Routledge
Taylor & Francis Group

Journal of Moral Education

ISSN: 0305-7240 (Print) 1465-3877 (Online) Journal homepage: http://www.tandfonline.com/loi/cjme20

Parole and the moral self: moral change mitigates responsibility

Javier Gomez-Lavin & Jesse Prinz

To cite this article: Javier Gomez-Lavin & Jesse Prinz (2019): Parole and the moral self: moral change mitigates responsibility, Journal of Moral Education, DOI: 10.1080/03057240.2018.1553153

To link to this article: https://doi.org/10.1080/03057240.2018.1553153

JOURNAL OF MORAL EDUCATION
https://doi.org/10.1080/03057240.2018.1553153

Routledge
Taylor & Francis Group

ARTICLE

Parole and the moral self: moral change mitigates responsibility

Javier Gomez-Lavin[a] and Jesse Prinz[b]

[a]Department of Philosophy, The University of Pennsylvania, Philadelphia, PA, USA; [b]Philosophy Program, The CUNY Graduate Center, New York, NY, USA

ABSTRACT

Recent studies demonstrate a moral self effect: continuity in moral values is crucial to ascriptions of identity in and over time. Since Locke, personal identity has been referred to as a 'forensic' concept, meaning that it plays a role in attributions of moral responsibility. If moral values are crucial to identity over time, then perceived changes in a person's set of values may reduce responsibility for past deeds. To test this, we examined the moral self effect in parole contexts. In this empirical article, we conducted two experiments, in which participants were significantly more likely to grant parole to agents who underwent a moral change as opposed to mere behavioral change. We conclude by discussing possible objections and implications of these philosophical results for the Lockean view of personal identity.

KEYWORDS

moral responsibility; personal identity; criminal justice; moral self; experimental philosophy

In Western philosophy, John Locke is credited with foisting personal identity—that is, what makes us count as the same person over time—onto the philosophical agenda. Locke's colorful exposé on the topic in the *Essay Concerning Human Understanding* (1700) introduced a method for analyzing the nature of personal identity, a positive or generative theory of identity and an account of why personal identity matters. In a nutshell, Locke argues that identity matters for moral purposes: it plays a central role in attributions of moral responsibility. At the same time, Locke's positive theory of personal identity does not make explicit reference to morality, or how morality may interact with or constitute identity. He applies a non-moral theory of identity to his theory of responsibility. A growing body of empirical work, which updates and naturalizes Locke's method for studying identity, suggests that Locke underestimated the role of morality in our conception of what makes someone count as the same person over time; in fact, moral values may themselves be essential or constitutive of personal identity (Hitlin, 2011). But this emerging theory, termed the moral self view, has not been brought to bear on Locke's contention that personal identity plays a role in responsibility. Here we take a step towards filling that important gap. We present new evidence that intuitions about the moral self also relate to responsibility. We will argue on both empirical and

CONTACT Javier Gomez-Lavin jglavin@sas.upenn.edu Department of Philosophy, the University of Pennsylvania, Philadelphia, PA 19104, USA

theoretical grounds that the moral self view can account for some aspects of responsibility attribution and that, indeed, it may be better poised to do so than the theory of personal identity that Locke favored. In making our case, we focus on a real-world context in which values may change: rehabilitation during periods of incarceration. Below we will present experimental results that bear on the relationship between moral identity and responsibility, along with a theoretical discussion to interpret our findings and bolster them with supporting arguments. Before we get underway, some background is necessary. In this section, we will briefly introduce Locke's method, theory and account of why theories of identity matter. Then we will present the moral self view that we have been developing and our strategy for relating this construct to responsibility.

The Lockean background

Locke's (1700, p. I.XXVII) seminal discussion of personal identity focuses on the question, what makes someone qualify as the same person over time? We all undergo various changes during the lifespan, such as physical and psychological growth, yet many of these changes do not seem to matter for personal identity. We normally regard a person as retaining their identity over time. Locke aims to identify the principles that guide our intuitions about what makes someone count as the same person. With non-human animals, continuity in body seems more important than continuity in mind. If a farmer owns a cow, it counts as the same animal as it progresses from young calf to adult. The farmer does not take notice of any psychological changes. It doesn't matter if the cow's personality remains constant or if it can reminisce about its past. In a Lockean frame, we keep track of animals by their bodies alone. With people, *per* Locke, things are different. We care a great deal about psychological continuity, and our social interactions afford many opportunities to learn about the personalities, memories and character of those we know. That is not to say that all psychological traits matter for identity. When we learn new facts, skills, or languages, for example, we don't become new people. So, Locke sets out to discover what kind of psychological continuity matters. Note that at this point in the discussion, Locke is not particularly interested in forming a thesis of *numerical identity* of any run of the mill objects, rather he wants to uncover the source of synchronic, temporally continuous identity of *persons* (cf. compare his earlier §§1–9 to §§11–12). He wants to know what we care about when we make judgments about identity. He is interested in real-world intuitions and practices. We follow him here.

Locke uses a method of thought experiments. He asks readers to imagine various changes and to consult their intuitions about whether these matter to identity over time. He has us imagine, for example, that someone's mind gets transplanted to another body. Here, readers are expected to have the intuition that they would still be the same person. From this Locke concludes that it is continuity of mind, not body, that matters for identity. But what *aspect* of psychological identity matters? Here Locke is not always easy to interpret. He uses the unhelpful phrase 'same consciousness' (Locke, 1700, p. I. xxvii.10). Though vague, this phrase is normally interpreted as referring to links of memory, which allow us to access the past through recollection. In one key passage, Locke imagines a case of someone who loses access to his memory entirely. He says this

is tantamount to a loss of self. It would be the same man, in some biological sense, but a different person (Locke, 1700, p. I.xxvii.20).

Having advanced this memory-based theory of what makes someone count as the same person over time, Locke turns to the question of why we need theories of identity. He answers that, 'person is a forensic term' (Locke, 1700, p. I.xxvii.26). By this, he means that we care about identity because we need to assign responsibility to people for past events. When we praise or blame, we want to make sure we are assigning responsibility to the right person. Locke seems to believe that his memory account satisfies this criterion. As Locke makes it explicit, '[Person] is a forensic term, appropriating actions and their merit. . .this personality extends itself beyond present existence to what is past, only by consciousness, whereby it becomes concerned and accountable; owns and imputes to itself past actions' (1700, p. I.XXVII.28). That is, it is in some large part *because* we have a robust continuity in our psychological lives, we are the kinds of creatures that can bear responsibility and other deontic relationship to our actions; as we can be blamed and praised for the pains and pleasures that we have brought about. He does, however, recognize that there are possible counter-examples. Consider someone who commits a crime while intoxicated with no memory of the event. Locke's theory seems to entail that we should not punish this person when he is apprehended at a later time, because the sober man cannot access his memories of the offending act. Locke demurs, stating 'human judicatures justly punish him; because the fact is proved against him, but want of consciousness cannot be proved for him' (Locke, 1700, p. I.xxvii.22). In other words, memory lapses are hard to prove in the courtroom so we should presume the sober man has knowledge of his drunken deeds. We are also concerned with a substantively *forensic* account of personal identity As such, if personal identity is a forensic term, if it is a term utilized for social purposes—independent of its possible metaphysics as a social construction—then we ought to be able to detect those downstream social effects. That is, if the folk think that someone's continuous, personal identity has substantively changed, will they continue to hold them responsible for past actions?

More recent philosophical treatments of personal identity continue to identify psychological continuity as an important mechanism for diachronic, personal identity; although unlike Locke's mnemonic account, contemporary theories offer a broader array of states that may contribute to identity. Particularly insightful is the work of Derek Parfit (1984), who modernizes a Lockean account to allow for degrees of continuity both with memories but also with other, even forward-facing, mental states such as intentions (207). As such Parfit explicitly mentions *character* as a candidate vehicle for these 'chains of strong connectedness' that constitute an individual's diachronic personal identity (207). Likewise, Parfit understands that personal identity has a forensic aspect, arguing that 'psychological continuity carries with it desert for past crimes;' however, as Parfit's account can admit of *degrees* of connectedness, he can likewise claim that responsibility and desert for past crimes may be tempered by the degree of connectedness to the past-self who committed said actions (325). Presaging our own work, Parfit argues that a convict's present punishment ought to be modulated by their psychological connectedness, of which character is 'more relevant' than memory, to their past, criminal self: 'When some convict is now less closely connected to himself at the time of his crime, he deserves less punishment. If the connections are

very weak, he may deserve none' (326). Explicitly terming this as a claim *about* 'reduced responsibility,' it is clear how Parfit's own theoretical work provides us with a ready hypothesis: If moral values are used to track psychological connectedness with past-selves, then significant changes of these values ought to be used forensically when making judgements of desert and responsibility.

We build on Locke's approach, but also suggest certain revisions. With respect to his methods, we find thought experiments useful in probing people's intuitions about identity, but recommend two improvements. First, Locke was writing before the advent of scientific psychology. Now, with experimental methods in the human sciences, and the power of statistical analysis, we need not rely on the intuitions of a single author. Any philosopher who reports her intuitions may be biased by culture, professional training and theoretical commitments. Psychological methods allow us to probe the intuitions of ordinary people, and statistics can help determine how strong those intuitions are with respect to any candidate dimension of identity (Knobe & Nichols, 2007). Second, Locke indulges in some very fanciful cases, such as a mind that migrates to another body. We are concerned that such cases may cloud intuitions rather than revealing them. If personal identity is something that matters in the real world, then thinking through changes that can actually take place may be a better way to probe what we care about in tracking identity.

As for Locke's theory, we favor another account, the moral self view, which we turn to now. With that on the table, we can turn to the question of identity and responsibility, which will be our primary focus in this work.

The moral self view

This work builds on a series of studies that we have recently conducted exploring intuitions of identity. The basic finding in that prior work is that moral continuity matters a great deal to people when it comes to identity over time, and, indeed, it matters more than many other aspects of psychological continuity, including memory (Strohminger & Nichols, 2014). Memory matters, as Locke's account would predict, but not nearly as much.

Our prior studies mostly use vignette methods. In these studies, we present participants with scenarios in which a fictional person's moral values have changed or some other trait has changed, and we ask whether the change impacts identity. The dependent variable is the question: Is this the same person before and after the change?[1] Answers are given on a scale (rating either degree or percent of change). We will not review all of our findings here, but a couple are especially relevant. Strohminger and Nichols (2014) asked participants to consider a number of circumstances where traits can change in the lifespan, including brain injury, drug use and age-related degeneration. Participants were asked to rate the degree of change in personal identity after changes of various kinds, including memory and identity. Moral changes were perceived as significantly more impactful for identity than any other trait that they measure. In each of five experiments, a change in morals was perceived as dramatically altering identity—always exceeding the midpoint of the scale. Memory loss was perceived as less impactful—always below the midpoint of the scale. Prinz and Nichols (2016) conducted similar experiments, in which people imagined a change in morals or loss of memory brought on by a head injury. Moral change was perceived as a loss of identity regardless of whether values went from bad to good or good to bad, and the effect held up

even in cases where the vignette describes someone who voluntarily chooses to change values for rational reasons. Memory loss was perceived to have only a modest impact on identity that fell well below the midline of the scales. Gomez-Lavin and Prinz (unpublished data) asked participants whether changes to identity brought on by moral change or memory loss were perceived as 'literal' or merely metaphorical changes. Changes associated with transformations in values were judged to be more real than metaphorical, and the reverse pattern was found for changes associated with memory loss.

This work suggests that ordinary intuitions forge a strong link between moral continuity and identity. Memory continuity (and other factors such as personality, cognitive abilities and agency) is regarded as less important. When considered with the previous studies cited, the literature on the moral self suggests that people consider continuity in moral values as the strongest psychological contributor to the preservation of a person's *qualitative* identity. This is the moral self view.

Moral self and moral responsibility: introducing the present research

Here we want to test how moral continuity and change bear on attributions of responsibility. With Locke, we agree that personal identity plays an important role in moral bookkeeping. When considering deeds in the past, we want to attribute responsibility to the right person in the present. If a perpetrator of misconduct undergoes a transformation that impacts identity, blame should be mitigated. We wanted to know whether moral change would be mitigating in this way. A positive answer to this question would help show that the moral self view can do some of the 'forensic' work that motivates a theory of personal identity according to Locke.

Following our methodological predilection for real-world cases, we chose to consider a kind of moral change that can take place in the real world: moral transformation through prison rehabilitation. We wanted to know: would someone whose values change in prison be held as responsible for past crimes as someone whose values had not changed?

In designing these experiments, we faced two immediate hurdles. First, we were concerned about retributivist impulses. When thinking about crime, people often form a strong desire to punish. This desire may cloud judgments about who deserves punishment; for example, we blame victims when no perpetrator can be identified (Hafer & Bègue, 2005). In pilot testing, we found that retributivist tendencies were very strong for violent crimes, and persisted in cases featuring non-violent offences, such as cheating on one's taxes. This retributivist tendency anchored participants' judgements about punishment despite their also judging that the individual had undergone a substantive change in identity. To avoid this, we decided to use implicit measures of responsibility, which, we thought, would be less vulnerable to retributivism: instead of asking directly about responsibility, we asked whether continued punishment is deserved and about candidacy for parole. We reasoned that early parole is granted to those who are now less blameworthy for their crimes. This introduced a second hurdle; namely, that early parole might also track beliefs about the deterrent impact of a prison sentence. That is, participants may judge that early parole is warranted merely because an offender served part of their prison sentence. So, we decided to compare two cases: prisoners whose values change and those who choose not to reoffend because they want to avoid future prison terms.

In selecting these comparison conditions, we elected not to pit moral change against memory change as we had in previous studies. That is, we do not consider a case where a prisoner forgets the crime he or she committed. In pilot work, we found that forgetting is not exculpating. This is consistent with Locke's injunction to punish those who commit crimes while drunk, but it is inconsistent with his memory-based theory of identity. Here, we assume, based on prior work, that moral continuity matters more than memory and we explore whether it plays some role in attributions of responsibility. We make two key predictions. First, with moral transformation, participants will judge that the character has gone through a qualitative transformation of identity (this would replicate the 'moral self effect' reported in the earlier studies). Second, characters who change *values* will be deemed worthier of parole than those who simply want to avoid being punished in the future (identity change reduces punishment, indicating reduced responsibility).

Experiment 1: present-tense parole decisions

The goals of our first experiment are two-fold: First, as a replication of the moral self effect in a real-world context against a non-mnemonic contrast class (i.e., a change of behavioral outlook as opposed to a loss of memory), and second, to examine the effects, if any, of moral change on attributions of desert and responsibility. Our hypotheses, as discussed earlier, are that we will successfully replicate the moral self effect, and that moral change as opposed to mere behavioral change will increase participants' willingness to parole a fictional prisoner.

Methods

Participants

We recruited 120 American adults (40% female) from Amazon's Mechanical Turk platform to participate in this experiment, with 55 participants randomly placed in our test condition, with the others placed in our control condition.[2]

Procedure and research design

Participants were directed to an online survey where we acquired their informed consent to participate and then they were randomly placed in one of our two conditions. Participants were then given brief instructions before reading either our control or treatment vignette, after which they were directed to answer a series of questions, including our dependent measures, manipulation checks, demographics and a comprehension check. The control vignette describes a fictional case of a male prisoner, 'James Wilson,' who after a series of petty crimes is convicted of vehicular manslaughter and sentenced to 10 years in prison. During his time in prison he begins to reflect on his behavior, and after serving eight years of his sentence his outlook has changed considerably. The vignette makes it clear that while criminal behavior does not appeal to him now that he knows the tough consequences, he still fully identifies with his past values—a point we emphasize throughout the story.

Our treatment vignette begins identically, with a fictional Mr. Wilson having reflected after eight years of his sentence for vehicular manslaughter. However, we make it clear that the character can no longer identify with his past behavior *and* values, emphasizing that his moral outlook and moral values have really changed. Importantly,

we do not describe the *direction* or valence of this change. That is, we do not describe the fictional Mr. Wilson as *improving* or *worsening*, but merely clarify that he has undergone a profound change. Vignettes are additionally reproduced in the Appendix to this article.

After completing the survey, participants were thanked for their participation and given a code to enter into Amazon's Mechanical Turk platform to access their payment ($0.25).

Measures

Once participants read and understood the vignette they proceeded to answer 12 questions, including four dependent measures, four manipulation checks, three demographic prompts and one comprehension check. All of the manipulation checks and three of the dependent measures were randomly presented to participants, with one additional dependent measure presented on a following page. Our dependent measures and manipulation checks are detailed below, alongside the anchors for their respective scales. Demographics asked for participants' self-identified gender, and religious and political outlooks. We had no prior predictions about individual-differences, as these have not occurred in prior research on the moral self effect. Furthermore, over 95% of participants correctly answered our comprehension check.[3]

(1) *Moral self effect replication*: 'Is Wilson the same person now as he was when he entered prison?' 7-point scale anchored at 1 – Very different and 7 – Exactly the same.
(2) *Desert*: 'Given his current state of mind, does Wilson deserve to remain in prison?' 7-point scale anchored at 1 – Definitely deserves more time, and 7 – Definitely deserves to be released.
(3) *Parole*: 'Should Wilson be granted parole?' 7-point scale anchored at 1 – Definitely grant parole and 7 – Definitely deny parole.
(4) *Manipulation check one; recidivism*: 'How likely is it that Wilson will feel an urge to commit a crime?' 7-point scale anchored at 1 – Very likely, and 7 – Very unlikely.
(5) *Manipulation check two; evasion*: 'How likely is it that he would get away with a crime if he were being monitored?' 7-point scale anchored at 1 – Very likely and 7 – Very unlikely.
(6) *Manipulation check three; character at time of arrest*: 'How would you describe Wilson's character at the time of his arrest?' 7-point scale anchored at 1 – Very bad person and 7 – Very good person.
(7) *Manipulation check four; character now*: 'How would you describe Wilson's character now?' 7-point scale anchored at 1 – Very bad person and 7 – Very good person.
(8) *Past-tense agreement*: Participants were first given a short prompt depending on their condition, 'After reflecting on Wilson's change of behavior [control condition]/change of values [treatment condition], the parole board decided to release him early.' They were then asked 'What do you think of the parole board's decision?' on a 7-point scale anchored at 1 – completely defensible and 7 – not at all defensible.

Results

Results were consistent with our predictions (see Figure 1 below). Participants judged the prisoner to have changed significantly more in the treatment (M = 2.05, SD = 1.24) as opposed to the control condition (M = 3.05, SD = 1.32, t(116.642) = 4.25, p < .001, d = 0.776). Prisoners in the treatment condition were seen as more deserving of release (M = 5.73, SD = 1.46) than those in the control condition (M = 4.97, SD = 1.57, t (116.9) = 2.74, p = .007, d = .939), and participants were more willing to grant these prisoners parole (M = 2.1, SD = 1.47) than in the control case (M = 3.03, SD = 1.61, t (117.25) = 3.27, p = .001, d = .597).[4] Participants also agreed more strongly with our past-tense measure in the treatment (M = 1.76, SD = 1.12) as opposed to control condition (M = 2.43, SD = 1.54, t(118) = 2.67, p = .009, d = .495).

Our manipulation checks revealed that prisoners in the treatment condition were judged to be less recidivistic (M = 5.8, SD = 0.99) than in the control condition (M = 4.67, SD = 1.41, t(118) = 5.04, p < .001, d = .936), less likely to evade future incarceration (M = 5.8, SD = 1.13) compared to control (M = 5.18, SD = 1.41, t (118) = 2.6, p = .01, d = .481), and, tellingly, prisoners who underwent a change of values were perceived as having a *better* character now (M = 5.47, SD = 1.1), than those who underwent merely a behavioral change (M = 4.42, SD = 0.86, t (101.5) = 5.77, p < .001, d = 1.06). Importantly, our only measure not to yield a significant difference between conditions related to participants' perception of the prisoners' character *at the time of arrest*, with both means falling towards the bottom-end of the scale (2.25 for our treatment condition and 2.52 for our control condition).

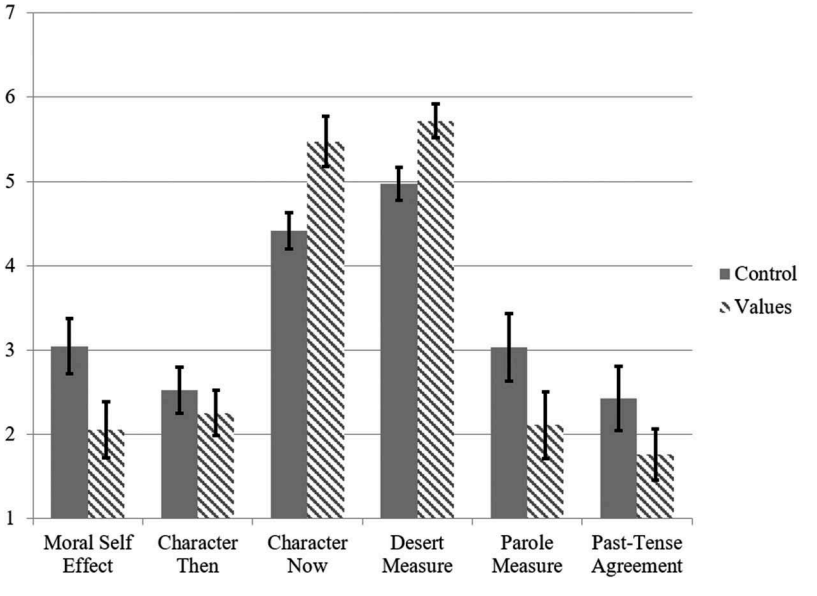

Figure 1. This graph depicts the means and standard errors for several of our measures in experiment one. *** p < .001, ** p < .01.

Mediation analysis and structural equation modeling

The foregoing analyses largely confirmed our hypotheses but left room for interpretation on one crucial question. Value change reduced participants' implicit-responsibility attributions as compared to our control condition, but it also resulted in lower estimates of recidivism; thus, we wanted to check whether it was the perceived identity change that drove down responsibility or merely the diminished likelihood to reoffend. To test that identity change impacts responsibility above and beyond reduced recidivism, we conducted a single mediation analysis as described by Hayes (2013). Participants' judgments of identity change (measure one) were used as the sole mediator, with our conditions serving as the independent variable and participants' willingness to parole the prisoners (measure 3) serving as our dependent measure. We performed our mediation via 5,000 bootstrapped samples in PROCESS (Hayes, 2013). We found a significant indirect effect of our conditions on parole judgments mediated by participants' perceptions of an identity change, $ab = 0.61$, (95% CI: 0.26, 1.0) (see Figure 2). Our mediator could account for approximately 66.4% of the total effect ($P_M = .664$). We performed a Sobel test that confirmed this partial mediation ($z = 3.49$, $p < .0005$). Finally, we verified that our parole measure did not itself serve as a stronger mediator of identity change ($P_M = .39$).

Furthermore, we prepared a simple structural equation model that characterizes the relationships between several of our measures: recidivism, moral self effect and desert, treated as exogenous variables on participants' parole scores. Here we see that both desert ($\beta = -0.33$, $p < .001$) and moral self effect ($\beta = 0.3$, $p < .001$) measures served as significant predictors of participants' parole scores, providing a model fit of $R^2 = .42$. Interestingly, recidivism scores were significantly correlated with both moral self effect scores and desert scores, but did not themselves serve as a predictor of parole scores (see Figure 3).

These findings led us to explore whether any effect of recidivism rate on parole scores might be mediated through our other measures (see Figure 2). Following a similar process to that described above we performed our mediation via 5,000 bootstrapped samples in PROCESS (Hayes, 2013). We found a significant indirect effect of recidivism on parole scores mediated by participants' perceptions of an identity change, $ab = 0.353$, (95% CI: 0.14, 0.56,) along with a significant indirect effect of said effect mediated by our desert measure, $ab^1 = -0.34$, (95% CI: −0.54, −0.14). These mediators could account for approximately 74% of the total effect ($P_M = .739$). We performed two Sobel tests that confirmed these partial mediations ($z = -2.9$, $p < .01$ for our MSE measure and $z = —3.1$, $p < .01$ for our desert measure).

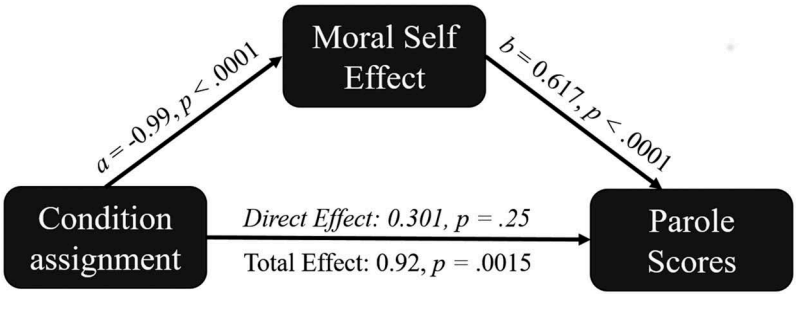

Figure 2. This diagram shows a significant partial mediation of our condition assignment on participants' parole scores by their ratings in our moral self effect measure. Coefficient values are unstandardized.

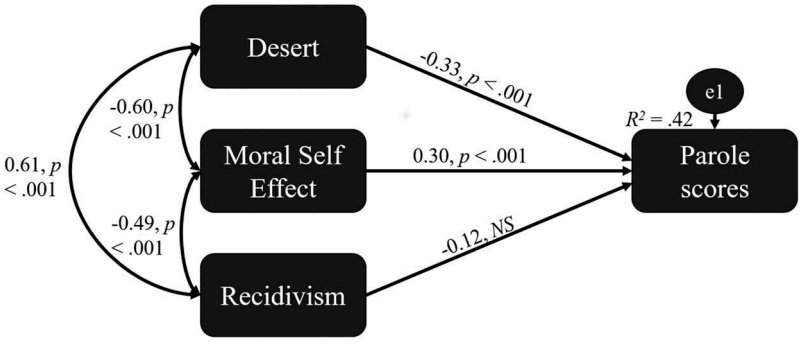

Figure 3. This diagram represents a simple structural equation model where three of our measures are treated as exogenous variables, two of which serve a significant predictors of participants' parole scores. Curved lines represent covariations, not regressions. Coefficient values are standardized betas.

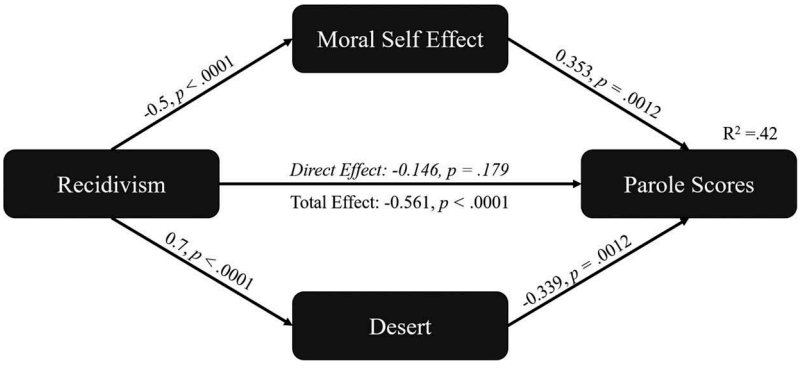

Figure 4. This diagram shows a significant dual partial mediation of participants' recidivism rates on their parole scores via our moral self effect and desert measures. Coefficient values are unstandardized.

Discussion

In experiment one we confirmed our two key predictions. We replicated the moral self effect in a prison scenario, designed to illustrate a real-world example of moral change, and we found that changes in moral self lead to reduced desire to punish and increased support for parole—implicit measures of responsibility. Our mediation analyses confirmed that perceived identity change was a factor influencing reduced responsibility. Identity change reduces responsibility above and beyond beliefs about recidivism, and the impact of recidivism on responsibility is mediated by change in identity.

Experiment two: past-tense parole decisions

Our second experiment seeks to both replicate and extend our previous findings. In particular, by measuring participants' agreement with past parole-board decisions, we can test whether participants will also disapprove of negative parole outcomes (i.e., denial of parole) in cases of moral change. This is likely a more stringent measure than our first

experiment allowed, as retributivist tendencies in the case of violent crime and implicit trust in authorities may make punishment seem justified. Thus, we have two central hypotheses, first that we will replicate previous results and that participants will disapprove of negative parole verdicts in our treatment condition as compared to control conditions.

Methods

Participants

We recruited 181 American adults (41.4% female) from Amazon's Mechanical Turk platform to participate in this experiment. This experiment incorporates a 2 × 2 between—subjects design, where we manipulate both the nature of the prisoner's change (e.g., a change in values—a 'Moral Change,' or a change in attitude—a 'Behavioral Change,' as with experiment 1) and the fictional parole-board's decision (accept or deny), resulting in four conditions. Participants were randomly distributed to our groups as noted in Table 1.

Procedure and research design

As with experiment one, participants were directed to an online survey where we acquired their informed consent to participate and then they were randomly placed in one of our four conditions. As before, participants were then given brief instructions before reading their condition-specific vignette, after which they were directed to answer a similar series of questions to experiment one.

The majority of our vignettes were identical to those used in experiment one. They describe the fictional case of 'James Wilson,' who is charged with vehicular manslaughter and sentenced to 10 years in prison, during which time he begins to change either his moral outlook or behavioral disposition. After reading this same brief description of Mr. Wilson as in Study 1 (see 1.1.2 for more details), participants were directed to read an additional paragraph which emphasized both that the parole board can sense the change that Mr. Wilson underwent, and that they know that prisoners are sometimes paroled after eight years. Here is where the second manipulation comes into play: the parole board either decides to grant or deny Wilson's parole request.

After completing the survey, participants were given a code to access their payment and thanked for their participation.

Measures

As with our previous experiment, once participants read and understood the vignette they proceeded to answer a series of questions including our dependent measure, a moral

Table 1. Random assignment of participants across conditions.

Independent Variables	Parole board accept	Parole board deny
Moral Change	$N = 43$	$N = 45$
Behavioral Change	$N = 46$	$N = 47$

Notes: Numbers of participants in each of our four conditions (e.g., moral change by parole board denial) are given

self effect replication measure, two manipulation checks and a short series of demographic questions. In this experiment, all of our measures were randomly presented to participants. Furthermore, given the similarities between the two experiments, we reduced the number of measures to narrow in on those that might resolve our hypotheses. They are listed below:

(1) *Moral self effect replication*: 'Is Wilson the same person now as he was when he entered prison?' 7-point scale anchored at 1 – Very different and 7 – Exactly the same.
(2) *Parole agreement*: 'What do you think of the parole board's decision?' 7-point scale anchored at 1 – Completely defensible, and 7 – Not at all defensible.
(3) *Manipulation check three; character at time of arrest*: 'How would you describe Wilson's character at the time of his arrest?' 7-point scale anchored at 1 – Very bad person and 7 – Very good person.
(4) *Manipulation check four; character now*: 'How would you describe Wilson's character now?' 7-point scale anchored at 1 – Very bad person and 7 – Very good person.

Results

Results were consistent with our predictions. We replicated both our moral self effect and the major findings of our first experiment. Participants judged the prisoner to have changed significantly more in the two moral-change conditions than in either behavioral condition (see Summary Table 2 and Figure 5). Participants also found the parole board's decision to *accept* the prisoner's parole petition to be significantly more defensible in the moral change condition than in the behavioral change condition (see Summary Table 5 and Figure 5). Similar to our results in experiment one, the prisoner's character at their time of arrest was thought to be uniformly bad, yielding no significant differences between the conditions (Summary Table 3 and Figure 5). Likewise, in the two moral change conditions, the prisoner's character now was deemed to be significantly better than in either behavioral condition (Summary Table 4 and Figure 5).

Furthermore, there was a statistically significant interaction effect between parole board decisions and the type of change on the combined dependent measures, $F(4,174) = 3.973$, $p = .004$, Wilks' $\Lambda = .916$. When analyzing the results of this MANOVA, it was clear that the interaction was driven by our measures' effects on participants' judgments of the acceptability of the parole board's decision. A two-way ANOVA confirmed this interaction (Figure 6): $F(1,177) = 13.81$, $p < .001$. Simple main effects analysis of the type of change confirmed that—across parole decisions—whether

Table 2. Summary t-test table for moral self effect measure.

Condition	*N*	Means (SD)	C1	C2	C3	C4
1) Moral change x accept parole	43	2.09 (0.95)	-	0.65 (NS)	4.88 ***	5.07 ***
2) Moral change x deny parole	45	2.24 (1.23)	-	-	3.9 ***	4.05 ***
3) Behavioral change x accept parole	46	3.28 (1.31	-	-	-	0.057 (NS)
4) Behavioral change x deny parole	47	3.29 (1.27)	-	-	-	-

Notes for tables. Numbers posted in the matrix reflect *t* values. * $p < .05$, ** $p < .01$, *** $p < .001$.

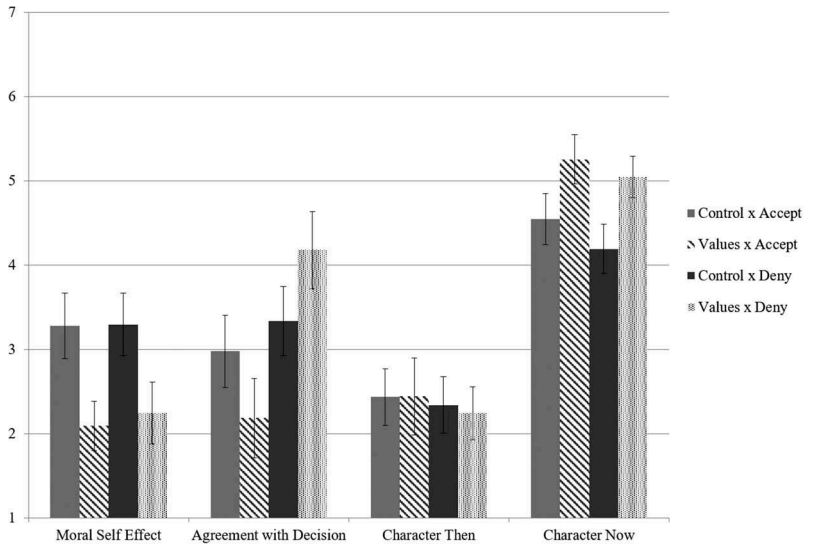

Figure 5. This graph depicts the means and standard errors for several of our measures in experiment two. Consult summary tables (Appendix) for t-values and significances between conditions.

Table 3. Summary t-test table for character at time of arrest measure.

Condition	*N*	Means (SD)	C1	C2	C3	C4
1) Moral change x accept parole	43	2.44 (1.49)	-	0.72 (NS)	0.03 (NS)	0.36 (NS)
2) Moral change x deny parole	45	2.25 (1.05)	-	-	0.83 (NS)	0.42 (NS)
3) Behavioral change x accept parole	46	2.44 (1.13)	-	-	-	0.4 (NS)
4) Behavioral change x deny parole	47	2.34 (1.15)	-	-	-	-

Notes for tables. Numbers posted in the matrix reflect *t* values. * $p < .05$, ** $p < .01$, *** $p < .001$.

Table 4. Summary t-test table for character after change measure.

Condition	*N*	Means (SD)	C1	C2	C3	C4
1) Moral change x accept parole	43	5.26 (0.95)	-	1.11 (NS)	3.39 ***	5.18 ***
2) Moral change x deny parole	45	5.04 (0.83)	-	-	2.56 *	4.49 ***
3) Behavioral change x accept parole	46	4.54 (1.03)	-	-	-	1.68 (NS)
4) Behavioral change x deny parole	47	4.19 (0.99)	-	-	-	-

Notes for tables. Numbers posted in the matrix reflect *t* values. * $p < .05$, ** $p < .01$, *** $p < .001$.

Table 5. Summary t-test table for Parole acceptability.

Condition	*N*	Means (SD)	C1	C2	C3	C4
1) Moral change x accept parole	43	2.19 (1.53)	-	6.1 ***	2.5 **	3.7 ***
2) Moral change x deny parole	45	4.18 (1.53)	-	-	3.86 ***	2.7 **
3) Behavioral change x accept parole	46	2.98 (1.44)	-	-	-	1.22 (NS)
4) Behavioral change x deny parole	47	3.34 (1.4)	-	-	-	-

Notes for tables. Numbers posted in the matrix reflect *t* values. * $p < .05$, ** $p < .01$, *** $p < .001$.

the prisoner underwent a moral or behavioral change mattered for participants' perceptions of the prisoners' current character ($F(1,179) = 39.71$, $p < .001$), and whether they were seen as having their identity changed ($F(1,179) = 30.07$, $p < .001$), but this cross-

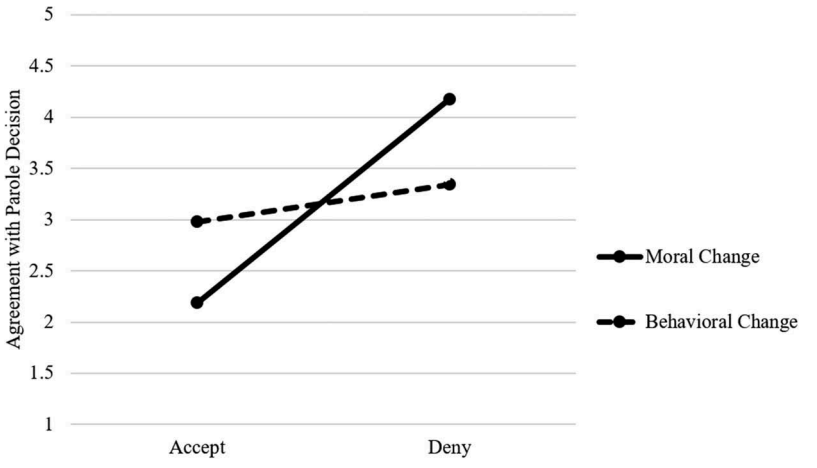

Figure 6. This graph depicts an interaction effect between our two treatments on participants' parole score means in experiment two. Note that the anchors for the agreement measure are 1 = the parole decision is completely defensible and 7 = the parole decision is not at all defensible.

conditions factor did not significantly affect judgments of the acceptability of the parole decision. Conversely, a single factor analysis of the parole board's decision demonstrated that this factor only affected participants' judgments of the acceptability of the parole decision ($F(1,179) = 26.03$, $p < .001$), with participants being less likely to accept a parole board's decision to deny parole regardless of the kind of change the prisoner experienced. At the same time, the interaction makes it quite clear that this general trend was greatly augmented for participants in the moral change conditions. Lastly, gender showed a small main effect for participants' ratings of the prisoner's character after their change ($F(1,178) = 8.68$, $p < .01$, $\eta^2 = .046$), with self-identified women describing the prisoner as having a better character after their transformation.

Discussion

Study two was designed to replicate and extend Study one. We replicated by showing, once again, that participants are more likely to endorse a positive parole decision in cases of moral change as compared to a control decision. As a more stringent measure, we wanted to know whether participants would go so far as to reject the authority of the parole board in cases where they delivered a negative parole decision. We found some tendency to overrule negative parole verdicts in both conditions, but it was much more pronounced in cases of moral change. This suggests that participants want to exonerate those who undergo moral change even if that goes against an official ruling. This indicates a strong link between moral identity and implicit measures of responsibility.

General discussion

In these experiments we replicated the moral self effect and showed, for the first time in this literature, a relationship between perceptions of moral identity and attributions of responsibility. We chose a kind of transformation that can take place in the real world:

retribution during incarceration. This keeps intuitions grounded in reality. Fearing that retributivist impulses would drive up explicit judgments of responsibility, we adopted implicit measures: are those who undergo moral change (as opposed to deterrence-based changes in desire to reoffend) judged to be deserving of early release? Do participants endorse parole for such individuals (experiment one) and reject negative parole decisions (experiment two)? We found affirmative answers to these questions, suggesting that perceptions of moral change are robustly exculpating. This may be taken as implicit evidence for a reduction in blame.

We want to conclude by articulating another way in which the moral self view outperforms the view attributed to Locke. Locke, recall, makes two central claims. First, he implies that memory is the basis of personal identity over time. Second, he says that the main purpose of tracking identity is to assign responsibility for past events. On reflection, there is a tension between these two Lockean theses. Notice that Locke has no built-in theory of why personal identity relates to responsibility. If personal identity were merely a matter of memory continuity, it would be unclear that identity is crucial for blame. Is blameworthiness really a function of recall? Is someone more blameworthy if they recall more details of a crime?

We think the moral self view has greater promise in explaining the link between personal identity and responsibility. Consider prison reform cases, like those we examine here. Moral rehabilitation is perceived as a change in character. Character, in turn, plays a crucial role in attributions of responsibility. This has been demonstrated by a growing body of research. When we judge someone to be morally responsible for a negative action, we attribute to them a bad character (Nadler & McDonnell, 2012; Pizarro & Tannenbaum, 2012) and when we attribute bad character, we judge people to be more responsible (Alicke, 1992; Nadler, 2012; Uhlmann, Zhu, & Diermeier, 2014). Together, such findings indicate that attributions of responsibility are, in part, character assessments. When we blame someone, we are not merely saying that they caused a bad outcome (that can happen innocently); we are saying they are the kind of people who are likely to do bad things—they are bad people. This makes sense conceptually. If people behaved randomly or inconsistently, there would be little reason to hold anyone responsible for anything. There would be no sense in which a one-off bad deed indicated anything about the doer of the deed, so negative attitudes towards that person and punitive interventions would serve little purpose. If, on the other hand, bad deeds tend to reveal something about character, then it makes sense to hold people accountable. Blame would be justified both as a deterrent (bad people are likely to be repeat offenders) and as retribution (we can rightly say that the bad deed issued from something about the person, rather than occurring randomly). Thus, moral identity is yoked to responsibility through the attribution of character. For criminals who change their values, change their ways, there is less reason, thereafter, to continue to hold them accountable. Our experimental findings suggest a direct relationship between moral identity, character and blame. This empirical result, when coupled with previous research on the moral self effect (Prinz & Nichols, 2016, Strohminger & Nichols, 2014) begins to fill the conceptual gaps left by the Lockean theory. It provides an opening for identity and its many roles in moral bookkeeping.

Naturally, our study and our conclusions are limited by our methods: These are not *real* parole cases and we are not asking people to serve on *real* parole boards.

Additionally, because of our previous findings that participants harbor strong retributivist tendencies, it is difficult to use these vignette experiments to directly query participants about responsibility judgements; hence our use of indirect measures. A promising direction of future work would be to analyze the large amount of data that already exists on parole board decisions; namely, to sift through the archival materials on previous parole board hearings and determinations, looking for signs implicating *moral* change or evaluation in the decision process. Of course, such a task is not simple, as accessing many of these records requires state-specific requests and may be blocked due to privacy concerns. Still, our initial findings indicate that moral change figures into people's judgments of qualitative, personal identity and that if qualitative identity is used for moral bookkeeping, then we ought to find proof in the archives.

Additionally, an alternative view may be that participants are judging that the character who undergoes the moral change deserves *less punishment* for their past crimes, but that they are still, in some robust sense, responsible for their past actions.[5] This concern is exacerbated by the fact that our measures did not explicitly disambiguate desert from responsibility, in large part due to our worry about the effects of retributivism. Our understanding is that judgments about early parole, and agreement with early parole-release, point to reduced responsibility for a criminal's prior action, especially as ratings for their character at the time of the crime are indistinguishable between our two conditions, while the prisoner's present character in cases of a moral change is perceived to be significantly better. However, future studies will have to devise measures to explicitly control for a possible dissociation between desert and responsibility.

Finally, some may be concerned about the consequences that our account of personal identity might have for educators and parents—those tasked with, ideally, raising morally conscientious agents. After all, to teach moral responsibility it may be best to hold a child responsible, despite their more plastic and flexible character.[6] Importantly, moral values are only a *part* of an agent's set of psychologically continuous states, although our present and past research suggests that they may be the most important set of states at play, as such, a child's personal identity may be tracked over time, despite their flexible character —and caregivers are especially motivated to see their children as the same person even across dramatic changes. But to the extent that they are engaged in character building, moral educators also implicitly recognize that children are changing persons. Do children become new people over development? In some sense, yes. They may also become persons in a new sense of the term as they acquire moral values. Similar appeals to additional machinery in the development and fostering of identity can be seen in other theories, such as Frankfurt's approach which involves the more cognitively demanding deployment of metacognitive capacities. When one says, 'My child has developed a good character,' 'my child' may refer in a way that focuses on organismic and social relations, which remain constant. But if we step back from that and ask whether a pre-moral two-year old is the same person as a post-moral teen, the issue is far from obvious. In some sense, yes, they are the same person, but in some very important sense, no, they aren't the same agent. Of course, more work, especially explicitly developmental work, must be done to ascertain whether caregivers are sensitive both to a child's set of moral or proto-moral values, and to what extent they truly *hold* the child responsible for a given

action—as opposed to engaging in a kind of responsiblization, or mock-responsibility, process (see for instance, Pettit & List, 2011, 157).

Building on prior work on moral identity, we sought here to explore the relationship between perceptions of moral continuity and attributions of responsibility. Examining intuitions about parole decisions, we found a relationship between moral change and punishment motivation: when criminals undergo moral change, people are more likely to judge that they deserve release. Positive parole decisions are endorsed, and negative decisions are rejected. These findings may be of some use in applied contexts for those who study parole board decisions and their assessment in the community. We would hope that theoretical accounts of personal identity can earn their keep by establishing real-world relevance. For now, we are content to report that our findings suggest that moral change can impact both perceptions of identity and implicit attributions of responsibility. Locke inaugurated research on personal identity, and he suggested that we need a theory that relates identity to practices of praise and blame. We think the moral self view sheds light on this relationship. If an adequate theory of identity should help account for attributions of responsibility, then moral identity is a good candidate. Moral identity is a constitutive element of character, and character, it turns out, is one of the things we are likely to be assessing when we hold people responsible.

Notes

1. Note, again, we are not examining participants' intuitions about *numerical identity*. Rather than coaching our participants on a series of philosophical theories of identity, we are letting them fill in—for themselves—an account of identity. What is relevant for our research is that our manipulation, that is, the relevant narrative changes to the *vignette*, are driving changes in participants' ratings of identity and responsibility. Further work by Berniūas and Dranseika (2016), who replicated Moral Self experiments in Lithuanian—a language that has explicit words for 'sameness' that correspond to qualitative and numerical identity—suggest that what participants are tracking in moral-self experiments are changes in *qualitative* identity. Thanks to an anonymous reviewer for highlighting this methodical concern.
2. As mentioned in the *Measures*, we did not request or collect other demographic information besides self-reported gender, alongside political and religious outlooks.
3. For the results reported below, the entire population is included as removing those individuals who failed the comprehension check did not significantly alter the results.
4. Notice that anchors were reversed between our desert and parole measures, in part to mitigate response perseveration.
5. Thanks to an anonymous reviewer for raising this suggestion.
6. Thanks to an anonymous reviewer for raising this concern.

Acknowledgements

We would like to acknowledge the contributions of our Self, Motivation and Virtue team members, Shaun Nichols and Nina Strohminger, who assisted with this work.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This work was supported by the Self, Motivation, and Virtue Project, funded by the Templeton Religion Trust;John Templeton Foundation [The Self, Motivation, and Virtue Project – "Invest];

Notes on contributors

Javier Gomez-Lavin is a Provost Postdoctoral Fellow at the University of Pennsylvania.

Jesse Prinz is a Distinguished Professor in the Philosophy Program at The CUNY Graduate Center.

ORCID

Javier Gomez-Lavin http://orcid.org/0000-0002-0476-8290

References

Alicke, M. D. (1992). Culpable causation. *Journal of Personality and Social Psychology*, *63*, 368–378.
Berniūas, R., & Dranseika, V. (2016). Folk concepts of person *and* identity: A response to Nichols and Bruno. *Philosophical Psychology*, *29*(1), 96–122.
Hafer, C. L., & Bègue, L. (2005). Experimental research on just-world theory: Problems, developments, and future Challenges. *Psychological Bulletin*, *131*(1), 128–167.
Hayes, A. F. (2013). *Introduction to mediation, moderation, and conditional process analysis: A regression-based approach*. New York: Guilford Press.
Hitlin, S. (2011). Values, personal identity, and the moral self. In S. J. Schwartz, K. Luyckx, & V. L. Vignoles (Eds.), *Handbook of identity theory and research* (pp. 515–530). New York: Springer.
Knobe, J., & Nichols, S. (2007). An experimental philosophy manifesto. In J. Knobe & S. Nichols (Eds.), *Experimental Philosophy* (pp. 3–14). Oxford: Oxford University Press.
Locke, J. (1700). *An essay concerning human understanding*. Oxford: Oxford University Press.
Nadler, J. (2012). Blaming as a social process: The influence of character and moral emotion on blame. *Law and Contemporary Problems*, *2*, 1–31.
Nadler, J., & McDonnell, M.-H. (2012). Moral character, motive, and the psychology of blame. *Cornell Law Review*, *97*, 255–304.
Parfit, D. (1984). *Reasons and Persons*. Oxford: Oxford University Press.
Pettit, P., & List, C. (2011). *Group Agency: The possibility, design, and status of corporate agents*. Oxford: Oxford University Press.
Pizarro, D. A., & Tannenbaum, D. (2012). Bringing character back: How the motivation to evaluate character influences judgments of moral blame. In M. Mikulincer & P. Shaver (Eds.), *The Social psychology of morality: Exploring the causes of good and evil* (pp. 91–108). Washington, DC: APA Press.
Prinz, J., & Nichols, S. (2016). Diachronic identity and the moral self. In J. Kiverstein (Ed.), *Handbook of the social mind* (pp. 449–464). London: Routledge.
Strohminger, N., & Nichols, S. (2014). The essential moral self. *Cognition*, *131*(1), 159–171.
Uhlmann, E. L., Zhu, L., & Diermeier, D. (2014). When actions speak volumes: The role of inferences about moral character in outrage over racial bigotry. *European Journal of Social Psychology*, *44*, 23–29.

Appendix

Study 1, Control Vignette

During his teens and early 20s, James Wilson was engaged in various forms of criminal activities. These ranged from shoplifting, to vandalism and minor drug offenses. Finally, in his late 20s, he was convicted of vehicular manslaughter, having killed another driver in a crash while texting behind the wheel. He was sentenced to 10 years in prison.

In prison, Wilson began to change, as he matured, he had time to reflect on the criminal urges that preoccupied him when he was younger. Now he has served eight of his 10 years, and his outlook has changed considerably. Though he can fully identify with his past values, criminal behavior has little appeal now that he knows the tough consequences. That is, Wilson's morals have not really changed, but he is now motivated to act in accordance with the law.

Study 1, Test Vignette

During his teens and early 20s, James Wilson was engaged in various forms of criminal activities. These ranged from shoplifting, to vandalism and minor drug offenses. Finally, in his late 20s, he was convicted of vehicular manslaughter, having killed another driver in a crash while texting behind the wheel. He was sentenced to 10 years in prison.

In prison, Wilson began to change, as he matured, he had time to reflect on the criminal urges that preoccupied him when he was younger. Now he has served eight of his 10 years, and he can no longer identify with his past behaviors and values. His moral outlook has transformed, and criminal activity does not interest him any longer. That is, Wilson's moral values have really changed, and he is now motivated to act in accordance with the law.