Personality and Social Psychology Bulletin
2015, Vol. 41(5) 629–642

DOI: 10.1177/0146167215571090
pspb.sagepub.com

Brutality Under Cover of Ambiguity: Activating, Perpetuating, and Deactivating Covert Retributivism

Katrina M. Fincher[1] and Philip E. Tetlock[1]

Abstract

Five studies tested four hypotheses on the drivers of punitive judgments. Study 1 showed that people imposed covertly retributivist physical punishments on extreme norm violators when they could plausibly deny that is what they were doing (attributional ambiguity). Studies 2 and 3 showed that covert retributivism could be suppressed by subtle accountability manipulations that cue people to the possibility that they might be under scrutiny. Studies 4 and 5 showed how covert retributivism can become self-sustaining by biasing the lessons people learn from experience. Covert retributivists did not scale back punitiveness in response to feedback that the justice system makes false-conviction errors but they did ramp up punitiveness in response to feedback that the system makes false-acquittal errors. Taken together, the results underscore the paradoxical nature of covert retributivism: It is easily activated by plausible deniability and persistent in the face of false-conviction feedback but also easily deactivated by minimalist forms of accountability.

Keywords

ambiguity, self-deception, impression management, retribution, punishment

Received August 3, 2014; revision accepted January 8, 2015

Psychologists have long suspected that public policy preferences are shaped by factors that people are either unaware of or unwilling to acknowledge. One of the earliest political psychologists, Harold Lasswell (1930/1986), advanced a pithy version of this view: "The distinctive mark of the homo politicus is the rationalization of the displacement of private motives in terms of public interests," (p. 262). This programmatic statement encouraged psychologists to adopt the working hypothesis that the reasons people offer for their policy preferences are often not the true drivers of those preferences—encouragement that no longer seems necessary given the widespread skepticism of introspective reports in contemporary research (Bersoff, 1999; Buehler, Griffin, & MacDonald, 1997; Dawson, Gilovich, & Regan, 2002; Klein & Kunda, 1992; Monin & Miller, 2001; Mullen & Skitka, 2006; Pezzo & Pezzo, 2007; Redlawsk, 2002; Rousseau & Tijoriwala, 1999; Tykocinski & Steinberg, 2005). Following through on this Lasswellian mission does however raise nontrivial methodological challenges. Penetrating veils of rationalizations in political life requires developing and validating research tools for disentangling true reasons from expressed reasons (Dawson et al., 2002; Monin & Miller, 2001; Norton, Vandello, & Darley, 2004).

The current article expands the detection-of-covert-motives agenda into the domain of punitiveness judgments, the willingness to impose penalties on norm violators. Past work in this area has focused on such suspect motives as racial animus (Sniderman, Brody, & Tetlock, 1991; Sniderman, Piazza, Tetlock, & Kendrick, 1991), system justification (Kay & Jost, 2003; Kay, Jost, & Young, 2005), and belief in a Just World (Lerner, 1980; Lerner & Lerner, 1981). Building on this work, the suspect motive in the current article is covert retributivism. Our working hypothesis is that in certain classes of situations, people want to impose retributive penalties on norm violators that exceed the limits that society places on punishment, and when that happens, people look for alternative covert ways of inflicting pain on norm violators.

Research at a psychological level of analysis has shown that when people see the perpetrator as culpable (Alicke, 2000; Alicke & Yurak, 1995; Malle & Guglielmo, 2012; Malle, Guglielmo, & Monroe, 2012; Weiner, Graham, Peter, & Zmuidinas, 1991; Weiner, Graham, & Reyna, 1997), their punishment recommendations are influenced by the

[1]University of Pennsylvania, Philadelphia, USA

Corresponding Author:
Katrina M. Fincher, Psychology Department, University of Pennsylvania, 3720 Walnut Street, Philadelphia, PA 19104, USA.
Email: katrinaf@sas.upenn.edu

retributivist goal of rectifying wrongs by imposing penalties proportionate to the harm done to victims (Carlsmith, 2006; Tetlock et al., 2007) as well as to society as a whole (Atran & Ginges, 2012; Ginges, Atran, Medin, & Shikaki, 2007; Haidt & Graham, 2009). This drive to punish is so strong that individuals will punish even when it is personally costly (Fehr & Gächter, 2002; Henrich et al., 2006) and when doing so they violate the norms of procedural justice (Skitka & Houston, 2001; Skitka & Mullen, 2002). The tenacity of retributivism—and its linkage to powerful affective reactions such as moral outrage (Goldberg, Lerner, & Tetlock, 1999)—makes it difficult to reduce to coldly cognitive utilitarian-deterrence calculations (Carlsmith, 2006; Darley & Pittman, 2003; Goldberg, Lerner, & Tetlock, 1999; McKee & Feather, 2008; Weiner et al., 1997). Proposed explanations for retributivism have included the desire to restore moral balance and reaffirm group norms (Durkheim, 1933; Heider, 1958; Tyler & Boeckmann, 1997; Vidmar & Miller, 1980; Wenzel & Thielmann, 2006) as well as micro-economic and evolutionary accounts that stress the adaptive value of sending out "don't-mess-with-me" signals (Fehr & Fischbacher, 2004; Sell, Tooby, & Cosmides, 2009).

It is useful however to bear in mind that retributivism does not trump all other considerations. Many factors come into play in shaping punitiveness judgments (Okimoto & Wenzel, 2011; Wenzel, Okimoto, Feather, & Platow, 2008). For instance, work on restorative justice suggests that people also care a great deal about restoring relationships between victims and perpetrators (Okimoto & Wenzel, 2009; Wenzel, Okimoto, Feather, & Platow, 2008). This focus on restoration and forgiveness implies a provocative mirror-image conjecture to the one underpinning the current studies: namely that there may also be situations in which state-imposed punishment exceeds what people think appropriate, and people engage in covert forms of leniency.

It is also useful to keep in mind that retributivism translates into punishment (as opposed to person-to-person revenge) only when there is political and institutional oversight of the translation process. Constitutional bans on cruel and unusual punishment underscore this point. "Punishment" refers to penalties imposed by *legitimate* third parties, such as judges, and implemented by agents of the state (Durkheim, 1933; Weber, 2009). Therefore, to understand the social psychology of punishment, we need to explore the interplay between the micro-level of analysis, the punitive preferences of individual citizens, and the macro-level of analysis, the societal norms that regulate "who can do what to whom for which transgressions."

One example of cross-level-of-analysis interplay arises when human nature changes more slowly than do the cultural-political norms regulating punishment. Work in historical sociology has revealed a rather dramatic trend toward restricting institutionally permissible forms of punitiveness, especially in what Henrich, Heine, and Norenzayan (2010) have labeled the WEIRD world (Western, Educated, Industrialized, Rich, Developed). With notable cultural-religious exceptions (Edwards, 1995; Ellsworth & Gross, 1994; Hartman, 1983), retributivism, as gauged by institutionally orchestrated displays such as public floggings and executions, has declined sharply over the last five centuries (Elias, 1969; Garland, 1991; Pinker, 2011; Spierenburg, 1984). In the early 21st century, violent retributivism has been officially condemned by the United Nation's Universal Declaration of Human Rights. These historical trends are often explained by invoking evolving societal norms toward violence and coercion, which have necessitated removing punishment from public view (Elias, 1969; Garland, 1991; Pinker, 2011)—and shifting the focus of punishment from the body to the mind (Foucault, 1977).

This analysis raises many questions, one of which is as follows: What happens when the retributive penalties that individuals want to impose exceed the level of punishment that society permits? Thus far, this question has attracted more sociological than psychological attention (Garland, 1991). For instance, Merton's (1949) classic analysis of delinquency suggests that when individual goals align with societal norms, people pursue their goals overtly—and when individual motivations do not align with societal norms, people pursue their goals covertly. Pursuing goals covertly requires attributional ambiguity—an environment in which multiple factors could potentially motivate a given behavioral response. Attributional ambiguity therefore creates plausible deniability about the true drivers behind one's conduct and thus enables specious reasoning when justifying questionable goals.

In theory, attributional ambiguity can enable people who want to be harsh to disguise their harshness—covert retributivism—or enable people who want to be lenient to disguise their leniency—covert forgiveness. In this article, we focus on covert retributivism which we suggest is likely when two preconditions have been met: (a) retributive motives exceed socially allowed punishment and (b) there is attributional ambiguity. For example, person could engage in covert retributivism by arguing that "I am not sentencing criminals to be assaulted or raped in prison. I am sentencing them to prison sentences which they deserve and it has unfortunately proven impossible to prevent incarcerated criminals from preying on other incarcerated criminals."

In the real world, it is often impossible to infer true intent from policy preferences because policies are multidimensional both in their content and consequences. But in the laboratory it is possible. Psychologists have created research paradigms that are well suited for uncovering covert motives in punitive practices—paradigms that rest on the attributional ambiguity premise that people are likelier to act on unacceptable impulses when they are placed in situations in which they can plausibly attribute their conduct to an acceptable alternative explanation (Langer, Fiske, Taylor, & Chanowitz, 1976; Norton et al., 2004; Snyder, Kleck, Strenta, & Mentzer, 1979). Applying this logic to covert

retributivism, people who feel that society is not tough enough on criminals will look for covert ways of inflicting additional pain and humiliation, such as prison violence or body cavity searches (Foucault, 1977; Van den Haag, 1991)—and when opportunities to do that arise in experiments, they will make use of them:

> **Hypothesis 1 (H1):** When punitive desires exceed societal constraints, people will covertly support violent punishment in proportion to the outrageousness of the offense.

The research literature offers little guidance on the pervasiveness or strength of covert retributivism. In this article, we focus on two unexplored moderating variables for the manifestation of this motive: (a) external accountability and (b) error feedback (i.e., information about the rates of false convictions and failures to convict).

Attributional ambiguity and accountability are intertwined constructs. On the one hand, attributional ambiguity makes it more difficult to hold decision makers accountable for their preferences. On the other hand, accountability may cause decision makers to worry that their true motives are detectable. Our working hypothesis is that when people feel they are under accountability scrutiny, they will quickly jettison covert retributivism. The reasoning is simple: Covert retributivism needs to remain covert. People who are making retributivist judgments under cover of attributional ambiguity should be sensitive to the possibility that others are aware that they are covertly satisfying socially unacceptable retributivist preferences.

However, it is unclear just how sensitive covert retributivists are to being monitored. Drawing on Lerner and Tetlock (1999) who postulated a continuum of accountability manipulations—from the minimalist to the maximalist—we chose to begin our search for checks and balances at the minimalist end. If covert retributivism can be attenuated by minimalist forms of accountability—by subtle hints that someone might be monitoring one's decisions, with no actual monitoring or requests for justifications or consequences for poor justifications—then there is no need to deploy more demanding forms of accountability. Covert retributivism may be easy to activate and perpetuate but also easy to deactivate, which leads to our second hypothesis.

> **Hypothesis 2 (H2):** Even minimal accountability manipulations will be sufficient to blow the cover off covert retributivism and induce punishment setters either to (a) acknowledge the true retributivist goals they are pursuing and/or (b) reconsider their non-retributivist rationales for pain-inflicting decisions.

This second hypothesis highlights the role of second-order accountability (i.e., "who is watching the watchers?") in containing covert retributivism. However, as long as covert retributivism is covert (i.e., difficult to detect by both those making decisions and those observing them), it may be difficult to learn from policy feedback. Put simply, to the degree an individual is unaware that she is trying to achieve a goal, she will be less likely to notice when she has gone too far in pursuit of that goal—and less likely to learn from feedback.[1]

One key trade-off confronting individuals making punishment judgments involves setting the thresholds for determinations of guilt. Threshold setting requires balancing the need to avoid making false-positive errors in criminal-justice systems (i.e., punishing the innocent) and false-negative errors (i.e., acquitting the guilty).

From a traditional game-theory perspective that assumes people are intuitive Bayesians who at least roughly update their policy preferences in response to environmental feedback, norm enforcers should learn fairly quickly, and equally quickly, from erroneous convictions and erroneous acquittals (Axelrod & Hamilton, 1981; Nowak, 2006; Nowak & May, 1993). False-conviction errors highlight the danger of harming the innocent, which should reduce enthusiasm for both covert and overt retributivism, whereas false-acquittal errors highlight the danger of the guilty escaping justice, which should increase enthusiasm for both covert and overt retributivism.

However, there are two distinct psychological reasons for hypothesizing that attributional ambiguity creates obstacles to learning. First, basic research on classical and operant conditioning has repeatedly shown that animals, humans included, have more difficulty learning new associations in noisy environments in which the true signal is embedded in distractor stimuli. For instance, classical conditioning researchers have demonstrated blocking and backward-blocking effects in which it is harder to learn associations between an action and a consequence when the consequence is associated with multiple actions (Kamin, 1969; Rescorla, 1970, 1988). Attributional ambiguity degrades the link between actions and consequences by creating multiple explanations for actions, and thus uncertainty about the true stimulus drivers of behavior. In essence, when sanctions are overt (e.g., people know that they are assigning norm violators to violent prisons precisely because they are violent), the feedback connection between having imposed a violent sanction and the consequences of such a sanction is transparent because there is no competing or distracting information. When there is attributional ambiguity, there is competing information (the socially acceptable pretexts for choosing violent prisons provided by the attributional ambiguity manipulation), and these competing cues can slow or even prevent learning. This suggests that impairments of learning occur due to backward blocking and should, therefore, apply equally, independent of the information about the consequences of such sanction decisions, all of which leads to the next hypothesis:

> **Hypothesis 3a (H3a):** Covert retributivists will find it equally hard to learn (adjust policy preferences) from

false-positive errors (sending the innocent to violent prisons) and false-negative errors (failing to incarcerate the guilty)—errors that are inevitable in any criminal-justice system.

By contrast, a motivated-reasoning perspective predicts asymmetric disruptions of learning (Kunda, 1990). In this view, people are often adept at maintaining false beliefs in spite of contradicting evidence—and are slow to learn lessons from history that they do not wish to learn. Applying motivated reasoning to punishment when retributivist motivation to punish is high suggests that people should learn quickly about false negatives (failures to convict the guilty) because this information will encourage people to become even more punitive (Goldberg et al., 1998), which is motivationally congruent. By contrast, people should be slow to learn news about false positives (wrongful convictions), that on logical grounds should encourage people to become more circumspect in meting out punishment.

Importantly, motivated reasoning can only occur where there is ambiguity about the plausibility of competing explanations (Kunda, 1999), therefore it should only impair learning for covert actions. Applying this to punishment suggests that ignoring news about wrongful convictions should be cognitively easier for covert than for overt retributivists because covert retributivists see no need to acknowledge that the punitive policies they are embracing are part of the punishment (e.g., "I endorse assignment to this prison because it has lower recidivism or escape rates, not because it is more violent"). Given that the physical component of the punishment has the psychological status of an unintended but unavoidable side effect, there is less psychological pressure to revise that preference in response to news about the risk of false conviction, all of which leads to the following hypothesis:

Hypothesis 3b (H3b): Covert retributivists will learn from failures to convict (misses), but not from false convictions (false positives).

This article reports five studies that tested these four hypotheses about the interplay between retributivist drives and the social context in which people make punishment recommendations. These studies examine what occurs when punitive drives exceed the limits of the punishments society allows. Study 1 tested the attributional ambiguity hypothesis that although people in early 21st century America declare an in-principle aversion to covert corporal punishment (violent prisons), they will often assign violators of important moral norms (e.g., murderous pedophiles) to violent prisons when they can disguise that preference under other rationales (H1). Studies 2 and 3 tested the power of minimalist-accountability manipulations to contain covert retributivism (H2). Studies 4 and 5 tested the learning-blockage and motivated-reasoning hypotheses (H3a and H3b) that participants will be (a) less likely to taper punishment recommendations in response to a false-positive miscarriage of justice when there is attributional ambiguity about the connection between those recommendations and the consequences than when sanctions are direct; (b) more likely to intensify punishment in response to a false-negative miscarriage of justice when there is attributional ambiguity than when sanctions are direct.

Study 1

Study 1 randomly assigned participants to conditions in which they role-played a judge who had the option of assigning defendants to violent prisons either covertly (under cover of attributional ambiguity) or overtly (no cover). We operationalized covert retributivism by adopting a social psychology paradigm widely used to explore covert racial and gender biases.

In the original task (Monin & Miller, 2001; Norton et al., 2004), participants selected and then justified the choice of a job candidate based on one of two resumes which had been pretested to be comparable, with each candidate possessing offsetting strengths and weaknesses. In a between-subjects design, the experimenter counterbalanced which resume was associated with the disadvantaged group's candidate and which with the favored group's candidate. This design made it possible to determine both whether discrimination occurred and whether the justification of choice was untrue. Like the casuistry task in the racial domain (Norton et al., 2004), our covert punishment task used two sets of stimuli and counterbalanced the crucial dimension across sets. As Figure 1 shows, the task used two prisons that had been pretested to be equally undesirable. We manipulated four features of the prisons: cost, security, recidivism, and rate of general education diploma (GED; high school equivalency) completion. We then varied (between subjects) which prison was associated with violent sanctions. Given that the prison linked to corporal harm was counterbalanced across participants, by aggregating across subjects we could measure (a) the degree to which participants collectively (though not individually) harbor violent preferences and (b) the degree to which participants collectively (though not individually) generate justifications for their preferences that align with the true (experimentally revealed) drivers of their preferences. We expected greater punitiveness when participants had the opportunity to keep their preferences to inflict physical harm covert (because they could justify pro-violent choices by invoking other features of the prison). "Greater punitiveness" can be gauged relative to a baseline or overt-sanction condition.

Method

Participants. Participants were 300 (167 male and 133 female) workers on Mechanical Turk who participated for

Prison A	Prison B
The prison is inexpensive for tax payers. No funds will have to be reallocated	The prison is expensive for tax payers. Funds will have to be cut from things such as schools programs to support the inmates
The prison has had a great number of security problems. 4 prisoners have successfully escaped in the past 10 years.	incarceration. The Prison is extremely secure. No prisoner has ever escaped
Violence Manipulation Which prison is violent is counterbalanced	
The recidivism rate is 67%. This means 67% of individuals who are released will return to prison.	The recidivism rate is 61%. This means 61% of individuals who are released will return to prison.
12% of inmates receive a GED while incarcerated.	7% of inmates receive a GED while incarcerated.

Figure 1. Prison stimuli used in the study.

US$1.00 payment. The sample was constrained to American participants. The average participant age was 32.8 years old (SD = 10.3), middle class (M = 2.4 on a 5-point scale), and slightly liberal (M = 4.3 on a 7-point scale).

Procedure. Study 1 tested punitive preferences under conditions of high and low ambiguity. All participants read about a crime and then completed a prison assignment task. Based on prior work on victim identifiability (Goodwin & Landy, 2014; Loewenstein & Small, 2007) and punishment (Lively et al., 2014), the crime was designed to provoke maximum moral outrage: the rape and torture of a 7-year-old girl, Suzie. Participants then made their punishment decision.

For the punishment decision, participants were randomly assigned to conditions in which they could either covertly or overtly sentence the perpetrator to a prison associated with a high risk of violence. Across conditions, the features of the punishment decision were identical; however, in the covert condition, the choice was embedded among other factors so that the participant's intentions were not clear. Participants sentenced the pedophile to one of two prisons: A or B. Each prison, A and B, had five attributes. Four of the attributes were basically irrelevant to covert retributivism: cost (expensive or cheap), security (0 or 4 escapes in last 20 years), recidivism rate (67% or 61%), and GED completion rate (12% or 7% of inmates). The levels of each of these four attributes were pre-scaled to create equally desirable prisons (e.g., variations in cost and security were roughly equally offset by the variations in recidivism and GED completion rates). In a between-subject design, embedded randomly among the other dimensions, was a fifth dimension: brutality within the prisons ("there is a great deal of gang violence within the inmate population and as a result there are many brutal beatings, occasional rapes, and related hospitalizations" vs. "there is little to no gang violence within the inmate population, and as a result, there are few beatings and no rapes or related hospitalizations").

The gauge of covert retributivism was the degree to which the presence of the fifth factor, prison violence, increased selection of that prison ("covert" retributivism because people could always argue that the other four features—and not violence—were driving their choices). After selecting a prison, participants were asked to select which of the five dimensions most influenced their choice.

In the overt condition, participants saw three pairs of prisons that varied along the dimensions of security, cost, and inmate brutalization, respectively. Participants selected a prison for the perpetrator from each pair of prisons. Descriptions of each factor were identical to descriptions in the covert punishment condition.

Results

Selection task. In the overt condition, 91.2% (124/136) and 89.7% (122/136) of participants preferred the safer and cheaper prisons, respectively—and 81.6% (111/136) preferred the less violent prison. This suggests that an overwhelming majority preferred the non-violent prison when it was transparent which factors were driving their preferences.

In the covert condition, we began by testing whether the prisons (controlling for which one was designated violent in counterbalancing) were equally matched: 49% (85/175) of participants picked Prison A over B, a non-significant difference, $\chi^2(1, 174) = .03$, $p = .87$. We then tested and found support for H1 that participants making choices under

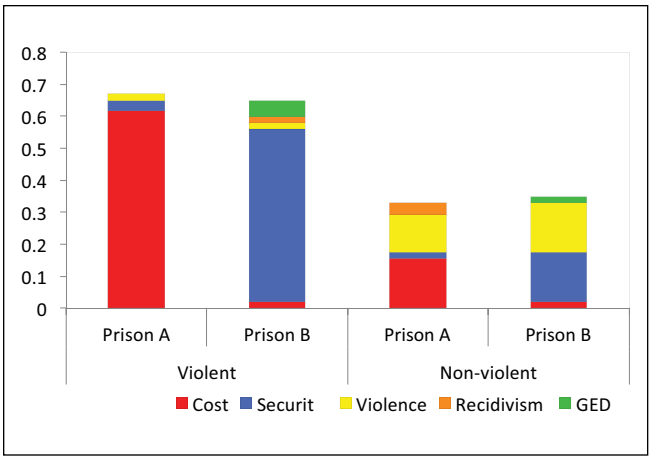

Figure 2. Participants' response to the prison task in Study 1.
Note. The bars represent the proportion of participants selecting Prison A or B as a function of the presence of violent sanctions. The colors represent justification for prison choices as a function of subjects endorsing option as the most important feature in decision.

attributional ambiguity would prefer the violent prison, $\chi^2(2, 173) = 24.02$, $p < .01$. When Prison A was violent, 68% (55/80) picked it, $\chi^2(1, 173) = 5.1$, $p = .024$. When Prison B was violent, only 32% (65/95) picked Prison A, $\chi^2(1, 173) = 5.76$, $p = .016$.

Justification of choices. Participants invoked cost and security as the key drivers of their choices. For Prison A—which was superior on the dimension of lower cost—71% of participants offered cost as the justification of their choices. For Prison B—which was superior on security—62% offered security as their justification. Most revealingly, only 5% of participants who selected the violent prison invoked inmate violence as a justification, whereas 35% of those who selected the non-violent prison invoked inmate violence as their justification. Endorsers of the violent option had significantly different patterns of choice justifications as a function of when Prison A or Prison B was more violent, $\chi^2(1, 119) = 46.28$, $p < .01$, whereas endorsers of the non-violent prison did not alter their justifications, $\chi^2(1, 54) = .43$, $p = .512$. This change in justifications, in conjunction with the change in preferences, suggests that endorsers of the violent prison were misattributing their preferences to features of prisons unrelated to violence. See Figure 2 for an illustration of the relationship between the selection task and choice justification.

Study 1 Discussion

Study 1 showed that people select violent sanctions when they had the opportunity to misattribute their selection to a socially acceptable rationale. People are much more willing to endorse violent sanctions covertly than overtly. It would appear that although retributivist motives are constrained by societal norms, people who are sufficiently morally outraged find alternative outlets for expressing these motives.

In dissecting covert retributivism, it is important to distinguish primary and secondary sanctions. Primary sanctions are those directly inflicted by the state, such as imprisonment, whereas secondary sanctions are foreseeable side-effects of primary sanctions, such as assignment to a violent prison. A follow-up survey experiment shows that people are wary of secondary sanctions that are explicitly spelled out. We asked 200 participants to endorse or reject the primary or secondary sanctions for the crime described in Study 1. As in Study 1, when secondary sanctions were explicitly foreseen, only 8.4% (18/200) selected the violent prison. However, when the same secondary consequence was transformed into a primary sanction—and inflicted by the state as corporal punishment—42% (89/200) of participants endorsed the use of this sanction.

These post-experimental data in conjunction with the results of Study 1 suggest that, for certain classes of crimes, respondents would prefer that the state administer beatings over other criminals doing the job. If the state would not do what respondents feel it should do, overt and covert responses differ. Overtly, participants dismiss secondary sanctions as wrong, while covertly participants respond as though criminals doing it is better than nothing at all. Consistent with H1, preferred punishments do sometimes exceed socially available punitive outlets and when this tension arises, people who oppose secondary sanctions when there is no attributional ambiguity become receptive to the idea when there is attributional ambiguity.

Study 2

Study 1 revealed an inconsistency between moralistic principles and punitive practice. We argue that this discrepancy will arise whenever attributional ambiguity creates plausible deniability that participants can use to obfuscate their underlying retributivist goal. However, Study 1 fails to specify the boundaries of covert retributivism. Study 2 tests H2a: Whether small accountability nudges can sensitize covert retributivists to the true reasons behind their choices and disrupt the rationalization process documented in Study 1. If true, people should begin to acknowledge the desire to inflict pain as the real driver behind their punishment selection. However, if people do not have introspective access to their retributivist motives—or are unwilling to budge from the socially desirable response options—the accountability nudge in Study 2 should have no effect.

Method

Participants. In all, 223 (121 female, 102 male) undergraduates at a northeastern university completed the study in individual computer terminals and were each paid US$10. The average age of participation was 20.1 (SD = 1.0).

Participants reported that their political attitudes were liberal (M = 5.6 on a 7-point scale). They were recruited through a behavioral lab on campus. The study was run for 1 week, and the sample size was determined by the number of sign-up volunteers.

Procedure. Participants sat at computer terminals separated by partitions, so that they could not see one another. The study used a 2 (Painful Sanction: A is painful or B is painful) × 2 (Accountability: minimalist, none) factorial design. The structure of the study was identical to the covert condition in Study 1 except, prior to choice justification, there was a self-reflection manipulation.

Participants read about the violent and gruesome murder of 10 children (a single identifiable victim was described). Because the selection task involved choosing a method of capital punishment, the perpetrator and his actions were designed to elicit maximally punitive sentiments. After reading the description, punitive sentiments were measured indirectly using assignment to execution facility.

Participants could assign the convicted felon to one of two execution facilities which were created by pairing variables pretested as important (cost and risk in transporting felon) and variables pretested as unimportant (nail discoloration and unpleasant scent), so that each execution method dominates on one variable. In these variable pairs, one option was clearly superior and one option was clearly inferior for each execution method. In a between-subject design, embedded randomly among the other dimensions, was a fifth dimension: pain felt during execution ("a moderate to great deal of pain" or "a small to moderate degree of pain"). This allowed participants to select the painful execution, but attribute it to a social acceptable rationale.

In the minimalist-accountability manipulation, after selecting an execution facility, but prior to justifying their choice, a research assistant turned on a computer camera, but did not set the camera to record. Participants knew the camera was not recording. Participants then justified their choice in the execution selection task using one of the five attributes that had been manipulated.

In the no accountability condition, the research assistant did not turn on the web camera and participants completed the execution selection task without interruption. During the debriefing, participants indicated the extent to which they experienced psychological discomfort from the study on a 0 to 100 scale.

Results

Although participants could select any of the five attributes as the primary reason for choosing Execution Method A or Execution Method B, the only attribute that we can say, with experimental precision, could be misattributed is "the degree of painfulness." Therefore, we recoded justification of choice as a binary variable which indicated if individuals used "the degree of painfulness" to justify choice. If individuals used "the degree of painfulness" to justify their choice, their response was coded as 1; if individuals used any other form of choice justification, their response was coded as 0. We then used binomial logistic regression to predict decision justification from execution choice, minimalist accountability, and painfulness of sanction.

The test entered three predictors (two manipulated independent variables: minimalist accountability and painfulness of sanction) and their interaction. Results suggest that the minimalist-accountability manipulation and the interaction between sanction painfulness and accountability were robust predictors of choice justification: bs = 3.3 and −1.7, standard errors (SE) = 1.1 and 0.84, Walds = 8.4 and 4.1, $p < .01$, $p < .05$. The main effect of the minimalist-accountability manipulation indicates that individuals used pain as a choice justification more frequently when there was even the hint of scrutiny that turning on the computer camera provided. The interaction indicates that in the minimalist-accountability condition, "the degree of painfulness" was most likely used as a justification when participants selected the painful execution; however, in the control condition, "the degree of painfulness" was likeliest to be invoked as a justification when participants selected the non-painful execution. The execution choice and condition were not even remotely significant, p = .44 and .45, as one would predict if randomization had been effective.

Study 2 Discussion

Study 2 revealed sharp boundary conditions on the effects of attributional ambiguity. Even minimalist-accountability manipulations can make people acknowledge violent punitive preferences they would otherwise conceal, which is not what we would expect either if people had no introspective access to their judgment process (and were unaware that they were doing anything impolitic) or if they were aware of what they were doing but felt they could plausibly insist that only socially desirable rationales were driving their choices.

Study 3

Study 2 suggests that even a slight accountability nudge can sensitize people to the true reasons behind their choices. But it did not demonstrate any effect on choice: that is, it demonstrates that people will only endorse covert sanctions when they see no hint that their decisions will be subject to any form of accountability scrutiny. Therefore, in Study 3, we tested H2b: that attributional ambiguity will not enable covert retributivism when people suspect any possibility of scrutiny (a minimalist form of accountability; Lerner & Tetlock, 1999).

Instead of the non-operational camera in Study 2, we created a different type of accountability nudge. We placed photos of human eyes in the header of the survey. Recent studies

have shown that people are surprisingly responsive to subtle cues of being watched, such as the presence of eye-like spots on the background of the computer on which they complete the task (Burnham & Hare, 2007; Haley & Fessler, 2005).

Method

Participants. In all, 206 (98 male and 108 female) students at a northeastern university completed the study in individual computer terminals and were each paid US$10. The average age of participation was 20.6 ($SD = 0.9$). Participants reported that their political attitudes tilted liberal ($M = 5.4$ on a 7-point scale). Participants were recruited through a behavioral lab on campus. The study was run for 1 week, and the sample size was determined by the number of students who signed up to participate.

Procedure. The study was a 2 (Prison type: A is violent or B is violent) × 2 (Accountability nudge: yes, no) factorial design. Participants were seated at computer terminals separated by partitions, so that they could not see each other. The design of the study was identical to the covert condition in Study 1, except the header of the survey varied across participants.

The content of the page header varied between subjects, and it was designed to manipulate participants' sense of being monitored by varying the presence of images of eyes. The page header consisted of a segment of 1 of 10 different modernist paintings. For half of the participants, the segment of these paintings used in the header was focused predominately upon the eyes. For the other half, the painting segments did not include eyes (or other signs of human presence), but were otherwise matched for the artist, the predominant colors, and the visual focus. An example pair of pictures is Vincent Van Gogh's 1889 sunflower and his 1889 self-portrait, both painted at St. Remey. Below the header image, the experimental task was listed. On the first page, participants read a description of a violent rape (used in Study 1). On the second page, participants completed the prison-selection task from the covert punitive condition in Study 1.

Results

As expected, the prisons were equally matched: 54% (122/225) of participants picked Prison A over B, a non-significant difference, $\chi^2(1, 225) = 1.60$, $p = .21$. Also, 56% (125/225) of participants picked the violent over the non-violent prison, a non-significant difference, $\chi^2(1, 225) = 2.78$, $p = .09$.

To capture the hypothesized interaction (H2b), we used a log-linear analysis. Participants who received the weak accountability nudge did not express covert punitiveness in response to attributional ambiguity, unlike those in the control group $G^2(3, 208) = 20.2$, $p < .001$. Whereas 69% of control participants selected the violent prison, only 42% of the minimalist-accountability participants selected the violent prison, $\chi^2(2, 225) = 15.61$, $p < .001$.

In the control condition, when violence was linked to Prison A, 71% picked Prison A. When violence was linked to Prison B, only 34% picked Prison A, a significant effect, $\chi^2(2, 113) = 14.89$, $p < .001$.

Conversely, in the minimalist-accountability condition, preferences were not significantly swayed by which prison was associated with violent outcomes. When violence was associated with Prison A, 47% picked Prison A. When violence was associated with Prison B, only 36% picked Prison A, a non-significant difference, $\chi^2(2, 112) = 3.95$, $p = .13$.

Study 3 Discussion

Consistent with H2b, the minimalist-accountability nudge reduced expressed preferences for the violent prison. Specifying exactly what is challenging. Impression management explanations (which focus on the views that others hold of the self) and intrapsychic explanations (which focus on preserving one's self-image) have long been known to be difficult, if not impossible, to disentangle (Tetlock & Manstead, 1985). For instance, reducing accountability to external audiences may make individuals less reflective and feel less accountable to internalized audiences. Conversely, lie-detection manipulations such as the bogus pipeline (Jones & Sigall, 1971) induce self-reflection and awareness and thus make individuals feel more accountable to both external and internal audiences (Tetlock & Manstead, 1985).

That said, it seems unlikely in this case that the behavioral changes induced by the minimalist-accountability manipulations are deliberate given that no reasonable person would consciously suppose that representations of human eyes on the computer screen were looking back at them. It also seems unlikely that minimalist-accountability inductions merely make people more self-aware and that this would induce already self-aware impression managers to rein in covert retributivism. Of the remaining explanatory options, we see one or a combination of the following two as the most plausible: (a) the effects of attributional ambiguity on punitiveness are the product of deliberate impression management but people unconsciously become more risk-averse impression managers in response to minimalist-accountability nudges which are associated with surveillance and potential punishment ("the guilty flee where no man pursueth" would be the Biblical version of this hypothesis); (b) the effects of both attributional ambiguity and minimalist nudges are mediated by processes outside of awareness and underscore the sophistication with which people can balance unconscious goals (Bargh & Chartrand, 1999; Hogan, 1982; Hogan & Blickle, 2013; Langer, Blank, & Chanowitz, 1978).

Regardless of the exact mix of impression management or intrapsychic processes underlying the behavioral change, the results of Study 3—in conjunction with those of Study 2—suggest that minimalist-accountability manipulations

(nudges) can both sensitize people to the reasons behind their choices as well as alter their actual choices.

Study 4

The first three experiments identified conditions under which people endorse covert sanctions toward isolated acts of criminality. In actual criminal-justice systems, however, decision makers assess a wide range of acts and occasionally get feedback on their mistakes. Experiment 4 focuses on the feedback effects of the error that the Anglo-American justice systems have traditionally assigned highest priority to avoiding: false-positive convictions of the innocent (Blackstone, 1875).

Both deterrence theories and retribution theories of justice suggest that learning of a false-positive conviction should incentivize those making punishment decisions to consider throttling back the severity of sanctions. For instance, rational-deterrence theory posits that lower conviction thresholds reduce the need for severe sanctions because would-be offenders are deterred by the high risk of punishment in itself (Polinsky & Shavell, 1998; Sunstein, Schkade, & Kahneman, 2000). Intuitive deterrence theorists should thus see false-positive convictions as a sign that conviction thresholds are low enough to reduce the need for severe (violent) punishment. In a similar vein, retributivist theories treat the (judicial or extra-judicial) infliction of harm on the innocent as a serious error and grounds for giving second thought to the severity of the sanctions being imposed, especially when those sanctions are seen as irreversible.

Experiment 4 explores responsiveness to false-positive errors by again asking participants to perform an overt or covert punishment-selection task before and after obtaining information about the false-positive propensities of the criminal-justice system.

Method

Participants. Participants were 242 (128 male and 114 female) workers on Mechanical Turk who participated for US$1.00 payment. The sample was 73% Indian, 16% American, and 11% other. The average age of participation was 35.7 (SD = 10.3), middle class (M = 2.9 on a 5-point scale), and very slightly liberal (M = 3.9 on a 7-point scale).

In the first part of the study, participants read a description of the crime and recommended a punishment (either covertly or overtly).[2] In the second part of the study, participants punished a second individual, for the same class of crime, however, had new information about the penal system's accuracy.

In the first phase of the study, participants read a description of a criminal conviction process that passed all the standard tests of procedural justice: The pre-trial investigation was fair (the investigators were diligent, unbiased, and followed the protocol exactly), the trial was fair (the judge and jury were impartial), the man had an outstanding defense attorney, and a jury of his peers found him guilty. The man was convicted of rape, assault, and kidnapping and sentenced to 5 years in prison. Participants had the opportunity to recommended extra punishment.

In the overt condition, participants were told that they had the option of adding an additional punishment to the sentence. Participants could choose between a prison term without beatings and a prison term that included a severe beating resulting in several weeks of hospitalization. This condition measured formal endorsement of corporal sanctions and therefore was substantially different from the overt condition in Study 1, which asked participants to select violent or non-violent prisons. In the covert condition, participants were given the opportunity to assign the perpetrator to one of two prisons, using the selection task from the covert condition of Study 1. As in Study 1, participants were randomly assigned to violent sanctions being associated with either Prison A or Prison B. To equate injuries across conditions, the violence in the prison was described as the "minimum of a severe beating resulting in several weeks of hospitalization."

In the second part of the study, participants learned that DNA testing had exonerated the man they had just either overtly or covertly punished. Participants then learned of another similar crime. The trial description included the same indications of procedural justice. Participants learned that the man had been found guilty and sentenced to 5 years in prison. Participants were again given the same opportunity to assign extra punishment. Participants viewed the same options they viewed in Part 1 of the study and made a second selection.

Results

In the initial selection task, a surprisingly large percentage, 70.7%, of participants in the overt condition preferred the violent over the non-violent sanction and 72.2% of participants in the covert condition of individuals preferred the violent sanction, a non-significant difference across groups, $\chi^2(1, 241) = .1$, $p = .92$. Overall, the violent sanction was preferred 71% (171/242) of the time.

To explore the main and interactive effects of conviction errors, past choice, and attributional ambiguity—and to test H3—we used a log-linear analysis. The significant three-way interaction suggests differential patterns of revision depending on attributional ambiguity, $G^2(3, 241) = 79.32$, $p < .001$. In the second punishment task, participants in the overt- and covert-sanctions condition reacted markedly differently when made aware of the earlier false-positive conviction. When the sanctions were overt, 90% of endorsers of the violent sanction changed their punishment recommendation, whereas only 20% of endorsers of the non-violent sanction changed their sanction, a significant difference, consistent with the hypothesis that learning should be straightforward when there are transparent linkages between decisions and outcomes, $G^2(1, 115) = 61.35$, $p < .001$. In contrast, when sanctions were covert, those opting for the

violent versus non-violent sanction were as likely to change as maintain their selection: 42% of those who endorsed the violent sanction revised their punishment recommendation, whereas 37% of those who selected the non-violent sanction revised their sanction, a non-significant differential, $G^2(1, 125) = 0.34$, $p = .76$.

Study 4 Discussion

Consistent with both H3a and H3b, people making attributional choices under attributional ambiguity did not become less punitive in response to news of false-positive convictions. But when there was attributional transparency, people did throttle back punitiveness judgments.

Study 5

Study 4 focused on the feedback effects of the error that our criminal-justice system prioritizes avoiding: false-positive convictions of the innocent. Study 5 explores whether the process observed in Study 4 was due to attributional ambiguity blocking learning (H3a) or to attributional ambiguity facilitating motivated reasoning, such as ignoring information incongruent with one's punitive preferences (H3b). Study 4 could not distinguish these possibilities because the information about the flaw in the criminal-justice system (a false-positive conviction) was always incongruent with the heightened retributive desires elicited by the crime. Study 5 added a condition in which the information would always be congruent: Learning about a recent false-negative acquittal should reinforce, not check, heightened retributivist desires (Goldberg et al., 1998). If the attributional-ambiguity-blocks-learning-in-general hypothesis were true, then learning should be equally impaired when desires are congruent or incongruent with error feedback. But if attributional ambiguity is facilitating motivated-reasoning processes (motivated learning), learning should only be inhibited when desires and information are incongruent.

Using a 2 × 2 factorial design, we examined the relative sensitivity to false-positive errors and misses by again asking participants to perform either an overt or covert punishment selection task and then providing them with news about either false-positive or false-negative error rates, and then the opportunity to revise their recommendation.

Method

Participants. Participants were 422 (278 male and 144 female) workers on Mechanical Turk who participated for US$1 payment. The sample was 69% Indian, 18% American, and 13% other. The average age of participation was 31.7 ($SD = 9.7$), middle class ($M = 2.4$ on a 5-point scale), and slightly liberal ($M = 4.1$ on a 7-point scale).

Procedure. The structure of Study 5 was identical to that of Study 4. In the first part of the study, participants read a description of the class of crime and recommended a punishment (either covertly or overtly).[3] In the second part of the study, participants were allowed to revise their recommendation based on new information about the accuracy of the penal system, which indicated either high false-positive conviction rates or high false-negative missed conviction rates.

Participants read a description of the judicial system that included multiple indications of a just system: fair pre-trial investigations, fair trials, strong public defenders, and jury-based trials. Participants then read a description and the sentencing guidelines (3-8) years for a class of offenses: rape. Participants then had the opportunity to recommend an extra punishment (overtly or covertly) for a subcategory of rape (child rape).

As in Study 4, in the overt condition, participants were told that they had the option of adding an additional punishment to the sentence. Participants could choose between a prison term without beatings and a prison term that included a severe beating resulting in several weeks of hospitalization. Again, this condition measured formal endorsement of corporal sanctions, and therefore was substantially different from the overt condition in Study 1. In the covert condition, participants were given the opportunity to assign the perpetrator to one of two prisons, using the selection task from covert condition of Study 1. To equate injuries across conditions, the violence in the prison was described as the "minimum of a severe beating resulting in several weeks of hospitalization."

In the second part of the task, participants learned that DNA testing had revealed a systematic bias in the legal system; either a 21% false positive (21% of those convicted were not guilty) or 42% misses (42% of those acquitted were guilty). Errors were asymmetrical based on pretesting that showed people saw the former error as roughly twice as serious as the latter error. Participants then could revise their punishment recommendation for child rapists; participants viewed identical text to Part 1 and made the selection again.

Results

In the initial selection task, 62.5% of overt-sanction participants and 69.2% of covert-sanction participants preferred the violent over the non-violent sanction, a non-significant difference, $\chi^2(1, 241) = 1.79$, $p = .18$. Overall, participants preferred the violent over the non-violent sanction, 65.9% (278/422), a significant difference, $\chi^2(1, 421) = 42.5$, $p < .001$. There was a significant effect of sanction overtness on responsiveness to error feedback, with more responsiveness in the overt-sanctions conditions, $\chi^2(7, 415) = 34.98$, $p < .001$.

To explore the effects of judicial errors, past choices, and attributional ambiguity, we used a log-linear analysis. Overall, there was a significant interaction between the overtness of sanctions, the direction of error feedback, and

initial selection, $G^2(4, 418) = 48.99$. This pattern emerged because individuals were less likely to revise selections when the sanctions were covert and when feedback indicated a greater danger of false-positive conviction errors (a result that is consistent with both H3a and H3b).

For overt sanctions, the interaction between support for violence and error feedback was highly significant, $G^2(3, 205) = 75.58$, $p < .001$. Participants preferred violent sanctions when misses were frequent and non-violent sanctions when false positives were frequent. For covert sanctions, the first-order interaction was significant, but less pronounced, $G^2(3, 211) = 15.56$, $p < .01$. Participants preferred violent sanctions when there were misses but there was no clear pattern of revision in response to false-positive convictions.

Replicating the results of Study 4, in the revision task when participants were made aware of false-positive convictions, participants in the overt- and covert-sanctions condition responded to such information very differently $G^2(3, 208) = 43.45$, $p < .001$. As in Study 4, when sanctions were overt and participants thought that there was a high false-positive conviction rate, 74% of endorsers of the violent sanction revised their sanction, whereas only 20% of endorsers of the non-violent sanction revised their sanction, a significant difference consistent with the hypothesis that, in the overt-sanction condition, retributivists back-off quickly in response to the risk of subjecting the innocent to physical abuse, $G^2(1, 95) = 29.57$, $p < .001$. Again, as in Study 4, when the sanctions were covert and participants thought that there was a high rate of false-positive convictions, those who endorsed violent over non-violent sanctions were roughly equally likely to change their minds: 40% of endorsers of the violent sanction revised their recommendation and 35% of endorsers of the non-violent sanction revised their recommendation, a result consistent with the hypothesis that those assigning offenders to violent prisons were not making the mental linkage to the risk of brutalizing the innocent, $G^2(1, 114) = 0.12$, $p = .735$.

Participants in the overt and covert conditions showed similar responses, $G^2(3, 207) = 1.04$, $p = .51$. Both groups were responsive to information about errors. When sanctions were overt and participants thought that there were many failures to punish the guilty, 29% of endorsers of the violent sanction revised their recommendation, whereas 86% of those who selected the non-violent sanction revised theirs, a significant difference consistent with the hypothesis that failures to punish the guilty conferred some normative license for explicit corporal punishment, $G^2(1, 111) = 39.99$, $p < .001$. When the sanctions were covert and participants thought that there was a high rate of failure to punish the guilty, 35% of endorsers of the violent sanction revised their recommendation, whereas 71% of endorsers of the non-violent sanction revised theirs, a significant difference, $G^2(1, 98) = 10.49$, $p < .005$. This pattern of results is consistent with H3b that those in the covert-sanction condition would only learn lessons from history that they were motivationally predisposed to draw.

Study 5 Discussion

These data suggest that covert retributivists are at risk of becoming prisoners of their punitive preconceptions who learn from false-negative errors to become more punitive, but don't learn from false-positive convictions to become less punitive. These data also have some bearing on the earlier discussion of the merits of impression management versus intrapsychic explanations for the effects of attributional ambiguity. If people were fully conscious impression managers, they would have no difficulty detecting the cues that are driving their judgments, so there should be no impairment of learning. Given that we observe impairment, this renders a pure impression management explanation less plausible.

General Discussion

Returning to Lasswell's proposition, the results of the present studies suggest that punishment decisions are at least partly shaped by the rationalization imperatives of "homo politicus." The displacement of private motives into punitiveness judgments may be particularly dangerous because covert retributivism has the potential to shape the implementation of institutional policies (Garland, 1993). Put more concretely, it is one thing to declare an opposition to prison violence and quite another to take affirmative steps to reduce violence in prisons.

We explored two hypothesized moderators of the effects of attributional ambiguity: minimalist-accountability nudges and error feedback. The former line of work suggests that attributional ambiguity effects are fragile and even subtle hints of surveillance reduce them. Studies 2 and 3 should offer some solace to anti-retributivists who worry that our findings imply that secondary sanctions, such as prison violence, are an incorrigible latent function of our penal corrections system.

Work on the effects of error feedback should however be more worrisome to anti-retributivists. Covert retributivism—as long as it is securely covert—can be stubbornly persistent. People working under the cover of attributional ambiguity were unwilling to reduce punitiveness in response to evidence of false-positive miscarriages of justice, but were willing to increase punitive judgments in response to false-negative miscarriages. This pattern of data is more consistent with the motivated-reasoning account than with the classical-learning "blockage" account, which predicts an across-the-board failure to modify judgments in response to error feedback cutting across both false positives and false negatives (Rescorla, 1988). From a motivated-reasoning perspective, it may just be too tempting for covert retributivists confronting false-positive errors to defend their self-images by simply continuing to use the attributional covers provided for them ("balancing all factors, I still recommend this prison"). It may also simply be too easy for covert retributivists confronting false-negative errors to satisfy their increased

need for retribution by relying even more extensively on violent prison assignments made under the cover of ambiguity.

These data suggest that the problem is not that society is just slow to learn what it is doing. Society may learn so selectively that it drifts toward ever greater reliance on covert punitiveness as a method of social control. These data also raise an intriguing question for follow-up work: How negative do systemic malfunctions need to become to break the grip of covert retributivism on policy reasoning?

Although the current studies focused on covert retributivism, the mirror-image phenomenon, covert forgiveness, is also a distinct theoretical possibility. In our framework, covert forgiveness requires two preconditions. The first precondition is that society is more punitive than people desire. Given prior work on restorative justice motives (Wenzel, Okimoto, Feather, & Platow, 2008; Okimoto & Wenzel, 2009) as well as work on the power of individuating information to "humanize" objects of judgment (Loewenstein & Small, 2007), excessive punitiveness likely occurs. For example, many U.S. citizens oppose the death penalty and widespread use of solitary confinement in prisons. The second precondition is that forgiveness is widely considered counter-normative, such as jury nullification in which jurors refuse to convict a clearly guilty person because they see the punishment as excessive.

Promising candidates for observing covert forgiveness involve applications of federally mandated sentencing guidelines and three strike laws, which require imposing long prison sentences for possessions of small quantities of narcotics or small acts of theft (sometimes even for first offenders). In addition to jury nullification, covert forgiveness could take the form of granting furloughs, conjugal visits, and favorable work duty assignments. The covert functions of individual discretion can thus cut in the direction of both covert forgiveness as well as covert punitiveness.

In closing, it is worth asking if secondary sanctions are, indeed, a latent function of our penal system—as the current data suggest—what is the appropriate reaction to this controversial empirical demonstration? One response might be to argue for prison reform and, in the interim, greater leniency. Weisberg and Mills (2003) pose the following judicial thought experiment in which an attorney advances the following argument for his client: "Your Honor, it is unfair and disproportionate to sentence my client to jail, since it will almost certainly subject him to violent attacks and sexual assault while incarcerated. Any sentence of incarceration effectively includes these 'secondary' sanctions."

Another response might be to argue for revisiting the constitutionality of physical punishment, especially if it is carefully calibrated. Moskos (2011) notes that offenders themselves might well prefer such penalties to incarceration, and sharply painful penalties might have more extreme retrospective disutility (Kahneman, Fredrickson, Schreiber, & Redelmeier, 1993), and hence greater deterrence value. The current article sheds no light, of course, on the appropriate societal response. It does, however, illustrate the power of experimentation to disentangle the manifest from the latent functions of societal practices as well as its power to highlight the ensuing risks of excessive rigidity and policy-lock-in effects.

Acknowledgment

We thank Greg Mitchell, Shefali Patil, Nick Rohrbaugh, and the associate editor Michael Wenzel for their constructive comments on earlier drafts of this manuscript.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research was supported by the Penn Integrates Knowledge Endowed Chair at the University of Pennsylvania.

Notes

1. One could argue that covert retributivism will be most disruptive of learning if people are unaware of the retributivist drivers of their judgments, as opposed to trying to conceal those drivers from others. If so, our findings of failure to learn from negative feedback tip the scales of plausibility in favor of the view that covert retributivists are not fully aware of what they are doing.
2. It is important to note that in this study, we are not only contrasting covert and overt sanctions. Because we are concerned with system learning, we are contrasting primary sanctions and secondary sanctions more generally. Primary sanctions are directly inflicted by the state, whereas secondary sanctions are foreseeable side-effects of state actions. For example, in the domain of corporal punishment, caning would be a primary sanction in Singapore but a foreseeable prison beating would be a secondary sanction.
3. Again, because we are concerned with system learning, we are contrasting primary sanctions and secondary sanctions.

Supplemental Material

The online supplemental material is available at http://pspb.sagepub.com/supplemental.

References

Alicke, M. D. (2000). Culpable control and the psychology of blame. *Psychological Bulletin*, *126*, 556-574.

Alicke, M. D., & Yurak, T. J. (1995). Perpetrator personality and judgments of acquaintance rape. *Journal of Applied Social Psychology*, *25*, 1900-1921.

Atran, S., & Ginges, J. (2012). Religious and sacred imperatives in human conflict. *Science*, *336*, 855-857.

Axelrod, R., & Hamilton, W. D. (1981). The evolution of cooperation. *Science*, *211*, 1390.

Bargh, J. A., & Chartrand, T. L. (1999). The unbearable automaticity of being. *American Psychologist*, *54*(7), 462.

Bersoff, D. M. (1999). Why good people sometimes do bad things: Motivated reasoning and unethical behavior. *Personality and Social Psychology Bulletin*, *25*(1), 28-39.

Blackstone, S. W. (1875). Commentaries on the Laws of England (Sharswood's Edition).

Buehler, R., Griffin, D., & MacDonald, H. (1997). The role of motivated reasoning in optimistic time predictions. *Personality and Social Psychology Bulletin*, *23*, 238-247.

Burnham, T. C., & Hare, B. (2007). Engineering human cooperation. *Human Nature*, *18*(2), 88-108.

Carlsmith, K. M. (2006). The roles of retribution and utility in determining punishment. *Journal of Experimental Social Psychology*, *42*, 437-451.

Darley, J. M., & Pittman, T. S. (2003). The psychology of compensatory and retributive justice. *Personality and Social Psychology Review*, *7*, 324-336.

Dawson, E., Gilovich, T., & Regan, D. T. (2002). Motivated reasoning and performance on the Wason selection task. *Personality and Social Psychology Bulletin*, *28*, 1379-1387.

Durkheim, E. (1933). *The division of labor in society* (G. Simpson, Trans.). New York, NY: The Free Press.

Edwards, L. P. (1995). Corporal punishment and the legal system. *Santa Clara Law Review*, *36*, 983-1093.

Elias, N. (1969). *The civilizing process, vol. I. The history of manners*. New York, NY.

Ellsworth, P. C., & Gross, S. R. (1994). Hardening of the attitudes: Americans; views on the death penalty. *Journal of Social Issues*, *50*, 19-52.

Fehr, E., & Fischbacher, U. (2004). Third-party punishment and social norms. *Evolution and Human Behavior*, *25*(2), 63-87.

Fehr, E., & Gächter, S. (2002). Altruistic punishment in humans. *Nature*, *415*(6868), 137-140.

Foucault, M. (1977). *Discipline and punish: The birth of the prison*. New York, NY: Random House LLC.

Garland, D. (1991). Sociological perspectives on punishment. *Crime and Justice*, 115-165.

Garland, D. (1993). *Punishment and modern society*. Chicago, IL: University of Chicago Press.

Ginges, J., Atran, S., Medin, D., & Shikaki, K. (2007). Sacred bounds on rational resolution of violent political conflict. *Proceedings of the National Academy of Sciences*, *104*, 7357-7360.

Goldberg, J. H., Lerner, J. S., & Tetlock, P. E. (1999). Rage and reason: The psychology of the intuitive prosecutor. *European Journal of Social Psychology*, *29*(56), 781-795.

Goodwin, G. P., & Landy, J. F. (2014). Valuing different human lives. *Journal of Experimental Psychology: General*, *143*, 778-803.

Haidt, J., & Graham, J. (2009). Planet of the Durkheimians, where community, authority, and sacredness are foundations of morality. In J. Jost, A. C. Kay, & H. Thorisdottir (Eds.), *Social and psychological bases of ideology and system justification* (pp. 371-401). New York, NY: Oxford University Press.

Haley, K. J., & Fessler, D. M. (2005). Nobody's watching? Subtle cues affect generosity in an anonymous economic game. *Evolution and Human Behavior*, *26*(3), 245-256.

Hartman, J. F. (1983). Unusual punishment: The domestic effects of international norms restricting the application of the death penalty. *University of Cincinnati Law Review*, *52*, 655-699.

Heider, F. (1958). *The psychology of interpersonal relations*. New York, NY: John Wiley.

Henrich, J., McElreath, R., Barr, A., Ensminger, J., Barrett, C., Bolyanatz, A., . . .Ziker, J. (2006). Costly punishment across human societies. *Science*, *312*(5781), 1767-1770.

Hogan, R., & Blickle, G. (2013). Socioanalytic theory. In N. D. Christiansen & R. P. Tett (Eds.), *Handbook of personality at work* (pp. 53-70). New York, NY: Routledge.

Jones, E. E., & Sigall, H. (1971). The bogus pipeline: A new paradigm for measuring affect and attitude. *Psychological Bulletin*, *76*(5), 349.

Kahneman, D., Fredrickson, B. L., Schreiber, C. A., & Redelmeier, D. A. (1993). When more pain is preferred to less. *Psychological Science*, *4*, 401-405.

Kamin, L. J. (1969). Predictability, surprise, attention, and conditioning. In B. A. Campbell & R. M. Church (Eds.), *Punishment and aversive behavior* (pp. 279-296). New York, NY: Appleton-Century-Crofts.

Kay, A. C., & Jost, J. T. (2003). Complementary justice: Effects of "poor but happy" and "poor but honest" stereotype exemplars on system justification and implicit activation of the justice motive. *Journal of Personality and Social Psychology*, *85*, 823-837.

Kay, A. C., Jost, J. T., & Young, S. (2005). Victim derogation and victim enhancement as alternate routes to system justification. *Psychological Science*, *16*, 240-246.

Klein, W. M., & Kunda, Z. (1992). Motivated person perception: Constructing justifications for desired beliefs. *Journal of Experimental Social Psychology*, *28*, 145-168.

Kunda, Z. (1990). The case for motivated reasoning. *Psychological Bulletin*, *108*(3), 480.

Kunda, Z. (1999). *Social cognition: Making sense of people*. Cambridge, MA: MIT Press.

Langer, E. J., Blank, A., & Chanowitz, B. (1978). The mindlessness of ostensibly thoughtful action: The role of "placebic" information in interpersonal interaction. *Journal of Personality and Social Psychology*, *36*, 635-642.

Langer, E. J., Fiske, S., Taylor, S. E., & Chanowitz, B. (1976). Stigma, staring, and discomfort: A novel-stimulus hypothesis. *Journal of Experimental Social Psychology*, *12*, 451-463.

Lasswell, H. D. (1986). *Psychopathology and politics*. Chicago, IL: University of Chicago Press. (Original work published 1930)

Lerner, J. S., & Tetlock, P. E. (1999). Accounting for the effects of accountability. *Psychological Bulletin*, *125*(2), 255.

Lerner, M. J. (1980). *The belief in a just world: A fundamental delusion*. New York, NY: Plenum Press.

Lerner, M. J., & Lerner, S. C. (Eds.). (1981). *The justice motive in social behavior: Adapting to times of scarcity and change*. New York, NY: Plenum Press.

Lively, P. C., Rozin, P., & Goodwin, G. P. (2014). *The lively sex bump*. Manuscript under revision.

Loewenstein, G., & Small, D. A. (2007). The scarecrow and the tin man: The vicissitudes of human sympathy and caring. *Review of General Psychology*, *11*, 112-126.

Malle, B. F., & Guglielmo, S. (2012). Are intentionality judgments fundamentally moral. In R. Langdon & C. Mackenzie (Eds.), *Emotions, imagination, and moral reasoning* (pp. 275-293). Hove, UK: Psychology Press.

Malle, B. F., Guglielmo, S., & Monroe, A. E. (2012). B lame is a moral judgment that has a cognitive and a social nature. We first focus on the cognitive side and introduce a theoretical model of blame. *Social Thinking and Interpersonal Behavior*, *313*.

McKee, I. R., & Feather, N. T. (2008). Revenge, retribution, and values: Social attitudes and punitive sentencing. *Social Justice Research*, *21*, 138-163.

Merton, R. K. (1949). *On sociological theories of the middle range [1949]*. New York, NY: Free Press.

Monin, B., & Miller, D. T. (2001). Moral credentials and the expression of prejudice. *Journal of Personality and Social Psychology*, *81*, 33-43.

Moskos, P. (2011). *In defense of flogging*. Basic Books.

Mullen, E., & Skitka, L. J. (2006). Exploring the psychological underpinnings of the moral mandate effect: Motivated reasoning, group differentiation, or anger? *Journal of Personality and Social Psychology*, *90*, 629-643.

Norton, M. I., Vandello, J. A., & Darley, J. M. (2004). Casuistry and social category bias. *Journal of Personality and Social Psychology*, *87*, 817-831.

Nowak, M. A. (2006). Five rules for the evolution of cooperation. *Science*, *314*, 1560-1563.

Nowak, M. A., & May, R. M. (1993). The spatial dilemmas of evolution. *International Journal of Bifurcation and Chaos*, *3*, 826-829.

Okimoto, T. G., & Wenzel, M. (2009). Punishment as restoration of group and offender values following a transgression: Value consensus through symbolic labelling and offender reform. *European Journal of Social Psychology*, *39*, 346-367.

Okimoto, T. G., & Wenzel, M. (2011). Third-party punishment and symbolic intragroup status. *Journal of Experimental Social Psychology*, *47*, 709-718.

Pezzo, M. V., & Pezzo, S. P. (2007). Making sense of failure: A motivated model of hindsight bias. *Social Cognition*, *25*, 147-164.

Pinker, S. (2011). *The better angels of our nature: The decline of violence in history and its causes*. New York, NY: Penguin UK.

Polinsky, A. M., & Shavell, S. (1998). Punitive damages: An economic analysis. *Harvard Law Review*, 869-962.

Redlawsk, D. P. (2002). Hot cognition or cool consideration? Testing the effects of motivated reasoning on political decision making. *The Journal of Politics*, *64*, 1021-1044.

Rescorla, R. A. (1970). Reduction in the effectiveness of reinforcement after prior excitatory conditioning. *Learning and Motivation*, *1*, 372-381.

Rescorla, R. A. (1988). Pavlovian conditioning: It's not what you think it is. *American Psychologist*, *43*, 151-160.

Rousseau, D. M., & Tijoriwala, S. A. (1999). What's a good reason to change? Motivated reasoning and social accounts in promoting organizational change. *Journal of Applied Psychology*, *84*, 514-528.

Sell, A., Tooby, J., & Cosmides, L. (2009). Formidability and the logic of human anger. *Proceedings of the National Academy of Sciences*, *106*, 15073-15078.

Skitka, L. J., & Houston, D. A. (2001). When due process is of no consequence: Moral mandates and presumed defendant guilt or innocence. *Social Justice Research*, *14*, 305-326.

Spierenburg, P. C. (1984). *The spectacle of suffering: Executions and the evolution of repression: From a preindustrial metropolis to the European experience*. Cambridge, England: Cambridge University Press.

Skitka, L. J., & Mullen, E. (2002). The dark side of moral conviction. *Analyses of Social Issues and Public Policy*, *2*, 35-41.

Sniderman, P., Brody, R., & Tetlock, P. (1991). *Reasoning and choice: Explorations in political psychology*. New York, NY: Cambridge University Press.

Sniderman, P. M., Piazza, T., Tetlock, P. E., & Kendrick, A. (1991). The new racism. *American Journal of Political Science*, *35*, 423-447.

Snyder, M. L., Kleck, R. E., Strenta, A., & Mentzer, S. J. (1979). Avoidance of the handicapped: An attributional ambiguity analysis. *Journal of Personality and Social Psychology*, *37*, 2297-2306.

Sunstein, C. R., Schkade, D., & Kahneman, D. (2000). Do people want optimal deterrence? *The Journal of Legal Studies*, *29*, 237-253.

Tetlock, P. E., & Manstead, A. S. (1985). Impression management versus intrapsychic explanations in social psychology: A useful dichotomy? *Psychological Review*, *92*(1), 59.

Tetlock, P. E., Visser, P. S., Singh, R., Polifroni, M., Scott, A., Elson, S. B., . . .Rescober, P. (2007). People as intuitive prosecutors: The impact of social-control goals on attributions of responsibility. *Journal of Experimental Social Psychology*, *43*(2), 195-209.

Tykocinski, O. E., & Steinberg, N. (2005). Coping with disappointing outcomes: Retroactive pessimism and motivated inhibition of counterfactuals. *Journal of Experimental Social Psychology*, *41*, 551-558.

Tyler, T. R., & Boeckmann, R. J. (1997). Three strikes and you are out, but why? The psychology of public support for punishing rule breakers. *Law & Society Review*, *31*, 237-265.

Van den Haag, E. (1991). Punishing criminals.

Vidmar, N., & Miller, D. T. (1980). Socialpsychological processes underlying attitudes toward legal punishment. *Law & Society Review*, *14*, 565-602.

Weber, M. (2009). *The theory of social and economic organization*. New York, NY: Simon and Schuster.

Weiner, B., Graham, S., Peter, O., & Zmuidinas, M. (1991). Public confession and forgiveness. *Journal of Personality*, *59*, 281-312.

Weiner, B., Graham, S., & Reyna, C. (1997). An attributional examination of retributive versus utilitarian philosophies of punishment. *Social Justice Research*, *10*, 431-452.

Weisberg, R., & Mills, D. (2003, October 1). Violence silence: Why no one really cares about prison rape. *Slate*.

Wenzel, M., Okimoto, T. G., Feather, N. T., & Platow, M. J. (2008). Retributive and restorative justice. *Law and Human Behavior*, *32*, 375-389.

Wenzel, M., & Thielmann, I. (2006). Why we punish in the name of justice: Just desert versus value restoration and the role of social identity. *Social Justice Research*, *19*, 450-470.