Nahmias, E. & Aharoni, E. (2017). Communicative Theories of Punishment and the Impact of Apology. In C. W. Surprenant (Ed.) Rethinking punishment in an era of mass incarceration. Routledge.

Communicative Theories of Punishment and the Impact of Apology

Eddy Nahmias and Eyal Aharoni[1]

1. Introduction

The incarceration system in the United States is broken. It does not effectively—much less efficiently—serve the forward-looking goals of punishment, which rely heavily on fear of incarceration. It does not rehabilitate criminals nor prevent recidivism effectively. Evidence suggests the current system fails to deter any more than shorter sentences or better conditions would (e.g., Cullen et al. 2011). And the incarceration system does not serve the backwards-looking retributive goals of punishment effectively either, in part because, as we discuss below, most people are not satisfied with punishment involving merely impersonal incarceration or the attendant suffering of wrongdoers, except in response to the worst crimes. Rather, we want punishment not only to *send* a message to wrongdoers but also for the message to be *received*, internalized, and acted upon. The U.S. penal system is not effective at leading wrongdoers to recognize the harm they have done, to want to repair the harm, and to change so that they will not repeat such harms. Criminal behavior may indicate that the offender does not sufficiently value the community's norms. Punishment might be most effective in achieving multiple aims if it expresses both this (backwards-looking) message about the offender's undervaluation of the community's norms and the (forward-looking) message that the offender must increase his valuation of those norms.

Here, we will argue that the *communicative theory of punishment* may best combine features of, and intuitions supporting, backwards-looking and forwards-looking theories of punishment, while also going beyond the expressivist theory's limited aim of expressing the community's condemnation of crime. We will then suggest some empirical predictions of such a theory and present a pilot study aimed to test some of these predictions. A background assumption of our discussion is that improving our broken incarceration system requires a better understanding of people's punitive psychology, such that proposed reforms to the system can serve the functions of public safety and rehabilitation without requiring deviations from people's punitive psychology that would risk a rejection of those reforms.

2. The Communicative Theory as a Unifying Theory of Punishment

The communicative theory of punishment states that the aim of punishment, including but not limited to incarceration and 'hard treatment', is to communicate to offenders (and to the rest of the community) both the message that they have violated the norms of their community and the demand that they must respond to this message appropriately. Punitive acts ideally express the condemnatory message in ways that will lead offenders to repent their crimes, to repair harms they have caused, and to reform so that they can be trusted not to repeat their wrongdoing (see Duff 2001). Though most often advanced as a *justificatory* theory of punishment, the communicative theory provides a framework for a *descriptive* account of our punitive psychology that helps explain the diversity of people's intuitions about just punishment, in part by explaining people's close attention to the mental states of offenders both leading up to crimes and in response to being caught. We will focus here on the descriptive advantages of the communicative theory, highlighting some justificatory features as they arise. Below, we describe traditional theories of punishment in terms meant to be as neutral as possible between their being descriptive or justificatory.

The traditional theories of punishment (except for ‘mixed theories’) are ‘one-directional’. Pure retributive theories are solely backwards-looking, stating that we punish in order to give offenders what they deserve for their past wrongdoing, typically meted out in terms of suffering proportionate to the harms they have willingly brought about. Such theories attend to past mental states of offenders to assess what punishment they deserve, since responsibility for outcomes depends in part on what the agent knew, intended, and controlled. But retributive theories do not attend to the offender’s mental states in response to punishment (except perhaps the amount of his suffering)—for instance, whether he better understands his community’s values. While some suggest that retributive punishment is most satisfying for victims and members of the community, the theory does not suggest a role for the responses of victims or the community to the punishment of offenders. And it may be, as we explain below, that people are most satisfied with punishment when any suffering it engenders serves to make the offender understand the nature of his crime and motivate him to change accordingly.

Most rival theories to retributivism are purely forward-looking, stating that we punish in order to bring about particular consequences, such as specific or general deterrence, incapacitation, or rehabilitation (if one *defines* ‘punishment’ as requiring offenders to suffer pain or setbacks through hard treatment solely because they deserve it, then these rival theories will be *alternatives to*—not theories of—punishment). Consequentialist theories attend less closely to the past mental states of offenders, since they are not assessing desert. Instead, they consider mental states only to the extent that they are relevant in determining what it will take to deter or incapacitate actual or potential offenders or whether an offender has been rehabilitated sufficiently to release him. And they are concerned with victims and community responses only to the extent that these responses are relevant to general deterrence or acceptance of the legal system's practices.

None of these one-directional theories adequately captures people's various interests in or attitudes towards punishment, including our obsession with the mental states of offenders—their intentions, beliefs, attitudes, and traits both leading up to their crimes and in response to their conviction and punishment. A more expansive theory is needed to unify the plurality of our punitive aims and intuitions. As Bedau and Kelly put it, "A strait-laced purely retributive theory of punishment is as unsatisfactory as a purely consequentialist theory with its counter-intuitive conclusions (especially as regards punishing the innocent). The practice of punishment, to put the point another way, rests on a plurality of values, not on some [single] value to the exclusion of all others" (2015). (See also the pluralistic purposes for sentencing outlined in the U.S. Model Penal Code (American Law Institute §1.02.) and recent suggested revisions that reference both retributivist aims and restorative justice.)

The communicative theory of punishment is 'multi-directional' in a way that can address some of these limitations of other theories. It is backwards-looking, like retributive theories, in that it requires that punishment be communicated only towards those who deserve it *because* of what they actually did. And it suggests close attention to the mental states of the offender before and during the crime to assess what punishment he deserves to have communicated to him. To know what punishment is appropriate, we must assess the offender's intentions, plans, and knowledge.

The communicative theory also includes elements of the expressive theory of punishment, which assesses the guilt of the offender and nature of the crime in order to express a corresponding level of moral condemnation to him and to the community for the offense (e.g., Feinberg 1965). But unlike the expressive theory, the communicative theory takes punishment to involve more than just one-way expression of the community's negative response to norm violations. The communicative theory goes further to *demand a communicative response* from the offender. And this responsive feature also explains its forward-looking aims. The punishment system commits the community (through its designated authorities) to carry out punitive acts that serve as directives or commands to the target offenders to carry out those directives.

Hence, the communicative theory suggests that punitive acts aim to achieve outcomes, many of which overlap with the standard forward-looking theories of punishment. Punishment aims to alter the offender's relevant mental states so that he increases his valuation of the community's norms (his respect for the law). Hence, it is essentially rehabilitative, though not primarily by inducing fear of punishment, but rather by helping the offender become reintegrated into the community once he has signaled that he better understands and values its norms. Such signals suggest that two crucial features of communicative punishment will be evidence of remorseful apology and costly compensation. Given the facts of human psychology, these demands of remorse, rehabilitation, and reparation will require that the offender experiences deprivation and suffering, and in some cases may be achieved best through the 'hard treatment' sometimes presented as the essential feature of punishment. However, the primary aim of punitive acts, on this view, is not merely or mainly the suffering suggested by retributive theories. Furthermore, current incarceration practices are poorly designed to encourage the outcomes suggested by the communicative theory. To demand of offenders that they reform, repent, or restore requires substantial reforms to the penal system so that it provides the means to allow that these communicative demands can in fact be met by convicted criminals.

Indeed, communicative punishment treats offenders as autonomous members of the community—albeit, ones who have not properly valued its norms—rather than aiming to exclude them or treating them as 'objects' to be fixed, or as 'outlaws' literally *out*side the community's legal norms. In this way, it avoids the potential dehumanization of criminals suggested by purely forward-looking theories of quarantine or rehabilitation (e.g., C.S. Lewis 1949). Rather than taking what P.F. Strawson (1962) calls the 'objective stance' towards offenders—aiming to manage, control, or fix them—communicative punishment treats offenders from within the 'participant stance' with the expectation that they have the cognitive capacities to respond to the demands communicated to them. (The connections between the communicative theory of punishment and Strawson's influential theory of moral responsibility have not been sufficiently explored, but see McGeer 2011.) In treating offenders as persons who are members of the community, communicative punishment might be more effective at changing their behavior over the long term than punishments targeted directly and exclusively at behavior modification. Offenders motivated to internalize social norms regulate themselves, whereas motivating them with extrinsic reinforcers will only work as long as those reinforcers are present.

Nonetheless, people will be inclined to take the objective stance towards some criminals, especially those who commit the most heinous crimes, who recidivate, who are perceived as lacking relevant psychological capacities, or who are perceived as members of out-groups or having little potential value to the community. Some of these offenders may in fact lack the capacities to receive or respond to the message communicated by punishment. It may be that the communicative theory must 'bifurcate' and allow for different types of treatment for some of these criminals, just as Strawson allows that it will be appropriate to take the objective stance towards some who wrong us (see Duff 2001: 166). However, the 'worst of the worst' pose challenges for every theory of punishment (e.g., forward-looking theories may require permanent quarantine; retributivism may suggest execution). But the communicative theory can, in practice, suggest ways to increase the scope of those offenders whom we are inclined to see as appropriate targets of inclusive communication rather than exclusory quarantine or dehumanizing treatment. This process will likely require encouraging people to see fewer criminals as irredeemable or as members of an out-group and more as members of the community with whom communication is possible, with the hope that it can lead to reintegration. Doing so will presumably require reforms to the way the legal system describes and treats crimes, criminals, and punishment.

3. The Psychological Plausibility of the Communicative Theory

The communicative theory of punishment seems to accord well with humans' punitive psychology. First, it is an extension of the attitudes and expressions used in our interpersonal relations when others treat us in ways that show a lack of proper valuation (or respect) for our relationship with them. When our friends, family members, or colleagues violate our normative expectations, we feel and express indignation or resentment, which communicates to them demands for expressions of remorse, potentially for reparative actions, and especially for commitment to better future behavior. Our ultimate aim is typically not to make them suffer, except insofar as they experience negative emotions associated with remorse or the setbacks required to make up for their wrongs. The aim of these emotions from the offender and the offended—from moral anger to remorse (what Strawson calls the 'reactive attitudes')—is ultimately to motivate us to repair relationships. To the extent that we are able to see criminal offenders as members of the community and aim to reintegrate them, the communicative functions of punishment are *extensions* of, not replacements for, our interpersonal punitive attitudes and responses.

Second, the communicative theory resonates well with evolutionary models of punishment. According to these models, punitive motivations evolved to protect members of the social group against those who threaten the norms that enable the group to thrive. Those who were unwilling to support the punishment of cheaters would have suffered a fitness cost at the hands of those cheaters, breaking down the benefits of participating in a social group. For example, if some members consistently reap the spoils of group hunting without contributing to the hunt, and without facing punitive sanctions, these cheating behaviors will be selected over cooperative behaviors. Hence, punitive responses would be selected for to allow the possibility of large-scale cooperation.

In the long term, potential cooperators are valuable enough to the community that efforts to try to reintegrate cheaters back into the community could be more beneficial than the costs of trying to ostracize or kill them. If so, then these pressures would likely select for punitive sentiments that motivate us to punish in a way that appeals not just to cheaters' fears of retaliation, but to their internal values. Specifically, these responses should signal to cheaters that they have undervalued the norms of the community and must update that valuation in order to reap the future benefits of community membership (Petersen et al. 2012; Sell 2005; Sznycer et al. 2016; Tooby et al. 2008; Tooby & Cosmides 2008). In this context, it would be useful to be able to accurately assess, and possibly increase, the capacity and willingness of offenders to internalize the social norms. Humans' strong interest in gauging offenders' intentions, previous behavior, and sincerity of apology seem, therefore, to accord with the punitive psychology suggested by communicative punishment. Ultimately, the forward-looking selective processes that allowed large-scale cooperation in our ancestors may have given us the backwards-looking reactive attitudes that motive punishment. But the goal of these punitive emotions and the actions they motivate is to appraise and communicate the need and possibility for reconciliation.

Furthermore, the communicative theory of punishment suggests that our punitive responses should be sensitive to whether the offender understands the message. Indeed, research supports this hypothesis. For example, Gollwitzer and colleagues (2009, 2014) found that participants were more satisfied with their punitive acts when they received evidence that the offender understood why he was being punished and even more so when he indicated even a minimal commitment to avoid future wrongdoing. These studies challenge some earlier studies (e.g., Carlsmith et al. 2002; Darley et al. 2000) suggesting that people are motivated to punish primarily by the retributive aim of harming the offender and to vent their anger (see Nadelhoffer et al. 2013). The Gollwitzer et al. studies, however, involved economic games, so the stakes were relatively low and the signals were minimal (e.g., "OK—I was greedy… and now I see what's wrong with that… I shouldn't be such a jerk in a situation like this"). Whether our desire to motivate such understanding of wrongdoing in the offender in these low-stakes contexts also applies to more serious wrongdoing has not been sufficiently studied. Thus, one motivation for our study is to test the effects of such responses to punishment, in the form of apologies, in the context of criminal wrongdoing.

4. Testing the Impact of Apology on Punishment Judgments

The communicative theory suggests a range of empirical predictions that would be useful to test. A few of these have been explored (e.g., research on restorative justice programs, reparations, and apologies; e.g., Schinkel 2014; Gold and Weiner 2000; Jehle et al. 2009; Peterson et al. 2012; Petrucci 2002; Rachlinski et al. 2012; Robinson et al. 2012; Scher and Darley 1997; O'Hara and Yarn 2002). Here, we focus on how information about relevant mental states of a criminal offender impacts punitive judgments, specifically information from which people can make inferences about the intentions and planning of the offender in carrying out the crime, about his character based on criminal history, and about the sincerity of his apology in response to his crime.

To determine whether the offender adequately values community norms, it is necessary to appraise his initial motivation and to rule out relevant excuses or justifications for the violation. For example, it would be useful to know if the offense was *intentional* in a way that indicates recognition of and disregard for the risks of harm to others. Those who commit the same act opportunistically, for instance, likely do not undervalue the community's norms as much, and therefore may be less likely to reoffend than someone who has planned their crime. Thus, punishing people whose violations were less intentional would be less efficient.

Similarly, to be sure that a punishment will be understood and will effectively reform the offender, we would need to know whether the offender is capable of internalizing the community’s norms. One way to do this is by evaluating the offender’s *offense history*. A consistent pattern of norm violations indicates greater future risk, in part because it reveals something about that offender’s characteristic (low) valuation of others. If punishment is aimed at communicating a message to the offender, then people should respond to evidence that the offender will be responsive to that message.

Another way to increase confidence that punitive messages will be internalized by offenders is to evaluate their commitment to changing their behavior. According to the communicative theory, one of the most credible ways for an offender to signal relevant changes of mental state and future behavior is with a sincere apology. Apologies can include various elements which provide evidence for different mental states and levels of sincerity. These elements include: (1) acknowledgment that one’s action violated a norm (i.e., was wrong) and of the harm it caused; (2) an expression of remorse for one’s action and its harmful outcomes; (3) a promise to avoid repeating such actions; and (4) an offer to make up for the harm caused by the action (e.g., compensation). Apologies might also include some explanation for why one did the action (though sincere apologies do not include attempts to excuse or justify one’s behavior), emotional expressions such as sorrow or shame, a request for forgiveness, or an expression of hope for an improved relationship (Scher and Darley 1997; Petrucci 2012; O’Hara and Yarn 2002).

These theoretical considerations lend themselves to specific predictions, namely that punishers will be responsive to information about criminal intent, criminal history, and the perceived sincerity of apology. We developed a pilot study to examine these three features of people’s punitive psychology and their influence on judgments about punishment in response to crime. While previous research has examined each of these features, we consider how they act in combination to influence people's punishment judgments.

To examine the influence of intent (1), we varied whether the criminal planned his crime (robbery) or whether he carried it out opportunistically. To examine the influence of past wrongdoing (recidivism) (2), we varied whether or not the criminal had been convicted and served a sentence for a similar crime. And to examine the influence of apology (3), we varied whether the criminal offers few or most elements of apology for his crime, suggesting an insincere versus a sincere apology. We manipulated these factors between subjects (high vs. low) and also within subjects (pre vs. post) and measured the effects on judgments of sentence severity and length. All our manipulations were designed to be relatively subtle in order to limit the influence of potential confounds and isolate the unique effects of the relevant features of each manipulation.

Notice that if people are punishing for expressivist or retributive reasons, it is not clear why they should or would lower punitive responses in response to apology or even offers of compensation. If the point of punishment is to express our condemnation for the crime or to cause the criminal to suffer in proportion to the harm of he caused, then the criminal's response should not be relevant to punishment, unless perhaps it is taken as evidence of diminished intent or as signaling diminished need to express condemnation. Pure retributive theories would also predict that people punish in response to what is deserved for the specific crime in question, and hence should not increase punishment in light of earlier crimes so long as those crimes had already been justly punished.

Furthermore, on a general deterrence theory, apology might be seen as problematic, since lowering punishment in response to apologies or reparations might signal to would-be perpetrators that they have an ‘escape valve’ if they are caught. This might be part of the reason that judges appear unlikely to lower sentences in response to apology (see Rachlinski 2012). On a specific deterrence or a rehabilitative theory, our reactions to apology should be based solely on our taking them to provide reliable information about recidivism. Since the communicative theory is also concerned with offenders’ responding to punishment by reforming and committing not to recidivate, it will be difficult to discern which of these theories best explains people’s responses to apologies unless we can discern whether offenders’ improved behavior is due primarily to fear of punishment or internalization of the community’s norms.

These predictions motivated the pilot study described below. Participants read a realistic case summary of a crime that has resulted in a conviction and were asked to make punishment recommendations, before and after receiving information about the offender’s prior crimes and the type of apology he offers.

5. Study Design

Participants were 601 undergraduates (67.2% female; 30% male) from Georgia State University participating for course credit in philosophy (n = 493) or psychology (n = 108). Although the punitive attitudes of GSU undergraduates are not expected to represent that of juries or judges, this population is valuable for the purpose of assessing folk psychological punitive attitudes and piloting new experimental methods. Furthermore, compared to many U.S. universities, the sample was demographically diverse with respect to age (*M* = 22.7, *SD* = 3.9; 57.7% between age 18-22), race (Black: 35.9%; White: 29.1%; Asian: 17.3%, Mixed/Other/Unknown/Prefer not to answer: 16.6%) and ethnicity (Hispanic/Latino: 13.7%; Mixed/Other/Unknown/Prefer not to answer: 7.9%).

The study used a 2 (intent) x 2 (criminal history) x 2 (apology) mixed factorial design, permitting measurement of differences in recommended punishment both between- and within-subjects. *Intent* was defined as the offender's level of criminal intent being "high" (planned crime) versus "low" (opportunistic crime). *Criminal History* was defined as the presence or absence of a prior criminal record (i.e., "second-time" versus "first-time" offender). *Apology* was defined as the presence of a "sincere" versus "insincere" verbal apology by the offender. All participants were presented with the information in the order *Intent, Criminal History, Apology*. Our primary dependent measure was a sentence length measure describing number of months in prison on an internally-developed, logarithmic scale from 1 year to 20 years and 2 months. The within-subjects comparison was achieved by delivering the sentencing measure three times, once after presentation of each of the three independent variables. This design feature, though it departs from actual legal procedure, enabled us to measure each participant's change in punitive sentiment as he/she learns more about the details of the offender's history and apology, thereby revealing something about the underlying cognitive updating process likely involved in punishment evaluations.

We tested the following hypotheses:

- **H1**: High intent will evoke more severe punishments than low intent (i.e., main effect of intent).

- **H2**: Within-subjects punishment will increase following presentation of prior criminal history (i.e., for second-time offenders). Likewise, second-time offenders will evoke more severe punishments than first-time offenders (i.e., main effect of criminal history).
- **H3**: Within-subjects punishment will decrease in the presence of a sincere apology. Likewise, the sincere apology will evoke lesser punishment than the insincere apology (i.e., main effect of apology).

The study design yielded eight vignette-based surveys, corresponding to each of the eight between-subjects conditions. (See Appendix for vignette stimuli.) The survey was administered online (Qualtrics) via personal computers. For each survey, after providing consent, participants were instructed to imagine that they are serving as a trial court judge for the sentencing phase of a criminal case. The participant's task was to consider the details of the case and produce a sentence recommendation. All vignettes were closely matched for length and reading level. The key dependent measure, sentence length, was introduced using the following instructions: "According to Georgia law, a person convicted of robbery is to be punished by imprisonment for not less than 1 nor more than 20 years. Given the following options for sentencing, how much time in prison should Mr. Jones receive for this offense?" For the second and third punishment measurements, participants were told that before their initial sentence recommendation had gone into effect, new facts about the case became available for consideration—i.e., information first about criminal history and then about apology.

In addition to our key dependent measure, several supplemental questions were examined to identify the factors most likely to influence and explain our primary effects. These included questions designed to capture punitive attitudes pertaining to the vignette, including deontological (e.g., the moral wrongfulness of the crime), consequentialist (e.g., the offender's dangerousness), and communicative (e.g., how much the offender understands the purpose of the punishment) attitudes measured on a 9-point scale. We also queried participants' philosophical attitudes about the purpose of punishment more generally, using a scale (adapted from Nadelhoffer, Heshmati, Kaplan & Nichols 2013), asking participants to rank the importance of six justifications for punishment such as giving offenders "what they deserve" and sending offenders "the message that they should feel genuinely remorseful and should make up for the harm they caused." We also included questions to assess the effectiveness of our manipulations such as the degree of agreement with the statement "Mr. Jones' apology was sincere." Finally, participants were asked to provide demographic information (e.g., gender, age, race and ethnicity, political orientation, religiosity).

6. Does apology reduce punishment?

This study was designed to determine whether apology mitigates punishment independently of criminal intent and criminal history. To answer this question, we examined two types of comparisons. First, we sought to evaluate the within-subjects changes in punishment before vs. after presentation of prior criminal history and before vs. after presentation of the (in)sincere apology. Then, we sought to understand whether the changes in sentences from pre- to post-criminal history and pre- to post-apology differed between experimental groups. We tested our hypotheses using a mixed Analysis of Variance with Intent (planned vs. opportunistic), Criminal History (1st time vs. 2nd time), and Apology (sincere vs. insincere) as between-subjects factors and Sentence Recommendation as a within-subjects factor. The sentence recommendation was collected at three points: (1) following the Intent, (2) Criminal History, and (3) Apology manipulations, respectively. We used planned comparisons to evaluate within-subjects changes in punishment independently for each level of our independent variables (corrected for multiple comparisons).[2]

(H1) Did high intent evoke more severe initial punishments than low intent? Yes. Prior to the introduction of criminal history or apology information, high intent (*M* = 4.23 years) evoked more severe punishments than low intent (*M* = 3.33 years), *p* < .01.

(H2) Did participants increase punishment following presentation of prior criminal history (i.e., for second-time offenders)? Yes. Punishment was greater following presentation of prior criminal history (*M* = 5.39 years) compared to that preceding criminal history information (*M* = 3.36 years), *p* < .001. Moreover, this pre/post change in sentence was significantly greater for the second-time offender (*MD* = 2.03 years) than the first-time offender (*MD* = -0.54 years) *p* < .001.

(H3) Did participants decrease punishment following presentation of a sincere apology?
Yes. Punishment was lower following the sincere apology (*M* = 4.02 years) compared to that preceding apology information (*M* = 4.54 years), at least under conditions of high criminal intent, *p* < .05. In addition, this decrease in punishment was greater for the sincere apology (*MD*

= -0.52 years) than for the insincere apology (*MD* = -0.02 years), but only in the high intent condition, *p* < .05. (See Figure 1.)

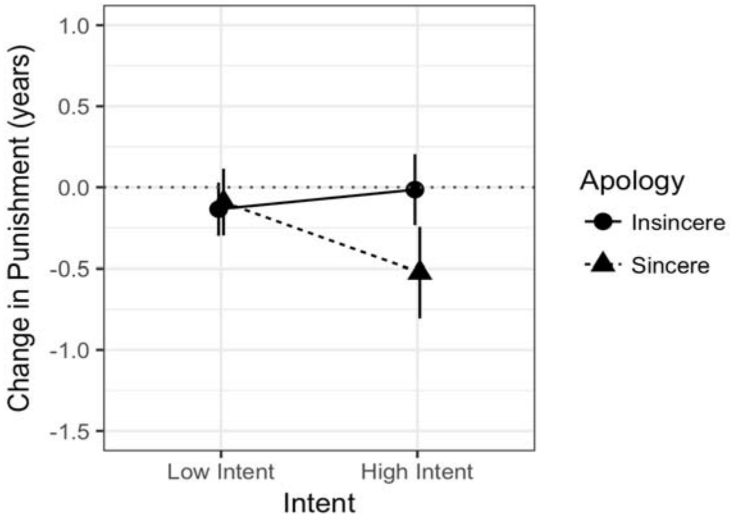

Figure 1. When the offense was committed with high criminal intent, a sincere apology significantly mitigated punishment relative to an insincere apology.

In addition to our primary hypotheses, several supplemental questions were examined to identify factors likely to clarify our primary effects. Most notably, participants assigned to the sincere apology condition were significantly more likely (*M* = 1.60) than those in the insincere condition (*M* = 0.32) to agree with the statement that: “Mr. Jones seems to understand why he needs to be punished” (*p* < .001). The same participants were less likely to agree that Mr. Jones will commit another crime after release (*p* < .05).

Furthermore, we asked participants to rank the importance of six justifications for punishment. On average, participants ranked specific deterrence and communication of a message as the two most important justifications. These two justifications, while statistically indistinguishable from each other (*p* = .44), were ranked proportionally higher than their nearest neighbor (*p's* < .001) according to a series of Wilcoxon Signed Rank tests with correction for multiple comparisons. (See Figure 2.)

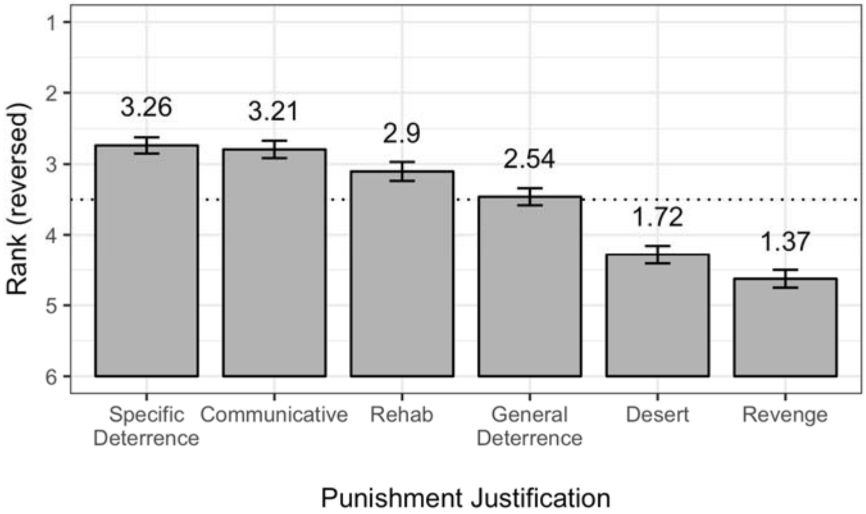

Figure 2. The two highest ranking punishment justifications were to deter offenders from reoffending and to send a message to offenders that they should feel remorseful and make up for the harm caused.

These results are consistent with the predictions of communicative theory that sincere apology should mitigate punishment by sending a credible message that the offender values the societal norms and will make efforts to compensate for the harms he has caused. However, the mitigating effect of apology is not large in comparison to the aggravating effect of criminal history. To the extent that apology communicates a commitment to play by the rules, the small effect should not be surprising alongside evidence of recidivism. That is, the presence of a criminal history could undermine the credibility of an apology in the way that "actions speak louder than words.” Yet, even with evidence of recidivism, we still observe a mitigating effect of sincere apology.

The intent by apology interaction was not predicted. We found that sincere apology was more effective in lowering sentencing judgments when intent was high. One possible explanation for this effect is that for the milder, low-intent offense, participants believed that there was less to apologize for. The absolute values of participants’ sentences for the low-intent crime also tended to be low, so these participants had less room on the scale to reduce their sentence. However, future research would be required to evaluate these explanations.

We limited the apology manipulation to a subset of established elements of apology. Specifically, the sincere apology uniquely included (1) an acknowledgement that the action was wrong and harmful, (2) an expression of remorse for causing the harm, (3) a promise to desist from such actions, and (4) a voluntary offer of costly compensation. However, the two apologies did not differ in the compensation outcome (the money was returned in both scenarios) or explicit request for forgiveness (no such request was made). These two elements may influence the mitigating effect of apology. We also presented minimal information about the emotions expressed in the sincere apology (or the victim’s emotions in response to apology), elements which have been shown to play a significant role in impacting punitive responses (Gold & Weiner 2000).

7. Future Research and Concluding Suggestions

We plan to consider several directions in future research. Extensions of this line of research should test the responses of those who actually make punitive judgments, including lawyers and ideally criminal trial court judges, permitting a direct comparison between experts' and laypersons' punitive intuitions. If the communicative theory of punishment is descriptively on track, it should be detectable even within the constraints that professional judges must oblige, such as federal sentencing guidelines. It would also be useful to test the effects of offender's apologies and offers of reparation made directly to the victim of the crime, and of the victim's accepting the apologies and offers and/or offering forgiveness versus not doing so (e.g., as occurs in some restorative justice procedures). The effects of more detailed descriptions of (or even video presentations of) emotional expressions by offender and victim would also be useful to explore.

The communicative theory of punishment assumes that punishment is designed to reform eligible individuals within the social community. As such, it is less clear about how we are likely to treat outsiders who violate the social norms of our ingroup. On one hand, we might not expect outgroup members to be familiar with our ingroup norms; on the other hand, we have less reason to invest in outgroup members in the first place. Thus, laypersons should be adept at distinguishing between ingroup and outgroup members, and should employ different types or amounts of punishment to achieve their different goals for outgroup members. It would hence be useful to test the effects on punitive judgments of participants' perception of offenders as members of the ingroup or outgroup.

The communicative theory of punishment has been neglected in discussions of penal reform, perhaps because it is not obvious what practical reforms it suggests, though theorists have discussed reforming and expanding the parole system and various restorative justice programs (e.g., Duff 2001; Schinkel 2014). To the extent that the theory suggests increased attention to the mental states of offenders before and during their crimes and after conviction of their crimes, it may face practical concerns about how best to assess these mental states with objective measures. Furthermore, there are procedural hurdles to the use of apologies in the legal system. Since a crucial element of apology is the offender's recognition and communication of his guilt, apology *before* conviction will be strongly discouraged by defense attorneys, except as part of a plea bargain, in which admission of guilt typically leads to sentence reduction—e.g., an automatic 2-3 level reduction in sentence (2012 United States Sentencing Commission §3E1.1). Apologies *after* conviction can seem—and can *be*—insincere if they are perceived as—or aimed at—an attempt to minimize punishment (see Rachlinski et al. 2012).

Relatively recent attempts to incorporate apology into restorative or reparative justice frameworks offer some practical guidance for reforms. However, these frameworks are typically presented as, and understood as, *alternatives* to punishment. One feature of the communicative theory of punishment is that punitive acts aim to encourage, or even force, offenders to apologize, to compensate, and to take on the work of rehabilitation. If we reform our broken incarceration system to more effectively induce offenders to understand the norms they have violated, to increase their valuation of them, and to make efforts to compensate for violating them, such reforms will not be *replacing* punishment. They may instead be moving us towards what we already believe punishment aims to do.

References

Bedau, H. and Kelly, E. (2015) "Punishment," *The Stanford Encyclopedia of Philosophy*, E. Zalta (ed.), http://plato.stanford.edu/archives/fall2015/entries/punishment

Carlsmith, K. M., Darley, J. M., & Robinson, P. H. (2002) "Why do we Punish? Deterrence and Just Deserts as Motives for Punishment," *Journal of Personality and Social Psychology* 83: 284–299.

Cullen, F. T., Jonson, C. L., & Nagin, D. S. (2011) "Prisons do not Reduce Recidivism: The High Cost of Ignoring Science," *The Prison Journal 91*(3 supp): 48S-65S.

Darley, J. M., Carlsmith, K. M., & Robinson, P. H. (2000) "Incapacitation and Just Deserts as Motives for Punishment," *Law and Human Behavior* 24: 659–683.

Duff, A. (2001) *Punishment, Communication, and Community,* New York: Oxford University Press.

Feinberg, J. (1965) "The Expressive Function of Punishment," *The Monist* 49(3): 397-423.

Funk, F., McGeer, V. & Gollwitzer, M. (2012) "Get the Message: Punishment is Satisfying if the Transgressor Responds to its Communicative Intent," *Personality and Social*

Psychology Bulletin 40(8): 986-997.

Gold, G. & Weiner, B. (2000) “Remorse, Confession, Group Identity, and Expectancies about Repeating a Transgression,” *Basic and Applied Social Psychology* 22(4): 291-300.

Gollwitzer, M. & Denzler, M. (2008) “What Makes Revenge Sweet: Seeing the Offender Suffer or Delivering a Message?” *Journal of Experimental Social Psychology* 45: 840–844.

Jehle, A., Miller, M. & Kemmelmeier, M. (2009) “The Influence of Accounts and Remorse on Mock Jurors’ Judgments of Offenders,” *Law and Human Behavior* 33: 393–404.

Lewis, C.S. (1949) “The Humanitarian Theory of Punishment,” *Twentieth Century: An Australian Quarterly Review* 3(3): 5-12.

McGeer, T. (2011) “Co-reactive Attitudes and the Making of Moral Community,” in R. Langdon and C. Mackenzie (eds.) *Emotions, Imagination, and Moral Reasoning*, New York: Taylor & Francis.

Nadelhoffer, T., Heshmati, S., Kaplan, D. & Nichols, N. (2013) “Folk Retributivism and the Communication Confound,” *Economics and Philosophy* 29: 235-261.

O'Hara, E. & Yarn, D. (2002) "On Apology and Consilience," *Washington Law Review* 77: 1121-1191.

Petersen, M. B., Sell, A., Tooby, J., & Cosmides, L. (2012) "To Punish or Repair? Evolutionary Psychology and Lay Intuitions about Modern Criminal Justice," *Evolution and Human Behavior 33*(6): 682-695.

Petrucci, C. (2002) "Apology in the Criminal Justice Setting: Evidence for Including Apology as an Additional Component in the Legal System," *Behavioral Science and the Law* 20: 337–362.

Rachlinski, J., Guthrie, C. & Wistrich, A. (2012). "Contrition in the Courtroom: Do Apologies Affect Adjudication?" *Cornell Law Review* 98: 1189-1243.

Robinson, P., Jackowitz, S. & Bartels, D. (2012) "Extralegal Punishment Factors: A Study of Forgiveness, Hardship, Good Deeds, Apology, Remorse, and Other Such Discretionary Factors in Assessing Criminal Punishment," *Vanderbilt Law Review* 65(3): 737-826.

Scher, S. & Darley, J. (1997) "How Effective are the Things People Say to Apologize? Effects of the Realization of the Apology Speech Act," Faculty Research and Creative Activity. Paper 26.

Schinkel, M. (2014) "Punishment as Moral Communication: The Experiences of Long-term Prisoners," *Punishment & Society* 16(5): 578–597.

Sell, A. (2005) *Regulating Welfare Tradeoff Ratios: Three Tests of an Evolutionary-computational Model of Human Anger*, Doctoral dissertation, University of California Santa Barbara.

Strawson, P. (1962) “Freedom and Resentment,” *Proceedings of the British Academy* 48: 1-25.

Sznycer, D., Tooby, J., Cosmides, L., Porat, R., Shalvi, S., & Halperin, E. (2016) “Shame Closely Tracks the Threat of Devaluation by Others, Even Across Cultures,” *Proceedings of the National Academy of Sciences* 113(10): 2625-30.

Tooby, J., & Cosmides, L. (2008) “The Evolutionary Psychology of the Emotions and their Relationship to Internal Regulatory Variables,” in M. Lewis, J. Haviland-Jones, L. Barrett Feldman (eds.) *Handbook of Emotions*, 3rd ed., New York: Guilford Press.

Tooby, J., Cosmides, L., Sell, A., Lieberman, D., & Sznycer, D. (2008) “Internal Regulatory Variables and the Design of Human Motivation: A Computational and Evolutionary Approach,” in A. Elliot (ed.) *Handbook of Approach and Avoidance Motivation*, New York: Taylor and Francis.

Appendix

Case Summary (differences between conditions are in bold)

Frank Jones has been convicted of one count of robbery. At trial, the following evidence was presented.

On the evening of March 13, 2016, thirty-year-old Mr. Jones entered Wilson's Convenience Store in Alpharetta, Georgia. Mr. Jones is seen on video surveillance walking into the store at 10:37pm. Mr. Jones took an item from the refrigerator and brought it to the counter. Mr. Jones handed a bill to the cashier, who was the only other person in the store.

The cashier opened the register and then walked into an adjacent workroom to get change. Mr. Jones waited approximately 15 seconds, looked around the empty store, then walked around the counter and put all the bills from the cash register into his pockets. The cashier then returned to the area behind the counter, standing in between Mr. Jones and the exit. Mr. Jones pushed the cashier out of his way, and she crashed into a glass display case, breaking the glass and losing consciousness. Mr. Jones then fled through the exit door. The cashier was found unconscious at 11:04pm and taken to the nearest hospital, where she received four stitches. Based on the store's electronic transaction history, the amount stolen was $674.

Low intent (opportunistic crime)	**High intent (premeditated crime)**
Mr. Jones' phone was investigated for any indication that the robbery was planned, **but no evidence of planning was found. Investigators found only one relevant text**	Mr. Jones' phone was investigated for any indication that the robbery was planned, **revealing a text message from earlier in the day in which he bragged to a friend about**

message from a friend of Mr. Jones asking him to pick up some milk from the store.	**his plans to "strong arm" the store when he knew there was only one cashier on duty.**

Additional Facts

Before you announce Mr. Jones' sentence, you are given additional facts about the case. First, you are informed that after he was found guilty of robbery, prosecutors discovered that…

Low Criminal History	**High Criminal History**
Mr. Jones has **never before been** convicted **or arrested for any other crimes in any U.S. state.**	Mr. Jones has **been previously convicted of another robbery in a different U.S. state and served his full prison sentence for that crime.**

At the end of the current trial, Mr. Jones was asked if he would like to make a statement to the victim in the courtroom. He provided the following statement:

Sincere Apology	**Insincere Apology**
I am **truly** sorry for **causing you all this pain. I understand that what I did was**	I am sorry **about what happened to you.** I didn't think it would turn out this way. I can't

wrong. There’s no excuse for my behavior. I can’t undo what I did, but **I promise I will do what it takes so that I will never do that kind of thing again.** Also, **I will work to repay you fully, not just for the money I stole, but also for the pain I caused.**	undo what I did, but **I definitely don’t want to go to prison.** Also, **you’ll get your money back because they got it out of my car when I was arrested.**

[1] Authorship is equal. We owe much gratitude to Julia Watzek who provided valuable assistance in survey design, analysis, and reporting.

[2] We also tested absolute differences in punishment between groups. However, unlike our within-subjects models, the corresponding between-subjects test yielded no significant differences between sincere vs. insincere apology. Given the lower sensitivity of between-subjects comparisons for small effects, we speculate that these tests may have suffered from Type II error. Additional measures were collected which are beyond the scope of the present research question and so are not reported here.