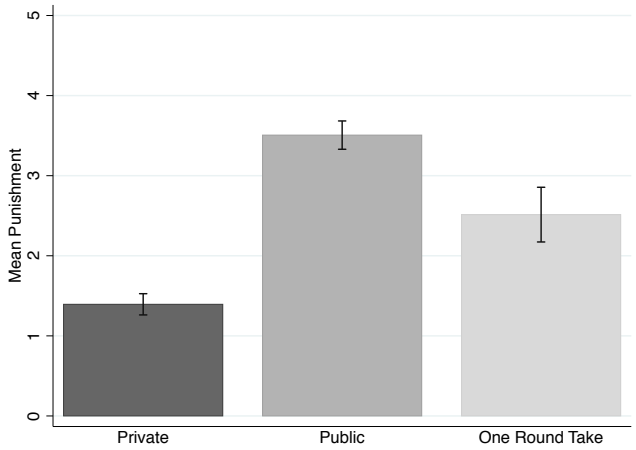

Figure 2: Mean punishment level chosen, round ≥ 5. Externalizing costs leads to large increases in punishment. These high levels continue even when punishment has no possible effect on future behavior in the One Round Take treatment. Error bars represent standard error of the mean.

holds when we control for round effects (column 2).

Column 4 of the table 2 presents regression results for our robustness check. 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.[18]

Taken together, our regression analyses confirm that cold glow does well in predicting responses of punishment decisions to cost structure and that indeed aggregate levels of punishment are above those consistent with Beckerian punishers. We note that other motives appear to exist, but cannot explain most of the variation in punishment. We now turn to see the effects of conditions on taking decisions.

[18]Another possible explanation for the difference in behavior between One Round Take and Public is that perhaps it is easier to ex-post rationalize punishment decisions in the former than in the latter.

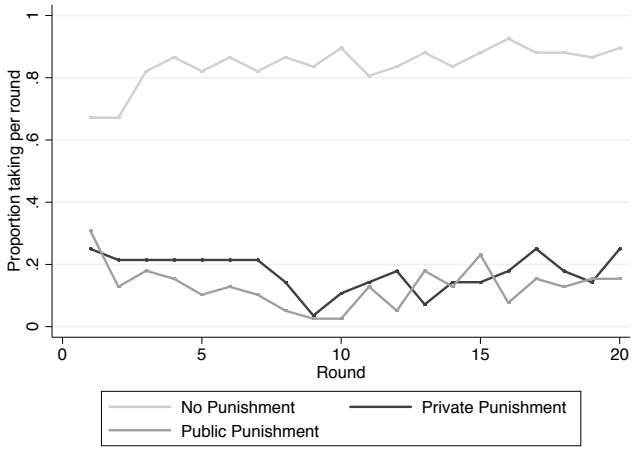

Figure 3: Experiment 1: Percent choosing take by condition. Though punishment levels are much higher in Public than in Private, it has no effect on realized levels of taking. However, potential takers do react to even the possibility of punishment: with no punishment possible taking levels are very high.

3.3.2 Taking Decisions

Figure 3 shows taking decisions by availability of punishment, and columns 5 and 6 of table 2 presents our regression results. Taking behavior is significantly higher in Punishment than in No Punishment conditions (column 5), which shows that general deterrence does matter: only 10% to 20% of participants who are able to take[19] choose to do so, even from round 1. However, there is no difference between the Public and Private conditions (column 6).

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. Thus, although general deterrence did lower taking levels, the extra punishment in the Public condition did not further reduce taking.

3.3.3 Differences in punishments: mechanisms

What drives differences across treatments in punishment choices? Table 3 displays the differences across the Private and Public treatments in the percentage of individuals choosing to use the punishment mechanism at all (columns 1) and conditional on choosing a positive punishment, the average subject levels of punishment (column 2). These results are statistically

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

suggestive but not significant at conventional levels ($p < .1$). We see that in the Public conditions, individuals are much more likely to opt into using the punishment technology and conditional on any punishment, punishments are also longer. Thus differences in levels of punishments come from both the extensive and the intensive margins. 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, 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.'

4 Experiment 2: Responses to Probability of Apprehension

Our second experiment asks how differences in probability of apprehension affect punishers' and potential norm-breakers' decisions. If punishers and norm-breakers don't symmetrically react to these changes, this can lead to socially wasteful levels of punishment, since probabilities enter into optimally deterring punishments. In particular, if norm-breakers respond to probabilities but punishers don't, low probabilities of apprehension can lead to excessively low levels of punishment. In addition, we compare ex-ante and ex-post punishment decisions.

4.1 Experimental Setup

We use a game to test how both sentences and potential norm-breaking respond to expected punishments.[20] 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 chooses 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.

[20]Experimental instructions are presented in the online appendix

Our treatments vary the probability that the taker will be found if he takes: in the "high probability" treatment, the taker is found with a probability 9/10; in the "low probability" treatment, with a probability 1/3.[21] 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.[22]

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 online 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).

4.2 Experiment 2: Results

4.2.1 Punisher Behavior

We now consider the behavior of punishers across conditions. The left graph 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 punishment choices. The mean punishment level is 4.0 units

[21]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, Darley, and Robinson (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 an incentive-compatible 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.

[22]Several recent studies have been undertaken to examine the validity of experimental data collected using AMT at stakes of $\sim$ 1 USD. They find that behavior on AMT matches well with standard laboratory results on economics games (Amir, Rand, and Gal (2012)) (Rand, Greene, and Nowak (2012)), and are based on samples that are more representative of the general population (Horton, Rand, and Zeckhauser (2011), Paolacci, Chandler, and Ipeirotis (2010)).

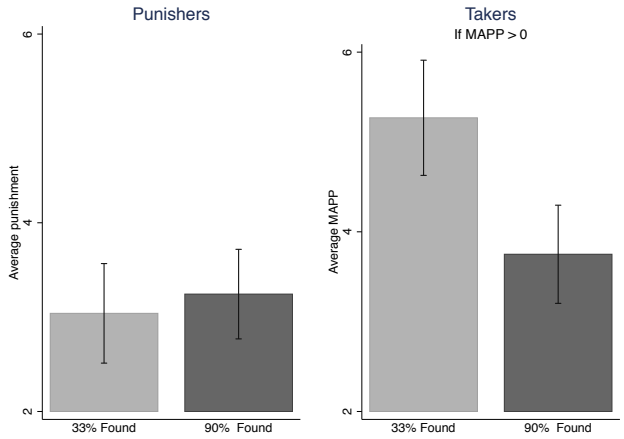

Figure 4: Experiment 2: While probability of capture has no effect on punishment decisions, it has a strong effect on takers' decisions. Error bars represent standard error of the mean.

(40 points) in the high probability condition and 4.1 unit (41 points) in the low probability condition, and the difference non-significant (see columns 1 - 3 of table 5). In the Beckerian model of punishment, probability of apprehension is a key feature in determining optimal sentences. Expected punishment is defined as punishment if caught, multiplied by likelihood that offenders are caught, and so changes in probability should be compensated by changes in punishment. Empirically, experiment 2 shows that this is, in fact, not the case: punishments do not differ across probabilities of apprehension.

4.2.2 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. Column 1 of 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 the right graph 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). Column 3 of table 6 shows our regression results, confirming there is a significant intensive margin response.[23] Unlike punishers, takers respond to the probability of being caught,[24] and so the punishment levels chosen are too low to deter many takers in the low probability condition.

4.3 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 control 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 5 and table 7 present 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[25].

We note that 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; but probabilities are again not factored in. Levels of punishment are however smaller when there is no possibility of deterrence (3.16 ex-post vs. 4.1 ex-ante), these differences being only significant at the 10 percent level. 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 deter-

[23]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 could be explained by higher risk aversion (Eckel and Grossman (2008)).

[24]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.

[25]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

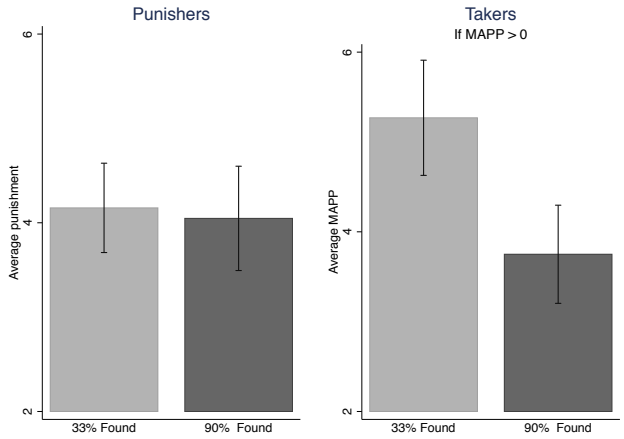

Figure 5: Experiment 2 Control Decisions: When punishment decisions are made ex-post, we see no main effect of probability of apprehension on punishment. Error bars represent standard error of the mean.

rence 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.

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

Columns 4 of table 5 and 5 of table 7 present 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. 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.

5 Experiment 3: Crowding Out

Our final experiment asks an individual level question motivated by our theory: to what extent do the sanction decisions individuals act as substitutes or complements to own levels of sanction? Our social level question asks how total levels of sanction inflicted change with the introduction of multiple punishers.

5.1 Main Experiment

In order to answer this question, we ran an experiment on AMT using a sample of 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.[26]

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.[27] 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 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 6 presents the average punishment choice of assigner

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

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

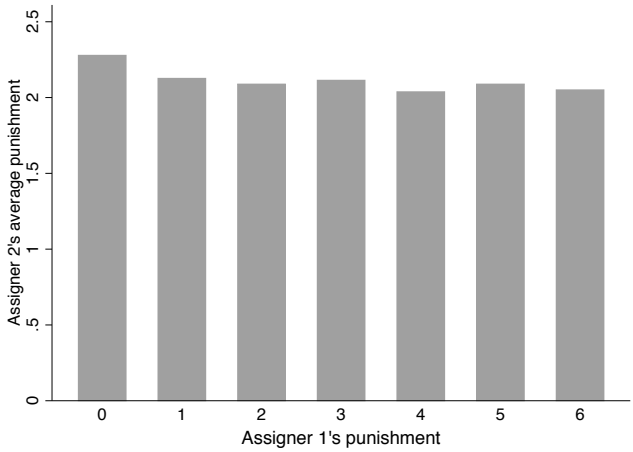

Figure 6: Experiment 3: Assigner 2's behavior. At the aggregate level we see no evidence of crowding out of assigner 2's punishments by assigner 1's punishments. Statistical significance calculated via clustered regressions due to correlation of decisions within individual, error bars omitted.

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, as confirmed in regression analysis (column 1 of table 8).

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[28], 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 (Column 2 of table 8).

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

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

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).

5.2 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 unit.

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 7 and column 3 of table 8 show the average levels of punishment chosen in the two conditions. Assigner choose slightly lower levels of punishment levels when $c = 40$ than when $c = 0$, but this difference is not statistically significant. 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 total realized sanctions are approximately 20 points in the $c = 0$ condition and 57 points in the $c = 40$ condition.

We find only a small effect on taker behavior, 78% of takers choose the cooperative action in the $c = 0$ condition and 85% of takers choose the cooperative action in the $c = 40$ condition. This difference is not significant and we attribute the small change to floor effects (recall that takers are caught 75% of the time in this control experiment).

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 we never observe perfect crowding out.

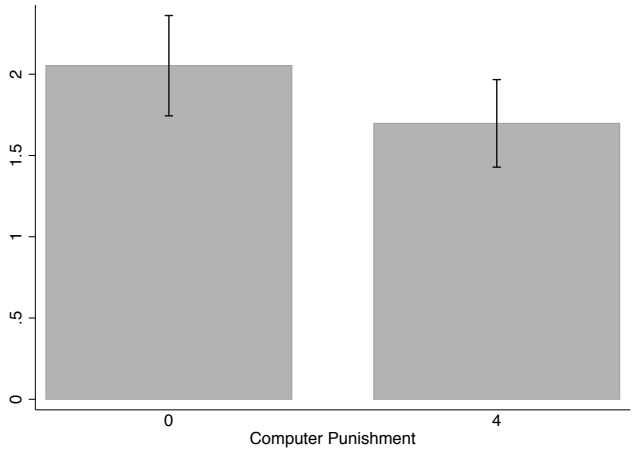

Figure 7: Study 2 Control Decisions. Higher levels of punishment by computer do not crowd out individual punishments. Error bars represent standard error of the mean.

6 Punishment Behavior in the Field

Psychological evidence shows that human punishment behavior is driven largely by blunt, affective motivations. Our lab experiments show that when aggregating these decisions outcomes may not coincide with Beckerian benchmarks. We now survey some evidence that suggests that cold glow motivations may have large effects for important outcomes in the criminal justice system.

Demand for punishment for private motives can affect aggregate outcomes through the behavior of elected officials. First, we note that if punishment of offenders 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 (2001) 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".

In addition to descriptive evidence, causal links have been identified: Berdejo and Yuchtman (2013) analyze changes in sentencing behavior of judges during election cycles. They find that judicial severity increases when judges are close to reelection and thus under political pressure from constituents, and sentences fall immediately afterwards.[29] These results

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

cannot be explained by differential work loads due to longer sentencing; variations in the month of nomination and election further 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 directly affect the behaviors of judges, and thus outcomes in the criminal justice system. We view that as a less likely place of influence, 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, Levav, and Avnaim-Pesso (2011)) which has put forth at least some evidence that judges are subject to predictable biases, so it is not impossible that cold glow is a partial motivator of judicial decisions.

In addition, there is evidence in law and economics arguing 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, Schkade, and Kahneman (2000) for two survey-based experiments). Punishers' insensitivity to probability of capture, an important input into optimal deterrence, is a behavior that cold glow punishers can display. Further understanding how "fair" punishment levels are determined is an important direction for basic science as well as practical considerations.

There has very little 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.[30] Taking our results to the field, Ouss (2014) investigates the 1996 California Juvenile Justice Realignment as a natural experiment. Prior to the realignment, sentences were chosen at the county level, but costs of incarceration were borne at the state level, and were therefore subsidized form the sentencing county's perspective. The realignment resulted in a discontinuous drop in use of juvenile corrections, but no discontinuous change in juvenile arrests, suggesting that subsidized corrections had been over-used. Similarly, Ater, Givati, and Rigbi (2014) 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, consistent with an imperfect factoring in of total costs of crime reduction when making arrest decisions.[31]

[30] 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.

[31] We note there are many other possible explanations for these results: police officers'

Imperfect crowding out of individual punishment preferences by prior punishments could play a role in labor markets. Having a criminal record impacts employability of an individual (Bushway, Stoll, and Weiman (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. Understanding the relative importance of these channels has important policy implications (for example, policies on shrouding criminal records).

Another interplay between psychological motivations underlying human punishment behaviors and groups' behaviors comes from recent discussions of 'mob-like' punishment on social media. Individuals who post 'socially unacceptable' materials on outlets such as Twitter have subsequently been routinely harassed, doxxed (ie. have their personal information such as addresses publicly revealed) and otherwise punished by a massive crowd. This has led in some cases to extreme emotional trauma, job loss and very high economic, social and psychological costs for the individuals involved.[32] While in the context of deterring motives for punishment this group response seems excessive (surely a smaller response would yield a similar level of deterrence) cold glow motivations make these events not so surprising. Technology makes each individual's act of punishment relatively costless, the public nature of the interactions makes anyone able to join in and the lack of crowding out of one's desire to punish by others' choices means that no natural mechanism exists to stop the cascade once it begins. Of course, individuals also often receive attention and social status in addition to cold glow for taking part in these situations, so a full explanation of why these incidents occur is of independent research interest. However, we argue that the behavioral economics of punishment are front and center in understanding these phenomena.

Individual decisions are a product of many factors: elections involve many non-judicial dimensions, jurors are prompted to depart from emotions,[33] and exact magnitudes of costs or probabilities of apprehension are generally not known by voters, juries or judges. In this way, our lab experiments are somewhat artificial. However, they allow us to study, in a controlled environment, punishment choices which are normally hard to observe in the field. We do not argue that experiments are a substitute for

effort provision might respond to costs, police evaluations could depend on number of arrests, etc.

[32]For an illustration of these online retaliations, see Jon Ronson's New York Times article "How One Stupid Tweet Blew Up Justine Sacco's Life".

[33]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."

traditional empirical analysis but rather a complement; experimental methods form an important part of a larger scientific portfolio. More research is needed but it seems clear that a richer understanding of human psychology can be highly valuable in aiding the understanding of important legal phenomena, we view the growing fields of behavioral and experimental law and economics as important contributors to this understanding.

7 Conclusion

Though many legal scholars and philosophers think of moral reasoning as driven by rational processes, 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 to make punishment decisions may sometimes collaterally result in social harmony, but in other domains can result in either highly inefficient over or under punishing. Our series of lab experiments show little evidence that standard rational motives such as deterrence or incapacitation, which underpin most economics of crime models, 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 is important to explain outcomes in settings relevant to law and economics, including aggregate outcomes in the criminal justice system. We have presented several possible channels through which cold glow can affect these aggregate outcomes.

We have primarily compared cold glow outcomes to 'material welfare' maximizing benchmarks and one may argue that if cold glow is indeed a parameter in a utility function, then cold glow itself could be a legitimate source of individual welfare. However, even if we take cold glow to be a legitimate source of welfare, problems can arise. For example, if individuals can select into the role of punisher, those with the 'strongest' cold glow might choose 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.

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 help design "rules of the game" that achieve normatively desired outcomes (Roth (2002)). These types of questions are especially important at the intersection of psychology, law, economics and institutional design (Hauser et al. (2014), Fudenberg and Peysakhovich (2014)): in the case of punishment institutions, effective rules of the game

will depend on the psychological motivations of the players. 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.

References

Amir, Ofra, David G. Rand, and Ya'akov Kobi Gal. 2012. "Economic Games on the Internet: The Effect of $1 Stakes." *PloS one* 7 (2):e31461.

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

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

———. 1993. "An experimental test of the public-goods crowding-out hypothesis." *The American Economic Review* 83 (5J):1317–1327.

Andreoni, James and Laura Gee. 2012. "Gun for hire: Delegated enforcement and peer punishment in public goods provision." *Journal of Public Economics* 96 (11-12):1036–1046.

Ater, Itai, Yehonatan Givati, and Oren Rigbi. 2014. "Organizational structure, police activity and crime." *Journal of Public Economics* 115:62–71.

Balafoutas, Loukas, Kristoffel Grechenig, and Nikos Nikiforakis. 2014. "Third-party punishment and counter-punishment in one-shot interactions." *Economics Letters* 122 (2):308–310.

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

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

Berdejo, Carlos and Noam Yuchtman. 2013. "Crime, punishment, and politics: an analysis of political cycles in criminal sentencing." *Review of Economics and Statistics* 95 (3):741–756.

Buckholtz, Joshua W, Christopher L Asplund, Paul E Dux, David H Zald, John C Gore, Owen D Jones, and Rene Marois. 2008. "The neural correlates of third-party punishment." *Neuron* 60 (5):930–940.

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

Carlsmith, Kevin M, John M Darley, and Paul H Robinson. 2002. "Why do we punish?: Deterrence and just deserts as motives for punishment." *Journal of Personality and Social Psychology* 83 (2):284 – 299.

Casari, Marco and Luigi Luini. 2009. "Cooperation under alternative punishment institutions: An experiment." *Journal of Economic Behavior & Organization* 71 (2):273–282.

———. 2012. "Peer punishment in teams: expressive or instrumental choice?" *Experimental Economics* 15 (2):241–259.

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

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

Cushman, Fiery, Anna Dreber, Ying Wang, and Jay Costa. 2009. "Accidental outcomes guide punishment in a ?trembling hand? game." *PloS one* 4 (8):e6699.

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

De Quervain, Dominique J-F, Urs Fischbacher, Valerie Treyer, Melanie Schellhammer, Ulrich Schnyder, Alfred Buck, and Ernst Fehr. 2004. "The neural basis of altruistic punishment." *Science* .

Duersch, Peter and Julia Müller. 2010. "Taking punishment into your own hands: An experiment on the motivation underlying punishment." *Working Paper* .

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

Fehr, Ernst and Urs Fischbacher. 2004. "Third-party punishment and social norms." *Evolution and human behavior* 25 (2):63–87.

Fehr, Ernst and Simon Gachter. 2000. "Cooperation and punishment in public goods experiments." *The American Economic Review* 90 (4):980–994.

Fehr, Ernst and Simon Gächter. 2002. "Altruistic punishment in humans." *Nature* 415 (6868):137–140.

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

Fudenberg, Drew and Parag A Pathak. 2010. "Unobserved punishment supports cooperation." *Journal of Public Economics* 94 (1):78–86.

Fudenberg, Drew and Alexander Peysakhovich. 2014. "Recency, records, and recaps: The effect of feedback on behavior in a simple decision problem." *Proceedings of the 15th ACM conference on Economics and Computation* .

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

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

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

Hauser, Oliver, David Rand, Alexander Peysakhovich, and Martin Nowak. 2014. "Cooperating with the future." *Nature* 511.

Henrich, Joseph, Jean Ensminger, Richard McElreath, Abigail Barr, Clark Barrett, Alexander Bolyanatz, Juan Camilo Cardenas, Michael Gurven, Edwins Gwako, Natalie Henrich et al. 2010. "Markets, religion, community size, and the evolution of fairness and punishment." *science* 327 (5972):1480–1484.

Herrmann, Benedikt, Christian Thöni, and Simon Gächter. 2008. "Antisocial punishment across societies." *Science* 319 (5868):1362–1367.

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

Houser, Daniel, Erte Xiao, Kevin McCabe, and Vernon Smith. 2008. "When punishment fails: Research on sanctions, intentions and non-cooperation." *Games and Economic Behavior* 62 (2):509–532.

Knoch, Daria, Alvaro Pascual-Leone, Kaspar Meyer, Valerie Treyer, and Ernst Fehr. 2006. "Diminishing reciprocal fairness by disrupting the right prefrontal cortex." *Science* 314 (5800):829–832.

Levitt, Steven D and Thomas J Miles. 2007. "Empirical study of criminal punishment." *Handbook of law and economics* 1:455–495.

McKelvey, Richard D. and Thomas R. Palfrey. 1995. "Quantal response equilibria for normal form games." *Games and economic behavior* .

Nikiforakis, Nikos. 2008. "Punishment and counter-punishment in public good games: Can we really govern ourselves?" *Journal of Public Economics* 92 (1):91–112.

Nikiforakis, Nikos and Hans-Theo Normann. 2008. "A comparative statics analysis of punishment in public-good experiments." *Experimental Economics* 11 (4):358–369.

Ostrom, E., J. Walker, and R. Gardner. 1992. "Covenants with and without a sword: Self-governance is possible." *The American Political Science Review* 86 (2):404–417.

Ouss, Aurélie. 2014. "Incentives Structures and Criminal Justice." *Working Paper* .

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

Paolacci, Gabriele, Jesse Chandler, and Panagiotis G Ipeirotis. 2010. "Running experiments on Amazon Mechanical Turk." *Judgment and Decision making* 5 (5):411–419.

Peysakhovich, Alexander, Martin Nowak, and David G. Rand. 2014. "Humans display a cooperative phenotype that is domain general and temporally stable." *Nature Communications* 5.

Peysakhovich, Alexander and Dave G. Rand. 2015. "Habits of Virtue: creating norms of cooperation and defection in the laboratory." *Management Science* .

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

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

Rand, David G, Joshua D Greene, and Martin A Nowak. 2012. "Spontaneous giving and calculated greed." *Nature* 489 (7416):427–430.

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

Roth, Alvin E. 2002. "The economist as engineer: Game theory, experimentation, and computation as tools for design economics." *Econometrica* 70 (4):1341–1378.

Sanfey, Alan G, James K Rilling, Jessica A Aronson, Leigh E Nystrom, and Jonathan D Cohen. 2003. "The neural basis of economic decision-making in the ultimatum game." *Science* 300 (5626):1755–1758.

Shavell, Steven. 1987. "A model of optimal incapacitation." *The American Economic Review* :107–110.

Singer, Tania, Ben Seymour, John P O'Doherty, Klaas E Stephan, Raymond J Dolan, and Chris D Frith. 2006. "Empathic neural responses are modulated by the perceived fairness of others." *Nature* 439 (7075):466–469.

Sunstein, Cass R, David A Schkade, and Daniel Kahneman. 2000. "Do People Want Optimal Deterrence?" *The Journal of Legal Studies* 29 (1):237–53.

Sutter, Matthias, Stefan Haigner, and Martin G Kocher. 2010. "Choosing the carrot or the stick? Endogenous institutional choice in social dilemma situations." *The Review of Economic Studies* 77 (4):1540–1566.

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

———. 2011. "Punish in public." *Journal of Public Economics* 95 (7):1006–1017.

Table 1: Summary of Experiments

	Punishment Technology	Probability of apprehension	Other punishment?	Punishment payment structure	Notes
Experiment 1: cost structures and punishment choices					
Social costs	Exclude from game	1/2	None	Punisher pays	
Private costs	Exclude from game	1/2	None	Group pays	
One Round take	Exclude from game	1/2	None	Group pays	Only 1 round of taking + punishment
Experiment 2: Probability of apprehension					
Low probability	Take away payoffs	1/3	None	Punisher pays	
High probability	Take away payoffs	9/10	None	Punisher pays	
Experiment 3: Crowding Out					
Response to other punishers' decisions	Take away payoffs	3/4	Other punisher	Punisher pays	
Response to computer decision	Take away payoffs	3/4	Random computer punishment	Punisher pays	

Table 2: Experiment 1 - Costs and availability of sanctions

	Public vs. private			Robustness Check	No vs. With Sanction	
	(1)	(2)	(3)	(4)	(5)	(6)
	All rounds	All rounds	Rounds 6-20	Costs vs. Deterrence	Sanctions	Costs
Outcome	**Punishment level**			**Punishment level**	**Taking behavior**	
Public	1.818*	1.809*	2.113**	2.082**		-0.0556
	(0.754)	(0.754)	(0.771)	(0.768)		(0.0728)
Round		-0.0406*				
		(0.0186)				
No Deterrence				-0.988		
				(0.780)		
Sanctions vs. None					-0.655**	
					(0.0371)	
Constant	1.734**	2.166**	1.394**	1.420**	0.841**	0.219**
	(0.455)	(0.530)	(0.457)	(0.458)	(0.0256)	(0.0625)
Observations	1067	1067	782	902	2407	1067

Results clustered at the subject level

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

Table 3: Experiment 1 - Individual differences in punishing, by condition

	(1) Number Opt Outs	(2) Average sanction, if > 0
Public	-3.151^{+}	1.408^{+}
	(1.777)	(0.809)
Constant	7.536**	3.374**
	(1.356)	(0.634)
Observations	67	57

One observation per subject
0 = chose not to have a sanction
$^{+}$ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$

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

	(1)	(2)	(3)	(4)
Public	1.887*			1.777*
	(0.778)			(0.750)
Punishment Chosen	0.579	-1.214^{+}	1.936*	
	(0.640)	(0.686)	(0.936)	
Stolen From				-0.369
				(0.287)
Constant	1.437*	2.358**	2.789**	1.896**
	(0.640)	(0.733)	(0.587)	(0.515)
Observations	1067	448	619	1067

Results clustered at the subject level
$^{+}$ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$

Table 5: Experiment 2: Punishment choice, by probability to get caught

	Punisher's choice			Target's opinion
	(1)	(2)	(3)	(4)
	Full Sample	Full Sample	If Punish = 1	Full Sample
Outcome	**Punish**	**Level**	**Level**	**Fair Level**
1 = High	-0.131	-0.135	0.592	0.724
	(0.0792)	(0.738)	(0.736)	(0.732)
1 = Female	-0.0300	-0.733	-0.703	-0.817
	(0.0794)	(0.739)	(0.739)	(0.789)
Constant	0.935**	4.505**	4.836**	4.600**
	(0.0688)	(0.641)	(0.617)	(0.586)
Observations	81	81	69	80

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: Taker's choice, by probability to get caught

	(1)	(2)	(3)
	Full Sample	Full Sample	If Take = 1
Outcome	**Take**	**MAPP**	**MAPP**
1 = High	0.114	-0.550	-1.700*
	(0.0988)	(0.721)	(0.814)
1 = Female	-0.227*	-2.127**	-2.035*
	(0.105)	(0.767)	(0.925)
Constant	0.724**	4.116**	5.896**
	(0.0785)	(0.573)	(0.665)
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: player 1 chose to take
$^{+}$ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$

Table 7: Experiment 2: Robustness Check. Punishment and taking choices with no deterrence, by probability to get caught

	Punisher's choice			Taker's choice	Target' opinion
	(1)	(2)	(3)	(4)	(5)
	Full Sample	Full Sample	If Punish = 1	Full Sample	Full Sample
Outcome	**Punish**	**Level**	**Level**	**Take**	**Fair Level**
1 = High	0.0355	0.202	0.112	-0.251*	0.168
	(0.0983)	(0.741)	(0.771)	(0.121)	(0.850)
1 = Female	0.151	-0.0533	-0.778	-0.221+	-0.267
	(0.0957)	(0.722)	(0.745)	(0.128)	(0.867)
Constant	0.727**	3.066**	4.189**	0.551**	5.456**
	(0.0900)	(0.679)	(0.716)	(0.108)	(0.771)
Observations	66	66	54	64	64

Standard errors in parentheses. High: found with a 90% chance; Low: found with a 33% chance
Take: player 1 chose to take
+ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$

Table 8: Experiment 3: 2nd punisher's choice, by 1st punisher (or computer) choice

	2 Punishers		Computer Control
	(1)	(2)	(3)
	Full sample	Crowd Out Types	Full Sample
Outcome	**Level**	**Level**	**Level**
Player 1 Penalty Choice	-0.0289	-0.569**	
	(0.0620)	(0.0585)	
1 = High Computer Penalty			-0.355
			(0.408)
Constant	2.199**	3.380**	2.053**
	(0.237)	(0.363)	(0.297)
Observations	553	196	81

Standard errors in parentheses. High Computer Penalty = 4; Low Computer Penatly = 0
+ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$