CHAPTER FOUR

PSYCHOLOGICAL ASPECTS OF RETRIBUTIVE JUSTICE

Kevin M. Carlsmith* and John M. Darley†

Contents

1. Retributive Justice, in Relation to Other Kinds of Justice 194
   1.1. Other systems for harm-based justice 195
   1.2. Retributive principles, criminal justice system practices, and policy issues 196

2. What Motivates the Desire to Punish? 197
   2.1. Methodological issues 197
   2.2. Verbal reports 198
   2.3. Behavioral measures 198
   2.4. Retribution 199
   2.5. Utility 200
   2.6. Specific deterrence 200
   2.7. Incapacitation 200
   2.8. General deterrence 201
   2.9. Carlsmith (2006) 203
   2.10. Darley et al. (2000) 204
   2.11. Carlsmith et al. (2002) 205
   2.12. Individual differences 206
   2.13. Connections with previous research 209

3. The Impulse to Punish as an Intuition 211
   3.1. The role of the reasoning system in moral judgments 213

4. Policy Implications 218
   4.1. Evidence for discrepancies between legal codes and community intuitions 219
   4.2. When do steps toward a crime cross the line to become a crime? 221
   4.3. The ex-anti function: Do people know the criminal law? 223

5. Conclusion 233

References 235

* Department of Psychology, Colgate University, Hamilton, NY 13346
† Department of Psychology, Princeton University, Princeton, NJ 08540
Abstract

Retributive justice is a system by which offenders are punished in proportion to the moral magnitude of their intentionally committed harms. This chapter lays out the emerging psychological principles that underlie citizens’ intuitions regarding punishment. We rely on experimental methods and conclude that intuitions of justice are broadly consistent with the principles of retributive justice, and therefore systematically deviate from principles of deterrence and other utilitarian-based systems of punishing wrongs. We examine the recent contributions of social-neuroscience to the topic and conclude that retributive punishment judgments normally stem from the more general intuitive-based judgment system. Particular circumstances can trigger the reasoning-based system, however, thus indicating that this is a dual process mechanism. Importantly, though, evidence suggests that both the intuitive and reasoning systems adhere to the principles of retribution.

The empirical results of this research have clear policy implications. Converging evidence suggests that the formal U.S. justice system is becoming increasingly utilitarian in nature, but that citizen intuitions about justice continue to track retributive principles. The resulting divide leads people to lose respect for the law, which means that they do not rely on the law’s guidance in ambiguous situations where the morally correct behavior is unclear. These are the dangers to society from having justice policies based jointly on the contradictory principles of retribution and utility, and we lay out an argument for enacting public policies more exclusively based on retributive principles of justice.

The field of psychological research on retributive justice, as compared to other kinds of justice, is of more recent empirical investigation and is correspondingly less well researched. Recently, though, it has attracted enough psychological interest so that one can glimpse the beginnings of a conceptual account on how people think about this sort of justice. Some psychological principles begin to emerge. In this chapter, we will summarize some of this research, with the customary emphasis on research conducted in our own laboratories, and drawing in work of others where relevant.

1. Retributive Justice, in Relation to Other Kinds of Justice

Retributive justice is roughly the question of how people who have intentionally committed known, morally wrong actions that either directly or indirectly harm others, should be punished for their misdeeds. Retributive justice is sometimes regarded as part of distributive justice, because distributive justice is in turn concerned with what people deserve for their actions. To some extent, of course, this is a question of nomenclature, but the great bulk of distributive justice research has concerned itself with what one positively deserves, in terms of a share-out of the rewards produced by
a group product. Distributive justice theories tend to assume positive outcomes and are concerned with whether those positive outcomes should be shared out equally, equitably in proportion to individual’s contributions to the group product, or on the basis of needs.

In keeping with the standard conventions in this area, we will reserve the term “distributive justice” to discuss reward allocations, and use “retributive justice” to refer to punishments that people deserve for their wrongdoing. As we will illustrate, the principles that distribute punishments are quite different from those used to distribute rewards.

1.1. Other systems for harm-based justice

There actually are several other systems for dealing with the perpetrators of harmful actions, and we need to position them in our account. We begin by discussing the Tort Law system.

The defining characteristic of the harms that these other systems deal with is that the perpetrator does not inflict the harms in question intentionally on others. The steam railroads of the 19th century were notorious inflictors of harm: burning bits of coal billowing out of locomotive smoke-stacks set farmers’ fields on fire, and quite a large number of livestock were slaughtered by speeding trains at unfenced rail crossings (Friedman, 1985). Cattle ranchers also made their own contributions to the infliction of damages on nearby farmers as their cattle broke through fences into farmer’s fields and consumed or trampled remarkably large amounts of farm product, or bulls impregnated carefully bred dairy cows. Mills and factories inflicted injuries on workers, and at least occasionally it was realized that the workers were due compensation for the medical costs of these harms. But it was not the case that the railroad owners, cattle ranchers, or mill owners intended to inflict harms on those injured.

The legal system that grew up to deal with these issues in the American context is the Tort Law system. Within that system, the United States’ treatment of these sorts of cases heavily relies on the concept of negligence. The thought here is that we all owe our neighbors a duty of care such that our activities do not inflict damages on them. If I have not acted up to a standard of care that a reasonable, prudent person would exhibit, and I am the proximate cause of the damage to you and your property, and the damage would not have occurred but for my imprudent actions, then I owe you compensation for the damages inflicted. The essential notion here is compensation, “making you whole” for the damages you have suffered. Similar issues arise in other situations such as medical malpractice, and liability for manufacturers for the harms caused by dangerous products they manufacture. That notion is to some extent similar to the notion of punishment, a topic to which we will return after we present
some of the psychological discoveries concerning what it is that people seek to do when they impose punishment.

1.2. Retributive principles, criminal justice system practices, and policy issues

One task of the psychologist is to discover how ordinary people conceptualize retributive justice principles. This is a task that is interesting in its own right, but it is also a genuinely important contribution to public policy issues because citizens have some right to have their opinions registered about what behaviors will be targeted by the criminal justice system. At a minimum, these opinions should be available to criminal code drafters; some would make the claim that they have some right to be represented in the criminal codes. We will return to this question after we examine some of the discoveries that have come from the psychological research.

In our research, the contours of the decision rules that citizens use in thinking about issues of culpability are often contrasted to the decision rules embedded in legal codes. One reason for this, of course, is the old but useful academic habit of having a “straw man” to spar with. However, as legal scholars have pointed out, what is called “common law” is likely to have a good many citizens’ views contained within it. This is because the common law, which was the base for much of the criminal law of the United States up to the first half of the 20th century, is based on judgments made by judges and jurists (stare decisis) as opposed to statutes passed by legislative bodies.

As a result, the criminal law, the part of the law that deals with retributive justice, has been to a considerable extent made by citizens in their role of jurists. Looking back over those juridical decisions, one is struck by how often the judges explicitly appealed to common sense notions of justice thought to be held by all citizens, or at least all righteous citizens. And as we know, the common law tradition is embedded in the legal systems of the United Kingdom, and thus in many former British colonies, including the United States. Therefore, many elements of our foundational criminal doctrines incorporate strong elements of “common sense,” if common sense is construed as the way that citizens of a society think about justice issues. Therefore often when one examines criminal legal doctrines, one is doing it not to erect straw men to oppose, but to discern how people who had to think very hard about questions of culpability and responsibility organized those thoughts.

There is also a case to be made for bilateral causality here. Our legal system commands respect from large groups of citizens, and thus is a credible source of what citizens’ moral principles should be (Tyler, 1990). Over the course of the many generations that people have been interacting with the legal system, the system has had the chance to influence citizens’ moral thinking through the vehicle of trials, films, courtroom dramas, and all the other ways that legal practice gets transferred to the popular
imagination. Such claims as “I have a right to a trial by a jury of my peers” or “a man is innocent unless proven guilty,” which initially seem products of moral intuitions, are more likely to be products of socialization into our particular culture.

Regardless of the origin, there are clear intuitions among citizens about what constitutes a just response to intentional violations of norms and laws. We turn next to these intuitions.

2. What Motivates the Desire to Punish?

Even a casual observer of human behavior comes away with the indisputable conclusion that punishment is ubiquitous within human communities. It is not clear, however, whether punishment is an emergent property of moral development, or rather is an extension of the human proclivity for tools. That is, does punishment stem from violations of moral strictures, or is punishment a technique for social control?

It is unlikely, of course, that a behavior as common and as important as punishment would depend solely upon a single motive. Reality almost certainly contains multiple motivational elements. Nevertheless, it is useful to identify the motives that dominate behavior. With regard to societal punishment, the rich traditions of moral philosophy, criminology, and legal theory provide a concrete starting point. Retributive justice, of course, plays a critical role in each of these areas, but so too does deterrence theory and incapacitation. We open this paper, then, seeking the origin of the desire to punish among individuals. When a person seeks to punish, what is he or she trying to achieve?

Moral philosophers have long struggled with the justifications that permit a society to punish its members. The debate is often simplified into two mutually incompatible schools of thought. Kant (1952/1790) argued that perpetrators deserve to be punished in proportion to their “internal wickedness,” and that the imperative to punish derived not from the future consequences of the punishment, but rather from a universal goal of giving people what they deserve. By contrast, Bentham (1962/1843), representing a utilitarian approach to ethics, argued that punishment, and indeed all action, ought to be assessed by the potential harm or benefit to society. Any decision to punish individuals, then, must weigh the potential harms to the individual (and, thus, to society) against the benefits to society, and punishment is moral only if society stands to benefit.

2.1. Methodological issues

Although the basic question of why citizens punish is straightforward, the route to an answer is not. There exist important methodological considerations, and variations in how the question is addressed have led to starkly
divergent conclusions. We will attempt to show that some of the confusion within the field stems from these different methodological approaches.

There are two fundamental approaches to this question: one can ask people why they punish, or one can observe people as they punish and infer the underlying motive. We will refer to the first as verbal reports and the second as behavioral measures.

2.2. Verbal reports

The advantage of verbal reports is that they are simple to measure, and they contain a relatively high degree of face validity. The problem with verbal reports is that they are frequently incorrect. Nisbett and Wilson (1977) first demonstrated this point in a series of experiments demonstrating that people had only limited insight into the motivations for their behavior. When asked why they made a particular choice, participants appeared to formulate an answer spontaneously and after-the-fact rather than introspecting and discovering the actual reason. Nisbett and Wilson concluded that although people do, indeed, have discrete motivations that follow clear rules, they have only limited insight into those motivations and are frequently unaware of them entirely, or are unable to articulate them accurately. Thus, verbal reports reveal little about the true internal states of the person, and are no more insightful about motivation than a stranger might be after observing the target. This general point holds true for punishment motivations as well, and is discussed in Carlsmith (2008) as well as later in this chapter.

2.3. Behavioral measures

A second approach, and one that we tend to favor, is to largely ignore the reasons that people furnish and to focus instead on their behavioral responses to various situations. Our general strategy has been to ask citizens to recommend punishments for a wide variety of carefully calibrated harms, and to draw inferences about their underlying motives for punishment. This approach allows us to create situations that cry out for severe punishment from one philosophical perspective, yet little or no punishment from other perspectives. The citizens’ responses (e.g., recommended sentence) clarify which perspective guides their intuition. This approach, frequently described as a policy-capturing approach (Cooksey, 1996), is useful because it can help formulate public policies that will most closely align with people’s intuitions of justice. We will say more about the limits of this approach toward the end of the chapter.

Our prototypical experimental design pits two perspectives against each other in a $2 \times 2$ arrangement. We create four versions of a vignette that describes an offense, and ask participants to recommend an appropriate punishment for one of those cases. For each perspective, perhaps
Retribution and Deterrence, we create a moderate and severe case. That is, we modify the vignette so that it is either relevant or irrelevant to each perspective. For example, the theory of retribution is concerned with the moral culpability of the perpetrator, and thus the vignette is adjusted such that the moral culpability is either high or low. A retributivist will adjust his or her punishment accordingly. By contrast, a committed deterrence theorist would be relatively unmoved by this variable, and far more sensitive to a deterrence-related variable such as the frequency of the particular type of crime. Across subjects and vignettes, one can see patterns of response, and from those infer which philosophical perspective is most relevant to punishment. This information can then be used to shape public policy so that it better matches public intuitions of justice and fairness.

The logic of these studies depends critically upon how these variables are instantiated, and work under the assumption that what is relevant for one perspective (e.g., moral culpability for “just deserts”), is irrelevant for another (e.g., deterrence). This is not to say that the perspectives never overlap in the factors that they consider relevant, but rather that the experimenters must choose those factors that are unique to each perspective. Several of our earlier papers (Carlsmith, 2006; Carlsmith et al., 2002) devoted substantial effort to defining and justifying our choice of manipulations, and we will summarize those efforts here because they are essential to understanding the results.

2.4. Retribution

Kant (1790/1952) best articulates the retributive justice stance by arguing that perpetrators ought to be punished in proportion to the moral offensiveness of their action. The goal of retributive justice is to give people their “just deserts,” and any future consequences of the punishment (e.g., utilitarian outcomes) are irrelevant to the assignment of punishment. For Kant, punishment needs no justification beyond the deservingness of the perpetrator.

The critical variable for retribution, then, is the moral outrage generated by the crime. More serious crimes deserve more serious punishments. At the same time, though, factors that mitigate or exacerbate the morality of the action also come into play. Thus, holding the severity of the crime constant—say, theft of $100—one would punish less severely if the money was needed to feed a starving child, and more severely if the money was needed to create the world’s largest margarita. Similarly, any circumstance that mitigates responsibility for the action would also mitigate the moral severity of the offense. Thus, negligent actions ought to be punished less severely than intentional actions, and unforeseeable harms ought to be punished less severely than foreseeable harms. In short, a theory of retributive justice will be concerned with those factors that increase or decrease the
moral outrage that the action generates. The theory is focused solely upon the action and seeks to balance moral magnitude of the offense with the punishment. Those who cry, “let the punishment fit the crime” are probably operating from a retributive stance. Similarly, the biblical decree of “an eye for an eye” reflects retributive imperative that crimes and punishments ought to be commensurate. Modern views of retributive justice seek proportionality rather than absolute parity. Thus, more serious crimes receive more serious punishments, but the punishments, on average, could be either more or less severe than the offense.

2.5. Utility

There are other justifications for punishment independent of retribution. Most notably, there is utility. One punishes an offender to reduce the likelihood of offenses in the future. This approach stems from the more general utilitarian movement in philosophy, in which all of one’s actions should be aimed towards improving the world. In brief, the morality of punishment rests solely upon whether it is likely to increase or decrease the total happiness in the world. Although punishment will almost certainly make the recipient less happy, it also has the potential to make society at large happier to the extent that it reduces future crime. If the “expected utility” of the proposed punishment is positive, then the punishment should take place. Within this general framework, there exist several versions of utility.

2.6. Specific deterrence

Here the assignment of punishment is for the purpose of deterring the specific offender from reoffending when he has completed the sentence just imposed. One could imagine a complex individualized calculation scheme that would attempt to determine the sentence that would be sufficient to deter exactly this individual in the future, but it is usually assumed that the standard sentence for the crime will deter this criminal from repeating the offense.

2.7. Incapacitation

Incapacitation, like specific deterrence, is designed to affect the behavior of the specific individual that has committed the crime, but does not do so through threat of punishment for future crimes. Instead, it renders any future offenses impossible, at least for a relatively long time period. This is accomplished most typically through incarceration in prison for a long duration sentence, although other means are also available. Sentences that
include capital punishment (for murderers), castration (for rapists), severing a hand (for thieves), disbarment (for lawyers), branding a scarlet “A” (for adulterers), and deportation (for nearly all classes of criminals), often reflect the goals of incapacitation. The goal of an incapacitationist is to prevent known criminals from reoffending, and thus the single most important criteria in deciding whether to sentence is the likelihood of recidivism. This construct is frequently evaluated by examining the criminal’s prior record, his or her impulsiveness, and/or other actions that suggest imminent criminal behavior. When a person is deemed to represent a threat to the future lawfulness of society, that person is incapacitated.

The inevitable tension generated by this approach is that society punishes not for the action committed, but for the probability that an action will be performed. Thus, when the theory operates perfectly, no crimes are committed, but not one person in jail has actually committed a crime.

2.8. General deterrence

This approach punishes one person in order to deter others from committing a similar crime. When the punishment is levied against an appropriately guilty perpetrator, and the punishment is proportionate to the offense, then the penalty can serve the utilitarian function of general deterrence, but also of specific deterrence and retribution. In its pure form, however, general deterrence need not serve these other functions. Indeed, it often makes sense to punish a perpetrator with excessive and disproportionate severity so as to make a clearer point to potential thieves. Although this violates the retributive tenets of fairness, it is clearly in society’s interest to punish one person severely if it will prevent dozens, hundreds, or thousands of crimes from occurring in the future. Extending this logic, one can see that society might do well to punish celebrities more severely than noncelebrities. Likewise, according to this theory, if the punishment will receive no publicity (e.g., a private punishment, or a punishment that nobody notices) then the punishment is immoral because it creates unhappiness with no offsetting deterrent benefit. By contrast, if the recipient is famous and will receive extensive press coverage, then the punishment ought to be as severe as possible because of the tremendous potential deterrent effect. Taken to its logical extreme, philosophers have argued that the key features of retribution—severity of the crime, responsibility, mitigating factors, etc.—are entirely irrelevant to a theory of deterrence. In fact, even guilt is irrelevant. One can imagine circumstances in which punishing an innocent would have such tremendous societal benefit (e.g., forestalling an imminent riot that would otherwise kill hundreds of innocents), that utility would demand punishment of the innocent for the greater good. Realizing the perceived injustice of this, most who advocate a deterrent approach to sentencing hold
that it can be used only on those guilty of crimes, and they also generally limit the duration of the sentence that can be imposed to one that does not exceed the bounds of what is justly inflicted for the crime the offender committed.

We have stated that the severity of the harm is relevant to retribution, yet irrelevant to utility. Readers may wonder whether severity of harm is truly orthogonal to utilitarian justifications of punishment. Intuitively it may seem that society has more need to deter serious crimes than petty crimes, and thus that utilitarians ought to carefully consider the danger posed by a particular crime. There are two responses to this. First, the effect of a given crime extends far beyond the initial victim, and includes all those who subsequently live in fear of future crime and those who become more likely to commit crimes as a result of witnessing an unpunished crime. Thus, the “cost” of punishment (borne by the offender) is almost certainly outweighed by the benefit to society, regardless of the punishment’s severity. Phrased differently, the cost to society for even minor crimes is so high, that a proportional response would always be severe. A utilitarian, however, would always limit the punishment to be just sufficient to deter future instances of the crime, and thus deservingness never enters the equation. Second, the logic of deterrence is that it will prevent future crimes merely by the threat of severe sanction. Indeed, if one believes in the efficacy of deterrence, then one need not be concerned about “overpunishing” minor crimes, because the threat of punishment alone will ensure that the punishment is never, in fact, carried out.

We have argued, as have numerous philosophers, that retribution and utility cannot be effectively integrated and that logic dictates an “either/or” approach. This strikes many readers as overly simplistic because it is easy to think “I want both: I want to stop future offenses and I want the perpetrator to receive a fair punishment.” There is certainly nothing wrong with this sentiment, and holding these desires simultaneously is not illogical. Indeed, it is frequently the case that a punishment can serve both functions effectively. Nonetheless, the two justifications are not isomorphic and frequently diverge on appropriate sentences. It is in these situations that one must make a choice, and that one motive will trump the other.

We pause here to note that there are other justifications derived from utilitarian theory that we have not studied and do not discuss in great detail. We omit them in large part because their primary supporters appear to us to be within academic circles rather than the citizenry. And because our primary focus in this paper and our research has been citizen intuitions of justice, they have been less important. These derivations include: rehabilitation, restoration, and restitution (but see pp. 208–209 for a discussion of restorative justice).

Armed with the methodological and philosophical background, we turn now to some of our empirical findings.
2.9. **Carlsmith (2006)**

A logical place to begin the search for motives is to examine the cognitive processes that lead to a particular behavioral outcome. In this case, we\(^1\) looked at what sort of information participants recruited when they were required to make punishment decision. Borrowing from the method of “behavioral process tracing” (Jacoby et al., 1987), Carlsmith created a paradigm that revealed whether participants were seeking information relevant to deterrence, incapacitation, or retribution when they were asked to assign criminal punishment.

Participants began the experiment knowing only that a crime had been committed, and that they were responsible for recommending a sentence. The experimenter presented them with different categories of information about the crime, each of which was uniquely relevant to deterrence, incapacitation, or retribution. Thus, for example, participants could select “frequency of crime in society” (a deterrence-related category), “likelihood of repetition” (an incapacitation category), and seven other possible bits of information that could be readily classified into the three categories. Once selected, they learned a particular detail of the case from that category, such as “crimes of this sort are extremely rare,” or “experts report that the perpetrator is very likely to repeat this crime given the opportunity.” The dependent measure was a count of frequency with which each category was chosen, and the order in which they were chosen. In short, people could choose to learn different types of information, and their choices could inform us of the underlying theory that guided their behavior.

Participants overwhelmingly chose the retributive items. On the first of five repetitions, 97% of participants selected a retributive piece of information. On the second and third repetitions, strong majorities (64% and 57%) also selected retributive items. As the retributive category ran out of content, people shifted their requests to the incapacitative category. Only when no other category was available did people seek out deterrence-related information.

In a follow-up study, participants were randomly assigned information about the crime from each of these categories, and asked to assign a punishment and to report their confidence in the correctness of those assignments. We found, not surprisingly, that people became more confident as they obtained more information. Importantly, though, those gains were also dependent upon the category from which they came. Each piece of retribution-related information provided a significantly larger confidence boost than did information from either the deterrence or incapacitation categories.

\(^1\) The pronoun "we" is used because although it is a single authored paper, it derived from the Carlsmith’s doctoral dissertation and Darley was deeply involved in the design and execution of the work.
This study reveals that when people are placed in the position of assigning punishment, they seek out information relevant to a retributive perspective and not to one of the utilitarian perspectives. Moreover, when they are forced to use utilitarian information, they feel less confident in the accuracy of their assignments than when they receive information pertaining to the perpetrator’s deservingness of punishment.

2.10. Darley et al. (2000)

The previous study showed which information people sought when punishing. This next study reveals people’s sensitivity to variations in the different types of information. Together, these studies reveal whether people seek to give perpetrators “what they deserve,” or whether they seek to punish solely for the purpose of preventing future crimes.

Participants in this study read 10 vignettes that described various crimes ranging from quite minor (e.g., stealing music CDs from a store) to quite severe (e.g., political assassination). The vignettes also varied on the prior history of the perpetrator: he was either described as a first time offender with no history of misbehavior, or as a repeat offender who had committed similar crimes in the past. In this way, we manipulated the moral severity of the crime (a retributive factor) and the likelihood of future offenses (an incapacitative factor). The dependent variable was recommended sentence. The design of this study allowed us to examine which factors elicited strong punitive responses, and thus which factors mattered to the punishers. In this way, we inferred which theory of punishment best captured people’s actual behavior, and it was not necessary to have people articulate or endorse a particular theory.

The results were quite clear. When it came to sentencing, people were highly sensitive to the severity of the offense, and largely ignored the likelihood that the person would offend again. Manipulation checks revealed that people clearly understood whether the perpetrators would be likely to reoffend, but that knowledge did not translate into sentence severity. Even for the most serious crimes that involved physical injury, the threat of future harm did not evoke incapacitative responses. Rather, people punished exclusively in proportion to the harm done and not for the harm that might be done in the future.

To further clarify our results, we asked the participants to review the vignettes a second and a third time, and to do so from a retributive perspective and an incapacitative perspective. We explained to them in some detail how these theories differed from each other, the goals of the theories, and how they operated in general. We then asked them to assign punishment again from these perspectives. Our purpose was twofold: first, we could reassure ourselves that people understood each perspective well enough to behave as an incapacitationist if they so desired (and, thus, that
our primary results were not merely the result of trying unsuccessfully to express their incapacitative desires), and second, to compare their “default” assignments to each of the “pure” perspectives, and thus more clearly reveal their goals of punishment.

As we predicted, people had no difficulty with the retributive perspective. Their sentences tracked the moral severity of the offense closely, and they completely ignored the incapacitative factors. When they adopted the utilitarian perspective, however, the results were more complex. They became highly sensitive to the future risk of the offenders as we had predicted, but they did not ignore the moral severity of the offense as the theory dictates. Even when participants were instructed to ignore the retributive factors, the moral severity of the crime intruded on their sentencing and remained a significant predictor of the sentence. This failure reveals, we think, the importance of the just deserts rationale to the ordinary person.

Finally, a comparison of the three adopted perspectives (default, retributive, and incapacitative) revealed that the first two were nearly indistinguishable from each other, and that both yielded responses quite different from the third. This finding, in conjunction with the earlier results, led to the conclusion that people spontaneously punish in a manner that is highly consistent with a theory of retributive justice and not in a manner consistent with the utilitarian goals of incapacitation.

2.11. Carlsmith et al. (2002)

In a subsequent companion study, we tested the retributive motive against the general deterrence motive. In the first study, we asked over 300 university students to read a vignette of criminal wrongdoing and to provide a recommended sentence for the perpetrator. Across participants, we manipulated the moral severity of the harm and the need to generally deter the particular crime. To further generalize our findings, we instantiated each of the manipulations in several ways. For example, one vignette manipulated the retributive factor by increasing or decreasing the harm committed (e.g., embezzling modest amounts of money from an employer versus illegally dumping toxic chemicals near a town’s water table to avoid expensive disposal fees). In another vignette, the harm was kept the same but the morality of the action was varied (e.g., stolen funds were used to benefit underpaid factory workers at the company’s overseas plant, vs. using the stolen funds to finance a lavish lifestyle and extensive gambling debts).

Again, the results were quite clear and mirrored our previous results. Regardless of how we instantiated the manipulations, people were highly sensitive to changes in the retributive factors, and ignored the deterrence factors. Figure 4.1 shows the path analysis conducted in Carlsmith et al. (2002),
2.12. Individual differences

In this study, we also measured individual differences to see if there are different types of people: those who punish for retribution and those who punish for deterrence. We explained the basic outlines of each theory to our participants, and asked them to choose the one that best described their personal perspective on punishment. We found that people varied widely on this question, and that it was quite easy to split people into three groups: retributivists, deterrists, and mixed perspectives. However, their actual behavior did not differ much across the three groups. Yet, those who self-identified as deterrists were slightly more sensitive to the deterrence manipulation (and gave more severe sentences for crimes that needed more general deterrence), but the effect was extremely small, $\eta^2 = .02$. Thus, although individual differences mattered, they were easily dwarfed by the more universal sensitivity to the retributive factors.

The third experiment within this paper further explored the disjunction between people’s stated goals of punishment and their actual behavior. As before, we asked people to self-identify themselves as retributivists or deterrists, and then had them evaluate one of our vignettes. People were again highly sensitive to retribution and ignored the deterrence manipulation, and individual differences only marginally affected their punitive responses. This time, however, we asked them to also engage in a resource
allocation task. They were given the option to allocate city resources toward catching the perpetrators of the crime (an action consistent with the retributive goals of punishment), or toward preventing future crimes of this sort (an action consistent with deterrence).

We found that people expressed strong support for preventing future occurrences of the crime, but that this desire was in no way related to the details of the crime. That is, people did not recommend more “prevention allocation” when the deterrence manipulation was high. Thus, what we find is that although people like and strongly endorse the outcome goals for utilitarian theories (e.g., reduced crime), they do not endorse the punishments that, according to the theories, would be required to achieve those goals. In other words, although people want punishment to reduce crime, they are not willing to incapacitate an individual merely because he or she is likely to commit a crime. Instead, they want to give a person what they deserve, and a person only deserves punishment after the fact. Likewise, people are unwilling to increase punishment when a sentence will be highly publicized, even though such an action would yield a greater deterrence benefit. Rather, people want the punishment to match the moral severity of the crime. In short, people want punishment to incapacitate and to deter, but their sense of justice requires sentences proportional to the moral severity of the crime (see also Carlsmith et al., 2007).

The preceding discussion implies a fundamental conflict within many individuals: they want and expect punishment to serve utilitarian goals, but when faced with a particular case they consistently choose punishments that serve retributive goals. This, in turn, suggests that people do not have a good sense of their own motives for punishment, and as a consequence may endorse public policies that will violate their own intuitions of justice.

This point was highlighted by Carlsmith (2008). The first experiment examined whether people could correctly articulate the motive that drove their decision to punish. Participants, who were broadly representative of the United States and other English speaking countries, evaluated a series of vignettes and assigned punishments in the usual manner. Each participant saw vignettes from each level of the retributive manipulation, and from each level of the utilitarian manipulation. By examining their pattern of responses to these vignettes, each participant’s sensitivity to the retributive and utilitarian factors was quantitatively assessed. Participants also self-identified their motive for punishing, both through single-item endorsements (e.g., “Which perspective is more important to you, deterrence or retribution”), and through multi-item scales designed by other researchers for the same purpose (Clements et al., 1998; McKee and Feather, 2006). If people know why they punish and are internally consistent, then these two measures ought to be highly related. That is, the behavior and the attitude ought to correlate. But the results showed that these two constructs had no relation to each other. Indeed, the reasons that people generated for their
just-completed punishment bore no relationship to the pattern of their responses across the four vignettes and across multiple measurement constructs (all \( r < .08, \text{ns} \)).

The second study examined the consequences of this disjunction between what people say and what people do when it comes to punishment. We hypothesized, and found, that people will endorse utilitarian policies in the abstract not realizing that they will reject those same policies in practice because the policies violate retributive tenets of punishment.

The inspiration for these studies came from the numerous observations of citizens enacting utilitarian legislation in one year, only to repeal that legislation shortly afterwards. California’s 3-strikes law is a case in point. Citizens overwhelmingly enacted this incapacitative law with a 70% majority in a statewide referendum. The law stated, simply, that after a third criminal conviction perpetrators go to jail for a very, very long time. What the citizens did not realize at the time, though, is that this law would deviate from their retributive intuitions. As a result of this law, offenders guilty of quite minor crimes (e.g., theft of a children’s video tape, or a Snickers Bar) faced sentences of 25 years to life. Once citizens witnessed this law in action, public support plummeted and surveys showed support to be under the 50% mark just a few years later.

The second study of Carlsmith (2008) mimicked this phenomenon. It first described the problem that schools have had with student drug use, and asked people (again, broadly representative) to evaluate several different approaches to the problem. Although the majority preferred proportional responses to offenses that would be called for by a retributive approach, a substantial minority (30%) supported a zero-tolerance approach based on utilitarian theory. When these same people discovered that these zero-tolerance rules violated proportionality, and thus their intuitive sense of fairness, support for the rules dropped precipitously to 6%. The paper concluded that people have only limited insight into the reasons for their punishment, and that this ignorance can lead them to endorse policies that, ironically, they will soon reject as unfair and unjust.

One more research finding is relevant to the present argument. “Restorative justice” is a phrase that refers to a set of justice practices that many who are dissatisfied with the more punitive aspects of our criminal justice system find attractive. These practices focus on justice procedures that bring about repair of the harms done to crime victims, restoration of the offenders to a law-abiding life style, and involve the community in the process. The debate within the restorative justice movement concerns whether retributive punishments are also permissible within the system; many, following Braithwaite’s lead, hold that they are not. We find a great deal to admire in the ideals of the restorative justice community, but given what we have reported about the intensity of the just deserts punitive impulse, we doubt that citizens will be willing to allow a purely restorative,
punishment-absent, treatment of at least reasonably serious crimes. In a scenario experiment (Gromet and Darley, 2006) testing this intuition, participants played the role of a judge assigning criminal cases to various court systems. One system was “purely restorative,” and it was made clear that the usual punitive options were not available. A second system, modeled on a procedure suggested by Barton (1999), allowed a mix of both restorative and punitive procedures. The third system was similar to the traditional court system in that prison sentences were the only corrective recourse.

The findings were quite clear. For crimes of very low moral magnitudes (e.g., Halloween mischief), about 80% of the respondents were willing to send the offender to the court that did not have punitive procedures at its disposal. However, for crimes of more moderate seriousness, just over 40% were sent to pure restorative courts, with a bit over 40% going to the mixed restorative/punitive procedure. For the serious crimes, such as attempted murder or rape only 10% sent to the pure restorative court, a bit over 65% sent to the mixed court, and about 25% sent to the court that had only punishments at its disposal. This was a small sample study, but if its results replicate on larger, more representative segments of the population, then the policy conclusion is that a majority of people will not tolerate a purely restorative justice system that does not allow for retributive punishments for moderately serious and more serious offenses. But if it is granted that there are actual gains for the victim and perhaps for the offender in restorative justice practices even if they take place in a context in which punitive sanctions are also likely to be imposed, then the “mixed” branch of the restorative justice system gains empirical support from these results. At the level of psychological theory, the conclusion might be that there is indeed a requirement that, when serious crimes are committed, punishment ensues; but most people would also allow for restorative practices as well. In a later experiment in the article, in which the respondents were told that the offender had successfully completed the victim-helping commitments he had made as part of the restorative agreement, across all the various offenses, they suggested imposing reduced prison sentences on the offender.

2.13. Connections with previous research

Our work on punishment motives is derived from the philosophical debate articulated by Kant, Bentham, and others. However, that particular debate can be usefully traced back to Plato (cf. Republic) and Aristotle (cf. Nichomachean Ethics), who discussed underlying principles for justice and fairness. Briefly, Aristotle believed that justice is achieved through a bottom-up process in which a series of fair decisions at the individual level inevitably and ultimately lead to a fair outcome for society. By contrast, Plato argued that fairness was best achieved by a top-down process
in which the fair outcomes for society were defined first, and individual cases were decided so as to achieve the broad outlines of justice. This distinction has been usefully updated and articulated by Brickman et al. (1981) in their discussion of micro- and macrojustice orientations. They suggest that people use different criteria for justice depending on which orientation they ascribe to, and that the criteria for each orientation inevitably conflict with each other. According to these authors, microjustice is concerned with the qualifying attributes of individuals. That is, it takes a myopic view of the individual without regard for the larger needs of society and examines the extent to which the particular individual deserves a particular outcome. By contrast, macrojustice is concerned with the appropriate order and well being of society. The deservingness of the individual takes a secondary role to that of society. Thus, microjustice is individuating whereas macrojustice is deindividuating.

Brickman et al. (1981) discuss these two orientations with regard to reward allocation rather than punishment, but we think that their points are broadly applicable and can be usefully mapped onto our discussion of utility and deservingness, a mapping that Miller and Vidmar (1981) foreshadowed in their chapter on the Psychology of Punishment Reactions. Miller and Vidmar presented a 2 x 2 table in which the basic motives for punishment (behavior control and retribution) were crossed with the target of the punishment (offender vs. other in the social environment). A utilitarian mindset is triggered when one is oriented toward macrojustice, and a deservingness or retributive mindset is triggered when one is oriented towards microjustice. Brickman et al. (1981) suggest that the orientations are mutable, but that microjustice is more psychologically salient across situations.

A basic social psychological question is raised by both Miller and Vidmar (1981) and Brickman et al. (1981) in their discussion of justice orientations. Namely, under what conditions might a person adopt one or the other orientation? We would suggest that a macrojustice orientation is produced in an individual when he or she is asked about whether punishments should be increased for a particular crime when the rate of that crime is perceived to be rising and thus threatens their sense of safety. The micro-orientation is triggered when a person is deciding what penalty is just for a person who has been convicted of a specific crime.

Recent work on Temporal Construal Theory by Liberman and Trope (Liberman and Trope, 1998; Trope and Liberman, 2000) also provides some guidance on this question. Briefly, these researchers show that temporally distant situations elicit abstract outcome-based orientations whereas temporally proximate situations elicit the opposite. We suspect, but do not yet have data to support, that temporally proximate decisions guide individuals to microjustice concerns and thus to applying principles of
deservingness and retribution. Temporally distant decisions guide individuals to macrojustice concerns focused on achieving a good outcome for society at large, and thus to applying principles of utility.

These two approaches are consistent with our findings that people are intuitive retributivists. We have found scant evidence for utilitarian orientations in behavior, but have found substantial support for it in self-reported goals of punishment (Carlsmith, 2008; Carlsmith et al., 2002). It seems probable that when an individual is faced with a specific sentencing task, it is temporally proximate and this triggers a microjustice orientation. Accordingly, the individual is highly sensitive to the extent to which the perpetrator merits punishment and assigns a correspondingly proportional sentence. By contrast, when the punisher is asked to report their general goals for punishment or to recommend sentencing policies and guidelines, he or she is cued to think at higher, more abstract, levels and thus adopt a macrojustice orientation. This in turn leads the punisher to rely on principles and sentencing strategies that will achieve societal goals even at the expense of proportional outcomes for individual offenders.

3. The Impulse to Punish as an Intuition

Increasingly, the evidence suggests that the decision to punish an individual is normally an intuitive rather than a completely reasoned decision. What does asserting that these judgments are intuitive mean here? In the last decades, judgment and decision-making researchers have demonstrated that people frequently make decisions via heuristic processes, in contrast to the conscious, deliberative processes that we think of as “reasoning processes.” Kahneman (2003) and others have suggested that it is important for psychologists to differentiate between judgments and decisions that are made by heuristic, intuitive methods and those made by conscious reasoning.

The processes of the intuitive system are identical to those of the perception system, in that they are rapid, can occur in parallel with other thought processes, and are automatic. However, their outputs are similar to the outputs of the reasoning process, in that they are likely to be conceptual representations, judgments, or decisions. As is well-known, people experience their perceptions as simple correct representations of “what is out there”; that is, people experience the perceptual world in the mode of a naïve realist. In turn, intuitions, like perceptions, are often taken by “the intuitor” to be as unproblematically correct as perceivers take their perceptions to be. Thus, there are a set of intuition-produced decisions, choices, and problem solutions that are experienced as summaries of the ways
“the world is,” because the person having the intuitions is unaware of the complex, potentially incorrect, cognitive processes that produced them.²

As we mentioned previously, when respondents receive a scenario in which some person commits a known morally wrong action, respondents experience a reaction of moral outrage that is a substantial predictor of the relative punishments that will be assigned to the perpetrator of the immoral action. We would suggest that this “feeling” of moral outrage is the conscious registration of the intuitive reaction to instances of moral wrongdoing. Kahneman (2003) comments that the assessment of the degree of “badness” represented in a stimulus is an intuitive system judgment: “Some attributes, which Tversky and Kahneman (1983) called natural assessments, are routinely and automatically registered by the perceptual system or by (the intuitive system) without intention or effort . . . . The evaluation of stimuli as good or bad is a particularly natural assessment. The evidence, both behavioral and neurophysiological, is consistent with the idea that the assessment of whether objects are good (and should be approached) or bad (and should be avoided) is carried out quickly and efficiently by specialized neural circuitry.”

Several psychological researchers have also concluded that the sorts of punishment decisions that we have been considering are products of intuitive reasoning. They demonstrate that people report quick judgments that certain acts are morally wrong, or deserving of a certain level of punishment, but are unable to report the reasoning that led them to that decision. Haidt’s (2001) well-known work on moral dumbfounding demonstrates that people quickly report that, for instance, an act of brother–sister sexual intercourse is wrong. But when Haidt’s experimenter pushes subjects to supply reasons why it is wrong, they often cite reasons, for instance the psychic damage that could be done to one of the participants, or that a genetically defective baby could result. However, those reasons cannot be true, as the experimenter points out, because elements in the story told to the participants rule out those possibilities. Subjects often end, having run out of reasons why the act is wrong, asserting, “it is just wrong,” thus falling back on their intuitions.

Other researchers, who have worked with what are called Trolley Car problems, report a similar effect. The core of the trolley car scenario, invented by Philippa Foot, describes a runaway trolley car that, if it continues, will kill five people who are on the track farther down the incline.³ But this outcome can be mitigated if you, the respondent, throw

---

² This section draws on work that JMD drafted for a forthcoming paper by Paul Robinson and John Darley (in preparation).
³ In Foot’s version of the problem, the five persons “have been tied to the track by a mad philosopher.” Foot was a philosopher, and allowed herself this in-joke. More recent researchers alter the story, and simply say that the five are, for instance, workers who will not see the run-away trolley soon enough to escape it.
a switch to divert the trolley to another track, in which case it will kill only one person. In an alternative scenario, the respondent must take direct action against an innocent to save the imperiled workers: “As before, a trolley is hurtling down a track towards five people. You are on a bridge under which it will pass, and you can stop it by dropping a heavy weight in front of it. As it happens, there is a very fat man next to you—your only way to stop the trolley is to push him over the bridge and onto the track, killing him to save five. Should you proceed?” The utilitarian considerations are identical here in the two cases: killing one person to save many. However, most people report that throwing the switch would be the moral action and they would do it, yet most people report that pushing the person onto the tracks would be immoral and they would not do it. When one set of researchers (Hauser et al., 2007), who had administered both scenarios to the same respondents, (a within subjects design), confronted their respondents with their discrepant decisions about what seemed a morally identical case, about 70% of the respondents could not give an explanation of why they differed on the two cases, echoing Haidt’s moral dumbfounding work.

Both sets of researchers assert that these judgments are intuitive ones: Haidt (2001, p. 814) concludes that moral judgments derive from “quick automatic evaluations (intuitions),” and Hauser et al. (2007, p. 125) suggest, “much of our knowledge of morality is . . . intuitive, based on unconscious and inaccessible principles.”

3.1. The role of the reasoning system in moral judgments

Provisionally, let us accept that it is often, perhaps always, the case that a person’s first reactions to a moral violation are intuitive in nature, and largely driven by just deserts concerns. The next task is to examine the role of the reasoning system in the final decisions. Kahneman sets the question: “In the context of an analysis of accessibility, the question of when intuitive judgments will be corrected is naturally rephrased: When will corrective thoughts be sufficiently accessible to intervene in the judgment?” (2003, p. 711).

As this suggests, it is possible for a person to learn to apply reasoning processes that can correct and override the conclusions presented by the intuitive system. As an example, many statistically trained psychologists have done so in the area of probability updating. Specifically they have learned that, relying on intuitions, they do a bad job of probability updating, and so know that they need to process the information through Bayes equations to come to the correct answer. However, statistically sophisticated researchers do not always draw on their knowledge when they should. In one of their earliest studies, Tversky and Kahneman (1971) demonstrated that...
statistically trained social scientists often failed to do this. As Kahneman comments (2003, p. 697), “remarkably, the intuitive judgments of these experts did not conform to statistical principles with which they were thoroughly familiar . . . We were impressed by the persistence of discrepancies between statistical intuition and statistical knowledge, which we observed both in ourselves and in our colleagues. We were also impressed by the fact that significant research decisions, such as the choice of sample size for an experiment, are routinely guided by the flawed intuitions of people who know better.” He goes on to conclude: “People are not accustomed to thinking hard and are often content to trust a plausible judgment that quickly comes to mind” (Kahneman, 2003, p. 699).

There are reasons for this. Intuitive judgments, like perceptual judgments, do quickly come to mind, and the fact that they are products of interpretive processes is not apparent to the person making the judgments. Because they are rapid, automatic judgments, they leave no reminders that they are interpretations, and thus provide no cues to the person that they might be incorrect. The intuitive system processes are like those used in our visual perception system, and psychologists have recognized that people are “naïve realists” with respect to what they perceive. They give percepts an “it couldn’t be otherwise” character. We are suggesting that the products of the intuitive system have a similar character.

Still, the reasoning system can be drawn on to override the conclusions of the intuitive system. As this suggests, the intuitive and reasoning system distinction creates yet another dual process theory in psychology in the area of human moral judgments. The question for those attempting a description of moral reasoning then becomes the question that faces all dual process theorists, which is to give some specification of the conditions that toggles the processor between experiencing intuitions and acting on them, as opposed to bringing reasoning processes to bear and following the reasoned conclusions, overriding the intuitions. For this article, a great deal rides on the answer to this question. If it is true that a person’s initial response to an incident of wrong doing is an intuitive, just deserts-driven one, yet this initial response is always or often overridden by a more reason-based decision, then we should quit analyzing the intuitive moral system and begin understanding the principles of the reason-based account.

We will argue that the kinds of moral reasoning that leads to the punishment judgments that comprise what we are attempting to model is almost always just deserts based. The first part of the argument is that moral reasoning often stops with the intuitive system result. This intuitive reaction is generally rapidly, automatically, and nonoptionally produced. And for the reasons we just mentioned, these intuitive reactions have a “truthy” character that makes them likely to be acted on without scrutiny.
There is a second reason why intuition often prevails. Moral judgments are often made by persons who are under what we now call “cognitive load.” Kahneman comments that:

As in several other dual-process models, one of the functions of System 2 (the reasoning system) is to monitor the quality of both mental operations and overt behavior (citations omitted). As expected for an effortful operation, the self-monitoring function is susceptible to dual-task interference. People who are occupied by a demanding mental activity (e.g., attempting to hold in mind several digits) are more likely to respond to another task by blurting out whatever comes to mind. (citation omitted)

To sum, punishment reactions are the product of a dual-process system in which the retributive desire is automatic, and the reasoning process that might override it is only selectively brought online.

But even if the reasoning process were triggered into action, would reasoning always override intuition? Previously we suggested that the intuitions that were driving reactions of moral outrage and relative severity of sentence assigned to crime vignettes were based on just deserts considerations. Is it the case that these just desert sentencing intuitions would be overridden if more deliberative reasoning processes were applied to the cases? In other words, is it the case that people will display more complex thinking about punishments for morally wrong actions if they were caused to reason about the cases? Two studies we reviewed previously suggest this is not the case, and that reasoning about punishments is often done from a just deserts perspective. First, in Carlsmith (2006) respondents were led to engage a reasoning process; they acquired an item of information that they selected, thought through how that affected their assessment of the appropriate punishment, and then chose another item to further refine their judgment. Results showed that, overwhelmingly, respondents acquired items giving themselves retributive information to inform their judgments. Thus, their conscious reasoning system sought retributive information as the relevant decision inputs. This is quite important. It suggests that the study respondents are not motivated to deploy a reasoning process that deviated from a just deserts-based approach when they had the chance to do so.

Second, recall that in Darley et al. (2000), participants were instructed to take either a retributive or utilitarian stance before assigning punishments. Participants had no trouble taking a retributive, just deserts stance, and the punishments they assigned while taking that stance did not differ from the punishments they assigned previously, when they were free to assign whatever punishments they thought appropriate. However, even when asked to punish according to a utilitarian stance, the participants were not able completely to do so. Their punishment patterns showed considerable reliance on just deserts, retributive considerations.

From these two studies, we conclude that although it is undoubtedly the case that people are capable of reasoning about appropriate punishment
assignments, these reasoning processes do not routinely contradict the retributive, desert-based judgments that are supplied by the intuitive system. In fact, reasoning processes often also seem to be driven by just desert considerations.

We next want to present brain-imaging research, but want to be quite clear on what it adds to the case we are presenting. First, and most impressively, it supports the claim that moral reasoning can involve dual processes, and adds the possibility that these dual processes take place in somewhat different brain locations. Second, it supports the idea that the first process is rapid, one that is generally computed when the moral case involves an individual committing direct harm to another. We interpret this to mean that it is an intuitive system product. There is also a slower process, involving brain areas associated with higher order reasoning. This slower process can sometimes override the first process decision.

These dual process results come from two neural imaging studies conducted by Joshua Greene and his collaborators (Greene et al., 2001, 2004). In these studies, the respondent was told a story, and the story ended with an action that the protagonist could take. The task of the respondent was to decide whether the action was inappropriate or appropriate. Concretely, the respondent might be told the trolley problem in which the protagonist threw the switch to save five and kill one.

In the first study, using trolley car and similar other problems, they found that a set of problems they call “personal violations” activated brain regions that previous research has associated with both emotion and social cognition activities. Other, “impersonal moral dilemmas” activated increased activity in areas associated with abstract reasoning and problem solving. The trolley car problem that posed the choice of letting the trolley continue on its tracks and killing five persons versus throwing a switch to shunt the trolley to a track on which it would kill only one, was an impersonal moral dilemma. It did not engage moral reasoning areas. Our previous “footbridge” example, in which the possible trolley-stopping action was pushing a specific other individual in the path of the oncoming trolley car, is a personal violation case: the action required is a direct personal one in which “I act directly to harm a specific other person.”

These personal violation cases generally drew quick reaction time decisions that the action in question was an inappropriate one to do and involved heightened brain activity in the emotion and social cognition areas (specifically, the medial prefrontal cortex, posterior cingulate/precuneus, and superior temporal sulcus/temperoparietal junction).

Most of the personal moral dilemmas were of a simple character, for instance, should a teenage mother kill her unwanted newborn infant. These were the cases decided rapidly, and overwhelmingly, the proposed response was judged “inappropriate.”
But a few cases elicited quite a different pattern of results. Here is one such case.

Enemy soldiers have taken over your village. They have orders to kill all remaining civilians. You and some of your townspeople have sought refuge in the cellar of a large house. Outside you hear the voices of who have come to search the house for valuables. Your baby begins to cry loudly. You cover his mouth to block the sound. If you remove your hand from his mouth, his crying will summon the attention of the soldiers who will kill you, your child, and the others hiding out in the cellar. To save yourself and the others, you must smother your child to death. Is it appropriate for you to smother your child in order to save yourself and the other townspeople?

Some respondents reported that in this case the personal violation action in question was appropriate rather than inappropriate. Generally those respondents realized that smothering the baby, although a horrible thing to do, was required to save the group and decided the action was appropriate. Our guess is that it was the final sentence of the scenario, “smother . . . in order to save . . .” that triggered the reasoning system into action. Other subjects thought for a long time, but decided the action was still inappropriate. (Both groups could be matched on reaction time.) Researchers interpreted the associated brain activation patterns as evidence that when participants responded in a utilitarian manner (judging personal moral violations to be acceptable when they serve a greater good) as reflecting both the involvement of abstract reasoning centers (regions in the dorsolateral prefrontal cortex) and also the engagement of cognitive control (specifically the anterior cingulate cortex, a brain region associated with cognitive conflict) in order to overcome the social-emotional responses elicited by the thought of smothering the baby.

On this account, dual processes make moral judgments: one process that is relatively rapidly produced is the product of social cognitive and emotional responses, and takes place nonoptionally. This is the intuitive system we discussed earlier. The second process involves abstract reasoning areas of the brain, ones that developed evolutionarily later than did the social cognitive and emotional brain areas, and is not always triggered into action. This, we suggest, is the reasoning system. Further, when this latter system is activated, its results are sometimes in conflict with the intuitions of the other system. The conflict is perhaps resolved by some assessment of the competing strengths of the two sets of signals. So the reasoning system that here is giving us the utilitarian result and overriding the impulse against killing a single other individual, is acting to monitor or limit the intuitive system result. It is this “override” response that is possible, but that we suggest is rare for the punitive judgments we are researching.
4. Policy Implications

The previously reviewed results seem to us to have some policy implications for various aspects of our justice system, and we now turn to what those might be. We argue that when people seek to follow the law, their impressions of what the law is, what the law allows or prohibits, is drawn less from their knowledge of actual laws than it is from their intuitions of what the law must be. And those intuitions are read off from their moral intuitions. That is, people tend to think that the legal codes are in accordance with their own moral intuitions. One reason that this is so is that these intuitions that give people their take on what the law holds come to mind automatically, and this gives the intuitions a character of certainty that resembles the certainty that people give to the products of their perceptual system.

This in turn means that, if legal codes differ markedly from citizens’ moral intuitions, then the legal system needs to make the considerable efforts necessary to inform citizens about what the law should be, and persuade them of the moral correctness of the court-held laws. Otherwise citizens will not agree with the criminal codes, and as we will demonstrate later, this will have negative consequences for their willingness to voluntarily obey the law.

This persuasion can be done, but it is difficult. It entails either engaging the reasoning system to override the intuitive system, or altering the intuitive processes to yield the desired answer. We suspect that both of these routes would be difficult, but little has been written on the topic because scholars have only recently conceptualized intuition and reasoning as dual process systems (cf. Oswald and Stucki, 2008). We will make a few preliminary observations.

For a reasoning system to override an intuitive judgment or decision, the individual must recognize that the intuition comes from a class of judgments that the individual has learned should be overridden. The question, then, is what percept triggers this override and brings the specific reasoning principles into operation. Sometimes a cue to engage the reasoning system comes from the form of the problem itself. One of us has (finally) learned to recognize when the form of a statistical question involves updating prior probabilities based on new information. This realization cues us that Bayesian updating calculations solve the problem, and further, that his intuitions about these calculations are generally wildly wrong. Reasoning ensues; intuition is overridden.

For this process to occur, however, the individual must recognize the fallibility of intuition, and this requirement may be particularly difficult when it involves the appropriate punishment for some specific case of wrongdoing. Those intuitions come with powerful associated emotions, and those emotions exert influence on both the reasoning and intuitive systems of judgment. Moreover, as we will show later in this paper, both
intuitions about punishment and reasoning about punishment appear to draw on principles of just desert. Thus, it is not clear that a person is likely to switch systems, or that switching systems are likely to alter the outcome.

The alternate process would involve changing the intuitive reaction. The nearest process to this that psychology is familiar with is emotional conditioning. The task here is to condition emotional reactions to certain stimuli. One common suggestion in the moral judgment literature is that emotions, such as anger, disgust, and contempt (that are the ones linked to moral judgments of the sort we are discussing), can be conditioned to respond—or not to respond—to a given stimulus. An experiment by Wheatley and Haidt (2005, experiment 2) has a condition that is wonderfully on point for our purposes. They showed highly hypnotizable respondents stories with no morally transgressive actions and that elicited no strong reactions from control groups. For subjects who were hypnotized to experience disgust at trigger words in the story, however, the story reliably elicited ratings that the actions were morally wrong!

If conditioning generally produces the emotional components of reactions to moral transgressions, then an insight about moral learning follows. As we know, emotional reactions to situations are often learned with few trials. This could explain how it is that children in a particular culture rapidly learn what counts as “wrongdoing” in their cultures. They do so because they experience the emotional reactions of adults as they witness the wrongdoing actions.

Sometimes, of course, the task of the person is to override or cancel the intuition that a particular action, which currently triggers the moral emotional response, is a wrong action. The emotional conditioning literature suggests, however, that eliminating the emotional component of the response is not an easy task. Habituation, such as happens for medical students dissecting corpses, may eventually come about, but via the dual process avenue we have suggested. The habituation may not be the disappearance of the emotional component of the response, but rather its weakening. This weakening allows it to be overridden by processes of reasoning that provide justification for the hitherto transgressive action.

4.1. Evidence for discrepancies between legal codes and community intuitions

In 1995, Paul Robinson⁴ and John Darley published Justice, Liability and Blame: Community Views and the Criminal Law. They report the results of 18 studies that searched for discrepancies between legal codes and

⁴ Paul Robinson is the Colin S. Diver Professor of Law at the University of Pennsylvania School of Law. His area of specialty is in criminal law, and particularly the empirical justifications for the doctrines from which criminal law is generated.
community views on what those codes should be. The method of the studies was standardized. Scenarios were created describing various actions that were possibly criminal in nature. It was the task of the respondent to assign whatever penalty he or she thought appropriate to the action. Respondents could decide that the action generated no liability—in other words the actor did not do anything that was criminal, or that the action generated liability and the condemnation associated with that, but did not require a prison sentence, or could assign a prison sentence that ranged in ascending steps from one day in prison through life in prison to the death sentence. The designs were usually within-subject designs, in that each respondent assigned penalties to a number of different scenarios. The within subject designs meant that subjects were made aware that they were being asked if they thought that the presented cases should be treated differently. (Although they were told that they would sometimes receive cases, to which it would be perfectly appropriate to assign the same penalties.) This we thought was the correct methodology, because we were interested in whether subjects’ judgments about cases differed under conditions of conscious comparison.

The scenarios were chosen to fit the definitions of crimes that were generated from the relevant laws (Robinson and Darley, 1995, Study 5, scenarios 2, 4). The issue investigated in this study was the correspondence between the legal rules about what constituted appropriate use of violence in self-defense and the citizens’ moral intuitions about this. The first scenario describes what the criminal code regards as a legitimate act of self-defense because the person cannot retreat from the threat. However, in the second scenario, shooting the other is not legitimate, because the threatened person is aware he could retreat and the codes generally demand that the person retreat when it is possible to do so. More precisely, according to the law, the first shooting is legitimate self-defense and receives no penalty but the second case is murder and should receive a high penalty. In fact the average liability assigned by our respondents to the second scenario was between 6 months and a year in prison and this was all that was assigned even though the respondents agreed with the statement that the person could have safely retreated. Further, 23% of the respondents assigned no liability and an additional 17% thought that the shooter should receive the censure of conviction, but serve no prison time. “Americans stereotypically ‘stand their ground’ and our subjects seem to want them to even when legal codes say they should not” (Robinson and Darley, 1995, p. 60). There has been a tendency for codes to shift toward approving this; several states have weakened the requirement to retreat when faced with a threat of violence.

---

5 This is a fair characterization of the law as it was stated at that time. Since then some states have changed their laws in the direction of not requiring retreat.
4.2. When do steps toward a crime cross the line to become a crime?

Legal codes have needed to consider at what point a person’s actions toward committing a crime become the crime of attempt and what the sentence should be for attempted crimes. Interestingly, the criminal codes have changed their stance on this question, and done so in a way that seemed to modern code drafters to make psychological sense.

Some brief history of the criminal law of the United States is needed here. The phrase, “common law,” used in the context of the law of the United States, refers to a patchwork of laws originally inherited from England in the period when the United States was a colony of England, and then modified by rulings made by judges or legislatively imposed by state congresses over the years. The criminal codes in effect in the United States until the middle of the 20th century were common law products.

In its treatment of attempt, the common law focused on how close the actor had come to completing the offense—his proximity to the offense, as it was called. The usual test for this was “dangerous proximity,” a test that would be fulfilled, for instance, if a burglar were on the property of a house he intended to rob.

In the 1960s, it was increasingly recognized that the common law criminal codes sometimes contained provisions that seemed odd, archaic, unjust, or contradictory. Further, criminal codes differed significantly by state. All of this seemed irrational and unjust to the prestigious American Legal Institute who spent some years designing the “Model Penal Code.” The code rationalized and systematized the chaotic patchwork of common law criminal codes. The orienting assumption that the code drafters took on was to create the code that a reasoning person in the modern era would judge that the penal code should be. Their efforts were successful: The model code they drafted has been adopted in whole or in part by 37 states.

The model penal code took a very different stance on defining criminal attempts than did the older common law codes. For example, the common law generated some counterintuitive results in its treatment of “impossible attempts.” It allowed a defense if an actor bought goods believing them to be stolen goods but it later turned out to be the case that they were not actually stolen. This counter-intuitive result came about because if the goods were not stolen, then it was impossible for the purchaser to buy stolen goods. The model penal code drafters found that absurd. What matters is that the actor believed that they were stolen, and thus what he did was buy what he thought were stolen goods. The model penal code drafters held that the actor’s liability was determined from the point of view of the actor, with the circumstances as the actor believed them to be. So he was liable for the offense of attempting to buy stolen goods.
Reasoning from this, the model penal code drafters thought through the question of what conduct counted as an attempted crime. They reasoned that the essence, the gravamen, of the offense, was that the actor had formed the settled intent to commit the crime. More specifically, they looked at how far the actor had moved toward the offensive conduct, and held that the commission of a substantial step toward the commission of the criminal offense was the key test. It showed that the actor had a willingness to commit the offense.

This is often called the subjectivist stance of criminalization, because it focuses on actors’ subjective culpability in attempting to bring about the criminal action. Coming close to bringing about the harm or evil of the offense was not required. All that was required was that the actor had progressed far enough to demonstrate a fully formed intention to commit the crime, and that was what taking a substantial step toward commission of the crime showed.

This subjectivist view gives little weight to whether the crime was committed or not, because that does not alter the actor’s culpability. Therefore, the Model Penal Code had an easy and consistent answer to what the penalty should be for the crime of attempted X; it should be equivalent to the penalty for the crime of committing X. The older common law, focused as it was on the actual occurrence of the harm as the offense, habitually gave some sentence discount to actors even when they were discovered in dangerous proximity to the crime, because the harm had not occurred. (Different offenses had different specified discounts for attempts, further depending on jurisdictions.)

Thinking about these formulations, we thought that the Model Penal Code drafters might have gone too far in the subjectivist direction from the point of view of the citizens. So in a study of citizens’ perceptions of the appropriate penalty for attempt offenses (Robinson and Darley, 1998), various aspects of the difference between the subjective and the objective perspectives were tested. In the first test, it was found that for the case of impossible attempts, respondents primarily agreed with the subjectivist perspective. The actor who smuggled talcum powder back from Columbia because he had been tricked into thinking it was cocaine received a penalty only slightly less that the actor that smuggled real cocaine, rather than the zero penalty that the objectivist stance would assign.

But in a second set of tests, the pattern of results matched the pattern that would be given by an old-fashioned common law objectivist. The completed crime scenario was the murder of a prosecutor who had prosecuted a family member of the murderer. For this, the murderer received a life sentence, with a few respondents giving the death sentence. The second scenario involved the same actor, now a would-be murderer, who shot at the prosecutor from close range but missed. This is known in criminal law circles as a completed attempt. He received a considerable sentence
reduction to the successful murder—just above 15 years in prison versus the life sentence given to the murderer. In yet other scenarios built around the same core, the actor fully intended to kill the prosecutor, but was halted even farther from the completion of the act. Sentences were correspondingly reduced. Further, as accomplices were introduced into the scenario, the actor’s sentences were again reduced. It seemed to us that the respondents were assigning punishment graded in accordance with their notions of the actor’s psychological distance from the murder act. (For an attempt to model this distance in terms of additive discounts from the prototype offense, see Robinson and Darley, 1998, pp. 436–440.)

What this tells us is that the respondents held an objectivist view of how offenses and attempted offenses should be graded—and they feel that graded offenses should have calibrated punishments assigned to them. But our respondents were staunchly subjectivist in punishing those who completed attempts at, for instance, smuggling when they had been tricked into smuggling talcum powder rather than cocaine. As the authors (Robinson and Darley, 1998, pp. 429–430) comment:

When we consider what our respondents took to be the minimum requirements for having committed a crime, the subjectivist perspective fares better. Even when the actions taken ended far away from the successful commission of the crime, and the actor assisted in the steps toward the crime rather than was the intended principal, substantial sentences were assigned. The respondents criminalized and punished many actions than a strict objectivist would if the objectivist held to the rule of no liability is to be assigned short of actions being committed that were “in dangerous proximity to the offense.” But, and again, they did not punish these offenses as if they were equivalent to the completed offense.

It is possible to speculate that, in our modern society, in which dangers can be inflicted at long range, and lethal weapons enable preparations to turn quickly to completions, the point at which people’s thinking about the “lines” people should not cross has moved further away from the completion of a dangerous action, toward not starting the action. If so, we would expect criteria for when actions can be taken in self-defense to move comparably. We did notice that the respondents who defended themselves and killed when they could have retreated did not receive the murder sentence that the law calls for.

4.3. The ex-anti function: Do people know the criminal law?

One further study was done on the subjective objective issue and the crime of attempt. First, it gives us a more fine grain picture of the way that the punishment for the attempted crime scales up as the attempt converges with the completed crime. Second, it gives us our first information on the important question of whether and to what extent citizens are aware of the contours of the criminal law.
The researchers (Darley et al., 1996) also presented attempt scenarios to respondents, describing increasing steps toward the commission of a crime, ending with the completed crime of murder in one instance and robbery in another. Again, as in the previous experiment, the dependent variable, the duration of the prison sentence assigned to the various scenarios, increased as the nearness of the actions to the complete commission of the crime increased. Taking an initial step toward the commission of the crime received a relatively low sentence; reaching the dangerous proximity stage just prior to the planned commission of the act received a sentence of about 60–70% of the duration of the sentence for the completed robbery or murder. Again, respondents are grading objectivists in that they assign punishments that are linked to the nearness of the action chain to its completion.

In this experiment, the researchers also asked the respondents to report how they thought the legal code of the state they lived in would treat the actions described in the scenarios. More specifically, they were asked to report the duration of the penalties that they thought were assigned to the various scenarios by the laws of the state in which they resided. Subjects resided in New Jersey, which is a Model Penal Code state, so the actual laws criminalized attempt at the substantial step level and, it will be remembered, assigned actors who took the substantial step toward the crime the full duration of the sentence they would receive for committing the crime.

This result can be seen in Fig. 4.2. The line representing the sentence duration that the respondents would assign to the various scenarios is far below the high and flat line representing the sentences that the legal code would actually assign. (This is the line that, starting at the substantial step scenario, assigns punishments equivalent to the punishment assigned for the attempt.) However, in this study, we remembered to ask the respondents what sentence they thought the state of New Jersey actually assigned. All answered. Their answers are shown in the third line on the figure, which is essentially identical to their answers as to what they thought the morally appropriate punishment should be. What is suggested by this high degree of overlap between the duration that the subject assigns to each scenario and the duration they think that the law assigns is that they are generating “what they think the law is” from their own moral intuitions about what would be the morally right sentence for each scenario. It strikes us that this tendency to assume correspondence between what the laws “must say” and one’s own moral intuitions may be generalizable; a point to which we will return.

On public policy terms this discrepancy between what the respondents think the law holds and what the law actually holds means that the respondents, many of whom were long-term residents of the state of New Jersey, are unaware of some very important doctrinal elements in the law of the state in which they live, are in this case blissfully unaware that actions that
they think would attract only trivial criminal sanctions would in fact draw sentences equal in severity to those visited on offenders who complete the actual crime.

Is this an atypical finding, or are citizens often ignorant of important elements of the criminal law? The importance of the question is this. The legal code in a complex society is designed to have several functions. First, it is designed to announce beforehand the rules by which citizens must conduct themselves, on pain of criminal punishment. This is called the ex ante function, and is one of the central constraints imposed on societies that wish to claim that they are fulfilling the “rule of law” requirements. A law that is hidden from the society that it is designed to govern is either no law, or if it is hauled out of hiding and imposed on some unlucky citizen, it is an unjust law. The criminal justice system relies on people knowing the law and knowing where the boundaries for their conduct lie. Ignorance does not excuse unlawful conduct; a fact summarized in the phrase “ignorance of the law is no excuse.”

In an empirical study, we (Darley et al., 2001) investigated the generality of this finding of ignorance. Are citizens aware of the lines drawn by the criminal codes of their states? Do people know what the law says? Before answering this, we need to specify what sets of laws we are talking about. It would be possible to rig our test of the question by ferreting out obscure laws about trivial matters. On the other hand, it would be possible to rig our test in the other direction by examining, for instance, whether people knew
that murder was against the law. People know murder is against the law not because they have scrutinized legal codes but because they cannot conceive that a criminal code would fail to criminalize murder.

For our research, we chose to see whether people are aware of the laws about such issues as being required to assist a stranger in distress, report a known felon, or retreat before using deadly force in defense of self or property. These laws struck us as being about important conduct, about situations in which many of us may find ourselves and in which, therefore, we look to the guidance of the legal code.

The state-by-state adoption of the Model Penal Code provided a natural experiment, because different states slightly modified different aspects of the code as they adopted it. Clearly, these state legislatures were scrutinizing the code, and deciding that certain of the proposed model code formulations were inappropriate and needed to be changed. So they changed what they enacted into law. But were the citizens of that state knowledgeable about what the law held in their state? Testing this question on these laws seemed to us to set up a fair test of whether people are aware of the contours of various important laws that were intended to govern their behavior.

We selected four states, each with a somewhat deviant law about what counts as criminal conduct. We asked selected residents of these states to read a series of scenarios. One scenario described an offense that is criminal in most states, but not in the deviant state, or an action that is criminal in the deviant state but not in the other states. Each state served as an experimental group with the deviant law for one scenario, and as one part of three-part control group for the other three laws.\(^6\)

Our question is whether the citizens of a state with a deviant law knew the content of that state’s law; therefore, the residents were asked for each of the four “crimes” to report what liability and punishment the law of the state in which they lived would assign.

In this study, we were also able to test our prediction that a respondent would infer that the law of the state was what the respondent thought it ought to be, and in turn the respondent’s general punitive attitudes would generate what the respondents thought the law ought to be.

Here, on more psychological terms, is the chain of reasoning that we suggest generates a respondent’s answer to the question of whether the state has made an action unlawful. First, general moral attitudes determine whether a person thinks a particular action is morally acceptable or unacceptable. Second, when people are asked their personal view on the criminality of the action, their own moral attitudes determine whether they perceive the action as criminal. Finally, when people are asked whether the state in which they reside criminalizes that action, they answer yes

\(^6\) In one case, two states held the deviant version of the law and two did not.
or no not because they know what the code says but because they assume that the state, in its moral wisdom, shares their personal moral views. Of course, some steps in this chain of reasoning may occur out of the awareness of the respondent. The inference is more likely to be spontaneous and automatic.

To test the knowledge of state law hypothesis, we asked respondents whether or not their state criminalized the four actions, and the severity of sentences that should be assigned to each action. To test for support for this “inferencing” hypothesis, we measured a few of the respondents’ punishment relevant attitudes and also whether they themselves thought the action in question should be criminal or not. We then compared the answer patterns across the four states.

The study was conducted by enlisting staff\(^7\) of universities in the selected states to be respondents. The laws in question were these: most states impose no duty to assist a person in trouble, even if it could be done without much inconvenience. Wisconsin imposes such a duty. North Dakota and Wisconsin require that a person with force must retreat if he can instead of using deadly force on the attacker. Texas and South Dakota do not require retreat. South Dakota requires that one report to the authorities the location of a person known to be a felon, no other state does. Finally, no state allows the use of deadly force on an offender who is taking only property, except Texas.

The results of the study are complex, but gave little support to the notion that the citizens of the state with the variant law accurately predicted that law. Respondents were first asked whether the action in question was criminalized by their state or not. For three of the four comparisons, the citizens of the deviant states had no reliable differences from the control states in their predictions of what was or was not the law. For instance, 36% of the Wisconsin respondents believed their state would punish a person who failed to assist a person in need—which of course was the right answer. However, 39% of the respondents from the three other states believed their state would punish although in fact their states would not punish. On whether there was a duty to retreat when attacked, similar results occurred. North Dakota and Wisconsin hold that a person cannot use force against another if the person can retreat, and 79% of their citizens said the duty to retreat existed. But 71% of the citizens in the states in which no such duty exists thought it did. Finally, South Dakota has a duty to report the location of a felon and 80% of its citizens thought so, but 76% of the citizens in states with no such duty thought their states did have the duty to report. To conclude, all this looks like citizens are guessing rather than knowing the laws of their state.

\(^7\) Staff rather than students because we wanted respondents who were long-term residents of the state and likely to have been so during their adult years.
The exception occurred on the question of whether it was lawful to use deadly force in defense of one’s property—specifically to shoot somebody who was running away with a computer monitor taken from the car in his garage. Seventy-seven percent of the Texas respondents thought that was legal, and they were right. In Texas, it was legal. Only 49% of the citizens of the other states thought so. Does this mean that the Texans knew the law of their state? Hold that question until we present more of the results.

What did predict whether respondents favored a deviant or standard law, regardless of their state of residence, were their background attitudes (Darley et al., 2001, p. 189 for these questions). Using structural equation modeling on each case, we determined that for the three cases the background attitudes were a strong determinant of the respondent’s personal judgment about what the appropriate sentence for the action should be, and that personal judgment in turn predicted what the respondent reported was the sentence that the state law would give. The actual law of the state did not predict what we would now call the respondent’s guesses about what the state law was.

With respect to the analysis of the Texas data, we said this: “As in the other models, people’s general attitudes toward the use of deadly force to protect property predicted the respondents’ personal sentence ($\beta = 0.58$), and their personal sentence predicted their estimates of the state’s sentence ($\beta = 0.52$). The presence of a state law that permits the use of deadly force was predictive of reduced sentencing ($\beta = 0.24$), and it was somewhat predictive of attitudes ($\beta = 0.15$). It is important to note, however, that there was no direct or mediated relationship between state law and personal sentence. The fit of this model was excellent (RMSEA < 0.001)” (p. 181).

Our conclusion was that respondents generally did not actually know what the law of their state held. Instead they guessed that their state had gotten the law right, which meant that they guessed that the state law agreed with what they thought was the morally appropriate answer. In the case of Texas, we thought that the “lenient opinions of the Texans were there first and influenced the passage of the correspondingly lenient laws” (p. 182).

For the moment, accept the conclusion that in many cases, the citizens of a state are not particularly aware of the exact content of the laws of their states, that is, are not aware of the laws of the jurisdictions that govern them. We should qualify this. They do get it right that prototypical criminal acts, such as actors of unjustified violence, or theft, are criminal. Common law codes were made by judges who held the moral intuitions of their communities, and those intuitions held that these acts were immoral and thus criminal penalties were assigned to them. Later, as codes became products of legislative enactments, legislatures criminalized these actions for the same reasons.

On these cases, it is not the case that the citizens consciously recognize that they are predicting what the codes hold. Instead, they simply hold that
of course the codes criminalize those morally wrong actions. Indeed, we suspect that the legitimacy that citizens grant to the law (Tyler, 1990) stems largely from the presumed (and occasionally erroneous) similarity between their personal moral codes and their State’s legal codes. However, if people, in a blanket way assume that they know the exact content of the criminal codes, based on the code’s exact congruence with their own moral codes, the research we have reviewed in this section, which has demonstrated that citizens intuitions about codes often get it importantly wrong, then they are living in a fool’s paradise.

One last fact is worth noting. We searched the newspapers of the deviant states via Lexis-Nexis for stories about why it was that the legislatures were considering adopting variant versions of standard laws. We thought that if information about contemplated laws was ever going to attract public interest, it was the debates about why the legislatures should adopt these laws that were different from the normal laws that would do so. We did not find any such articles, using all the key words we could think of. If we take newspapers to be a proxy for coverage in the public media, then apparently the public media is not a medium to count on for transmission of legal codes to the general public.

This may be unfortunate. Among legal philosophers, there is considerable agreement that the Rule of Law is a precondition for the just society. One central tenet of the rule of law is that laws must be promulgated. The notion here is that they must be available to the public so that the public can know the laws and use them to guide their behavior. To serve their ex ante function, that is, the law codes must be promulgated.

What, specifically, does this mean? Various dictionaries signal two possible threads of the definition of promulgate. Merriam-Webster’s online dictionary says, “to make (as a doctrine) known by open declaration.” The American Heritage® Dictionary of the English Language (Fourth Edition, 2000) says “To make known (a decree, for example) by public declaration; announce officially.” Both of these call to mind images of a public pronouncement of the new law or code, more specifically, attempts to communicate the substance of the law to the public. However, Merriam-Webster’s also gives us this: “to put (a law) into action or force.” What might that mean? Apparently, concerning the content of the country’s laws, it is this second thread that governs. What it means with respect to laws is that they be published in some “official register,” generally State publications on the criminal, or administrative laws of the state.

The official register no doubt meets the legal definition of “promulgated” but from an intelligent social policy perspective, it certainly fails. That is, if our account of how people think it is that they know the laws of their state—as read off their moral intuitions, the “register as promulgation” notion certainly fails any sensible test. As our examples of laws, the legislatures have chosen to modify makes clear, it is exactly the parts of the law
that are morally unclear and under debate that are the ones on which legislative modifications are likely. No one is debating the law against murder, or against theft. What is debated, inevitably, are the specifications of the margins of what counts as theft, or duty, or exactly when violence in self-defense is allowable. So it seems truistic that it is exactly this aspect of the laws, on which controversy exists, that large segments of the populace will get wrong. For the legislative community to fail to grasp the chance to explain why it is they draw code lines where they do is a terrible failure of responsibility.

Recent shifts in grounds for making actions “against the law” make the situation worse. The existence of contradictions between code and community views has been increased by the recent tendency to blur the line between criminal offenses and violations of regulatory and other civil codes. The criminal justice system deviates from the community’s intuitions about appropriate criminal laws in two ways: by failing to “criminalize” actions that the community thinks are morally wrong, and by criminalizing actions that the community regards as morally acceptable. Depending on which of these intuitions is violated, the community may respond with different specific actions. If the law “allows” actions that the community regards as condemnable, then the community may seek to mobilize informal sanctions on these activities, including inflicting vigilante penalties as severe in magnitude as death. If, on the other hand, the law criminalizes actions that the community, or a segment of the community, thinks are morally permissible, then it is likely that those actions will continue to be practiced by people but will “go underground.” This in turn generates the rise of entrepreneurs who profit by facilitating people’s indulgence in those activities, the establishment of venues such as “speakeasies” in which those activities are practiced, and bribes to legal authorities willing to overlook the existence of these actions and venues.

But in addition to these undesirable reactions that attempt to engage or escape control mechanisms for these actions, we suggest that when the criminal justice system is seen as out of tune with community sentiments a less obvious but more common and troublesome reaction occurs—the loss of moral credibility that the justice system standardly possesses in a reasonable society. The justice system loses its relevance as a guide to good conduct.

Initially, the reactions may be limited to conclusions about the idiocy of the specific legal rule that offends the community’s morality, but as the apparatus of the social control agents of the government, such as police and the courts are mobilized to enforce the senseless laws, or as those forces seem to stand by passively as moral offenses are committed, there develops a generalized contempt for the system in all its aspects, and a generalized suspicion of all of its rules.
Others have recognized this as well. Deviations from what is perceived as deserved punishment may have consequences for the respect in which the criminal justice system is held. As Robinson and Darley (manuscript in preparation) state:

Just as the institution of the criminal law may be brought into disrepute by the too easy attribution of criminality in situations where the label criminal is generally thought inappropriate, so also may the institution be undercut if it releases as non-criminal those society believes should be punished. This does not mean that the criminal law may not be a means of educating the public as to the conditions under which moral condemnation and punishment is inappropriate. It does mean that the results may not depart too markedly from society’s notions of justice without risking impairment of the acceptability and utility of the institution.

Vigilante justice is an extreme version of a possible reaction. A host of other less dramatic—but more common—forms of resistance and subversion also can be committed when citizens perceive injustice. Jurors may disregard the jury instructions and convict or acquit according to their personal moral judgments rather than the legal rules. Police officers may not arrest or prosecutors prosecute if they think that the law in question criminalizes an innocent action or minor transgression. Or alternatively, police may informally punish actions that they think the legal codes fail to criminalize, or trump up charges against the offenders. Witnesses may lose any incentive to offer their information and testimony.

Tom Tyler’s (1990) groundbreaking research on legitimacy supports this. He has found that Americans are likely to obey the law when they view it as a legitimate moral authority. In turn, they are likely to regard the law as a legitimate moral authority when they regard the law as being in accord with their own moral codes. As Tyler concludes, “the most important normative influence on compliance with the law is the person’s assessment that following the law accords with his or her sense of right and wrong” (p. 64).

Two recent experiments (Greene, 2003; Nadler, 2005) provide evidence for the beginnings of this rejection process in individuals who discover contradictions between the legal codes in force and their own moral intuitions of justice. In these studies, participants read about a case in which there was a mismatch between the laws in existence and the moral intuitions of the participants. The mismatch was one that many participants found shocking. It was derived from a real world instance in which one young man dragged a young girl into a semi-private space and raped and killed her. A friend, aware of the series of actions, took no steps to intervene or report the action to authorities that might have intervened. The watcher and the friend spent the next two days gambling...
in casinos, later, the watcher bragged to friends about the crime. Because there was no law against the watcher’s actions in the state in question, no legal action was brought against him.  

Participants made aware of this and other unjust cases rated themselves significantly less likely to cooperate with police and less likely to use the law to guide their behavior after reading these intuition-violating cases. More specifically, participants who had read cases in which the legal system behaved in ways counter to their moral intuitions rated themselves more likely to take steps aimed at changing the law (including replacing legislators and prosecutors and breaking the law while taking part in demonstrations), less likely to cooperate with police, more likely to join a vigilante or watch group, and less likely to use the law to guide behavior. Overall, participants appeared less likely to give the law the benefit of any doubt after reading cases where the law was at odds with their intuitions. In the Nadler study, learning about a case in which a similar mismatch occurred on one law “caused participants to report a willingness to flout unrelated laws commonly encountered in everyday life as well as willingness of mock jurors to engage in juror nullification” (Nadler, 2005, p. 1399).

John Coffee (1992, pp. 1881–1882) makes the point that the actual frequency of such intuition-violating laws is irrelevant. Actual use of the criminal sanction might remain rare, but it is the threat of its use that must be chiefly considered in evaluating the degree of freedom within a society. To be sure, some may justify pervasive use of the criminal sanction based on simple cost-benefit reasoning: the loss to those imprisoned is less than the harm thereby averted through specific and general deterrence. Yet this analysis depends on a myopic social cost accounting. Even if the deterrent effect gained under such a system of enforcement exceeded the penalties actually imposed, additional costs need to be considered, including the fear and anxiety imposed on risk-averse individuals forced to live under the constant threat of draconian penalties. These citizens would bear not only the risks of false accusation and erroneous conviction, but also the constant fear that they might commit an unintentional violation. Ultimately, if we measure the success of the criminal law exclusively in terms of the number of crimes prevented, we could wind up, in Herbert Packer’s memorable phrase, “creating an environment in which all are safe but none is free.”

---

9 For a narrative describing the case facts and a discussion of the issues arising in it, see Paul H. Robinson, Criminal Law: Case Studies and Controversies, Section 15, "The Act Requirement and Liability for an Omission" (2005). Many respondents found the absence of a law criminalizing this conduct unbelievable. In their study of community intuitions of justice, Robinson and Darley (1995, Study 4) found that their respondents generally punished actors who failed in a duty to rescue a person in distress, if the intervention could be achieved without serious inconvenience or danger to the potential rescuer, as it could have been in the case described here.
5. Conclusion

Penal sentencing practices are much contested in our society, and one message from this chapter is that it is time for empirical psychology to speak-up. In the same way that psychology has made contributions to so many other aspects of justice policy, so too must it make research contributions to the punishment policy “contest.”

What do we suggest those contributions might be? In our society at least, crime descriptions mobilize a reaction that we have characterized as “moral outrage.” This is an intuitive rather than a reasoned reaction, one that brain-imaging studies suggest contains both emotional and cognitive elements. When presented with specific scenarios describing concretely specified crimes, a person’s intuitive reactions are driven primarily by just desert concerns. Utilitarian considerations, by contrast, are much less influential. As is often the case with intuitive judgments, it is possible that the judgment maker may mobilize reasoning processes to correct and override the intuitive judgment. May mobilize, but does not always or even often mobilize: And even when reasoning processes are mobilized, they frequently rely on just deserts considerations. In some of the studies we reported, the respondents used a reasoning process to arrive at their decisions, and these reasoning processes were also driven by just deserts considerations.

More research, with larger and more representative samples of citizens, needs to be done before we can claim these generalizations are descriptive of the responses to crimes by citizens of our society. Fascinating research will test whether the findings are consistent across societies—whether just deserts reactions are “human nature” or the results of cultural socialization. But for a moment, let us assume that they are general at least within our society, and comment on the policy relevant implications.

To begin, we need to say more about what the just deserts reaction is and is not. It is best characterized as a “proportional just deserts” stance. Von Hirsch has been the most consistent advocate of punishment according to a just deserts rule, and Wasik and Hirsch (1990) define proportional just deserts as the system in which punishment “severity should be allocated according to the blameworthiness of the criminal conduct” (p. 509). This is elaborated as proportional “to the conduct’s harmfulness or potential harmfulness and to the offender’s degree of culpability in committing the conduct.” Several comments are important to make here. First, there is no requirement that the punishment be “equal to the crime,” only that it be proportionate to it. Different societies have evolved toward having wider or narrower ranges of prison sentence durations, which are the normal mechanism for varying punishment severity. The United States, as we documented, has a wide range of durations including life without parole and the death
penalty, and assigns high-severity punishments far more than do other Western societies. The Scandinavian countries generally use a more restricted range of punishments. But all of them can employ proportionate punishment systems. What is required is that the most severe punishments the society gives are assigned to what the society regards as the most blameworthy offenses. This can be stated quite precisely on psychometric terms. The rank ordering of the moral wrongness of crimes should correlate with the rank ordering of punishments used by the society. To make what we are saying quite clear, a person or a society can decide never to give a death penalty, even to murderers, and still be doing proportional just deserts if it assigns the murderer the most severe prison sentence the person or the society allows. Proportional justice is not an “eye for an eye,” *lex talionis*, system.

Looking back on our discussion of when it is that the citizens of a society move toward disrespecting their justice system, it too can be stated on psychometric terms. Contempt will develop when the sentencing practices of the society are importantly out of synchrony with the citizens’ rank orderings of the blameworthiness of crimes.

We suggested that this was what was developing in the United States. We reviewed cases in which the criminal justice system failed to criminalize actions that the citizens found morally appalling, and did criminalize actions that the citizens found undeserving of the criminal sanction. We reviewed some preliminary evidence that when citizens discovered these discrepancies, they began to lose respect for the justice system, began to lose confidence in the legal system as a guide to moral behavior. We suggested that this was not a necessary result of community-code discrepancies. It should be possible to convince citizens of the reasoned conclusion that the criminal code “had gotten it right,” but that the criminal justice system was failing in its duty to persuade citizens that its criminalization patterns were the correct ones.

Many critics of a proportional just deserts system assert that the necessary consensus on blameworthiness cannot be found within any society, or at least within our society. They are simply empirically incorrect. We earlier discussed the correlations that have been found in samples in our society, and found them remarkably high.9

There is a simple formulation of what we wish to say. For a number of reasons, the forces acting to shape the criminal laws of our society are moving toward creating laws that are more and more distant from the moral judgments of its citizens about which behaviors are immoral and ought to be criminalized, and which are moral and should not be subject to criminal sanctions. As we, as well as others have documented, the psychological costs

---

9 We freely acknowledge that researchers such as Haidt (Haidt and Hersh, 2001) and Jost (Jost et al., 2003) have identified important topics in which delineated groups of people disagree passionately about the moral trespass of certain actions (e.g., abortion). But these strike us as notable exceptions to the otherwise consensual view within society of what constitutes a crime and harm doing.
of this is that the law loses credibility as an honest guide to moral behavior. Worse, it draws contempt on itself when, for instance, it fails to punish callous disregard for victims in distress, or in other cases demands punishment for actions that seem innocent to citizens. The “rule of law” is threatened when the rules of law violate citizen intuition.

REFERENCES


