The Anchoring Effect and Moral Judgements

Maya Samantha Rodriguez

Advised by Paul Bloom, Brooks and Suzanne Reagan Professor of Psychology, and Matthew Jordan, PhD

Submitted to the faculty of Cognitive Science in partial fulfillment of the requirements for the degree of Bachelor's of Science

Yale University
April 22nd, 2019

Abstract

Anchoring is a cognitive phenomenon where exposure to incidental numbers can change people's numerical judgements. Considerable work has explored how anchoring can affect people's judgements, valuations, and to some degree behaviors. However, little work has examined how anchoring can affect people's attitudes outside of numerical judgements. We propose that others' judgments can act as anchors, and influence people's attitudes in the moral domain. To test this, we conducted a series of studies in which judgments from others acted as anchors for moral judgements about just punishments in a mock jury scenario. We found that severe or lenient anchors from others influence people's moral judgements, leading to more severe or lenient judgments, respectively. We also tested whether punishment anchors affected the intensity of people's moral emotional responses, and found that, contrary to many models of moral judgment, moral emotions and moral judgments can dissociate.

Keywords: anchoring, moral psychology, punishment, decisionmaking

Contents

The Anchoring Effect and Moral Judgements

Judgements about good and bad, right and wrong, shape how we interact in a social context and evaluate others around us. We aim to associate with those who do good and avoid or punish those whose actions we judge as bad. Moral evaluations can have a profound impact on our social world. Our social world can also end up shaping how we make these evaluations. Information from others, whether it is in the form of prescriptive norms about what is right or wrong to do in a certain context, or judgements about how good or bad certain actions are, can end up influencing the moral judgements we form. Outside information can alter our judgments in other domains as well. Our numerical and statistical judgements can be altered and influenced by a number of biases. Here we explore the effects of one of these, the anchoring effect. We hypothesize that there may be a similarity between how the presence of extreme numbers can affect our numerical judgements and how extreme opinions may similarly influence our moral ones.

The anchoring effect describes a phenomenon observed by psychologists and behavioral economists in which people's judgements are influenced by incidental numbers (Kahneman & Jacowitz, 1995). People may be influenced by incidental numbers when they are asked to provide an estimate of a value or quantity they do not know. For example, Frederick and Mochon (2012) found that, when asked to estimate an unknown value, in this case the weight of a giraffe, people gave consistently lower numbers when first asked to estimate a low anchor (the weight of a wolf) and consistently larger numbers when first asked to estimate a high anchor (the weight of a whale) than they do when asked to estimate the weight of a giraffe when given no anchor at all. When reminded of larger values, people estimated that a giraffe weighed more whereas, when reminded of smaller numbers, people estimated less. Why might it be the case that incidental numbers can have so much influence on the judgments people form?

There are multiple competing theories for the mechanism of the anchoring effect. One is *selective accessibility*. Selective accessibility proposes that anchoring occurs because, by priming either large or small numbers, other numbers of that size are more easily accessible when the participant is asked to make the target judgement (Strack & Mussweiler, 1997). To illustrate, selective accessibility would dictate that thinking of a small animal before being asked to judge how much a giraffe weighs causes the mind to conjure a smaller representation of a giraffe because smaller things have been primed and are more easily accessible.

Another account of anchoring is *scale distortion theory*. Scale distortion theory explains that providing an anchor changes how the respondent views where the target object falls on a scale (Frederick & Mochon, 2012). Thinking of a very small object causes the scale to shift down by making larger numbers seem very large and smaller numbers seem more reasonable, while the opposite happens for larger numbers. To continue the earlier example, thinking of a small animal, in this case, would cause large weights to seem massive, making a smaller but not impossibly small weight for a giraffe seem like a more reasonable response. Evidence has been found in support of both scale distortion and selective accessibility models of anchoring (Frederick & Mochon, 2012).

Anchors can either be self-generated (thought of by the participant) or provided by the experimenter (Furnham & Boo, 2011). They work best when they have some degree of informational relevance to the judgement task at hand (Furnham & Boo, 2011). For example, the weight of a small animal can act as an anchor for the weight of a larger animal, but the number of toes it has may not (Frederick & Mochon, 2012). However, completely task-irrelevant numbers have also been found to act as anchors in some situations as well (i.e. Ariely et al. 2003). Anchors that seem beyond the boundaries of what is plausible, or those that are too extreme, have a limited or no effect compared to plausible anchors (Furnham & Boo, 2011). That

said, anchors have effects across a range of situations. Anchoring has been observed in cases of guessing general knowledge or probability of outcomes, such guessing the weights of animals, percentage of UN countries located in Africa, or the likelihood of a nuclear war (Furnham & Boo, 2011). Anchoring can also have an effect on consumer behavior and the valuation of items. Ariely et al. (2003) found that something as arbitrary as writing down a social security number before being asked about willingness to pay for certain objects could move valuations of those objects up or down. Higher social security numbers, which are completely arbitrary values, caused items to be valued higher, demonstrating one impactful consequence anchoring can have on behavior.

One poignant illustration of the potential real-world consequences of anchoring is the impact it can have in a legal context. Enough & Mussweiler (2001) found that anchoring could affect judges' sentencing decisions in criminal cases. Criminal court judges were given cases and told that the prosecutor had recommended either a high or low sentence. As expected, judges delivered significantly different severe or lenient sentences based on the sentencing anchor that they were given (Enough & Mussweiler, 2001). This effect was replicated with law students, finding that participants followed the same trend when a high or low sentencing recommendation was given by a freshman computer science major as when it was given by a prosecutor, even though they rated the recommendation as less relevant (Enough & Mussweiler, 2001). This demonstrated that the source of the information does not matter; people will still anchor towards information from arbitrary and unreliable sources. Given demonstrated links between severity of sentencing decisions and recidivism in those convicted, this finding also demonstrates one area where this cognitive phenomenon can have very serious and impactful consequences on people's lives (Gendreau, Cullen, & Goggin, 1999).

Anchoring's impacts are made even more consequential by the fact that experts are still somewhat susceptible to anchoring effects (Northcraft & Neale, 1987). The anchoring effect is not easily prevented or mitigated. People will still demonstrate the effect even when given an additional incentive for accuracy, and it will still occur if participants are warned about it and try actively to avoid it (Furnham & Boo, 2011). Anchoring is a highly pervasive and difficult to avoid aspect of cognition that can influence decision-making to strong consequences.

One yet unexplored area where these anchoring effects may influence judgements is within the domain of moral cognition. Moral judgements are immediate, intuitive reactions concerning the "goodness" or "badness" of an individual's action. While there are different models of how moral judgements are made and formed, and the cognitive processes underpinning these judgements, many agree that moral judgements are largely based on immediate emotional reactions to situations, as opposed to reasoned consideration (Haidt, 2012). While rational deliberation can play a role in moral judgment formation, especially in impersonal dilemmas that are less emotionally salient, both behavioral and neurological evidence point towards automatic emotional processes dominating moral decision-making (Greene, Sommerville, Nystrom, Darley, & Cohen, 2001; Greene & Haidt, 2002). According to the social intuitionist model, moral judgements are the product of an immediate and automatic emotional reaction to a dilemma, and the reasons people typically give for having certain moral views follow after as a result of trying to justify the initial emotional intuition (Haidt, 2012). That is, the reasons people give—at least on the social intuitionist model—are rationalizations. While it seems that the largest factor in forming moral judgements are these immediate intuitive reactions, many other factors may influence these judgements as well.

Influence from others can also impact moral judgements. In fact, social persuasion, in which judgments from others can act to shift people's immediate intuitions, is explicitly

discussed in the social intuitionist model (Haidt, 2012). Other experimental evidence has demonstrated that specific forces of social influence like conformity have been shown to have a pronounced effect on how people make and express numerical judgements like length and distance (Asch, 1956). Conformity has been found to influence moral judgements as well. Using Asch's famous paradigm of having a subject make judgements in the presence of confederates presenting information that runs contrary to intuitions, Kundu and Cummins (2013) found that participants show effects of conformity when making moral judgements as well. Participants rated traditionally impermissible scenarios as more permissible when confederates said that they were permissible and traditionally permissible scenarios as less permissible when confederates said that they weren't permissible.

Another way that social information can influence moral judgements is in the form of social norms. Some aspects of morality are relatively universal, like the principles that harm is bad and fairness is good, while others are more culturally specific and are shaped by social context (Graham et al, 2013). One area where this social aspect of morality becomes clear is in the domain of social norms. Norms are typical behaviors and generally accepted customs within a particular group, and they can be closely linked to moral judgements (Sunstein, 1996).

Regardless of their specific content, people typically have negative reactions when norms are broken. Some norms are clearly moralized and align with our folk conceptions of morality, such as norms of reciprocity and cooperation, which serve important societal functions like enforcing cooperation among groups and are often punished when they're violated (Fehr & Fischbacher, 2004). Some norms are clearly moralized but seem more arbitrary, for example, religious traditions about sexual practices, food consumption, and dress can all be moralized but are highly culturally specific (Graham et al, 2013). The reason why these specific norms are moralized may be related to the fact that the violation of them induces disgust (Nichols, 2002).

Some norms are not necessarily moralized, like in America cars drive on the right side of the street and Wall Street employees usually wear suits to work, but despite the lack of obvious moralization there is usually some kind of negative reaction or unpleasant emotion when a norm like this is violated. For example, showing up to your job at Goldman Sachs one day in pajamas would be unusual and inappropriate even though it is hard to argue doing so is morally wrong.

People will follow prescribed moral norms even when those norms are completely arbitrary (Pryor, Perfors, & Howe, 2019). One study constructed for participants an arbitrary moral "norm" about other's actions in a moral in moral situations where the "right" action to take was somewhat ambiguous, telling them that in a previous version of the study they were participating in, an error in randomization had assigned 75% of participants to a certain decision condition. This is notable in that in this study the norm was both completely arbitrary and recognizably task-irrelevant (because it was based upon a randomization error and not other participant's decisions). Despite this, participants would report that they would conform to this norm, even though they were aware it was completely random (Pryor, Perfors, & Howe, 2019). This finding demonstrates another subtle and automatic way that social information can influence moral judgements.

Research has demonstrated the effects that external information in the form of anchoring in various cognitive domains from judgement formation to consumer behavior, but little research has examined the potential effects that anchoring could have in the moral domain. Similarly, to how people's morals can be shaped by information from others in the form of norms moral opinions may be moved by the presence of others' moral judgments, even if those judgments are extreme. This could be one mechanism by which people end up coming to change their moral opinions. Exposure to more extreme opinions could cause people to form judgments closer to

those extremes. This could imply that extreme moral opinions once introduced are helpful and somewhat necessary in driving social change.

This also introduces a new element to traditional models of moral decision making. Others' judgments could be an additional informational component of a "starting point" where individuals form their decisions from. This additional information would not change the direction of the judgement but may change how severe or strong it ends up being by altering the baseline for the original judgement.

Based on this work we have conducted a series of studies testing the hypothesis that information from others can act as an anchor for moral judgements. According to our predictions, the presence of severe judgments from others will cause people's opinions to become more severe, and the presence of lenient judgements will cause other's judgements to become more lenient. We first establish the existence of an effect and then test a possible mechanism underlying our documented effect.

Study 1

Methods

Participants.

Using Amazon Mechanical Turk, we recruited 300 participants (51% male, mean age = 36.6).

Design.

Participants were first told that for the purposes of this study, they were to imagine that they were a member of a jury deciding a sentence for a convicted defendant and read one vignette about the defendant and their hypothetical crime. The jury scenario provided a valid, real-world scenario where people 1) have to weigh on the moral wrongness of certain actions and 2) where they receive information about other people's moral judgements, making the scenario one that was both plausible to participants and one that obscured the aims of the study. The vignettes were one sentence descriptions of crimes stating the defendant had already been found guilty. The crimes provided were spousal abuse, animal abuse, tax evasion, robbery, and vandalism.

Participants were randomly assigned to one of three "information" conditions where they were given information about the sentencing recommendations of the other members of the jury: (1) high information, (2) low information, or (3) no information. In the 'high information' condition, punishments recommended by the other jurors were high, including one that was absurdly high (1 year in prison, life in prison, 1 year in prison, and death penalty). In the 'low information' condition, punishments recommended by the other jurors were low (a $150 fine, a $5000 fine, a $150 fine, and no punishment). In the 'no information case', which served as the control, participants were not given any information about other juror's recommendations.

Measure.

For this study, we developed an 9-point ordinal scale of punishment ranging from "no punishment" to "death penalty" (*1- no punishment, 2 - $150 fine, 3 - $5000 fine, 4 – 6 months probation, 5 – 6 months house arrest, 6 – 1 month in prison, 7 – 1 year in prison, 8 – life in prison, 9 – death penalty)*. The punishment judgement was the dependent variable. We took people's punishment judgements as a measure of moral condemnation.

Results

We used ordered probit regression to examine punishment judgments as a function of condition (high, low, or no information). We also controlled for which transgression (i.e., spousal abuse, tax evasion) by including 'vignette' as a fixed effect. We found that participants in the high information condition opted for more severe punishments ($M = 5.70$, $SE = 0.16$) than those who received no information ($M = 5.11$, $SE = 0.17$), $\beta = -0.48$, $SE = 0.15$, $p = 0.002$), as predicted. Those who received no information condition gave more severe punishments than those in the low information condition ($M = 4.00$, $SE = 0.17$), $\beta = -0.56$, $SE = 0.16$, $p < 0.001$. We did not find any evidence that others' judgments interacted with vignette.

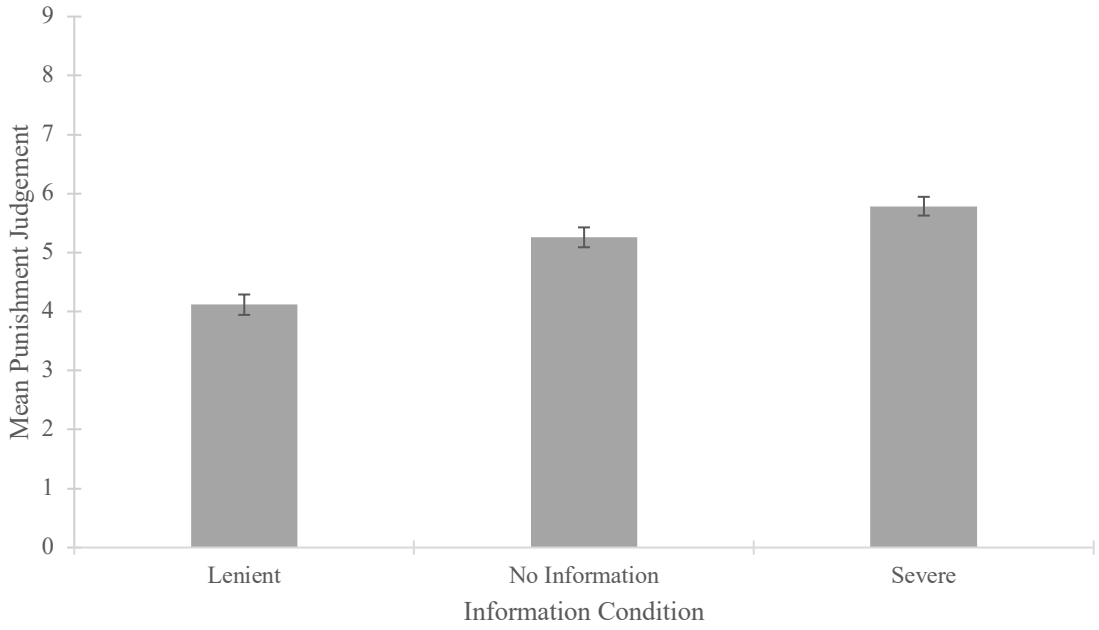


Fig 1. Mean punishment judgements by condition in Study 1. Scale numbers correspond to: 1- *no punishment, 2 - $150 fine, 3 - $5000 fine, 4 – 6 month's probation, 5 – 6 months house arrest, 6 – 1 month in prison, 7 – 1 year in prison, 8 – life in prison, 9 – death penalty*

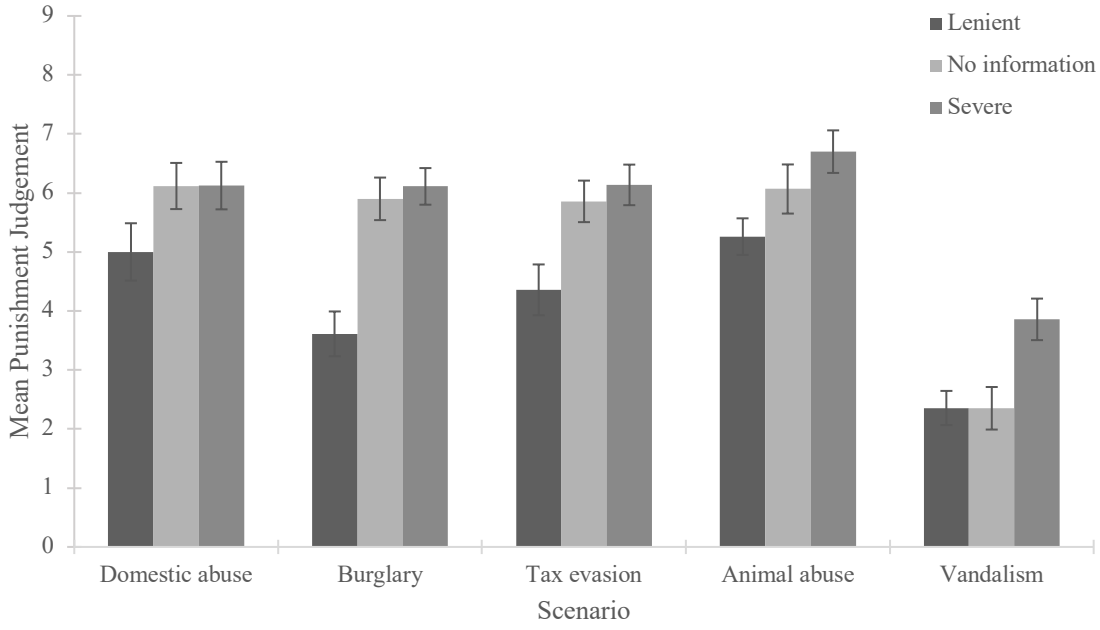


Fig 2. Mean punishment judgements by condition by scenario in Study 1

A significant increase in punishment judgements in the severe information case from the no information case and a significant decrease in punishment judgements in the lenient case from the no information case supports the presence of an anchoring effect which caused judgements to shift towards the severe or lenient values provided.

Study 2

Study 1 established the existence of an effect of information from others on people's punishment judgements. Following this, we considered the possibility that we see a shift in people's punishment judgments as a function of jury information because either the high or low punishment information may truly impact people's experienced moral emotions, which in turn

impact their punitive judgments. Previous work has demonstrated links between moral outrage, and moral emotions like anger and people's desire to punish (Bastian, Denson, & Haslam, 2013; Nelissen & Zeelenberg, 2009). Given the link between these two factors and punishment, we hypothesized that one reason why we saw punishment judgements changing across conditions was that something about exposure to the information in each condition was changing people's experience of those two factors. That is, when participants learned that other jurors wanted to severely punish the transgressor, the participant may have consequently experienced a more negative emotional reaction to the transgression itself. In Study 2 we attempted to test this by collecting an emotional measure and seeing if it varied between the information cases.

Methods

Participants.

Using Amazon Mechanical Turk, we recruited 457 participants (49% male, mean age = 36.3).

Design.

Our design for Study 2 was similar to our design for Study 1 with a few alterations. We made two key changes. First, the crime vignettes were narrowed down to three scenarios – animal abuse, domestic abuse, and racist vandalism – because those were the cases we felt were the most emotionally salient and likely to evoke a strong response. Second, the scale was also altered slightly. The juror recommended punishments in the high information case were lowered to reflect alterations in the scale.

Measures.

Before asking for the main DV – punishment judgements – we also asked a series of four questions in randomized order measuring participants' *moral outrage*. Participants were asked to rate how much they agreed with the following statements on a 7-point Likert scale. The

statements asked were: (1) “I feel a compelling need to punish the defendant”, (2) “I feel a desire to hurt the defendant”, (3) “I believe the defendant is evil to the core”, and (4) “I feel morally outraged by what the defendant did to the victim.”

We also introduced a series of *emotional measures* looking at emotions commonly associated with moral judgements – anger, disgust, and contempt. Participants were asked to indicate the degree they felt each emotion when considering the crime vignette they had read on a 7-point Likert scale.

We updated our punishment scale for this and subsequent versions of the study. The low endpoint remained “no punishment” while the high endpoint was changed to be “life in prison”, and 5 years in prison was added as an intermediate between 1 year in prison and life in prison. The updated scale was as follows (*1- no punishment, 2 - $150 fine, 3 - $5000 fine, 4 – 6 months probation, 5 – 6 months house arrest, 6 – 1 month in prison, 7 – 1 year in prison, 8 – 5 years in prison, 9 – life in prison)*.

Results

We began our analysis by examining punishment judgments as a function of condition (high information, low information, no information). Just as in Study 1, we found that those who saw others gave more severe punishments offered more severe punishments ($M = 6.30$, $SE = 0.15$) than those who received no information ($M = 5.63$, $SE = 0.15$), $\beta = -0.37$, $SE = 0.12$, $p = 0.002$, who offered more severe punishments than those who saw others gave more lenient punishments ($M = 4.81$, $SE = 0.14$), $\beta = -0.42$, $SE = 0.12$, $p < 0.001$.

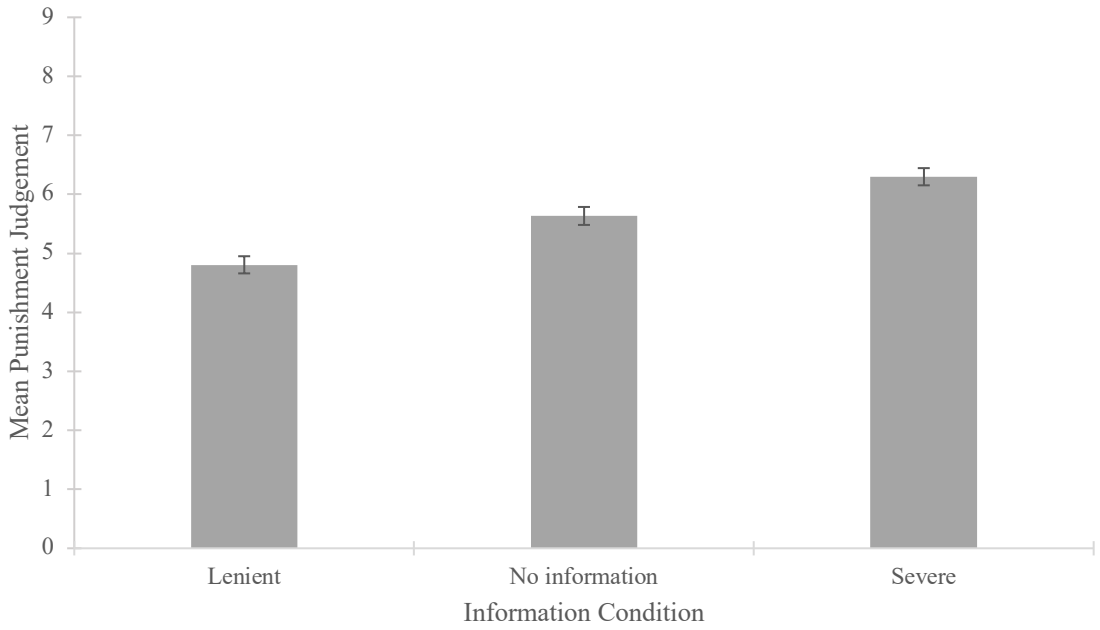


Fig 3. Mean punishment judgements by condition in Study 2. Scale numbers correspond to: *1- no punishment, 2 - $150 fine, 3 - $5000 fine, 4 – 6 months probation, 5 – 6 months house arrest, 6 – 1 month in prison, 7 – 1 year in prison, 8 – 5 years in prison, 9 – life in prison*

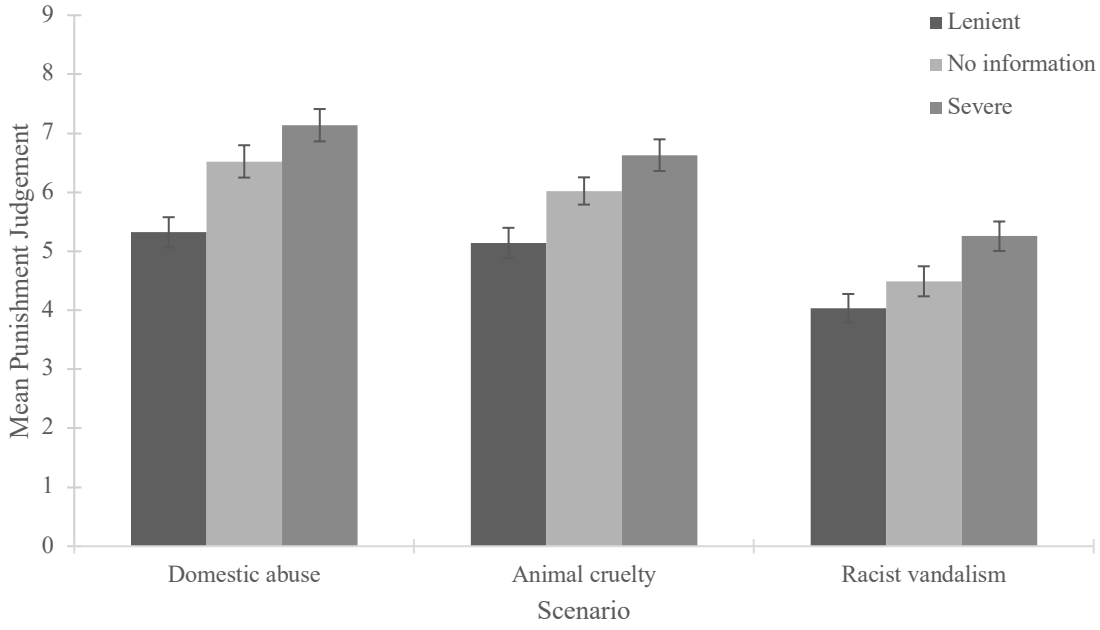


Fig 4. Mean punishment judgements by condition by scenario in Study 2

The moral emotion items were highly reliable ($\alpha = 0.80$), so we combined them into a composite measure to determine the effect of information condition on emotional response. We did not find evidence that information condition affected emotional response, despite the fact that overall emotional response was highly related to punishment decisions, $\beta = 0.27$, $SE = 0.02$, $p < 0.001$.

We replicated the effect observed in Study 1 by again demonstrating significant differences between punishment judgements across the different information conditions. The observed correlation between the emotional response and punishment judgements is evidence that those who were more morally outraged chose harsher punishments. The lack of a relationship between information condition and emotional response, however, indicates that emotions were not changing overall in response to observing others' severe or lenient judgements.

Study 3

Because Study 2 replicated the effect found in Study 1 but did not find a change in emotions by condition, in Study 3 we attempted to induce this more directly by including an emotional manipulation in the form of emotionally valenced vignettes, which we predicted would make judgements more severe or lenient depending on the content of the manipulation.

Methods

Participants.

Using Amazon Mechanical Turk, we recruited 455 participants (55% male, mean age = 36.4).

Design.

Study 3 introduced an additional manipulation aimed at altering the participants' emotions. At the same time that they were presented with the information about juror's punishment decisions, participants were given vignettes that they were told described how the other jurors had reached their decision. These vignettes were either high emotionally valenced – emphasizing the emotions that the other jurors felt such as empathy for the victims and outrage towards the defendant (for example the following was presented for the high emotion case in the spousal abuse scenario "In reaching our decision, we considered the impact the crime had on its victim as well as the consequences the sentence would have for the defendant. We were outraged when we considered how terrified the defendant's spouse was, and angry that the defendant would hurt someone that loved and trusted him. The defendant must pay for his actions"), or low emotionally valenced – emphasizing that in reaching the decision jurors had placed weight on the importance of impartiality and following sentencing guidelines. Full vignettes used are included in Appendix A.

Participants were randomly presented with either a high emotion vignette, a low emotion vignette, or no emotional information in addition to a randomly paired punishment information judgement.

Unlike previous versions, this study was run within subjects, where participants were presented with all three crime scenarios in random order and a randomized mixture of high, low and no punishment information vignettes and high, low and no emotional manipulation vignettes accompanying the scenario.

Measures.

The measures were identical to Study 2

Results

Using mixed effect ordered probit models, we examined punishment decisions as a function of others' judgments, the emotionality of the justification they read, and the interaction of these two variables. We included subject-level random effects and transgression fixed effects. Again, we found evidence that others' judgments affect punishment decisions, such that those who see others gave severe punishments also offered more severe punishments ($M = 6.53$, $SE = 0.09$) than those who received no information ($M = 6.15$, $SE = 0.09$), $\beta = -0.23$, $SE = 0.08$, $p = 0.004$, who offered more severe punishments than those who saw others gave more lenient punishments ($M = 5.37$, $SE = 0.09$), $\beta = -0.51$, $SE = 0.08$, $p < 0.001$. However, we found no evidence that the emotionality of the punishment justification affected punishment decisions, βs < 0.07, SEs > 0.08, ps > 0.350.

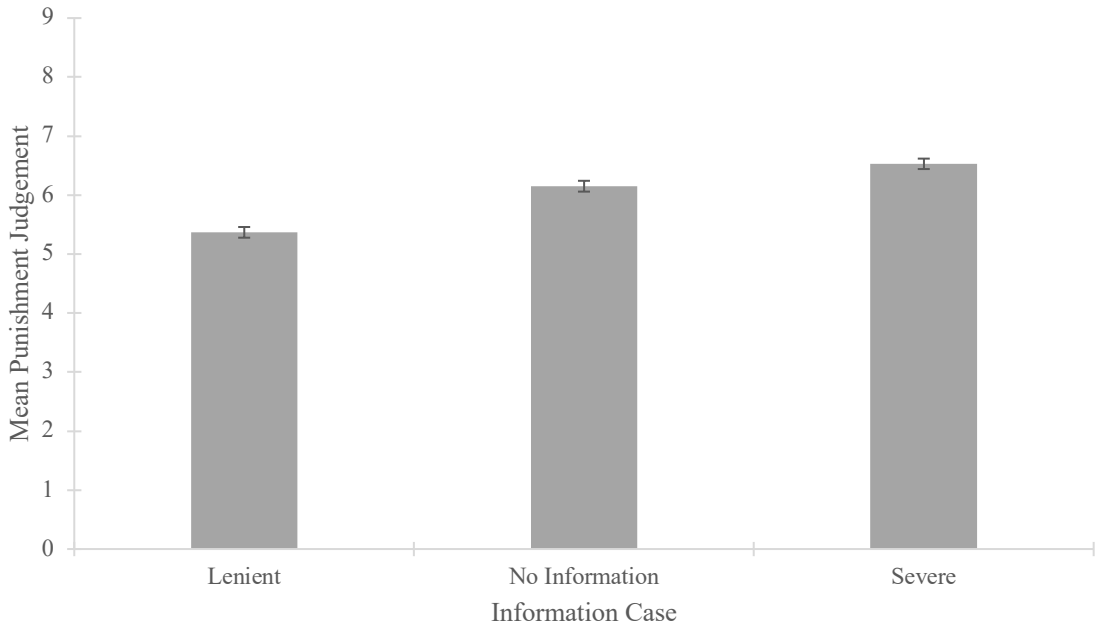


Fig 5. Mean punishment by condition in Study 3.

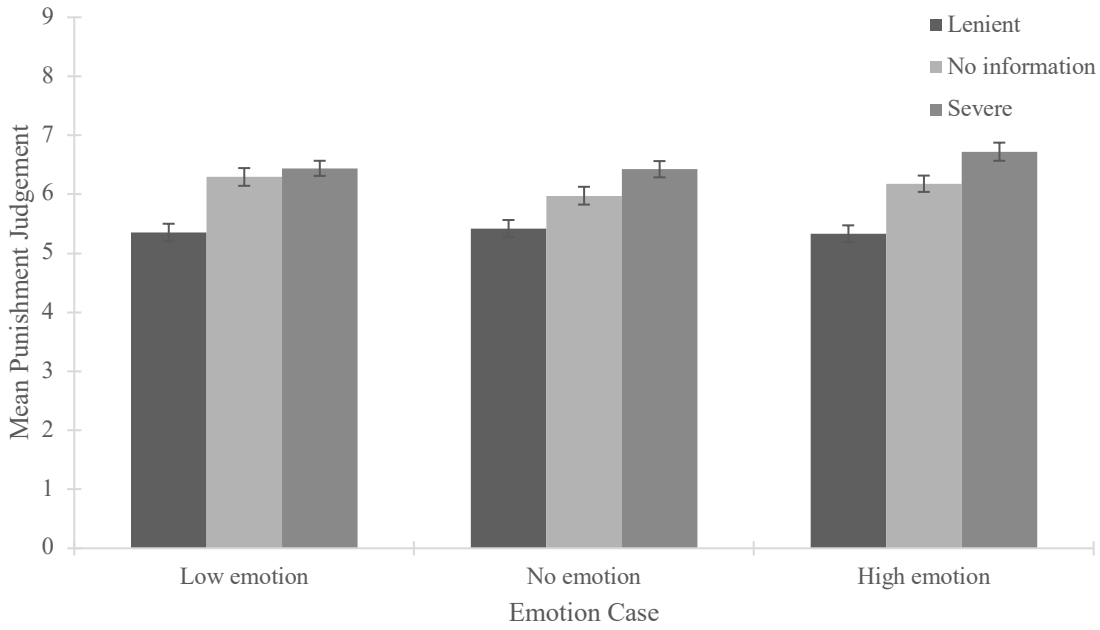


Fig 6. Judgements by emotion by condition in Study 3

The moral emotion items were highly reliable ($\alpha = 0.83$), which allowed us to create a composite emotional response score. We did not find evidence that either the emotionality of the punishment justification or others' judgments affected participants' emotional responses, despite the fact that emotional responses were highly predictive of punishment decisions, $\beta = 0.34$, $SE = 0.02$, $p < 0.001$. Emotional response did not interact with either emotionality or others' judgments.

We again replicated the initial effect of Studies 1 and 2 by demonstrating significant differences between punishment judgements across the different information conditions. However, despite again finding that reported emotions were linked positively to punishment judgements, we did not find an effect of the emotional manipulation we attempted. There were no significant differences in punishment judgements across the emotional manipulation conditions, indicating that the emotional manipulation did not affect judgements. Additionally, we did not find significant differences in reported emotional responses across the different

emotional manipulation conditions, indicating that the attempted emotional manipulation was unsuccessful in altering reported emotions.

Discussion

The results of our three studies support our overall hypothesis that anchors can influence people's moral judgements. Study 1 demonstrated the effect, finding that people chose significantly lower punishments in the presence of more lenient anchors than they did without any information, and more severe punishments in the presence of high anchors than they did without any information. This suggests that the presence of these anchors causes judgements to shift.

Study 2 replicated the results of Study 1, further establishing the reliability of the effect. One possible mechanism for this change is that the presence of this additional information altered people's experienced moral emotions, with the proposed emotional mechanism acting as an intermediary step between hearing another's judgements and forming one's own. Study 2 attempted to test this mechanism by introducing measures of emotion and moral outrage. We found that those who experienced stronger moral emotions and more outrage offered more severe punishments, consistent with previous findings on how moral outrage and emotions are linked to punitive behaviors. However, we did not find that these factors were different across the information conditions, only that the variance among individuals within information conditions was linked to severity of judgements.

Study 3 attempted to test the proposed emotional mechanism more directly through an attempted manipulation of people's emotions. While Study 3 replicated the original effect of Studies 1 & 2, there was no effect of the emotional manipulation we attempted on either people's emotional responses or their judgements. This suggests that the emotional manipulation we attempted was not effective, and also could suggest that emotions are not the mechanism for the effect which we observed and that the effect acts through a different pathway.

Evidence towards the former explanation comes from the fact that the emotional measures were not significantly different across the different emotion conditions. We predict this may have occurred because participants were given information about the emotions experienced by other juror members. Reading about the emotions of someone else in a similar situation to one's own may not cause one to feel more strongly or to mirror those emotions. Our emotional manipulation may have been ineffective because the emotions people were reading about were of other jurors, not the victims or the defendants. This would be consistent with work on victim impact statements and perspective taking demonstrating that hearing about the emotions and experiences of victims and perpetrators can have an influential effect on increasing punitive behavior (Myers & Greene, 2004).

One alternative explanation of the current findings is that no anchoring is occurring, people are simply mirroring the information they have been given. People may not have a particularly strong sense of what an appropriate punishment is for the given transgressions. When they are given information about other's judgements, it is possible participants may just copy that information because it conserves the cognitive effort of having to form one's own judgement. When no other information is available, they make the decision using some other heuristic, whether it's guessing randomly, simulating what seems appropriate or simply choosing the middle of the scale.

We do not think this explanation can explain our data. Specifically, in every condition, the mean punishment judgements were lower than the most lenient punishment information provided in the severe case and higher than the most severe punishment information provided in the lenient case, and in both cases far away from both the average and mode of the judgements they were told other people had made. This provides evidence that the outside information was shifting people's judgements up or down from a no information baseline, but only by a small

amount. If people were simply copying what others had said, we would expect their judgements to match at least one of the values provided by the information vignettes, but this was not the case. This indicates that there was an effect of our manipulation beyond simple copying.

The effect observed is similar to previous work done on anchoring but there were some important differences. First, multiple "anchors" were provided and while all the values were in a similar direction (either high or low) the values themselves were varied. Most work on anchoring has been done around a single value. Anchoring is also typically demonstrated with numerical cognition, and while numbers were provided, what we were ultimately asking for were not people's numerical judgements due, but rather ordinal punishment judgements that increased in severity but do not track perfectly onto an interval scale the way most previous research done on anchoring has.

There may have been additional complications arising from the ordinal scale. While we specifically chose this measure because we felt it was more valid and emotionally salient than, for example, asking people for a number of months in prison or dollar amount for a fine, it was not an interval measure and the "distance" between each of the points on the scale may not have been consistent. Additionally, while we intended for the scale to serve as a proxy for moral condemnation or moral disapproval, it may not have tracked these things directly. Because we did not collect any other traditional measures of moral condemnation, we cannot be sure that our scale tracks onto moral judgements perfectly. However, our scale did track fairly closely to the other measures of moral condemnation that we collected, specifically outrage and desire to harm the defendant, so we feel confident in saying it was reflective of moral judgements to a certain degree. Further research may want to include multiple measures of moral disapproval including more direct ones.

What we have found is a new and consequential area of research into people's moral cognition. The results of our study have implications for both practical real-world issues and our understanding of moral decision-making on a broader level.

We choose the jury scenario in this study because it was familiar and easy to explain to participants. There are some obvious implications for this line of research to jury decision-making and the criminal justice system. While in many criminal cases judges, not juries decide sentences, this anchoring effect, which has also been demonstrated in judges and their sentencing decisions and has also now been demonstrated in jury scenario has real consequences on people's lives (Enough & Mussweiler, 2001). There are considerable ethical implications given the impact a longer prison sentence can have on someone's life and the arbitrary nature of these factors influencing these decisions. Further research could possibly guide criminal justice policy recommendations to reduce these anchoring effects in judges' decisions and jury deliberations, including changing how sentencing recommendations are presented or requiring jury members to either pre-commit to a decision or announce their decisions concurrently so that individuals are not influenced by previous answers.

More broadly this line of research provides a possible pathway for general moral shift. If people's moral opinions are influenced by high or low anchors from others, maybe the presence of extreme moral opinions are necessary for the general morals of a culture to shift in one direction or the other. To give an example, public opinion on sexual harassment has changed quite dramatically in the last half a century. What was acceptable 50 or even 10 years ago is dramatically different from what we currently allow in 2019. Under our anchoring prediction, it is possible that this general cultural shift in what was morally acceptable occurred because with the rise of recent social movements and popularized scandals, people were suddenly hearing that certain actions previously condoned were morally unacceptable, and the presence of these more

extreme information caused opinions to shift towards that direction. This would imply that sometimes extreme opinions, while not necessary, are capable of changing people's moral opinions, something thought to be fairly hard to do.

This could also explain the phenomenon of group conformity and groups collectively showing worsening moral behavior**.** If members of a group begin to express support for typically impermissible moral behaviors, like harming certain groups of people or denying rights to individuals, this moral anchoring effect would predict that those around them would also experience a moral shift in that direction, because of anchoring towards that more extreme opinion. This could account for how large groups of people come to moralize atrocities.

Our work also contributes to the broader puzzle of understanding how we come to form moral judgements. We have shown one way in which moral decisions are permeable to influence from other people's opinions. Our findings suggest that judgements of badness can be a function both of individual intuitive reactions and outside information. Intuitive emotional reactions may be the largest part of the equation, however other factors may influence the baseline by which judgements are made in subtler ways. We have found what we believe to be one such influencing factor.

Future areas of research could go in several directions. If emotions are not the mechanism causing the moral shift we discovered, one next step would be to look for what is. Possible explanations could include explanations for numerical anchoring, such as scale distortion theory. Another compelling direction would be to test if the effect holds when the anchor is outrageous or extreme. For example, if all members of the jury chose "life in prison", would participants still anchor up towards that value, or would that information just be discarded all together because it seems unreasonable? One further direction would be to examine the role of actions signaling a moral opinion and not the expression of the opinion itself. Perhaps the direct expression of an

opinion isn't needed but even observing people acting as though they hold a certain moral belief would trigger the shift. This direction could help explain how this effect influences our world day to day, since people often behave in ways that signal their moral opinions even when they do not vocalize them.

Ultimately what we have shown is an application in which broader cognitive forces can shape our moral judgements. The anchoring effect can influence our moral decisions as well as other kinds of valuations. We've here demonstrated a way in which this effect, as well as the forces of social information and consensus, can influence our moral judgements.

Author Contributions

Rodriguez and Jordan developed and designed the study with input from Professor Bloom. Rodriguez developed the vignettes and implemented the survey flow for all versions of the study, with assistance from Jordan. Jordan uploaded the study and implemented the data collection on mTurk. Jordan ran the statistical analyses and Jordan and Rodriguez together wrote the Methods and Results sections. Rodriguez wrote the Introduction and Discussion sections with input from Jordan and Bloom.

Acknowledgements

Thank you to:

Paul Bloom – for introducing me to Moral Psychology and for invaluable feedback and support in the completion of this project.

Matthew Jordan – for introducing me to this project and for a tremendous amount guidance and support throughout the thesis process.

Julia Marshall - for introducing me to research, 3 years of academic mentorship, and helping me figure out Qualtrics.

Alexa, Gracie, Emily, Matti and the rest of the Mind and Development Lab for their time and feedback on this project.

The Cognitive Science department - for allowing me to dedicate 4 years of college to a very specific academic passion.

Morse College, Head Panter-Brick and Dean Gleason, Mark Sheskin, and my peer reviewers.

The Spotify Classic Rock playlist - for writing inspiration.

And my friends and family for the endless support, especially –

Andi – without whom I would've been a Psych major,

Caroline – for hours in the library writing together,

Danielle and Rob – for understanding my enthusiasm for this project and reflecting it back towards me.

References

Ariely, D., Loewenstein, G., & Prelec, D. (2003). "Coherent arbitrariness": Stable demand curves without stable preferences. *The Quarterly journal of economics*, *118*(1), 73-106.

Asch, S. E. (1956). Studies of independence and conformity: I. A minority of one against a unanimous majority. *Psychological monographs: General and applied*, *70*(9), 1.

Bastian, B., Denson, T. F., & Haslam, N. (2013). The roles of dehumanization and moral outrage in retributive justice. *PloS one*, *8*(4), e61842.

Critcher, C. R., & Gilovich, T. (2008). Incidental environmental anchors. *Journal of Behavioral Decision Making*, *21*(3), 241-251.

Enough, B., & Mussweiler, T. (2001). Sentencing Under Uncertainty: Anchoring Effects in the Courtroom 1. *Journal of applied social psychology*, *31*(7), 1535-1551.

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

Frederick, S. W., & Mochon, D. (2012). A scale distortion theory of anchoring. *Journal of Experimental Psychology: General*, *141*(1), 124.

Furnham, A., & Boo, H. C. (2011). A literature review of the anchoring effect. *The journal of socio-economics*, *40*(1), 35-42.

Gendreau, P., Cullen, F. T., & Goggin, C. (1999). *The effects of prison sentences on recidivism* (pp. 4-5). Ottawa, Ontario: Solicitor General Canada.

Graham, J., Haidt, J., Koleva, S., Motyl, M., Iyer, R., Wojcik, S. P., & Ditto, P. H. (2013). Moral foundations theory: The pragmatic validity of moral pluralism. In *Advances in experimental social psychology* (Vol. 47, pp. 55-130). Academic Press.

Greene, J. D., Sommerville, R. B., Nystrom, L. E., Darley, J. M., & Cohen, J. D. (2001). An fMRI investigation of emotional engagement in moral judgment. *Science*, *293*(5537), 2105-2108.

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

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

Herzog, S. M., & Hertwig, R. (2009). The wisdom of many in one mind: Improving individual judgments with dialectical bootstrapping. *Psychological Science*, *20*(2), 231-237.

Inbar, Y., Pizarro, D. A., & Bloom, P. (2009). Conservatives are more easily disgusted than liberals. *Cognition and emotion*, *23*(4), 714-725.

Jacowitz, K. E., & Kahneman, D. (1995). Measures of anchoring in estimation tasks. *Personality and Social Psychology Bulletin*, *21*(11), 1161-1166.

Kundu, P., & Cummins, D. D. (2013). Morality and conformity: The Asch paradigm applied to moral decisions. *Social Influence*, *8*(4), 268-279.

Lorenz, J., Rauhut, H., Schweitzer, F., & Helbing, D. (2011). How social influence can undermine the wisdom of crowd effect. *Proceedings of the national academy of sciences*, *108*(22), 9020-9025.

Myers, B., & Greene, E. (2004). The prejudicial nature of victim impact statements: Implications for capital sentencing policy. *Psychology, Public Policy, and Law*, *10*(4), 492.

Nelissen, R. M., & Zeelenberg, M. (2009). Moral emotions as determinants of third-party punishment: Anger, guilt and the functions of altruistic sanctions. *Judgment and Decision making*, *4*(7), 543.

Nichols, S. (2002). Norms with feeling: Towards a psychological account of moral judgment. *Cognition*, *84*(2), 221-236.

Northcraft, G. B., & Neale, M. A. (1987). Experts, amateurs, and real estate: An anchoring-and-adjustment perspective on property pricing decisions. *Organizational behavior and human decision processes*, *39*(1), 84-97.

Pryor, C., Perfors, A., & Howe, P. D. (2019). Even arbitrary norms influence moral decision-making. *Nature Human Behaviour*, *3*(1), 57.

Strack, F., & Mussweiler, T. (1997). Explaining the enigmatic anchoring effect: Mechanisms of selective accessibility. *Journal of personality and social psychology*, *73*(3), 437.

Sunstein, C. R. (1996). Social norms and social roles. *Colum. L. Rev.*, *96*, 903.

Vul, E., & Pashler, H. (2008). Measuring the crowd within: Probabilistic representations within individuals. *Psychological Science*, *19*(7), 645-647.

Appendix A – Vignettes & Full Questions Asked

1. Crime Scenarios

"In this study, imagine you are on a jury and need to make a decision about what punishment someone should receive for committing a crime. We will describe what the defendant has been convicted of on the pages that follow. Your task is to determine what punishment they should receive."

Domestic abuse: "In this case, the defendant has been convicted of domestic abuse for beating their spouse." (Used in Studies 1, 2, and 3)

Burglary: "In this case, the defendant has been convicted of burglary for breaking into someone's house and stealing some electronics." (Used in Study 1 only)

Tax evasion: "In this case, the defendant has been convicted of tax evasion after not paying taxes for two decades." (Used in Study 1 only)

Animal cruelty: "In this case, the defendant has been convicted of animal cruelty for neglecting, then killing a pet dog." (Used in Studies 1, 2, and 3)

Vandalism: "In this case, the defendant has been convicted of vandalism for spray painting an abandoned building in a remote part of town." (Used in Study 1 only)

Racist Vandalism: "In this case, the defendant has been convicted of vandalism for spray painting racist obscenities on a black church." (Used in Studies 2 and 3)

2. Information Cases

Lenient information: "Below is some information about what punishment the other jurors thought the defendant should receive. Please read this information carefully.

Juror 1 said the defendant should receive a $150 fine.
Juror 2 said the defendant should receive a $5000 fine
Juror 3 said the defendant should receive a $150 fine.
Juror 4 said the defendant should receive no punishment."

Severe information: "Below is some information about what punishment the other jurors thought the defendant should receive. Please read this information carefully.

Juror 1 said the defendant should receive 1 year in prison.
Juror 2 said the defendant should receive 5 years in prison
Juror 3 said the defendant should receive 1 year in prison.
Juror 4 said the defendant should receive life in prison."

3. Moral Outrage Measures (used in Studies 2 and 3)

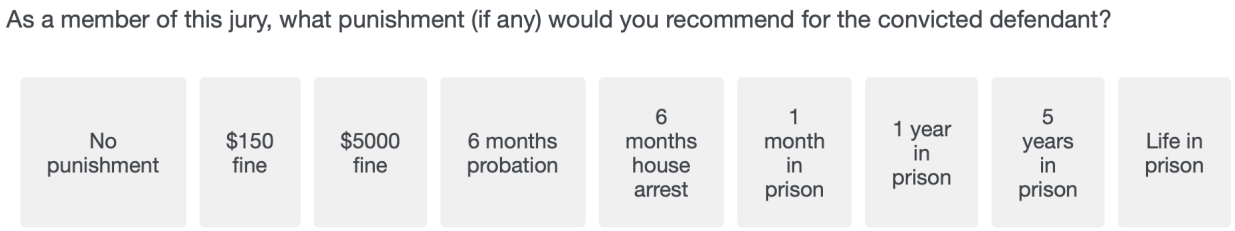
“Please rate how much you agree with the following statements
I feel a compelling need to punish the defendant.
I feel a desire to hurt the defendant.
I believe the defendant is evil to the core.
I feel morally outraged by what the defendant did to the victim.”

Questions were presented in a random order and scored on a scale of 1 (not at all) to 7 (very much)

4. Emotional Measures (used in Studies 2 and 3)
“Please indicate the degree to which you felt each of the emotions below when considering the crime you just read about.
To what extent did you feel **disgusted?**
To what extent did you feel **angry?**
To what extent did you feel **contempt?**”

Questions were presented in a random order and scored on a scale of 1 (not at all) to 7 (extremely so)

5. Punishment Scale



6. Emotion Vignettes (used in Study 3)

High Emotion Domestic Abuse Case: "In reaching our decision, we considered the impact the crime had on its victim as well as the consequences the sentence would have for the defendant. We were outraged when we considered how terrified the defendant’s spouse was, and angry that the defendant would hurt someone that loved and trusted him. The defendant must pay for his actions."

High Emotion Animal Abuse Case: "In reaching our decision, we considered the impact the crime had on its victim as well as the consequences the sentence would have for the defendant. We were outraged when we considered how terrified the defendant’s dog must have been, and angry that the defendant would abuse and kill a pet that depended on and trusted him. The defendant must pay for his actions."

High Emotion Racist Vandalism Case: "In reaching our decision, we considered the impact the crime had on its victim as well as the consequences the sentence would have for the defendant.

We were outraged when we considered how disrespected and threatened the churchgoers must have felt, and angry that the defendant vandalized a religious institution that has a lot of value in the community. The defendant must pay for his actions."