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

Aurélie Ouss[†], Alexander Peysakhovich[‡]

October 31, 2013

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. In calculating optimal levels several parameters are key: total social costs, total social benefits and the probability that offenders are apprehended. However, levels of punishment often are determined by aggregates of 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 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. We show how these individual choices can then lead to inefficiencies in final social outcomes, such as levels of punishment chosen and total costs incurred. 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, 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. However, it has an important instrumental role: it produces cooperation and social order. Since Becker (1968) there has been a large theoretical interest in the economics of crime and punishment.[1] The Beckerian framework focuses on levels of punishment which ensure optimal deterrence, where marginal (material) costs of punishment equal the marginal (material) benefits of decreased crime.[2] 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 aggregates of many individual decisions. For example, voters change laws (directly or via representatives), individuals lobby lawmakers, juries of peers determine guilt or sentences.[3] 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?[4] Second, when levels of punishment are chosen by individuals are socially optimal outcomes reached?

Researchers in the behavioral sciences have recently become interested 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 et al. (1992), Fehr and Gachter (2000)) 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.[5] 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

[1] Empiricists have taken an interest in this topic as well, see Levitt and Miles (2007) for a recent review.

[2] Deterrence is not the only function of punishment discussed in the economics of crime literature. In particular, incapacitation and specific deterrence (Shavell (1987)) are often mentioned. However, all of these economically-motivated analyses have a common thread: they each view punishment as a means to ensure social cooperation.

[3] This is the case in some tort cases in the US, or in criminal trial courts in France.

[4] In the economic framework, *aggregate* parameters are of interest. We focus on the simplest case of the theory and study three parameters: probability of detection, total social costs and total social benefits.

[5] Though there is considerable evidence that what actions constitute a cooperation norm violation appear to vary by society (Henrich et al. (2010), Herrmann et al. (2008)).

of altruism in which individuals receive utility from the act of being cooperative itself and not the 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 Gachter (2002)).[6] Finally, research in moral psychology hints that this motive is very blunt.[7] 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 social cost 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 mechanisms,[8] 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 criminals.[9] 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.[10] As our

[6] Recent research in neuroscience (Sanfey et al. (2003), Knoch et al. (2006)) suggests that both affective and controlled processes are important for punishment behavior but the exact nature of this interplay remains unknown.

[7] 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.

[8] There is a substantial literature that looks at the differential effectiveness of different punishment mechanisms in various games (Xiao and Houser (2011), Houser et al. (2008), Andreoni and Gee (2012), Sutter et al. (2010), Nikiforakis (2008), Casari and Luini (2009)). In our analysis, however, we fix a mechanism and vary parameters of the environment as opposed to fixing 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.

[9] We develop a formal model of cold glow punishments in the online appendix.

[10] We choose these three facets of punishment as they have particular relevance for

main contribution we explore cold glow punishment in a series of lab experiments. Our experiments are designed to not only to look for individual motives and but also to relate individual decisions to aggregate outcomes. This paper contributes to the new and growing fields of experimental and behavioral law and economics, by showing how lab experiments can help understand important judicial behaviors, as part of a larger methods portfolio. 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.[11] Our design allows us to not only ask whether individual behavior responds to particular parameter changes but also whether aggregate outcomes are 'optimal' in some sense.

Our first experiment looks at how punishment choices respond to cost structures: punishment decisions are made by individuals, and we vary whether the costs of implementing the punishment are borne by the individuals 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 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 total share provided. 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 pun-

important field behaviors. We survey existing empirical evidence in section 5.

[11] Many papers consider the addition of punishment to public goods games (Ostrom et al. (1992), Fehr and Gachter (2000)), prisoner's dilemmas and 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)). We add to this literature by employing simple third party punishment games where the calculation for deterrence maximizing punishments is transparent.

ishment 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 - ie. those in which the punishee's initial action directly affects the punisher's material welfare. Anderson and Putterman (2006) find a 'demand curve' for punishment in public goods games by varying prices. Casari and Luini (2012) find that punishment of a non-cooperator in a public goods game by another person is not a substitute good for one's own punishment. Finally Duersch and Müller (2010) find that individuals are willing to pay for the right to be the person who administers a level of punishment. 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.

The rest of the paper presents each of our experiments in turn (sections 2 - 4), before discussing how these findings relate to punishment decisions in judicial cases (section 5). Section 6 concludes.

2 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, leads to higher demand for punishments. 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?[12]

2.1 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

[12] 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.

= 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 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.[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 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.[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 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.

[13] 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.

[14] Appendix 1 presents the experimental instructions

2.2 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 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, only their assigner's choice of penalty rounds. General threats therefore cannot be emitted. Specific deterrence could be a consideration, which we explore formally in the appendix, making different assumptions on takers' behaviors. The main result is that in all cases, 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 positive exclusion can be rationalized by a specific deterrence motives, since taking is only possible in the first round of this treatment condition.

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.

2.3 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.

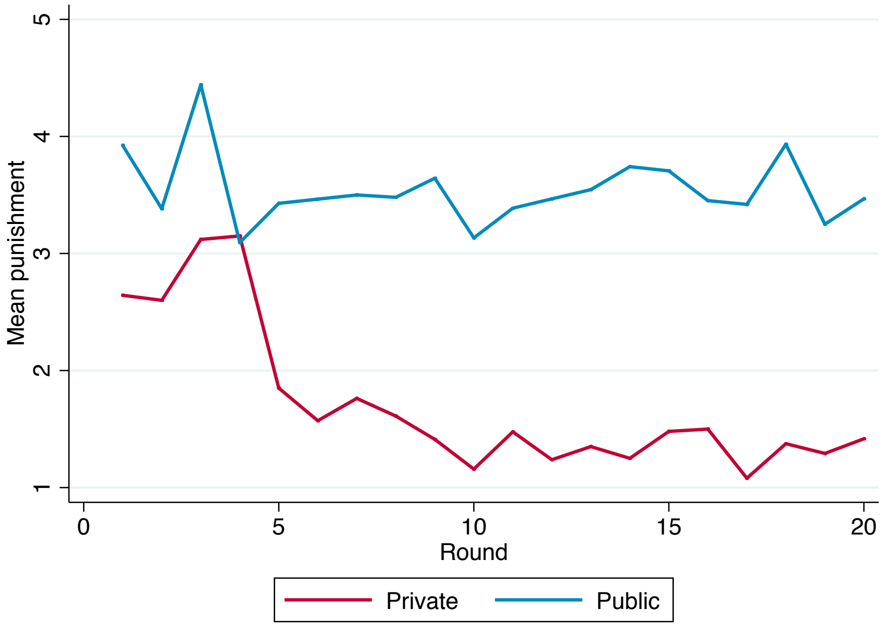


Figure 1: Mean punishment level chosen by round. There is a learning effect, but after a short time punishment levels in Private drop substantially below those in the Public condition.

2.3.1 Punishment Decisions

We first look at punisher's 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 down. Furthermore, the average levels of punishment chosen in the Public condition far exceed than the levels in

[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.

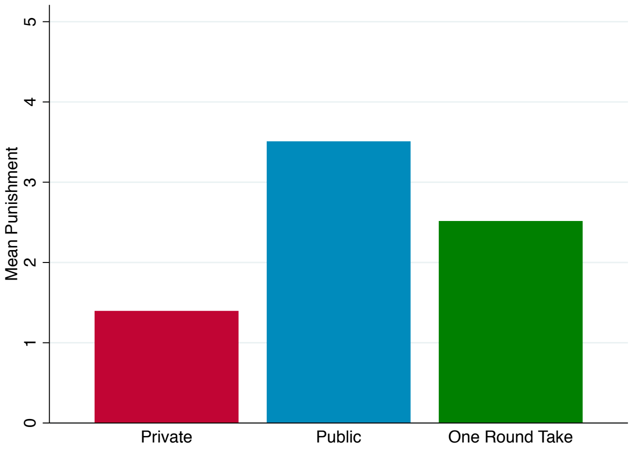


Figure 2: Mean punishment level chosen, round $\geq$ 5. Externalizing costs leads to large increases in punishment. These high levels continue even when punishment has no possible effect on future behavior in the One Round Take treatment.

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

Column 4 of the table 1 presents regressions 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.[17]

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.

2.3.2 Taking Decisions

Figure 3 shows taking decisions by availability of punishment, and columns 5 and 6 of table 1 presents our regression results. Taking behavior is significantly higher in Punishment and No Punishment conditions (column 5), which shows that general deterrence does matter: only 10% to 20% of participants who are able to take[18] 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 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 low levels of punishment, since probabilities enter into

[17] 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.

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

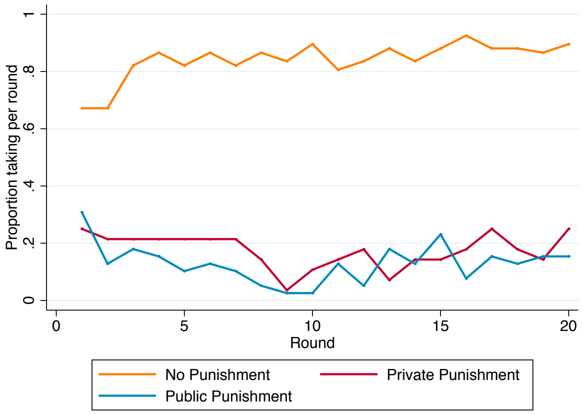


Figure 3: Experiment 1: Percent choosing take by condition. Though punishment levels are much higher in Public than in Private, it has no effect on realized levels of taking. However, potential takers do react to even the possibility of punishment: with no punishment possible taking levels are very high.

optimally deterring punishments. In addition, we compare ex-ante and ex-post punishment decisions.

3.1 Experimental Setup

We use a game to test both how sentences are chosen, and how potential norm-breakers respond to expected punishments.[19] 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.

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.[20] All players are informed of all rules at the beginning of the game.

[19] Experimental instructions are presented in the appendix

[20] Some studies in psychology have investigated the effects of probability of apprehen-

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.[21]

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).

3.2 Experiment 2: Results

3.2.1 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 (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 2). This behavior is inconsistent with that of a Beckerian punisher.

sion 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.

[21] 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 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)).

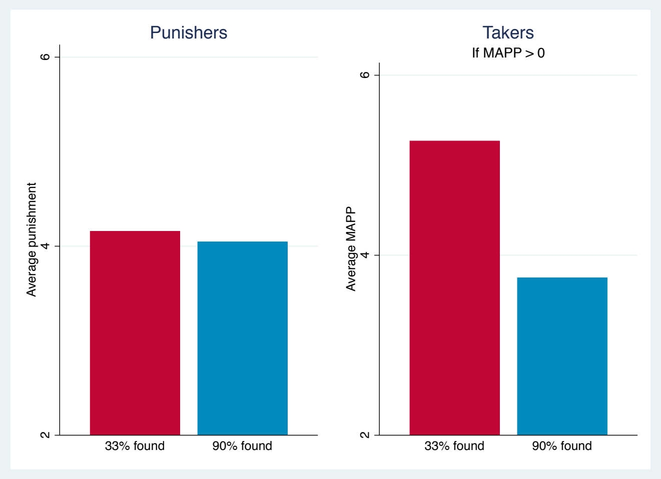


Figure 4: Experiment 2: While probability of capture has no effect on punishment decisions, it has a strong effect on takers' decisions.

3.2.2 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 3 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 3 shows our regression results, confirming there is a significant intensive margin response.[22] Unlike punishers, takers respond to the probability of being caught,[23] and so the punishment levels chosen are too low to deter

[22] 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)).

[23] This also allows us to control away a lack of attention or understanding by partici-

many takers in the low probability condition.

3.3 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 4 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[24].

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 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), but these differences are 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 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.

3.4 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 2 and 5 of table 4 present our regression analysis. Unsurprisingly, targets want higher punishments than assigners: this

pants as the result of the null effect on punishment decisions as individuals are randomly assigned into roles.

[24] 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

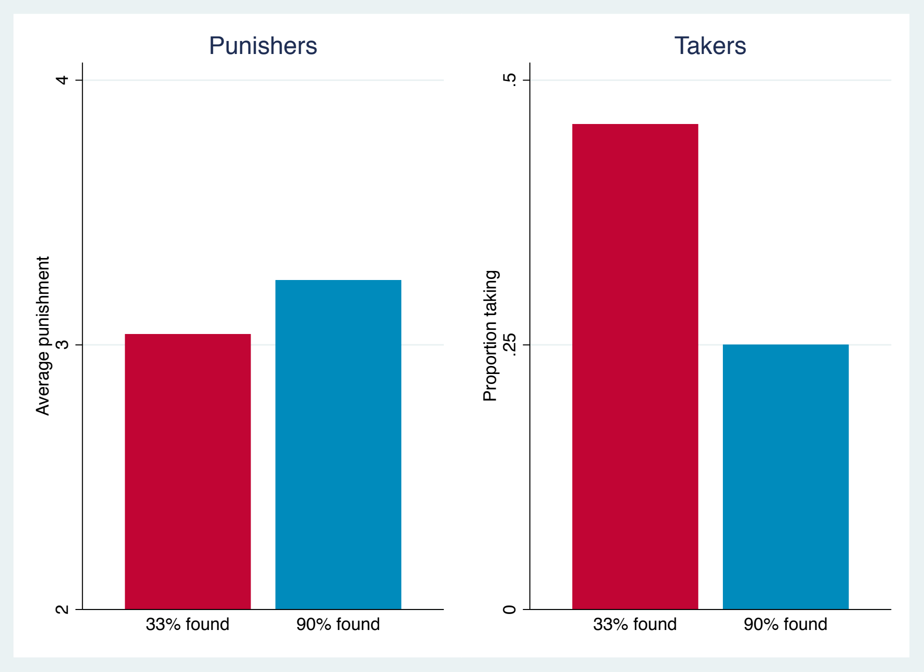


Figure 5: Experiment 2 Control Decisions: When punishment decisions are made ex-post, we see no main effect of probability of apprehension on punishment.

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.

4 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.

4.1 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.[25]

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.[26] 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 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 5).

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

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

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

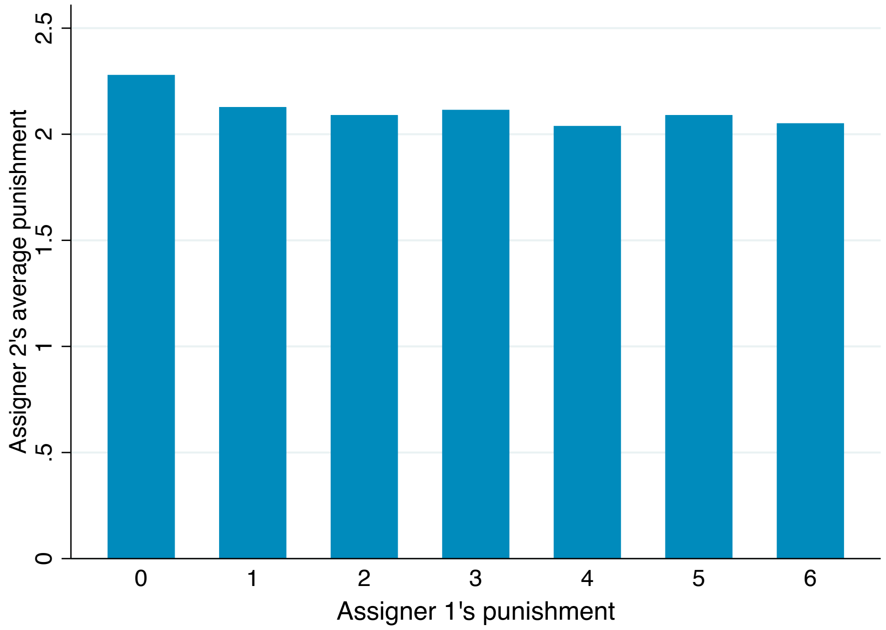


Figure 6: Experiment 3: Assigner 2's behavior. At the aggregate level we see no evidence of crowding out of assigner 2's punishments by assigner 1's punishments.

whose sanction choices increases in assigner 1's choice (crowd-in types[27], 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 5).

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 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).

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

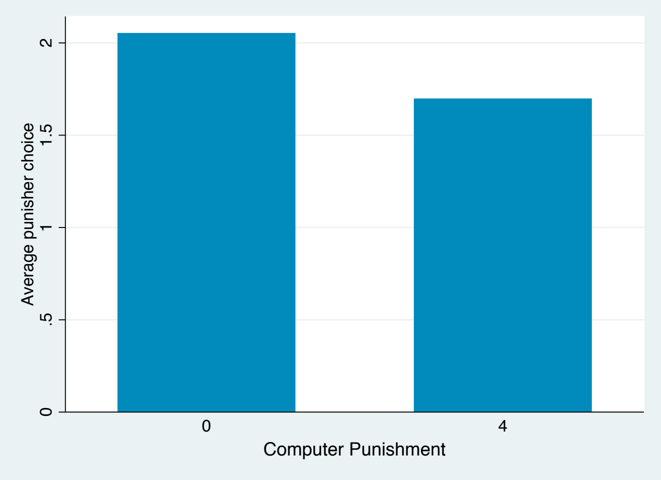


Figure 7: Study 2 Control Decisions

4.2 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 5 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.

5 Punishment Behavior in the Field: Criminal Justice

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 (2009) 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.[28] These results 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

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

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 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 et al. (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 et al. (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 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.[29] To our knowledge, the only paper to this date 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.[30]

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 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. Understanding the relative importance of these channels has important policy implications (for

[29] 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.

[30] 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.

example, policies on shrouding criminal records).

We acknowledge that individual decisions are a product of many factors: elections involve many non-judicial dimensions, jurors are prompted to depart from emotions,[31] 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 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.

6 Conclusion

Though many legal scholars and philosophers think of moral reasoning as driven by rational processes, the 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 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. To gain a fuller understanding of legal phenomena, 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.

[31] 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.”

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 help design "rules of the game" that achieve normatively desired outcomes.[32] These types of questions are especially important at the intersection of psychology, law and economics: in the case of punishment institutions, effective rules of the game will depend on the psychological motivations of the players. Under the assumptions that individuals punish for public goods motives (theories of deterrence, incapacitation) punishment could be under provided due to free-riding motivations. Thus 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.

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

References

Amir, O., Rand, D. and Gal, Y. (2012), 'Economic games on the internet: The effect of $1 stakes', *PloS one* **7**(2), e31461.

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

Andreoni, J. (1993), 'An experimental test of the public-goods crowding-out hypothesis', *The American Economic Review* pp. 1317–1327.

Andreoni, J. and Gee, L. (2012), 'Gun for hire: Delegated enforcement and peer punishment in public goods provision', *Journal of Public Economics* .

Ater, I., Givati, Y. and Rigbi, O. (2012), 'Organizational structure, police activity and crime: Evidence from an organizational reform in jails', *Working Paper* .

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

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

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

Baron, J. and Ritov, I. (2009), 'The role of probability of detection in judgments of punishment', *Journal of Legal Analysis* **1**(2), 553–590.

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

Berdejo, C. and Yuchtman, N. (2009), 'Crime, punishment and politics: An analysis of political cycles in criminal sentencing', *Unpublished manuscript, Harvard University* .

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

Camerer, C., Ho, T. and Chong, J. (2004), 'A cognitive hierarchy model of games', *The Quarterly Journal of Economics* **119**(3), 861–898.

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

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

Casari, M. and Luini, L. (2012), 'Peer punishment in teams: expressive or instrumental choice?', *Experimental Economics* **15**(2), 241–259.

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

Cornes, R. and Sandler, T. (1994), 'The comparative static properties of the impure public good model', *Journal of Public Economics* **54**(3), 403–421.

Costa-Gomes, M., Crawford, V. and Broseta, B. (2003), 'Cognition and behavior in normal-form games: An experimental study', *Econometrica* **69**(5), 1193–1235.

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

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

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

Duersch, P. and Müller, J. (2010), 'Taking punishment into your own hands: An experiment on the motivation underlying punishment'.

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

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

Fehr, E. and Gachter, S. (2000), 'Cooperation and punishment in public goods experiments', *The American Economic Review* **90**(4), 980–994.

Fehr, E. and Gachter, S. (2002), 'Altruistic punishment in humans', *Nature* **415**(6868), 137–140.

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

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

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

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

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

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

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

Henrich, J., Ensminger, J., McElreath, R., Barr, A., Barrett, C., Bolyanatz, A., Cardenas, J. C., Gurven, M., Gwako, E., Henrich, N. et al. (2010), 'Markets, religion, community size, and the evolution of fairness and punishment', *science* **327**(5972), 1480–1484.

Herrmann, B., Thöni, C. and Gächter, S. (2008), 'Antisocial punishment across societies', *Science* **319**(5868), 1362–1367.

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

Houser, D., Xiao, E., McCabe, K. and Smith, V. (2008), 'When punishment fails: Research on sanctions, intentions and non-cooperation', *Games and Economic Behavior* **62**(2), 509–532.

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

Knoch, D., Pascual-Leone, A., Meyer, K., Treyer, V. and Fehr, E. (2006), 'Diminishing reciprocal fairness by disrupting the right prefrontal cortex', *Science* **314**(5800), 829–832.

Levitt, S. and Miles, T. (2007), 'Empirical study of criminal punishment', *Handbook of Law and Economics* **1**, 455–495.

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

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

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

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

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

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

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

Rand, D., Greene, J. and Nowak, M. (in Press), 'Spontaneous giving and calculated greed', *Nature* .

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

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

Sanfey, A. G., Rilling, J. K., Aronson, J. A., Nystrom, L. E. and Cohen, J. D. (2003), 'The neural basis of economic decision-making in the ultimatum game', *Science* **300**(5626), 1755–1758.

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

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

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

Sutter, M., Haigner, S. and Kocher, M. G. (2010), 'Choosing the carrot or the stick? endogenous institutional choice in social dilemma situations', *The Review of Economic Studies* **77**(4), 1540–1566.

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

Xiao, E. and Houser, D. (2011), 'Punish in public', *Journal of Public Economics* **95**(7), 1006–1017.

Table 1: 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 2: 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 3: 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 4: 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 5: 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

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

A Models of Third Party Punishment

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.

A.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.

A.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.[33] 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

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

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 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

[33] 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.

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 these decisions differ.

A.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.

Our assumptions about cold glow can be justified empirically: individuals have been shown to often sacrifice personal payoffs to reduce the payoffs of players behaving selfishly in games such as the public goods game (Ostrom et al. (1992)) or the dictator game. This occurs even in third party punishment (Fehr and Fischbacher (2004)), and when punishments can't be used to 'teach a lesson'.[34][35]

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.

[34] 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.

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

A.3.1 Ex-Post Behavior

First, we introduce a basic model of social preferences which depend on the *action* taken by another player.[36] We start with simple three player model, and then extend it to N players.

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. 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:

[36] 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)).

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[37] while Fehr and Fischbacher (2004) find that less fair (more inappropriate) divisions in a dictator game are punished more by third parties.

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.[38]

Four Players We now move to the case where several individuals observe the taker's behavior and can choose to punish him, in order to discuss how

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

[38] 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.

multiple punishers' choices interact. Suppose now there 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:

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

P's utility function is now as follows:

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

Keeping the old assumptions on the shape of the λ function, there exists a unique $s^*_P(\beta, s_T, s_{P_2})$ describing the original punisher's optimal choice. Our model allows for several types of interactions in punishment 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 for the Beckerian punisher), a punisher only cares about is the overall level of punishment. When crowding out is imperfect, the punisher also cares about how *he* changes the taker's utility.[39]

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 knowledge on P's part about of how wrong T's action was. With this uncertainty, other players' actions serve as a signal and so can lead to crowding in of punishments.[40]

Note, again, that we do not make any assumptions on the behavior of *levels* of λ in s_{P_2}, 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.

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

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

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.

A.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.

Thus the new order of the game is: 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 to take, 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 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.[41]

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

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

[41] 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.

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.[42] 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.

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-

[42] This also means that our model nests a decision-maker who cares only about the deterrence aspects of punishments by setting $\frac{\partial \lambda}{\partial s_P}$ 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.

making. We characterize the relative sizes of some of these motives in experiment 2.

A.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_1 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 representative 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.[43] 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

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

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.

B 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 above $\overline{s}$.

□

C Experiment 1: Punishment Predictions, by Punishment Motives

C.1 Incapacitation

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, there can be no incapacitation motives.

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.[44]

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.

C.1.1 Specific Deterrence

We present different assumptions on takers' behaviors and derive punishment choices:

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,[45] which is exactly

[44] 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.

[45] 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

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.[46]

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, 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.

D 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

therefore assume that loss from exclusion for 1 round is equal to 4.

[46] 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.

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 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.

Appendix: Experimental Instructions

Appendix 1: General Instructions

Please read the following instructions carefully. If you have any questions, do not hesitate to ask us. Aside from this, no communication is allowed during the experiment.

Instructions

This is a computerized experiment on decision-making. You will be paid for participating and the amount you earn will depend on the decisions that you make.

The full experiment should take about 60 minutes. At the end of the experiment, you will be paid privately and in cash for your participation.

All information collected in this experiment will be anonymous and neither the experimenter nor other participants will be able to link your identity to your decisions. In order to maintain this privacy, please do not reveal your decisions to any other participant.

We consider ourselves bound by the promises we are making to you in this protocol, we will do everything we say and there will be no surprises or tricks.

If you think that something weird is going on, it is probably a bug in the software – please let an experimenter know so that we can get it fixed.

We are interested in individual choices so please **remember that there are no right or wrong answers**.

Payment

In this experiment you will earn Monetary Units (MUs) through the decisions that you make. At the end of the experiment, these MUs will be converted into dollars at a rate of 50 MU per dollar. You will start the experiment with some initial allocation of MU in your account and you will earn additional MU from decisions made by yourself and other players.

In addition to any money you earn from your decisions you will also receive a $10 show-up fee.

General Instructions

This experiment will consist of individual parts called *interactions*. At the start of each *interaction* you will be randomly matched in a group with other individuals, at the end of each *interaction*, existing groups will be disbanded and new groups will be formed. Each *interaction* will be split into sub-intervals called *rounds*.

Each round will consist of several decisions that will impact your experimental earnings.

At the beginning of each *interaction* a *group pot* will be created which will contain some initial MU. Decisions that group members make during the game can change the amount of MU in the group pot. *At the end of each interaction, the total contents of the group pot will be divided evenly amongst all group members and added to their personal account of MU*.

The instructions on the next page explain the rules of the game for the first *interaction* of the experiment – each other *interaction* may have slightly different rules and you will be informed of the rule changes at the start of each *interaction* via your computer screen. At all times, every individual in the room will always be playing with the same set of rules.

Appendix 2: Instructions, No Punishment Condition

One Round in the First Interaction

The first interaction will consist of 20 rounds.

Each *round* will consist of 3 parts: SOLVE, TAKE, PAYOFF. The round will proceed as follows:

PART 1 - SOLVE: You will see a screen with two numbers, a blank text box and a DONE button. You will be asked to enter the sum of the two numbers into the box and press DONE. You will have up to 60 seconds to make your decision.

If you enter the correct sum, *you will get 4 MU*.

If you enter the wrong sum (or enter nothing), *you will get 0 MU*.

PART 2 - TAKE: You will see a screen with two buttons, TAKE or DON'T TAKE. If you choose TAKE, one random player in your group will be selected, you will receive 2 MU and they will lose 3 MU from their personal account. You will have up to 30 seconds to make your decision.

If you choose TAKE, there is a 50% chance you will be FOUND OUT, this will only occur if you choose TAKE. You will be informed if you are FOUND OUT. Being FOUND OUT will not affect your payoffs and will not personally identify you to other players.

PART 3 - PAYOFF: On the next screen you will see the results of the round including how many MU you gained or lost and whether you were selected by a player choosing to TAKE.

When everyone in your group presses OK, a new round will begin. If you do not press OK within 30 seconds, the computer will press OK automatically.

Appendix 3: Instructions, Private Punishment Condition

You will now play another interaction. Just like the first interaction, this interaction will consist of 20 rounds. Each round have the same basic structure as before, SOLVE, TAKE and PAYOFF, however, there will now be one extra part.

After you are shown your PAYOFF but before any players are FOUND OUT, you will be asked to enter an amount of Penalty Rounds to be possibly assigned to an individual who is FOUND OUT in this round. You may enter any number between 0 and 10.

After everyone enters in the amount of Penalty Rounds they would assign, each individual who chose to TAKE is FOUND OUT with 50 percent probability. For each individual who is FOUND OUT, one other individual from the group is chosen at random to be their ASSIGNER.

Each FOUND OUT individual is assigned the number of Penalty Rounds entered by their ASSIGNER. Note that you can never be your own ASSIGNER, so whatever amount of Penalty Rounds you enter can only be assigned to a different player who is FOUND OUT in this round.

An individual who is assigned Penalty Rounds will not be allowed to participate in the interaction for that many rounds. That is, they do not get to SOLVE, TAKE or ASSIGN Penalty Rounds. They may, however, lose MU to individuals choosing TAKE.

Each Penalty Round will cost the ASSIGNER 2 MU from their personal account. If the interaction ends before the Penalty Rounds are up, the ASSIGNER will not be charged for the extra rounds. You will only be charged for Penalty Rounds you choose to assign if you are picked to be an ASSIGNER.

An Example: suppose you enter that you would assign 10 Penalty Rounds. If you are picked as an ASSIGNER, you will be charged 20 MU (if there are more than 10 rounds left in the game) and the individual who was FOUND OUT will not be able to participate in the game (SOLVE, TAKE or ASSIGN) for 10 rounds. If you are not picked as an ASSIGNER, you will not be charged.

Appendix 4: Instructions, Public Punishment Condition

You will now play another interaction. Just like the first interaction, this interaction will consist of 20 rounds. Each round have the same basic structure as before, SOLVE, TAKE and PAYOFF, however, there will now be one extra part.

After you are shown your PAYOFF but before any players are FOUND OUT, you will be asked to enter an amount of Penalty Rounds to be possibly assigned to an individual who is FOUND OUT in this round. You may enter any number between 0 and 10.

After everyone enters in the amount of Penalty Rounds they would assign, each individual who chose to TAKE is FOUND OUT with 50 percent probability. For each individual who is FOUND OUT, one other individual from the group is chosen at random to be their ASSIGNER.

Each FOUND OUT individual is assigned the number of Penalty Rounds entered by their ASSIGNER. Note that you can never be your own ASSIGNER, so whatever amount of Penalty Rounds you enter can only be assigned to a different player who is FOUND OUT in this round.

An individual who is assigned Penalty Rounds will not be allowed to participate in the interaction for that many rounds. That is, they do not get to SOLVE, TAKE or ASSIGN Penalty Rounds. They may, however, lose MU to individuals choosing TAKE.

Each Penalty Round that ends up being assigned will cost 5 MU from the Group Account. If the interaction ends before the Penalty Rounds are up, the Group Account will not be charged for the extra rounds.

An Example: suppose you enter that you would assign 10 Penalty Rounds. If you are picked as an ASSIGNER, the Group Account will be charged 50 MU (if there are more than 10 rounds left in the game) and the individual who was FOUND OUT will not be able to participate in the game (SOLVE, TAKE or ASSIGN) for 10 rounds. If you are not picked as an ASSIGNER, your decision will not have any effect on the Group Account.

Appendix 5: Instructions, One Round Take Condition

You will now participate in another interaction. Most of the rules in this interaction will be the same as before, however it will only last 11 rounds. Individuals will NOT be able to use the TAKE option in rounds 2 through 11. The first round will have the same basic structure as before: SOLVE, TAKE and PAYOFF. It will also have one extra part.

After you are shown your PAYOFF but before any players are FOUND OUT, you will be asked to enter an amount of Penalty Rounds to be possibly assigned to an individual who is FOUND OUT in this round. You may enter any number between 0 and 10.

After everyone enters in the amount of Penalty Rounds they would assign, each individual who chose to TAKE is FOUND OUT with 50 percent probability. For each individual who is FOUND OUT, one other individual from the group is chosen at random to be their ASSIGNER.

Each FOUND OUT individual is assigned the number of Penalty Rounds entered by their ASSIGNER. Note that you can never be your own ASSIGNER, so whatever amount of Penalty Rounds you enter can only be assigned to a different player who is FOUND OUT in this round.

An individual who is assigned Penalty Rounds will not be allowed to participate in the interaction for that many rounds. Each Penalty Round that ends up being assigned will cost 5 MU from the Group Account.

An Example: suppose you enter that you would assign 10 Penalty Rounds. If you are picked as an ASSIGNER, the Group Account will be charged 50 MU and the individual who was FOUND OUT will not be able to participate in the game for 10 rounds. If you are not picked as an ASSIGNER, your decision will not have any effect on the Group Account.

Appendix 6 : Experiment 2 Instructions, Responses to Probability of Apprehension

Thanks for being part of our research!

We are interested in the way that individuals make decisions in social situations. You must be at least 18 years old to participate in this study, if you are not, please close this window and exit the HIT.

Your participation will take no more than 5 minutes.

In this study, you will be anonymously put into a group with two other people recruited from Mechanical Turk and the three of you will participate in an interaction.

During the interaction you will earn points based on your decisions and the decisions of other individuals in your group.

We will pay everyone who participated 1 cent per 2 points they earned in the game via the bonus system on Mechanical Turk. You can earn up to 50 extra cents for participating!

This is in addition to the base compensation of 30 cents that you receive from just accepting the HIT.

You will be informed of your bonus payment once everyone signed up for the HIT has finished making decisions. You can expect this bonus payment within ten (10) days of your participation in this study.

Instructions

Everyone starts in groups of three players.

The other two people in your group will be real individuals also recruited from Mechanical Turk.

Every player will start the interaction with a balance of 80 points.

At the start of the interaction, each person will be given a role: one person will be the assigner, one person will be player 1, one person will be player 2.

The interaction will proceed in order:

1) First, the assigner will make a decision

2) Player 1 will make a decision

3) Player 2 will be a passive observer

The decisions are:

1) The assigner will choose a **penalty unit level**, this can be any number between 0 and 10.

2) Player 1 will see the penalty level that the assigner imposed and choose between **"Take"** and **"Don't Take."**

a) If player 1 chooses **"Take"**, then **player 1 will gain 20 points** and **player 2 will LOSE 30 points**.

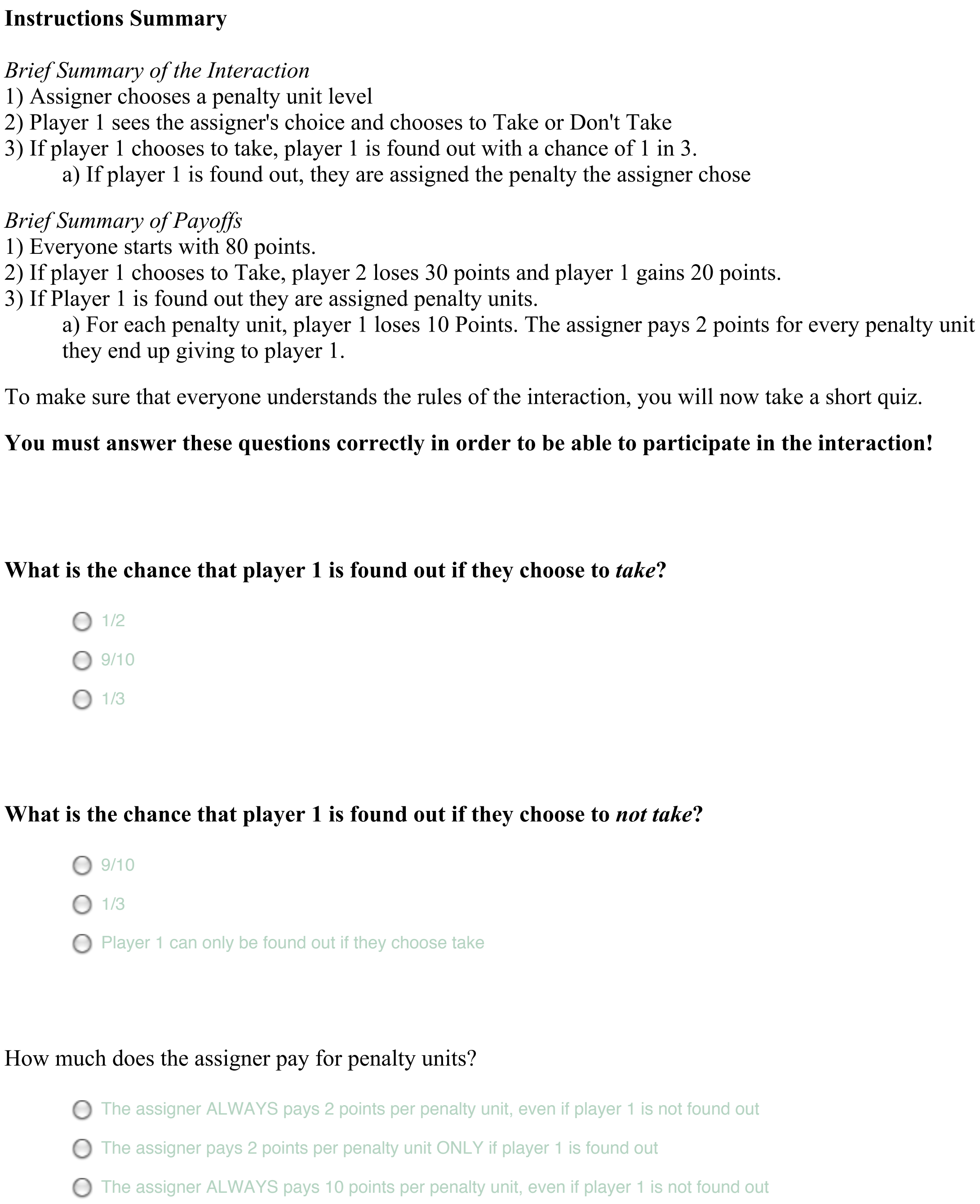
b) If player 1 chooses **"Don't Take"** everyone will keep their 80 points.

3) If player 1 chooses "Take" they may be **found out**. This will happen in **1 out of 3** cases**.**

a) If player 1 is found out, they will be given the penalty chosen by the assigner. For every penalty unit that player 1 receives, player 1 will LOSE 10 points.

b) If player 1 is found out, the assigner will be charged 2 points per unit they give to player 1.

c) The assigner will not be charged if player 1 is not found out or does not take.



Instructions Summary

Brief Summary of the Interaction
1) Assigner chooses a penalty unit level
2) Player 1 sees the assigner's choice and chooses to Take or Don't Take
3) If player 1 chooses to take, player 1 is found out with a chance of 1 in 3.
    a) If player 1 is found out, they are assigned the penalty the assigner chose

Brief Summary of Payoffs
1) Everyone starts with 80 points.
2) If player 1 chooses to Take, player 2 loses 30 points and player 1 gains 20 points.
3) If Player 1 is found out they are assigned penalty units.
    a) For each penalty unit, player 1 loses 10 Points. The assigner pays 2 points for every penalty unit they end up giving to player 1.

To make sure that everyone understands the rules of the interaction, you will now take a short quiz.

You must answer these questions correctly in order to be able to participate in the interaction!

What is the chance that player 1 is found out if they choose to *take*?

○ 1/2

○ 9/10

○ 1/3

What is the chance that player 1 is found out if they choose to *not take*?

○ 9/10

○ 1/3

○ Player 1 can only be found out if they choose take

How much does the assigner pay for penalty units?

○ The assigner ALWAYS pays 2 points per penalty unit, even if player 1 is not found out

○ The assigner pays 2 points per penalty unit ONLY if player 1 is found out

○ The assigner ALWAYS pays 10 points per penalty unit, even if player 1 is not found out

Player 1

You will be player 1 in your group. You start with 80 points.

You can now choose to Take or Not Take.

If you take, you will gain 20 points, player 2 will lose 30 points.

If you choose to take, there is a 9 out of 10 chance you will be found out and given the assigner&#39;s penalty.

Each penalty unit the assigner gives you will cause you to lose 10 points from your final earnings.

You will make the following decision: you enter your maximum acceptable possible penalty. This is the penalty level such that...

...if the assigner chooses a penalty unit level above this, you would prefer not to take

...if the assigner chooses this penalty unit level or below this, you would prefer to take.

Because you do not always get caught, you can think of this as entering the maximum level of risk you would be willing to bear to use the take option. Also...

...if you would prefer to never take, you should choose a maximum acceptable possible penalty of 0.

...if you would prefer to take no matter what, you should choose a maximum acceptable possible penalty of 10.

Use the buttons below to select your maximum acceptable possible penalty.

○ 0 Penalty Units (0)
○ 1 Penalty Unit (1)
○ 2 Penalty Units (2)
○ 3 Penalty Units (3)
○ 4 Penalty Units (4)
○ 5 Penalty Units (5)
○ 6 Penalty Units (6)
○ 7 Penalty Units (7)
○ 8 Penalty Units (8)
○ 9 Penalty Units (9)
○ 10 Penalty Units (10)

Player 2

You will be player 2 in this interaction.

You start with 80 points. If player 1 chooses to take, you will lose 30 points.

If they choose to not take, you will keep your full 80 points.

As player 2, you will not make any decisions, however we are interested in what decisions you think other people should make.

What do you think the "fair" level of penalty units is for the assigner to choose? (Reminder, each penalty unit causes player 1 to lose 10 points from their final earnings if they choose to take and are found out, which happens in 9 out of 10 times they choose to take)

______ I think the fair level of penalty is.... (1)

You will receive your bonus payment within ten (10) days!

We would like to ask you two brief demographics questions.

Please enter your age (in number of years) below:

Please enter your gender.

○ Male (1)
○ Female (2)

Control experiment, 9/10 probability to be found

Instructions

Everyone starts in groups of three players.

The other two people in your group will be real individuals also recruited from Mechanical Turk. Every player will start the interaction with a balance of 80 points.

At the start of the interaction, each person will be given a role: one person will be the assigner, one person will be player 1, one person will be player 2.

The interaction will proceed in order:

1) First, player 1 will make a decision

2) The assigner will make a decision

3) Player 2 will be a passive observer

The decisions are:

1) Player 1 will choose between

a) If player 1 chooses to “Take”, then player 1 will gain 20 points and player 2 will lose 30 points.

b) If player 1 chooses “Don’t Take”, everyone will keep their 80 points.

2) If player 1 chooses Take, they may be found out. This will happen 9 out of 10 times.

3) If player 1 is found out, the assigner will be able to give player 1 between 0 and 10 penalty units. For every penalty unit, player 1 will lose 10 points and the assigner will be charged 2 points.

Instructions Summary

This a brief summary of the game:

1) Player 1 chooses to Take or Don't Take

2) If player 1 chose to take, he is found with a chance of 9 in 10.

3) If player 1 is found, the assigner chooses a number of penalty units to give to player 1.

Payoffs

1) Everyone starts with 80 points.

2) If player 1 chooses to Take, player 2 loses 30 points and player 1 gains 20 points.

3) For every penalty unit player 1 is given, they will lose 10 points, in addition the assigner must pay 2 points per penalty unit they give out.

To make sure that everyone understands the rules of the interaction, you will now take a short quiz.

You must answer these questions correctly in order to participate in the interaction.

What is the chance that player 1 is found out if they choose to take?

○ 1/2 (0)
○ 9/10 (1)
○ 1/3 (0)

What is the chance that player 1 is found out if they choose to not take?

○ 9/10 (0)
○ 1/2 (0)
○ Player 1 can only be found out if they choose take

Player 1

You will be player 1 in your group.

You start with 80 points.

You can now choose to Take or Not Take.

If you take, you will gain 20 points, player 2 will lose 30 points.

If you choose to take, there is a 9 out of 10 chance you will be found out and given the penalty the assigner chooses.

Each penalty unit the assigner gives you will cause you to lose 10 points from your final earnings.

Use the buttons below to indicate whether you wish to take.

○ Don't Take (0)
○ Take (1)

Assigner

You will be the assigner in your group.

You start with 80 points.

You can now choose how many penalty units you wish to assign to player 1 if they choose to take and are found out (which will happen in 9 out of 10 cases when they chose to Take).

Each penalty unit will cause player 1 to lose 10 points from their final total if they are found out and cost you 2 points.

Use the buttons below to select a number of Penalty Units between 0 and 10.

○ 0 Penalty Units (0)
○ 1 Penalty Unit (1)
○ 2 Penalty Units (2)
○ 3 Penalty Units (3)
○ 4 Penalty Units (4)
○ 5 Penalty Units (5)
○ 6 Penalty Units (6)
○ 7 Penalty Units (7)
○ 8 Penalty Units (8)
○ 9 Penalty Units (9)
○ 10 Penalty Units (10)

Player 2

You will be player 2 in this interaction.

You start with 80 points.

If player 1 chooses to take, you will lose 30 points. If they choose to not take, you will keep your full 80 points.

As player 2, you will not make any decisions, however we are interested in what decisions you think other people should make.

What do you think the "fair" level of penalty units is for the assigner to choose? (Reminder, each penalty unit causes player 1 to lose 10 points from their final earnings if they choose to take and are found out, which happens in 9 out of 10 times they choose to take)

______ I think the fair level of penalty is.... (1)

Appendix 7 : Experiment 3 Instructions, Crowding Out

You are eligible to participate! Thanks for being part of our research!

We are interested in the way that individuals make decisions in social situations. You must be at least 18 years old to participate in this study, if you are not, please close this window and exit the HIT.

Your participation will take no more than 10 minutes.

In this study, you will be anonymously put into a group with other people recruited from Mechanical Turk and four of you will participate in an interaction.

During the interaction you will earn points based on your decisions and the decisions of other individuals in your group.

We will pay everyone who participated 1 cent per point they earned in the interaction via the bonus system on Mechanical Turk. You can earn up to $1.30 extra for participating!

This is in addition to the base compensation of 50 cents that you receive from just accepting the HIT.

You will be informed of your bonus payment once everyone signed up for the HIT has finished making decisions. You can expect this bonus payment within ten (10) days of your participation in this study.

Instructions

Everyone starts in groups of four players.

The other three people in your group will be real individuals also recruited from Mechanical Turk.

Every player will start the interaction with a balance of 100 points.

At the start of the interaction, each person will be given a role: one person will be assigner 1, one person will be assigner 2, one person will be player 1, one person will be player 2.

The interaction will proceed in order:
1) First, assigner 1 will make a decision
2) Player 1 will make a decision
3) Player 2 will be a passive observer
4) Assigner 2 will make a decision

The decisions are:
1) First, assigner 1 will choose a **penalty unit level**, this can be any number between 0 and 6.

2) Player 1 will see the penalty level that the assigner imposed and choose between **"Take"** and **"Don't Take."**
a) If player 1 chooses **"Take"**, then **player 1 will GAIN 30 points** and **player 2 will LOSE 40 points**.
b) If player 1 chooses **"Don't Take"** everyone will keep their 100 points.

3) If player 1 chooses "Take" they may be **found out**. This will happen in **3 out of 4** cases.
a) If player 1 is found out, they will be given the penalty chosen by assigner 1. For every penalty unit that player 1 receives, player 1 will LOSE 10 points.
b) If player 1 is found out, assigner 1 will be charged 2 points per unit they give to player 1.
c) Assigner 1 will not be charged if player 1 is not found out or does not take.

4) If player 1 chooses "Take" and is **found out**, assigner 2 will also get to make a decision.
a) Assigner 2 will see how many penalty units assigner 1 gave to player 1 and will be able to give player 1 between 0 and 6 additional penalty units.
b) Assigner 2 will be charged 2 points per unit they give to player 1.
c) A player 1 who chose to take and is found out will receive the **sum of the penalty units given by assigner 1 and assigner 2, for each penalty unit, player 1 will lose 10 points.**

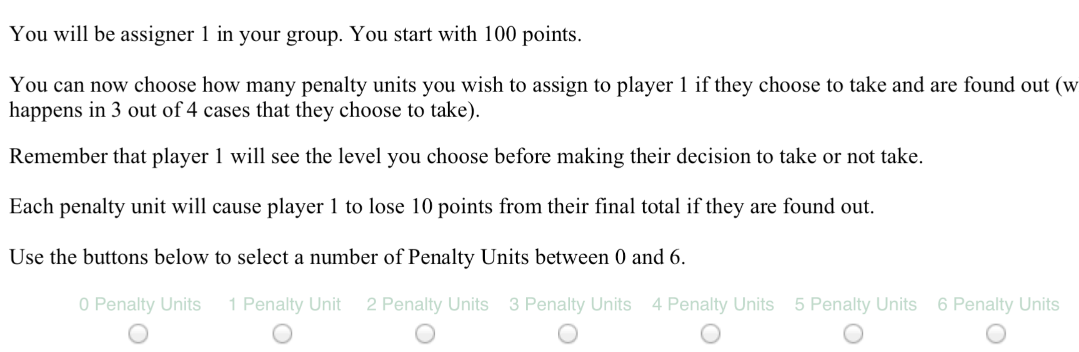
Instructions Summary

Brief Summary of the Interaction
1) Assigner chooses a penalty unit level
2) Player 1 sees the assigner's choice and chooses to Take or Don't Take
3) If player 1 chooses to take, player 1 is found out in 3 out of 4 cases
a) If player 1 is found out, they are assigned the penalty the assigner chose
4) If player 1 is found out, assigner 2 will see how many penalty units assigner 1 gave and can give player 1 additional penalty units

Brief Summary of Payoffs
1) Everyone starts with 100 points.
2) If player 1 chooses to Take, player 2 loses 40 points and player 1 gains 30 points.
3) If Player 1 is found out they are assigned penalty units by both assigners.
a) For each penalty unit, player 1 loses 10 Points. Each assigner pays 2 points for every penalty unit they end up giving to player 1.

To make sure that everyone understands the rules of the interaction, you will now take a short quiz.

You must answer these questions correctly in order to be able to participate in the interaction!



Assigner

You will be assigner 1 in your group. You start with 100 points.

You can now choose how many penalty units you wish to assign to player 1 if they choose to take and are found out (which happens in 3 out of 4 cases that they choose to take).

Remember that player 1 will see the level you choose before making their decision to take or not take.

Each penalty unit will cause player 1 to lose 10 points from their final total if they are found out.

Use the buttons below to select a number of Penalty Units between 0 and 6.

0 Penalty Units	1 Penalty Unit	2 Penalty Units	3 Penalty Units	4 Penalty Units	5 Penalty Units	6 Penalty Units
○	○	○	○	○	○	○

Player 1

You will be player 1 in your group. You start with 100 points.

You can now choose to Take or Not Take.

If you take, you will gain 30 points, player 2 will lose 40 points.

If you choose to take, there is a **3 out of 4 chance you will be found out and given assigner 1's penalty. Remember, that if you are found out, assigner 2 will also have a chance to give you penalty units.**

Each penalty unit assigners give you will cause you to lose 10 points from your final earnings.

You can make a decision that depends on the action that assigner 1 takes. You will do so as follows: you enter your **maximum acceptable possible penalty from assigner 1.** This is the penalty level such that...

...if assigner 1 chooses a penalty unit level above this, you would prefer not to take

...if assigner 1 chooses this penalty unit level or below this, you would prefer to take.

Because you do not always get caught, you can think of this as entering the maximum level of risk you would be willing to bear to use the take option. Also...

...if you would prefer to never take, you should choose a **maximum acceptable possible penalty from assigner 1** of 0.

...if you would prefer to take no matter what, you should choose a **maximum acceptable possible penalty from assigner 1** of 6.

Use the buttons below to select your **maximum acceptable possible penalty from assigner 1**.

0 Penalty Units	1 Penalty Unit	2 Penalty Units	3 Penalty Units	4 Penalty Units	5 Penalty Units	6 Penalty Units
○	○	○	○	○	○	○

Player 2

You will be player 2 in this interaction. You start with 100 points.

If player 1 chooses to take, you will lose 40 points and player 1 will gain 30 points. If they choose to not take, you will keep your full 100 points.

As player 2, you will not make any decisions, however we are interested in what decisions you think **other** people should make.

What do you think is the "fair" level of penalty units that player 1 should receive? (Reminder, each penalty unit causes player 1 to lose 10 points from their final earnings if they choose to take and are found out, which happens 3 out of 4 times they choose to take)

	0	1	2	4	5	6	7	8	10	11	12
I think the fair level of penalty is....											

Assigner 2

You will be assigner 2 in your group. You start with 100 points.

You can now choose how many penalty units you wish to assign to player 1 if they choose to take and are found out (which happens in 3 out of 4 cases that they choose to take).

Each penalty unit will cause player 1 to lose 10 points from their final total if they take and are found out.

Use the buttons below to select a number of Penalty Units between 0 and 6.

Remember that assigner 1 has already made a choice to give penalty units to player 1 before you do, so you will now be asked to make a choice for each action assigner 1 could have taken.

If player 1 is found out and **assigner 1 chooses to give 0 Penalty Units**, I will give player 1...

0 Penalty Units	1 Penalty Unit	2 Penalty Units	3 Penalty Units	4 Penalty Units	5 Penalty Units	6 Penalty Units
○	○	○	○	○	○	○

If player 1 is found out and **assigner 1 chooses to give 1 Penalty Units**, I will give player 1...

0 Penalty Units	1 Penalty Unit	2 Penalty Units	3 Penalty Units	4 Penalty Units	5 Penalty Units	6 Penalty Units
○	○	○	○	○	○	○

If player 1 is found out and **assigner 1 chooses to give 2 Penalty Units**, I will give player 1...

0 Penalty Units	1 Penalty Unit	2 Penalty Units	3 Penalty Units	4 Penalty Units	5 Penalty Units	6 Penalty Units
○	○	○	○	○	○	○

If player 1 is found out and **assigner 1 chooses to give 3 Penalty Units**, I will give player 1...

0 Penalty Units	1 Penalty Unit	2 Penalty Units	3 Penalty Units	4 Penalty Units	5 Penalty Units	6 Penalty Units
○	○	○	○	○	○	○

If player 1 is found out and **assigner 1 chooses to give 4 Penalty Units**, I will give player 1...

0 Penalty Units	1 Penalty Unit	2 Penalty Units	3 Penalty Units	4 Penalty Units	5 Penalty Units	6 Penalty Units
○	○	○	○	○	○	○

If player 1 is found out and **assigner 1 chooses to give 5 Penalty Units**, I will give player 1...

0 Penalty Units	1 Penalty Unit	2 Penalty Units	3 Penalty Units	4 Penalty Units	5 Penalty Units	6 Penalty Units
○	○	○	○	○	○	○

If player 1 is found out and **assigner 1 chooses to give 6 Penalty Units**, I will give player 1...

0 Penalty Units	1 Penalty Unit	2 Penalty Units	3 Penalty Units	4 Penalty Units	5 Penalty Units	6 Penalty Units
○	○	○	○	○	○	○