Punishing Without Looking for Reputational Gain

Jillian J. Jordan
Nour S. Kteily
Punishing Without Looking for Reputational Gain

Jillian J. Jordan
Harvard Business School

Nour S. Kteily
Harvard Business School
Punishing without looking for reputational gain

Jillian J. Jordan1*, Nour S. Kteily2

1 Harvard Business School, Soldiers Field, Boston, MA 02163
2 Kellogg School of Management, Northwestern University,
   2211 Campus Dr, Evanston, IL 60208

*Corresponding author: Jillian J. Jordan
Email: jijordan@hbs.edu
Phone number: (651) 895-3771
Mailing address: Soldiers Field, Boston, MA 02163

Keywords: Opposing perspectives; Outrage culture; Signaling; Ideology; Moralistic punishment

Link to OSF page with all data, scripts, experimental materials, and pre-registrations:
https://osf.io/3es2k/?view_only=e272c31aa7304eca841f084d582187fb
Abstract
Critics of “outrage culture” allege that “virtue signaling” drives people to punish alleged wrongdoers without due consideration. But do people actually “punish without looking” for reputational gain? And if so, is this because unquestioning punishment looks particularly virtuous? We examined punishment without looking across three studies of Americans (total $n = 7,952$), in which “Actors” chose whether to sign real punitive petitions about politicized issues (“punishment”), after deciding whether to read articles opposing these petitions (“looking”). To manipulate reputation, we paired Actors with co-partisan “Evaluators”, varying whether Evaluators observed (i) nothing about Actors’ behavior, (ii) (only) whether Actors punished, or (iii) whether Actors punished and whether they looked. We found that Evaluators financially rewarded Actors who did (vs. did not) punish. And correspondingly, making punishment observable to Evaluators (i.e., moving from our first to second condition) drove Actors to punish more overall. Furthermore, because some of these individuals did not look, making punishment observable also increased rates of punishment without looking. Thus, reputation can encourage unthinking punishment. But do punishers who eschew opposing perspectives look particularly virtuous? No: Evaluators preferentially rewarded Actors who punished with (vs. without) looking. And correspondingly, making looking observable (i.e., moving from our second to third condition) drove Actors to look more overall—and to punish without looking at comparable or diminished rates. We thus find that reputation fuels unthinking punishment, but simply as a byproduct of encouraging punishment in general. Indeed, rather than fueling reflexive decisions, spotlighting punishers’ decision-making processes can actually encourage reflection.
Introduction

Consider the following narrative: Following an accusation of wrongdoing, a so-called “outrage mob” quickly appears. Initially, just a few voices speak out to condemn the accused. But before long, more and more people “pile on,” hastily choosing to mete out punishment in a race to signal their virtue while paying little attention to evidence or opposing perspectives.

According to critics who bemoan what they call “outrage culture”, this narrative is an all-too-accurate representation of the current state of societal discourse. Indeed, in recent years, a growing number of people—from academics, to media personalities, to members of the business community, to everyday citizens—have expressed concern that society has become too quick to enact moralistic punishment without due consideration, yielding outcomes that are disproportionate, unfair, and sometimes even downright cruel (1–5).

Consistent with this perspective, even highly-educated professors can, by their own admission, make the mistake of punishing without sufficient reflection. In February 2022, 38 Harvard faculty members signed an open letter condemning the University’s choice to sanction John Comaroff, an Anthropology Professor who was accused, by several students, of sexual harassment. In their letter, the 38 signatories expressed “dismay” at the sanctions, vouched for Comaroff’s character, and endorsed an account of events put forward in a press release from Comaroff’s lawyer (6).

Shortly after their open letter was published, 73 other Harvard faculty penned a new letter criticizing their colleagues for rushing to judgment on the basis of one-sided information (7). After more details about the allegations against Comaroff came out, 35 of the 38 original signatories ultimately retracted their signatures, acknowledging that they were “lacking full information about the case” (8). Following the retraction, an op-ed in the Harvard Crimson (9) accused the original signatories of failing to “critically engage with the source of [the description of events on which their letter was based]— the accused’s lawyer! — and its inherent bias” and argued that “the letter’s uncritical engagement with limited perspectives, voiced in such decisive and unambiguous terms, was hasty and uninformed.”

Thus, some have argued that punishment in the absence of careful consideration can have negative social consequences. From this perspective, individuals should respond to allegations of wrongdoing by carefully evaluating the evidence and considering different perspectives, and should join punitive efforts only after reaching a high threshold of certainty that punishment is deserved.

Yet social commentators have also emphasized that, when merited, condemnation and punishment play the essential social function of holding wrongdoers accountable and discouraging future wrongdoing (10, 11). Consider how contemporary social movements like #MeToo and #BlackLivesMatter have used outrage to fuel moral progress by holding perpetrators accountable for transgressions—like sexual assault and racism—that many believe have too often been committed with impunity (12–15). From this perspective, if people are too hesitant to join justified punitive efforts (e.g., because their certainty threshold is excessively high), it may be impossible to generate the momentum required to create needed change. Thus, one might see a trend towards punishment with less consideration of opposing perspectives as negative or positive for society, depending on the context and one’s orientation towards the moral issue at hand.

What forces cause people to punish without evaluating opposing perspectives? One proposal is that reputation motives (16–23) are a key contributor. In particular, some critics have alleged that we are facing an epidemic of “virtue signaling”, whereby people are increasingly
driven to showcase their moral credentials for reputational gain, and that virtue signaling can fuel potentially unmerited punishment (24). In a 2019 op-ed (25), for example, Haidt and Rose-Stockwell discuss “what happens when people use moral talk to enhance their prestige in a public forum” and argue that “nuance and truth are casualties in this competition to gain the approval of the audience…Context collapses. The speaker’s intent is ignored.” Moreover, others have argued that the influence of virtue signaling may be amplified by increasingly homogenous social networks—sometimes referred to as “echo chambers” (26, 27)—that preferentially expose us to audiences who share our ideological perspectives and are especially likely to reward our outrage (25, 28–30).

This commentary suggests that reputation may drive people to punish without careful consideration, especially when surrounded by ideologically extreme audiences. Yet despite the frequency with which critics have raised alarm, these claims have not been investigated empirically. Do people actually “punish without looking” for reputational gain?

The link between punishment and reputation

There is clear evidence for a link between punishment and reputation in general. People frequently act to punish wrongdoers (31–36)—including sometimes via collective group efforts (37–39)—and doing so can confer reputational benefits. In particular, punishment can be socially rewarded (30, 40, 41) and serve as a signal of trustworthiness (42–50) by conveying that the punisher is unlikely to themself transgress (42, 43). And consequently, reputation motives can drive people to punish. People are more likely to punish in contexts where their decisions are observable to others (51, 52), and where punishment has greater signaling value (44, 45).

Although less well empirically established, there are also good theoretical reasons to expect that the general link between punishment and reputation might be amplified by certain types of audiences. In contemporary society, collective punishment frequently engages with politically divisive moral issues (e.g., sexual assault, police brutality, racism, or abortion) and is coordinated, often via online platforms, by groups of ideologically-aligned individuals (who tend to see these issues similarly). In such contexts, it stands to reason that punishment might confer larger reputational benefits—and thus reputation might do more to encourage punishment—when the reputational audience is more ideologically convicted with respect to the moral issue at hand. Indeed, a recent study revealed that people expect more ideological audiences to react more positively to moralistic punishment (52).

Reputation as a potential driver of punishment without looking

While reputation clearly drives punishment, it is much less clear whether reputation drives people to punish without looking. In this paper, we investigate this question, and consider two distinct reasons that reputation might have this power.

First, reputation could cause people to punish without looking as a specific reputational strategy. In particular, people might expect punishment to look especially virtuous if it is enacted unquestioningly. And consequently, in contexts where people expect their decision-making process to be observable to others, reputation motives might drive people to punish specifically without looking.

Why might punishers expect to accrue particularly large reputational rewards if they choose to ignore opposing perspectives? For one, evidence suggests that rising ideological segregation and polarization has left people increasingly intolerant of those who disagree with them about politicized issues (53, 54). Such partisan climates both reflect and promote sharp
intergroup divides, and could heighten expectations of loyalty to the partisan ingroup. Therefore, when people have the opportunity to join punitive efforts in politicized domains, we might expect them to be penalized for engaging with opposing perspectives, given that such perspectives are likely to come from out-group members.

Moreover, considering opposing perspectives before punishing might be judged negatively insofar as it conveys skepticism that punishment could be merited, and therefore disloyalty towards the underlying moral cause. Punishers who eschew opposing perspectives, on the other hand, might benefit from appearing especially committed to the relevant cause, and consequently reap greater reputational benefits from their punishment—especially in the eyes of more ideologically convicted (and thus committed) audiences.

Affirming this logic, theoretical modeling reveals that individuals who cooperate without attending to the costs of doing so can be preferentially chosen as interaction partners (55). Furthermore, cooperative decisions that are faster, more intuitive, and less calculating tend to be evaluated more positively (18, 56–59) and seen as more diagnostic of an individual’s true preferences (60). And, correspondingly, subjects faced with the opportunity to cooperate make less calculating decisions when their decision process (and not just their ultimate decision) is observable to a reputational audience (56). Thus, cooperating without looking can serve as a specific reputational strategy, and we might expect punishment without looking to function similarly.

Still, punishment is more morally complex than cooperation. Whereas one is unlikely to cause harm via cooperation, punishment is unfair—and thus morally bad—when it is not deserved. Thus, in the context of punishment, individuals who consider opposing perspectives might actually enjoy the reputational benefit of seeming fair-minded, insofar as they are seen as less likely to enact unmerited punishment. Moreover, considering opposing perspectives might signal competence: thoughtfulness, deliberation, and reflection are often extolled as virtuous qualities, including in the philosophical tradition of enlightenment thinking (61), as advice for managerial decision-makers (62, 63) and by modern proponents of “rationality” (64, 65). Consequently, while punishment without looking might function as an especially strong signal of loyalty to a moral cause, it may also carry reputational liabilities: punishers who eschew opposing perspectives could appear less fair-minded or competent. And this logic might be especially likely to hold in the context of more ideologically moderate (and less convicted) reputational audiences.

Thus, while punishers might expect eschewing opposing perspectives to make them appear especially virtuous, there is also basis to doubt this proposition. Importantly, however, even if punishment without looking does not confer especially strong reputational benefits (relative to punishment with looking), reputation might still function to encourage punishment without looking.

In particular, there is a second way that reputation could drive punishment without looking: as a byproduct of its power to fuel punishment in general. If reputation generally encourages punishment, and some individuals who are reputationally-driven to punish happen to choose not to look, then reputation could give rise to elevated rates of punishment without looking. And there are several reasons that punishers might choose not to look, even if they do not expect unquestioning punishment to appear particularly virtuous.

Specifically, considering opposing perspectives can be time-consuming and effortful, and require engaging with sources (e.g., from out-group members) that one finds distasteful. These costs might discourage looking, especially for punishers who are confident that punishment is
merited (and see opposing perspectives as unlikely to change their minds). Furthermore, punishers might avoid considering opposing perspectives as a consequence of motivated reasoning: given an incentive to punish (e.g., potential reputational gain), some punishers might not wish to risk having their minds changed by countervailing arguments. For all these reasons, when reputation incentivizes punishment in general, some individuals who are induced to punish might choose to do so without looking—even if they do not expect punishment without (vs. with) looking to look especially good.

In sum, then, reputation could drive punishment without looking in two distinct ways. First, reputation could fuel punishment without looking as a specific reputational strategy, insofar as people believe that punishers who eschew opposing perspectives will appear particularly virtuous. And second, reputation could encourage punishment without looking as a byproduct of encouraging punishment in general. Here, we investigate the power of reputation to drive punishment without looking through these two channels.

Paradigm Overview

To this end, across three preregistered studies of Americans (total n = 7,952), we sought to empirically quantify and measure punishment without looking using an incentive-compatible paradigm with meaningful stakes. We invited Actor subjects to sign real punitive petitions about hot-button politicized moral issues (“punishment”), with or without first reading articles expressing perspectives opposing the petitions (“looking”). Furthermore, to incorporate reputation, we invited Evaluator subjects to decide how to allocate financial resources to Actors, and varied across conditions the information that Evaluators had about Actors’ punishment and looking decisions.

More specifically, to operationalize punishment, we gave Actors the chance to sign real punitive petitions from the website Change.org. We selected petitions that engaged with politicized moral issues, and assigned subjects who identified as Democrats vs. Republicans to engage with distinct petitions. In particular, Democrats read petitions calling for the firing of either (i) University of Central Florida professor Charles Negy, on the basis of “abhorrent racist comments he has made and continues to make on his personal Twitter account,” or (ii) Los Angeles Police Department Chief Michel Moore, on the basis of Moore’s statement that the death of George Floyd (an African-American man who was murdered by a police officer in 2020) “is on protestors and looters hands.” These petitions align with the broader #BlackLivesMatter movement, which protests incidents of police brutality—and racial injustice more generally—against Black people, and is preferentially supported by liberal (vs. conservative) Americans. Thus, we expected Democrats to anticipate that signing these petitions would be reputationally rewarded by members of their ingroup (i.e., other Democrats).

In contrast, Republican subjects read a petition calling for the removal of “Blue Lives Murder” merchandise from Amazon, which the petition characterized as “hatred merchandise”. This petition aligns with the broader #BlueLivesMatter countermovement, which advocates that individuals who kill law enforcement offers should be sentenced under hate crime statutes, and is preferentially supported by conservative (vs. liberal) Americans. Thus, we expected Republicans to anticipate that signing this petition would be reputationally rewarded by their ingroup (i.e., other Republicans). We note that we employed two different petitions for Democrats (but not Republicans) because the Negy petition closed before we finished collecting data for this paper.

To operationalize “looking,” we gave Actors the chance to consider opposing perspectives before deciding whether to sign the petitions. In particular, Actors had the chance to
read real news articles with headlines suggesting that punishment might be unmerited. For example, for the Negy petition, one article espousing an opposing perspective was titled “UCF professor behind tweets deemed racist says he is subject of ‘witch hunt’”; for the Moore petition, one countervailing article was titled “Despite criticism, LAPD Chief Michel Moore maintains support in political circles” (Moore petition); and for the Amazon petition, one opposing article was titled “Spring woman says ‘Blue Lives Murder’ shirts for sale on Amazon are hate speech. Activist disagrees.”

Finally, to operationalize reputation, we assigned Evaluator subjects to receive an endowment of money, and decide how much to share with the Actor in an economic “Dictator Game” (DG). Actors therefore faced financial incentives to be seen positively by their Evaluators, who always belonged to their political in-group (i.e., we paired Democrat Actors with Democrat Evaluators, and Republican Actors with Republican Evaluators). Critically, to manipulate reputation, we varied whether Evaluators could observe (i) nothing about Actors’ behavior, (ii) Actors’ punishment (but not looking) behavior, or (ii) Actors’ punishment and looking behavior.

In Study 1, we investigate Evaluators’ judgements of Actors, in order to shed light on the reputational consequences of punishment without looking. First, we seek to confirm: do Evaluators in our paradigm create reputational incentives for punishment in general by reacting more positively to Actors who do (vs. do not) chose to punish, as the “byproduct” hypothesis assumes? And second, we ask: do Evaluators create reputational incentives for punishment specifically without looking by reacting more positively to punishers who do not (vs. do) consider opposing perspectives, as the “specific reputational strategy” hypothesis predicts? Furthermore, to probe whether increasingly polarized or ideologically extreme reputational audiences may enhance the power of reputation to fuel punishment without looking, we also investigate how the answers to these questions depend on the Evaluator’s ideology with respect to the moral domains of the petitions.

Next, in Studies 2-3, we turn to investigating Actors’ behavior across reputation conditions, in order to shed light on whether and how reputation might drive Actors to punish without looking. First, to directly test the “byproduct” hypothesis, we ask: does placing reputational scrutiny on punishment in general (by making punishment behavior observable to Evaluators) give rise to heightened rates of punishment without looking? And second, to directly test the “specific reputational strategy” hypothesis, we ask: does additionally placing reputational scrutiny on looking (by making punishment and looking behavior observable to Evaluators) further encourage punishment without looking? Furthermore, for both questions, we consider the answers in contexts where Actors are paired with less ideological Evaluators (Study 2) vs. more ideological Evaluators (Study 3).

Results

Study 1

In Study 1, we assigned focal subjects to the role of Evaluator. In Study 1a (n = 629 Evaluators who identified as Democrats), subjects were randomly paired with Democrat Actors who either had the chance to sign the Negy (n = 308) or Moore (n = 321) petitions. In Study 1b (n = 600 Evaluators who identified as Republicans), all subjects were paired with Republican Actors who had the chance to sign the Amazon petition. We note in Study 1a, the Negy and Moore petitions produced identical patterns of results (as documented in SI Section 1.1); thus, here we report results collapsing over petition.
We measured Evaluators’ impressions of Actors on several dimensions. First, subjects received an endowment of money ($50\$) and decided how much, if anything, to share with the Actor; we take as our DV the percentage of this endowment that was shared. Second, subjects rated the positivity of their overall impression towards the Actor, and the Actor’s loyalty, fairness, and competence, on 0-100 scales. In particular, we were interested in impressions of loyalty with respect to the moral causes underlying the petitions; we thus asked subjects to rate the extent to which the Actor was a “a loyal supporter of Black Lives Matter” (in Study 1a) or a “a loyal supporter of Blue Lives Matter” (in Study 1b).

To shed light on Evaluators’ impressions of different Actor behaviors, we employed the “strategy method” from behavioral economics (66). Subjects sequentially considered a series of different Actors (who made different punishment and looking decisions), decided how much to share with each one (in the event that they were ultimately matched together), and rated each one on the above dimensions.

Specifically, subjects began by considering Actors about whom they had looking and punishment information. First, in random order, subjects evaluated Actors who did vs. did not look (described as spending an above vs. below amount of time considering opposing perspectives), before deciding to punish. Then, in corresponding order, subjects evaluated Actors who did vs. did not look, before deciding not to punish. Next, subjects considered Actors about whom they had punishment information but not looking information; in random order, they evaluated Actors who did vs. did not punish. Finally, subjects evaluated an Actor about whom they had no information. (To compute study bonuses, we paired Evaluators with Actors, and implemented the Evaluator sharing decision that corresponded to the Actor’s actual behavior and the observability condition the pair was assigned to.)

For all studies in this paper, we adhered closely to our preregistered analysis plans with some minor deviations; see SI Section 3 for a discussion of these deviations.

**Did Evaluators create reputational incentives for punishment in general?**

We begin by investigating whether Evaluators in our paradigm created reputational incentives for punishment in general, as the “byproduct” hypothesis assumes. To this end, we investigate how Evaluators reacted to Actors who did vs. did not punish, when given no information about whether the Actor chose to look (Figure 1A). Among Democrats, we find that Evaluators formed more positive overall impressions of punishers than non-punishers ($b = 24.08 [21.55, 26.61], t = 18.69, p < .001$) and also shared significantly more money with punishers ($b = 15.80 [13.79, 17.81], t = 15.44, p < .001$). Furthermore, subjects rated punishers as more loyal ($b = 36.82 [34.59, 39.05], t = 32.47, p < .001$), fair ($b = 19.23 [17.06, 21.41], t = 17.38, p < .001$), and competent ($b = 17.42 [15.37, 19.46], t = 16.71, p < .001$, $n = 629$). Similarly, Republicans rated punishers more positively overall ($b = 22.10 [18.77, 25.42], t = 13.05, p < .001$), shared more money with punishers ($b = 15.32 [12.81, 17.82], t = 12.02, p < .001$), and rated punishers as more loyal ($b = 26.91 [23.49, 30.34], t = 15.45, p < .001$), fair ($b = 19.86 [17.15, 22.58], t = 14.37, p < .001$), and competent ($b = 18.02 [15.45, 20.58], t = 13.80, p < .001$, $n = 600$). Thus, across all DVs, subjects reacted positively to punishers, demonstrating that Evaluators in our paradigm indeed created reputational incentives for Actors punish in general. Moreover, we observe similar effect sizes among Democrats and Republicans, suggesting that our paradigm created psychologically similar contexts across these groups.

We also investigate the potentially moderating role of Evaluator ideology. Do Evaluators who are more ideologically convicted (with respect to the moral domain of the petition) create
**stronger** reputational incentives for punishment? Yes: we find that Evaluators react more positively to punishers (vs. non-punishers) if they are more (vs. less) politically partisan, supportive of Black Lives Matter (among Democrats) or Blue Lives Matter (among Republicans), and supportive of the relevant petition. Illustrating this pattern, Table 1 reports how each of these dimensions of ideology interact with the contrast between punishers vs. non-punishers to predict overall positivity ratings, and Figure 1 separately plots results among subjects who identify as “strong” vs. “weak” Democrats or Republicans.

**Figure 1. Evaluations of punishment, and punishment with vs. without looking.**

A) Evaluators rated punishers more positively than non-punishers. Green violin plots illustrate estimated kernel densities of Evaluator positivity ratings for Actors who did vs. did not choose to punish, given no information about looking. B) Evaluators rated punishers who chose to look more positively than punishers who declined to look. Red violin plots illustrate kernel densities of Evaluator positivity ratings for Actors who punished as a function of whether or not they first considered opposing perspectives. Each violin plot also shows the median (white circles), mean (black triangles), interquartile range (thick lines), and upper- and lower-adjacent values (thin lines). We separately plot evaluations from Evaluators who identified as weak partisans (left plots), (ii) Evaluators who identified as strong partisans (middle plots), and (iii) all Evaluators (right plots).
The moderating role of ideology on evaluations of punishment, and punishment with vs. without looking. We consider the moderating role of three dimensions of Evaluator ideology on their evaluations (i.e., overall positivity ratings) of (i) punishers vs. non-punishers (middle column) and (ii) punishers who did vs. did not look (right column). The middle column focuses on preferences for punishment; for each ideology dimension, we report simple effects of punishment among more vs. less ideological Evaluators, and the interaction between our ideology measure and punishment. The right column focuses on preferences for punishment with (vs. without) looking; for each ideology dimension, we report simple effects of looking among more vs. less ideological Evaluators, and the interaction between our ideology measure and looking. For our first ideology dimension, a binary measure of partisanship strength, we report simple effects among Evaluators who identified as strong vs. weak partisans in a forced choice question. For our other two ideology dimensions, continuous measures of support for Black Lives Matter/Blue Lives Matter and petition support, we represent effects among more/less ideological Evaluators by reporting simple slopes at +1/-1 SD from the mean on the relevant ideology dimension (or at the scale maximum, in cases where +1 SD from the mean exceeds this value). We find that more ideological Evaluators reacted relatively more positively to punishers (vs. non-punishers) and relatively less positively to punishment with (vs. without) looking.

**Did Evaluators create reputational incentives to punish specifically without looking?**

Next, we turn to asking: did Evaluators create reputational incentives to punish *specifically without looking*, as the “specific reputational strategy” hypothesis predicts? To address this question, we investigate how Evaluators reacted to Actors who chose to punish with vs. without first considering opposing perspectives (Figure 1B).

Among Democrats, we find that Evaluators formed more positive overall impressions of punishers who *did* (vs. *did not*) look ($b = 10.46 [8.60, 12.32], t = 11.07, p < .001$). They also shared more money with punishers who looked ($b = 4.39 [2.94, 5.83], t = 5.96, p < .001$), and rated them as fairer ($b = 16.30 [14.25, 18.35], t = 15.61, p < .001$) and more competent ($b = 11.10 [9.39, 12.81], t = 12.73, p < .001$). Interestingly, however, they rated punishers who looked as less loyal supporters of Black Lives Matter ($b = -5.55 [-7.14, -3.95], t = -6.82, p < .001$).
Republicans showed the same patterns, rating punishers who looked more positively overall ($b = 4.55 [2.79, 6.30], t = 5.08, p < .001$), sharing more money with them ($b = 1.72 [.29, 3.14], t = 2.36, p = .018$) and rating them as fairer ($b = 8.62 [6.86, 10.38], t = 9.62, p < .001$) and more competent ($b = 5.20 [3.52, 6.87], t = 6.10, p < .001$). Furthermore, Republicans rated punishers who looked as less loyal supporters of Blue Lives Matter ($b = -6.90 [-8.88, -4.91], t = -6.83, p < .001$).

Thus, Evaluators in our paradigm did not react especially positively to punishers who declined to consider opposing perspectives. Interestingly, Evaluators did see punishing without looking as an especially strong signal of loyalty: they rated punishers who considered opposing perspectives as less loyal than punishers who eschewed these perspectives, highlighting a reputational disadvantage of looking. Yet looking also had distinct reputational advantages: Evaluators rated punishers who considered opposing perspectives as fairer and more competent. And, in our paradigm, the advantages of looking outweighed the disadvantages, such that Evaluators rated punishers who looked more positively overall and sent them more money. Consistent with this proposal, in SI Section 1.2 we report non-preregistered mediation analyses that reveal (i) negative indirect effects of looking on overall positivity via diminished ratings of loyalty, and (ii) countervailing positive indirect effects of looking via heightened ratings of fairness and competence, that were significantly larger in magnitude.

We also note that our design allows us to examine Evaluators’ responses to punishers who did vs. did not look in a between-subjects analysis (because Evaluators encountered these two Actor profiles first, and we randomized their order). These between-subject analyses were also preregistered, and produce results that are similar in direction and magnitude to our within-subject analyses. However, in line with the reduced statistical power they afford, they are less consistently significant; see SI Section 1.3 for details.

Our within-subject data also allow us to conduct a set of non-preregistered analyses of the preferences of individual Evaluators. For each individual, we calculate a preference score for punishment with vs. without looking (computed as the difference in evaluations of these two Actor profiles) across each of our DVs; Figure 2 plots the distribution of scores. For our overall positivity DV, 61.69% of Democrat subjects showed a preference for punishment with looking (i.e., they rated a punisher who looked more positively than a punisher who did not look), while 18.76% rated both punishers equally and 19.55% favored the punisher who did not look. Among Republicans, 48.50% of subjects showed a preference for punishment with looking, 24.33% showed no preference and 27.17% showed a preference for punishment without looking. Thus, Evaluators were indeed most likely to prefer punishers who did consider opposing perspectives. Still, some individuals showed no preference, or preferred punishers who did not look.

We also again investigate the role of Evaluator ideology. Do more ideological Evaluators react any more positively to punishers who choose to eschew opposing perspectives? We find that the answer is yes. As illustrated by Table 1 and Figures 1-2, more ideological Evaluators showed relatively weaker preferences for punishment with (vs. without) looking. Of note, however, even strong partisans did not prefer punishment without looking to punishment with looking. As reported in Table 1, strong partisan Democrats still formed significantly more positive overall impressions of punishers who did (vs. did not) look, while strong partisan Republicans showed no significant difference in their overall positivity towards punishers who did vs. did not look.
Figure 2. Relative evaluations of punishment with vs. without looking, across individuals and dependent variables. For each individual Evaluator, we compare evaluations of Actors who did vs. did not look before deciding to punish. We then plot distributions of relative evaluations across each of our five dependent variables, with positive values corresponding to a higher rating for punishment with looking, and negative values corresponding to a higher rating for punishment without looking. Evaluators who identified as weak partisans are plotted in orange, and Evaluators who identified as strong partisans are plotted in blue.

**Study 1 Synthesis**

In sum, Study 1 reveals that Evaluators in our paradigm created reputational incentives for punishment in general: they preferred punishers to non-punishers. Importantly, however, Evaluators did not create reputational incentives for punishment specifically without looking. Instead, they preferred punishers who did look before deciding to punish. The results of Study 1 are therefore compatible with the possibility that reputation could drive punishment without looking as a byproduct of general reputational incentives for punishment. Yet they cast doubt on the hypothesis that reputation drives punishment without looking as a specific reputational strategy.

The preference that Evaluators showed for punishment with looking is all the more notable in light of the fact that Evaluators preferred punishers to non-punishers. In particular, this pattern suggests that Evaluators did not simply respond positively to the consideration of opposing perspectives because they were unsure about their own attitudes towards the petitions. To the contrary, Evaluators strongly rewarded Actors who ultimately chose to sign the petitions, sending them more money and rating them more positively than their counterparts who declined to sign. And yet despite this positive general evaluation of punishment, Evaluators still conferred the largest reputational benefits on punishers who first considered countervailing arguments.
Study 1 also reveals that more (vs. less) ideological Evaluators hold relatively stronger preferences for punishment, and weaker preferences for punishment with looking. In this way, Study 1 highlights the important role that audience ideology may play in shaping reputational effects. And it suggests that while Evaluators, on average, hold preferences both for punishment and punishment with looking, there may be an inverse relationship between these preferences.

Consistent with this suggestion, we note that supplemental analyses (using our overall positivity DV) indeed show that Evaluators with stronger preferences for punishment in general have weaker preferences for punishment specifically with looking (Democrats: $r = - .26, p < .001$, Republicans: $r = - .20, p < .001$). Interestingly, however, even Evaluators who prefer punishers to non-punishers tend, on average, to prefer punishers who do (vs. do not) look (Democrats: $b = 8.08 [5.79, 10.37], t = 6.93, p < .001$; Republicans: $b = 2.42 [.07, 4.78], t = 2.02, p = .044$), underscoring the relatively robust preference Evaluators hold for punishers who first consider opposing perspectives.

Finally, we note that our Study 1 data also allow us to investigate evaluations of non-punishers who did vs. did not look. In these analyses, which we report in SI Section 1.4, we consistently find that both Democrat and Republican Evaluators react more positively to non-punishers who did (vs. did not) look, across all DVs. We thus find that Evaluators create reputational incentives for Actors to consider opposing perspectives regardless of whether or not they ultimately choose to punish. This finding further supports our conclusion that looking can have positive reputational consequences.

Studies 2 and 3

In Studies 2-3, we shift to investigating how reputation shapes punishment without looking behavior. To this end, we assigned focal subjects in Studies 2-3 to the role of Actor. And we informed all subjects that they would have the chance to punish (by signing one of our punitive petitions), could consider opposing perspectives before doing so, and would be paired with a co-partisan Evaluator who would decide how much money to share with them.

To manipulate reputation, we assigned each Actor subject to one of our three observability conditions. In Nothing Observable, subjects learned that the Evaluator would not find out whether they punished or whether they looked. In Punishment Observable, subjects learned that the Evaluator would find out whether they punished, but not whether they looked. And in Both Observable, subjects learned that the Evaluator would find out whether they punished and whether they looked. After assigning subjects to their conditions, we measured looking (which we defined in Studies 2-3 as clicking at least one link to an opposing perspective article) and punishment (which we defined as clicking the link to the petition and self-reporting having signed it), allowing us to identify subjects who punished without looking.

We use this design to ask two key questions. First, by comparing Nothing Observable to Punishment Observable, we can ask: how does placing reputational scrutiny on whether subjects punish in general (by making punishment decisions observable to Evaluators) influence rates of punishment without looking? And by answering this question, we can evaluate the hypothesis that reputation gives rise to punishment without looking as a byproduct of creating general reputational incentives for punishment. Second, by comparing Punishment Observable to Both Observable, we can ask: how does additionally placing reputational scrutiny on whether subjects look (by making looking observable to Evaluators) influence rates of punishment without looking? And by answering this question, we can evaluate the hypothesis that reputation gives rise to punishment without looking as a specific reputational strategy.
Studies 2-3 also investigated how audience ideology might shape the influence of reputation on punishment without looking. To this end, we paired Actors with relatively less ideological Evaluators in Study 2 than Study 3. In particular, in Study 2, subjects were paired with Evaluators who identified as “a weak [Democrat/Republican], who only leans towards the party”. In Study 3, however, subjects were paired with Evaluators who identified as “a strong [Democrat/Republican] who strongly supports [Black Lives Matter/Blue Lives Matter]”.

We anticipated that, by featuring a more ideological Evaluator, Study 3 might create a more fertile environment for reputation to drive punishment without looking, for two reasons. First, more ideological Evaluators might create stronger reputational incentives for punishment in general, given that more ideological Evaluators in Study 1 showed stronger preferences for punishers over non-punishers. And second, more ideological Evaluators might be more likely to encourage Actors to use punishment without looking as a specific reputational strategy, given that more ideological Evaluators in Study 1 reacted relatively more positively to punishers who declined to look (despite not showing an absolute preference for punishment without looking).

Studies 2 and 3 were otherwise almost identical, but also varied with respect to the petitions they featured and their sample sizes. In Study 2, we recruited $n = 1808$ Democrats (for Study 2a, which featured the Moore petition) and $n = 1796$ Republicans (for Study 2b, which featured the Amazon petition). In Study 3, we recruited $n = 1974$ Democrats (for Study 3a, which featured the Negy petition) and $n = 1145$ Republicans (for Study 3b, which again featured the Amazon petition). See Methods for a discussion of why our sample sizes differed across studies. We also note that our Study 3a data was collected in two batches; however, all key results held significantly within just the first batch, and all results survive corrections for “peeking” between batches. See SI Section 2.1 for more information.

**How does placing reputational scrutiny on punishment in general influence rates of punishment without looking?**

We begin by evaluating our “byproduct” hypothesis. We thus ask: does placing reputational scrutiny on whether Actors choose to punish (by making punishment observable to Evaluators) influence rates of punishment without looking? To answer, we compare our Nothing Observable and Punishment Observable conditions, which differed only with respect to whether punishment was observable (Figure 3A). We present here our results for this comparison in each of Studies 2 and 3.

**Study 2.** We start with Study 2. Before analyzing rates of punishment without looking, we first consider overall rates of punishment (independent of looking). Did making punishment observable increase punishment in general? We find suggestive evidence that the answer is yes: Among Democrats, overall rates of punishment were marginally significantly higher in Punishment Observable (23%) than Nothing Observable (19%), $b = .04 [-.01, .09], t = 1.69, p = .091, n = 1222$. Similarly, Republicans were marginally more likely to punish in Punishment Observable (27%) than Nothing Observable (22%), $b = .05 [.002, .09], t = 1.89, p = .060, n = 1214$. These results are consistent with the proposal that Actors (correctly) anticipated that Evaluators would reward them for punishing. In other words, we find suggestive evidence that Actors perceived general reputational incentives to punish, as the byproduct hypothesis assumes.

Next, we turn to directly testing the byproduct hypothesis by asking: did making punishment observable (and thus activating these general reputational incentives for punishment) increase rates of punishment without looking? Among Democrats, we find that the answer is no: rates of punishment without looking were comparable across the Punishment Observable (13%)
and Nothing Observable (12%) conditions, \( b = .01 [-.02, .05] \), \( t = 0.68 \), \( p = .497 \), \( n = 1222 \). In contrast, however, Republicans were significantly more likely to punish without looking in Punishment Observable (19%) than in Nothing Observable (14%), \( b = .05 [.01, .10], t = 2.57, p = .010, n = 1214 \).

Thus, in Study 2—which featured relatively less ideological Evaluators—we find suggestive evidence that placing reputational scrutiny on punishment in general can increase punishment overall. And critically, among Republicans but not Democrats, we find that placing reputational scrutiny on punishment can also increase punishment without looking specifically.

**Study 3.** Turning to Study 3, we find strong evidence that making punishment observable increased punishment in general. Among Democrats, overall rates of punishment were significantly higher in Punishment Observable (30%) than Nothing Observable (19%), \( b = .12 [.07, .16], t = 4.98, p < .001, n = 1319 \). Similarly, Republicans were more likely to punish in Punishment Observable (25%) than Nothing Observable (15%), \( b = .10 [.05, .16], t = 3.59, p < .001, n = 763 \).

We also find strong evidence that making punishment observable increased punishment without looking. Among Democrats, we find significantly higher rates of punishment without looking in Punishment Observable (16%) than Nothing Observable (10%), \( b = .06 [.02, .09], t = 3.13, p = .002, n = 1319 \). Similarly, Republicans were more likely to punish without looking in Punishment Observable (20%) than in Nothing Observable (12%), \( b = .08 [.03, .14], t = 3.14, p = .002, n = 763 \).

Thus, in Study 3—which featured relatively more ideological Evaluators—we find strong evidence, among both Democrats and Republicans, that placing reputational scrutiny on punishment can increase both punishment overall and punishment without looking specifically.

Together, then, we find support for the hypothesis that reputation may drive punishment without looking by incentivizing punishment in general, and consequently giving rise to increased rates of punishment without looking as a byproduct. Furthermore, we found stronger evidence for this proposal in Study 3 than in Study 2—a pattern that is consistent with our finding from Study 1 that more ideological Evaluators create stronger reputational incentives for punishment in general. However, any comparisons between Studies 2 and 3 remain merely suggestive, because (i) we did not randomly assign Actors to Study 2 vs. 3 and (ii) for Democrats, Studies 2 vs. 3 featured different petitions.
Figure 3. Making punishment observable drives punishment without looking, while making looking observable does not. A) Drawing reputational scrutiny to punishment (by making punishment observable) can drive Actors to punish more overall, and also to punish specifically without looking. For each study, we plot the proportion of subjects who chose to (i) punish and (ii) punish without looking, across the Nothing Observable vs. Punishment Observable conditions. B) Drawing reputational scrutiny to looking (by making looking observable) does not drive punishment without looking, and actually increases rates of looking overall. For each study, we plot the proportion of subjects who chose to (i) look and (ii) punish without looking, across the Punishment Observable vs. Both Observable conditions. Error bars are 95% CIs.

How does placing reputational scrutiny on looking influence rates of punishment without looking?

Next, we evaluate our “specific reputational strategy” hypothesis. We thus ask: does placing reputational scrutiny on whether subjects look (by making looking observable to Evaluators) influence rates of punishment without looking? To answer, we compare our Punishment Observable and Both Observable conditions, which differed only with respect to whether looking was observable (Figure 3B).

Study 2. We start with Study 2. Before analyzing rates of punishment without looking, we investigate overall rates of looking (independent of punishment). Did making looking observable influence rates of looking in general? We find strong evidence that the answer is yes: Among Democrats, overall rates of looking were significantly higher in Both Observable (52%)
than Punishment Observable (35%), $b = .17 \ [.11, .22], t = 5.88, p < .001, n = 1206$. Similarly, Republicans were more likely to look in Both Observable (38%) than Punishment Observable (24%), $b = .14 \ [.08, .19], t = 5.13, p < .001, n = 1197$. These results imply that Actors (correctly) anticipated that Evaluators would reward them for choosing to consider opposing perspectives (even in a context where, as reported previously, Evaluators also rewarded Actors who ultimately chose to punish). In this way, our results suggest that placing reputational scrutiny on looking is unlikely to encourage punishment without looking—and thus cast further doubt on the specific reputational strategy hypothesis.

Next, we directly test this hypothesis by asking: did making looking observable increase rates of punishment without looking? For both Democrats and Republicans, we find that the answer is no. Among Democrats, we find significantly lower rates of punishment without looking in Both Observable (9%) than in Punishment Observable (13%), $b = -.04 \ [-.08, -.01], t = -2.35, p = .019, n = 1206$. And among Republicans, we find no significant difference between rates of punishment without looking in Both Observable (16%) and Punishment Observable (19%), $b = -.03 \ [-.08, -.01], t = -1.53, p = .127, n = 1197$.

Thus, in Study 2—which featured relatively less ideological Evaluators—we find no evidence that placing reputational scrutiny on looking increases rates of punishment without looking. In fact, among both Democrats and Republicans, making looking observable increased rates of looking overall. Furthermore, making looking observable directionally decreased rates of punishment without looking, and this difference was significant among Democrats.

**Study 3.** Turning to Study 3, we again find that making looking observable encouraged looking in general. Among Democrats, overall rates of looking were significantly higher in Both Observable (49%) than in Punishment Observable (41%), $b = .08 \ [.02, .13], t = 2.72, p = .007, n = 1284$. Similarly, Republicans were more likely to look in Both Observable (28%) than Punishment Observable (15%), $b = .13 \ [.07, .19], t = 4.52, p < .001, n = 779$.

Further mirroring Study 2, Study 3 also provides no evidence that making looking observable encourages punishment without looking. Among Democrats, we find no significant difference between rates of punishment without looking in Both Observable (13%) and Punishment Observable (16%), $b = -.03 \ [-.07, .01], t = -1.41, p = .158, n = 1284$. Similarly, Republicans show no significant difference between Both Observable (19%) and Punishment Observable (20%), $b = -.01 \ [-.07, .04], t = -0.45, p = .651, n = 779$. Thus, in Study 3—which featured relatively more ideological Evaluators—placing reputational scrutiny on looking increased rates of looking and did not influence rates of punishment without looking.

Together, then, Studies 2-3 provide no support for the hypothesis that reputation drives punishment without looking as a specific reputational strategy. Making looking observable consistently encouraged Actors to consider opposing perspectives, and diminished or did not influence their propensity to punish without looking.

Furthermore, making looking observable had directionally larger positive effects on looking, and negative effects on punishment without looking, in Study 2 than in Study 3. This pattern is consistent with our Study 1 finding that more ideological Evaluators create weaker reputational incentives for punishment with (vs. without) looking. It is notable, however, that even the quite ideological Evaluators that we featured in Study 3 did not motivate Actors to use punishment without looking as a specific reputational strategy.
Studies 2-3 Synthesis

In sum, Studies 2-3 provide clear evidence that reputation can drive punishment without looking as a byproduct of general reputational incentives for punishment that arise from making punishment observable. But they provide no evidence that reputation drives punishment without looking as a specific reputational strategy in contexts where looking is also observable—even when the reputational audience is strongly ideological with respect to the relevant moral domain. We do find, however, that more (vs. less) ideological audiences may facilitate punishment without looking, both by (i) creating stronger reputational incentives for punishment in general and (ii) doing less to encourage looking and discourage punishment without looking. In all of these ways, the results of Studies 2-3 align closely with our results from Study 1.

Finally, we note that in the SI, we report how making punishment observable influenced looking (SI Section 2.2), and how making looking observable influenced punishment (SI Section 2.3), in each of Studies 2-3. These analyses were also preregistered, but are not included in our main text analyses above because they are less relevant to our key theoretical questions.

Discussion

Critics of so-called “outrage culture” allege that virtue signaling drives people to join collective punitive efforts without careful consideration. And while a growing body of work suggests that moralistic punishment can confer reputational benefits, the more specific proposal that reputation drives people to eschew opposing perspectives and unthinkingly hop on the punitive bandwagon—especially when embedded in ideologically-homogenous “echo chambers”—has remained empirically unexamined.

To test these ideas, we designed a paradigm to quantify “punishment without looking” in the context of real punitive petitions about consequential, politicized moral issues. Across three studies, we find clear evidence that people do sometimes choose to support punitive efforts without first considering opposing perspectives, and that reputation can encourage this behavior. Critically, however, we find that reputation fuels punishment without looking not as a specific reputational strategy, but rather as a byproduct of general reputational incentives for punishment.

Supporting the conclusion that reputation can drive punishment without looking by incentivizing punishment in general, subjects in our studies (i) conferred reputational benefits on punishers (Evaluators in Study 1) and (ii) were more likely to punish when their punishment behavior was observable to a reputational audience (Actors in Studies 2-3). And critically, making punishment observable to a reputational audience also increased rates of punishment specifically without looking (among Republican Actors in Study 2 and both Democrat and Republican Actors in Study 3). Thus, we find evidence that reputation can drive people to punish, and that some people who are reputationally driven to punish chose to do so without considering other perspectives. Consequently, reputation can drive punishment without looking as a byproduct.

Yet we found no evidence that reputation can drive punishment without looking as a specific reputational strategy. Interestingly, Evaluators in Study 1 did see punishing without (vs. with) looking as an especially strong signal of loyalty, supporting the proposal that considering opposing perspectives might undermine a punisher’s perceived commitment to an underlying moral cause. Critically, however, declining to look also had reputational costs. Ignoring opposing perspectives made punishers appear less fair and competent to Evaluators—a pattern of results that aligns with the perspective that punishment has to the potential to become morally bad (and
thus unfair) when it is not justified, and that thoughtfulness, deliberation, and reflection are socially-valued and competence-relevant traits. On net, then, Evaluators conferred larger reputational benefits on punishers who *did* consider opposing perspectives before punishing. And correspondingly, in Studies 2-3, making looking observable to a reputational audience *did not* encourage Actors to punish without looking. Instead, making looking observable did not influence or even *decreased* rates of punishment without looking, and reliably *increased* looking overall.

Together, then, our results suggest that reputation does *not* drive people to signal their virtue via unthinking punishment—but it *does* drive people to signal their virtue via punishment in general, and some of the resulting punishment is unthinking. In this way, our results cast doubt on the perspective that punishers ignore conflicting viewpoints as a way to curry favor with like-minded audiences. Rather, punishers who decline to look seem to simply prefer not to consider opposing perspectives.

Why might punishers—including those who are driven to punish for reputational gain—prefer not to look? One consideration may be that evaluating opposing perspectives takes time and effort, and requires engaging with dissimilar others. Punishers may prefer to avoid paying these costs. Alternatively, punishers may prefer not to consider opposing perspectives because they are motivated to conclude that punishment *is* merited (e.g., because they anticipate a reputational upside to punishment) and do not wish to have their minds changed (for related theorizing in the context of motivated “solution aversion”, see ref 60). Our studies were not designed to definitively discriminate between these two mechanisms, and doing so is a worthy subject for future research. Nevertheless, in SI Section 2.2, which reports how making punishment observable influenced rates of looking in Studies 2-3, we discuss potential implications of this analysis for the question of whether reputational incentives for punishment discourage the consideration of opposing perspectives.

Our findings have implications for how one might intervene to foster productive punishment behavior. As described in the introduction, considering opposing viewpoints before punishing may have the positive social consequence of reducing undeserved punishment. Yet as people raise their personal thresholds for joining a collective punitive effort, its probability of successfully holding wrongdoers accountable may be diminished. In contexts where we believe that the societal threshold for punishing is too low (i.e., people are too willing to punish without careful consideration), we may wish to intervene on our social environments to encourage looking. And under the theory that unthinking punishment is used as a specific reputational strategy, one might conclude that we need to (1) discourage people from caring about their reputations and/or (ii) change attitudes about the merits of considering opposing perspectives.

Our results challenge these prescriptions. We find that people (correctly) expect considering opposing perspectives to look *good*. In other words, people do not think that declining to look will appear virtuous, and expect reading articles that oppose a punitive petition to boost their reputations—even in a context where they also perceive reputational incentives to ultimately sign the petition. This finding contradicts the narrative that people have become so intolerant of those who disagree with them that one can be penalized for open-mindedness. And it suggests that, rather than merely viewing virtue signaling as a *problem*, it may be fruitful to *leverage* virtue signaling as a tool to encourage looking—for example, by giving people opportunities to broadcast their engagement with opposing perspectives.

Our results also highlight the importance of the ideological composition of audiences and/or social networks in shaping the nature of reputational influence. In Study 1, we found
evidence that more ideological audiences create strong stronger reputational incentives for punishment in general, and weaker reputational incentives for punishment with (vs. without) looking (Study 1). And correspondingly, Studies 2-3 suggest that, in the presence of more ideological audiences, (i) making punishment observable does more to encourage punishment overall and punishment without looking specifically and (ii) making looking observable does less to encourage looking and discourage punishment without looking. Still, because subjects were not randomly assigned to Study 2 vs. 3 (and among Democrats, Studies 2 vs. 3 featured different petitions), these patterns remain merely suggestive; future work should attempt to shed further light on the relationship between audience ideology and the behavioral tendency to punish without looking.

Relatedly, future work should further investigate the inverse relationship that we observed between the reputational audiences that most strongly rewarded punishment vs. looking. Interestingly, despite the fact that both punishment and looking were both, on average, reputationally rewarded by the audiences in our studies, the more a given Evaluator rewarded punishment, the less they tended to reward punishers who did (vs. did not) choose to look. For this reason, while punishing specifically without looking was not an optimal reputational strategy in the context of our studies, our results do provide a framework for thinking about when punishment without looking might be evaluated most positively. In particular, our results suggest that some audiences—perhaps those who are extremely ideological and especially likely to prize loyalty over fairness or competence—might create reputational incentives to punish specifically without consideration for countervailing perspectives. Extremist groups (e.g., hate groups) might be one such example.

Future research should also evaluate the generalizability of our results, including across cultures, moral domains, and social contexts. A more “naturalistic” investigation of punishment without looking on social media platforms would be of particular interest. It would also be valuable to consider different operationalizations of punishment and of looking. For example, one might expect the reputation value of looking to increase for more severe forms of punishment (e.g., physical aggression, or the unilateral decision to fire somebody) but to lessen for milder punishments (e.g., condemnatory gossip in a private, dyadic conversation).

Similarly, with respect to looking, there are many different ways that somebody might gather additional information before deciding whether to punish. While we focused on the consideration of opposing perspectives, future studies could also investigate adjacent looking behaviors—such as seeking out basic facts about an allegation of wrongdoing, verifying sources, asking additional questions, or evaluating additional perspectives that are not clearly countervailing (i.e., seeking out confirmatory evidence that punishment is merited).

Of note, relative to these other looking behaviors, one might expect the consideration of opposing perspectives to be evaluated relatively negatively by in-group members, given that it could be seen as particularly disloyal or undermining of a perceived commitment to punishing. It is thus perhaps especially striking that looking before punishing did not, on net, have reputational costs in our studies. That being said, the amount of looking might matter too. For example, if an individual considers many opposing perspectives, or demands what is seen as an unreasonable amount of evidence that punishment is actually merited, we might expect their looking behavior to seem so disloyal that it becomes a net reputational negative. Future research should investigate this interesting possibility.

In conclusion, humans have a taste for moralistic punishment and sometimes join collective punitive efforts without considering opposing perspectives. Our results suggest that
reputation motives contribute to such “punishment without looking” by incentivizing people to punish in general. Critically, however, our results fail to support the hypothesis that people use unthinking punishment as a specific virtue signaling strategy, and instead highlight that engaging with opposing viewpoints can have reputational advantages.

**Materials and Methods**

**Studies 1a-b**

Here, we provide more details about our Materials and Methods; we also note that all data, scripts, experimental materials, and pre-registrations are available on the OSF (https://osf.io/3es2k/?view_only=e272c31aa7304eca841f084d582187fb), and full experimental stimuli for all studies are documented in SI Section 4.

We recruited a target \( n = 600 \) Evaluator subjects from Prolific Academic for each of Studies 1a (which recruited Democrats; final \( n = 629, \) 46% male, average age = 34 years) and 1b (which recruited Republicans; final \( n = 600, \) 45% male, average age = 37 years). As in all studies in this paper, subjects began by reporting their gender and political party identification, and then completing two attention checks; following our preregistrations, we did not analyze data from subjects who failed these checks.

Next, we introduced subjects to the Dictator Game involving the Actor. We told subjects that they (i.e., the Evaluator) would be “Player 1” and another subject (i.e., the Actor), of their same political party affiliation, would be “Player 2”. Furthermore, subjects learned that they would receive 50 cents, and decide how much, if anything, to share with Player 2. Subjects then answered three comprehension questions about the Dictator Game structure. For all studies in this paper, our main text analyses include all subjects, regardless of comprehension; see SI Section 1.5 for preregistered secondary analyses that restrict to subjects who correctly answered all questions (and produce very similar results).

Next, we introduced subjects to the Actor’s punishment and looking decisions. We told subjects that Player 2 had a chance to sign the relevant punitive petition, and then provided subjects with a brief description of the petition’s aims and a screenshot of the petition. Next, we told subjects that Player 2 had the chance to consider opposing perspectives before deciding whether to sign the petition. In particular, we told subjects that Player 2 received links to specific articles providing opposing perspectives (and provided a screenshot of an example headline) and also that Player 2 could search the Internet for other opposing perspectives. We also told subjects that Player 2 could choose not to consider opposing perspectives, and that Player 2s varied in the amount of time they spent considering opposing perspectives.

Next, we collected subjects’ evaluations of different Actors. To begin, we asked subjects to consider a Player 2 who chose to sign the petition, and either spent an above- or below-average amount of time (manipulated between subjects) considering opposing perspectives before doing so. In other words, we described a punisher who either did or did not look. We told subjects that we would match Player 1s with Player 2s at the end of the Study, and that they might be matched with this particular Player 2. Subjects then decided how much money to share with this Player 2, in the event that they were ultimately paired together (between 0 and 50 cents in 5-cent increments). Afterwards, they used 0 to 100 sliding scales to rate the Player 2 on overall positivity, and then, in random order, loyalty towards Black Lives Matter or Blue Lives Matter, competence, and fairness.

Next, we told subjects that on subsequent pages, they would make sharing decisions about other potential Player 2s with whom they might be paired. Then, across six pages, subjects...
evaluated six other Player 2s (with six other behavioral profiles). The second Player 2 was a punisher who did or did not look (whichever the subject did not evaluate first). Next, subjects evaluated non-punishers who did, or did not, look (in an order that corresponded to the order in which they evaluated punishers). Subjects then evaluated non-punishers and punishers, without receiving any information about their looking behavior (in random order). Finally, subjects evaluated a Player 2 without receiving any information about their decisions; evaluations of this Player 2 were not analyzed but rather used to compute bonuses.

Finally, we presented subjects with a post-experimental survey in which they provided their own evaluation of the petition, described how they made their study decisions, provided thoughts about their impression of the study, rated their previous participation in related studies and beliefs about whether the petition and Player 2 were real, and answered a set of demographic and survey questions.

Studies 2a-b

For Study 2a, we recruited a target \( n = 1800 \) Democrat Actor subjects from MTurk (final \( n = 1808, 39\% \) male, average age = 38 years). For Study 2b, we recruited a target \( n = 1800 \) Republicans Actor subjects, first from MTurk and then from Prolific Academic, once it became clear that the number of Republicans on MTurk who remained naïve to our paradigm was insufficient to achieve our full preregistered sample (final \( n = 1796, 52\% \) from MTurk and 48% from Prolific, 46% male, average age = 40 years).

After subjects reported their genders and political party identifications and completed two attention checks, we introduced them to the Dictator Game involving the Evaluator. We told subjects that they (i.e., the Actor) would be Player 2 and another subject (i.e., the Evaluator) of their same political party affiliation would be Player 1, receive 50 cents, and decide how much to share with them. Subjects then answered three comprehension questions about the Dictator Game structure. If any questions were initially answered incorrectly, subjects were required to re-answer the questions until all were correct.

Next, we introduced subjects to their punishment and looking decisions. We began by presenting subjects with a brief description, and screenshot, of the relevant petition, and telling them that they would later have the chance to sign it. Subjects in Nothing Observable learned that Player 1 would not find out whether they chose to sign; subjects in Punishment Observable and Both Observable learned that Player 1 would find this out.

Afterwards, we told subjects that they would have the chance to consider opposing perspectives before deciding whether to sign. We explained that we would provide links to specific opposing perspective articles (and provided a screenshot of an example headline) and told subjects that they could also search the Internet for other opposing perspectives, or choose not to consider any opposing perspectives. Subjects in Nothing Observable and Punishment Observable learned that Player 1 would not find out how long they spent considering opposing perspectives; subjects in Both Observable learn that Player 1 would find this out. Subjects then answered three comprehension questions about what Player 1 would vs. would not learn about their punishment and looking decisions (and again were required to correct any initially incorrect answers).

We next told subjects that on the subsequent screen, they would have the chance to consider opposing perspectives, and reminded them about whether their time spent considering such perspectives would be observable to Player 1. Then, subjects advanced to a screen in which we provided headlines for and links to two articles providing opposing perspectives, and
reminded subjects than they could search for other articles. We told subjects to advance the screen whenever they wished, and were ready to decide whether or not to sign the petition.

When subjects were on this “looking” screen, we tracked their link-clicking behavior and, per our preregistration, defined “looking” as clicking at least one link to an opposing perspective article. Importantly, we did not tell subjects that their link-clicking would be tracked, in order to keep this looking measure more “honest”. See SI Section 2.5 for secondary analyses of Studies 2-3 that define looking as either (i) time spent on the “looking” screen or (ii) the continuous number of opposing perspective article links clicked. These alternative analyses were preregistered secondary analyses for Study 3, and produce very similar results.

We next told subjects that on the subsequent screen, they would decide whether to sign the petition. We also told subjects that we would ask them, if they signed the petition, to show us that they actually signed by reporting information about the screen that Change.org displayed to them after signing. (We also assured subjects that we would not collect their identifying information if they chose to sign). Then, on the subsequent screen, we gave subjects a link to the petition and asked them self-report whether or not they signed (and report information about the completion screen if so). We also asked subjects to report their continuous commitment to supporting the petition; per our preregistrations, we do not analyze this DV.

When subjects were on this “petition” screen, we tracked their link-clicking behavior and, per our preregistration, defined “signing” (i.e., punishing) as clicking the link to the petition and self-reporting signing. Again, we did not tell subjects that their link-clicking would be tracked. We chose not to incorporate the reported completion screen information into our preregistered definition of “signing”, given that (i) pilot testing revealed that almost everyone who met the above criteria did report this information correctly, but (ii) there is some room for ambiguity in categorizing subjects’ reports as “correct” (given that some responses were close to but not completely accurate).

Finally, we presented subjects with a post-experimental survey in which they reported their own personal moral support for the petition, the ways they spent their time on the “looking” screen, their previous familiarity with the petition, the ways that they made their choices in the study, the extent to which they were motivated by reputation, their beliefs about whether the petition, articles, and Player 1 were real, and a set of demographic questions.

**Studies 3a-b**

For each of Studies 3a-b, we preregistered and recruited an initial target $n = 1200$ Actor subjects from MTurk. Thus, we set smaller initial target sample sizes for Study 3 ($n = 1200$) than Study 2 ($n = 1800$). This decision reflects that, in Study 2 but not Study 3, we hypothesized that our “Both Observable” condition might significantly decrease rates of punishment without looking, relative to our “Punishment Observable” condition. However, we anticipated that this effect would be quite small if it existed; thus, we wanted a relatively larger sample size in Study 2 to provide power to detect such a decrease.

For Study 3a (which recruited Democrats), upon reaching our initial target sample of $n = 1200$, all of our key results were significant. However, while finishing data collection for Study 3b (which recruited Republicans, who are under-represented on MTurk), we decided to direct Democrats to Study 3a, increasing its sample size. After making this decision, we registered an amendment to our Study 3a preregistration, increasing our target to $n = 2000$ subjects. We also preregistered a plan to correct, in our analyses, for the fact that we “peeked” at first batch of data before deciding to collect a second batch; all significant Study 3a results survive these
corrections. See SI Section 2.1 for analyses of Study 3a by batch, and results across both batches that correct for peeking. Our final sample sizes were $n = 1974$ for Study 3a (47% male, average age = 40 years) and $n = 1145$ for Study 3b (44% male, average age = 43 years).

The method for Studies 3a-b was otherwise identical to the method for Studies 2a-b, except that (i) we described the Evaluator as more ideological in Studies 3a-b than in Studies 2a-b and (ii) Study 3a featured the Negy petition while Study 2a featured the Moore petition.

Acknowledgements
We thank Matthew Higgins and Emmet Sullivan for helpful research assistance.

References
2. J. Ronson, So you’ve been publicly shamed (Riverhead Books, 2016).
13. A. Romano, The second wave of “cancel culture”: How the concept has evolved to mean different things to different people. Vox (2021) (April 25, 2022).


47. Y. Horita, Punishers may be chosen as providers but not as recipients. *Letters on Evolutionary Behavioral Science* 1, 6–9 (2010).


52. J. J. Jordan, N. S. Kteily, People punish moral transgressions for reputational gain, even when they personally question whether punishment is merited. *Available at PsyArXiv* (2022).


59. V. Capraro, J. Kuilder, To know or not to know? Looking at payoffs signals selfish behavior, but it does not actually mean so. *Journal of Behavioral and Experimental Economics* 65, 79–84 (2016).


Supplementary Information

for

Punishing without looking for reputational gain

1. Supplemental analyses of Studies 1a-b ................................................................. 2
   1.1 Analyses of results for the Negy vs. Moore petitions in Study 1a ..................... 2
   1.2 Mediation analyses .......................................................................................... 2
   1.3 Between-subjects analyses .............................................................................. 3
   1.4 Analyses of evaluations of non-punishers who did vs. did not look ................. 4
   1.5 Analyses of perfect comprehenders .................................................................. 4
       1.5.1 Evaluations of punishers vs. non punishers ........................................... 5
       1.5.2 Evaluations of punishers who did vs. did not look ................................. 5

2. Supplemental Analyses of Studies 2-3 ............................................................... 5
   2.1 Analyses of Study 3a by batch, and corrections for “peeking”.......................... 5
   2.2 Analyses of how making punishment observable influences looking ............ 6
   2.3 Analyses of how making looking observable influences punishment ............. 7
   2.4 Analyses of perfect comprehenders .................................................................. 7
       2.4.1 Effects of making punishment observable .............................................. 8
       2.4.2 Effects of making looking observable ..................................................... 8
   2.5 Analyses of alternative specifications of looking ............................................. 9
       2.5.1 Looking time (natural-log transformed seconds) ..................................... 9
       2.5.2 Number of articles ................................................................................... 10

3. Discussion of preregistration deviations ............................................................ 10
   3.1 Studies 1a-b .................................................................................................... 10
   3.2 Studies 2-3 ..................................................................................................... 11

4. Appendix: Experimental stimuli ........................................................................... 12
   4.1 Studies 1a-b .................................................................................................... 12
   4.2 Studies 2-3 ..................................................................................................... 26
1. Supplemental analyses of Studies 1a-b

1.1 Analyses of results for the Negy vs. Moore petitions in Study 1a

In Study 1a, subjects were randomly paired with Actors who either had the chance to sign the Negy \( (n = 308) \) or Moore \( (n = 321) \) petition. As reported in the main text, the Negy and Moore petitions produced identical patterns of results; thus, our main text results collapsed over petition. However, per our pre-registration, in Table S1 below, we report our key Study 1 analyses separately for each petition, and also report interaction results that compare results across the two petitions. We find that each of our key results hold significantly within each of the two petitions. However, for some results, the effect size differs significantly across petitions, as reflected by a significant interaction.

<table>
<thead>
<tr>
<th>Dependent Variable</th>
<th>Evaluations of punishers vs. non-punishers (Coefficients reflect preference for punishment)</th>
<th>Evaluations of punishment with vs. without looking (Coefficients reflect preference for punishment with looking)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall positivity</td>
<td>Negy Petition: ( b = 26.81 ) [23.17, 30.45], ( t = 14.51, p &lt; .001 ) &lt;br&gt;Moore Petition: ( b = 21.45 ) [17.93, 24.98], ( t = 11.99, p &lt; .001 ) &lt;br&gt;Interaction: ( b = -5.36 ) [-10.41, -1.31], ( t = -2.08, p = .038 )</td>
<td>Negy Petition: ( b = 8.56 ) [5.87, 11.25], ( t = 6.26, p &lt; .001 ) &lt;br&gt;Moore Petition: ( b = 12.29 ) [9.72, 14.85], ( t = 9.44, p &lt; .001 ) &lt;br&gt;Interaction: ( b = 3.73 ) [0.73, 6.74], ( t = 1.98, p = .048 )</td>
</tr>
<tr>
<td>Money shared</td>
<td>Negy Petition: ( b = 17.99 ) [14.96, 21.02], ( t = 11.68, p &lt; .001 ) &lt;br&gt;Moore Petition: ( b = 13.71 ) [11.05, 16.38], ( t = 10.17, p &lt; .001 ) &lt;br&gt;Interaction: ( b = -4.28 ) [-8.30, -0.26], ( t = 2.09, p = .037 )</td>
<td>Negy Petition: ( b = 3.70 ) [1.42, 5.98], ( t = 3.19, p = .002 ) &lt;br&gt;Moore Petition: ( b = 5.05 ) [3.24, 6.85], ( t = 5.49, p &lt; .001 ) &lt;br&gt;Interaction: ( b = 1.35 ) [-1.56, 4.25], ( t = 1.91, p = .363 )</td>
</tr>
<tr>
<td>Fair</td>
<td>Negy Petition: ( b = 21.60 ) [18.39, 24.82], ( t = 13.22, p &lt; .001 ) &lt;br&gt;Moore Petition: ( b = 16.96 ) [14.02, 19.89], ( t = 11.38, p &lt; .001 ) &lt;br&gt;Interaction: ( b = 4.65 ) [-8.99, -0.31], ( t = -2.10, p = .036 )</td>
<td>Negy Petition: ( b = 15.73 ) [12.97, 18.50], ( t = 11.19, p &lt; .001 ) &lt;br&gt;Moore Petition: ( b = 16.84 ) [13.81, 19.87], ( t = 10.94, p &lt; .001 ) &lt;br&gt;Interaction: ( b = 1.11 ) [-2.98, 5.20], ( t = 1.53, p = .124 )</td>
</tr>
<tr>
<td>Competent</td>
<td>Negy Petition: ( b = 19.62 ) [16.64, 22.60], ( t = 12.95, p &lt; .001 ) &lt;br&gt;Moore Petition: ( b = 15.30 ) [12.49, 18.11], ( t = 10.71, p &lt; .001 ) &lt;br&gt;Interaction: ( b = 4.31 ) [-8.40, -2.22], ( t = 2.07, p = .039 )</td>
<td>Negy Petition: ( b = 10.18 ) [7.62, 12.74], ( t = 7.82, p &lt; .001 ) &lt;br&gt;Moore Petition: ( b = 11.99 ) [9.69, 14.28], ( t = 10.27, p &lt; .001 ) &lt;br&gt;Interaction: ( b = 1.81 ) [-1.62, 5.24], ( t = 1.04, p = .300 )</td>
</tr>
<tr>
<td>Loyal</td>
<td>Negy Petition: ( b = 36.39 ) [33.12, 39.66], ( t = 21.65, p &lt; .001 ) &lt;br&gt;Moore Petition: ( b = 37.24 ) [34.19, 40.28], ( t = 24.05, p &lt; .001 ) &lt;br&gt;Interaction: ( b = 0.85 ) [-3.61, 5.31], ( t = 0.37, p = .708 )</td>
<td>Negy Petition: ( b = 4.69 ) [-7.00, -2.37], ( t = -3.99, p &lt; .001 ) &lt;br&gt;Moore Petition: ( b = 6.37 ) [-8.59, -4.16], ( t = -5.66, p &lt; .001 ) &lt;br&gt;Interaction: ( b = -1.69 ) [-4.88, 1.51], ( t = -1.04, p = .300 )</td>
</tr>
</tbody>
</table>

Table S1. Results of Study 1a by petition.

1.2 Mediation analyses

As described in the main text, Evaluators in Studies 1a-b rated punishers who did (vs. did not) consider opposing perspectives more positively overall, and also rated them as more fair and more competent. Interestingly, however, they rated punishers who looked as less loyal. This pattern of results suggests that (i) looking has reputational advantages and disadvantages, and (ii) the disadvantages are outweighed by the advantages, such that looking is a net reputational positive. Here, we support this proposal with mediation analyses.

Before describing our mediation analyses, we note that two of our mediators, fairness and competence, were highly correlated with each other (among Democrats, \( r = .86, p < .001 \); among Republicans, \( r = .87, p < .001 \)). In contrast, loyalty was less strongly correlated with the other mediators (correlations with fairness: Democrats, \( r = .63, p < .001 \); Republicans, \( r = .51, p < .001 \); correlations with competence: Democrats, \( r = .59, p < .001 \); Republicans, \( r = .49, p < .001 \)).

In light of this correlation structure, we consider two distinct multiple mediation models: one that takes fairness and loyalty as mediators, and one that takes competence and loyalty as mediators. This approach avoids placing fairness and competence together in a single model, which would be uninformative in light of their high collinearity. In each model, we take overall positivity as our dependent variable, the contrast between punishers who did vs. did not look as our independent variable, and the relevant pair of mediators as our mediating variables.
In Table S2, for each of Studies 1a-b, we report the total effect of looking on overall positivity, as well as analyses from each of our two mediation models. In particular, for each mediator within each model, we report (i) the A path (i.e., the effect of looking on the mediator), (ii) the B path (i.e., the effect of the mediator on overall positivity, controlling for looking and the other mediator), and (iii) the indirect effect of looking on overall positivity via the mediator. Additionally, for each model, we report the direct effect of looking on overall positivity.

For each of Studies 1a-b, we consistently observe significant positive indirect effects for fairness and competence, and significant negative indirect effects for loyalty. Furthermore, the negative indirect effects of loyalty are consistently significantly smaller in magnitude than the positive indirect effects of fairness and competence (as revealed by the 95% CIs reported in Table S1), consistent with the positive total effect of looking on overall positivity.

### Table S2: Mediation analyses of Studies 1a-b.

<table>
<thead>
<tr>
<th></th>
<th>Democrats (Study 1a) n = 629</th>
<th>Republicans (Study 1b) n = 600</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total Effect</td>
<td>b = 10.46 [8.60, 12.32], t = 11.07, p &lt; .001</td>
<td>b = 4.55 [2.79, 6.30], t = 5.08, p &lt; .001</td>
</tr>
<tr>
<td>Effects for Fair</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Path B: b = .75 [0.70, 0.81], z = 27.59, p &lt; .001</td>
<td>Path B: b = .78 [0.74, 0.82], z = 36.77, p &lt; .001</td>
<td></td>
</tr>
<tr>
<td>Indirect effect: b = 12.27 [10.50, 14.05]</td>
<td>Indirect effect: b = 6.74 [5.32, 8.16]</td>
<td></td>
</tr>
<tr>
<td>Effects for Loyal</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Path A: b = -5.55 [-7.14, -3.95], z = -6.83, p &lt; .001</td>
<td>Path A: b = -6.90 [-8.88, -4.92], z = -6.83, p &lt; .001</td>
<td></td>
</tr>
<tr>
<td>Path B: b = .12 [0.06, 0.17], z = 4.36, p &lt; .001</td>
<td>Path B: b = .08 [0.04, 0.12], z = 4.35, p &lt; .001</td>
<td></td>
</tr>
<tr>
<td>Indirect effect: b = -.65 [-1.00, -0.30]</td>
<td>Indirect effect: b = -.56 [-.86, -.26]</td>
<td></td>
</tr>
<tr>
<td>Direct effect</td>
<td>b = -1.16 [-2.67, .35], z = -1.51, p = .132</td>
<td>b = -1.63 [-2.98, -.28], z = -2.37, p = .018</td>
</tr>
<tr>
<td>Effects for Competent</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Path A: b = 11.10 [9.39, 12.81], z = 12.74, p &lt; .001</td>
<td>Path A: b = 5.20 [3.53, 6.86], z = 6.10, p &lt; .001</td>
<td></td>
</tr>
<tr>
<td>Path B: b = .78 [0.73, 0.83], z = 30.40, p &lt; .001</td>
<td>Path B: b = .78 [0.73, 0.84], z = 28.86, p &lt; .001</td>
<td></td>
</tr>
<tr>
<td>Indirect effect: b = 8.63 [7.19, 10.07]</td>
<td>Indirect effect: b = 4.07 [2.73, 5.41]</td>
<td></td>
</tr>
<tr>
<td>Effects for Loyal</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Path A: b = -5.55 [-7.14, -3.95], z = -6.83, p &lt; .001</td>
<td>Path A: b = -6.90 [-8.88, -4.92], z = -6.83, p &lt; .001</td>
<td></td>
</tr>
<tr>
<td>Path B: b = .13 [0.07, 0.18], z = 4.43, p &lt; .001</td>
<td>Path B: b = .06 [0.02, 0.10], z = 3.18, p &lt; .001</td>
<td></td>
</tr>
<tr>
<td>Indirect effect: b = -.70 [-1.07, -.33]</td>
<td>Indirect effect: b = -.41 [-.70, -.13]</td>
<td></td>
</tr>
<tr>
<td>Direct effect</td>
<td>b = 2.53 [1.15, 3.91], z = 3.58, p &lt; .001</td>
<td>b = .89 [-.41, 2.19], z = 1.34, p = .180</td>
</tr>
</tbody>
</table>

1.3 Between-subjects analyses

As described in the main text, the designs of Studies 1a-b allow us to examine Evaluators’ responses to punishers who did vs. did not look in a between-subjects analysis (because Evaluators encountered these two Actor profiles first, and we randomized their order). These between-subject analyses were also pre-registered, and produce results that are similar in direction and magnitude to the within-subject analyses we report in our main text. However, in
line with the reduced statistical power they afford, they are less consistently significant. Below, we report the between-subject results for each of our dependent variables.

Among Democrats (n = 629), in our between-subjects analyses, Evaluators formed more positive overall impressions of punishers who did (vs. did not) look (b = 9.92 [6.49, 13.35], t = 5.68, p < .001). They also shared directionally more money with punishers who looked, although this difference was not significant (b = 3.44 [-.77, 7.65], t = 1.60, p = .109). And they rated punishers who looked as significantly fairer (b = 15.83 [12.28, 19.38], t = 8.76, p < .001) and more competent (b = 10.32 [7.01, 13.63], t = 6.12, p < .001). In contrast, however, they rated punishers who looked as less loyal supporters of Black Lives Matter (b = -7.44 [-10.56, -4.32], t = -4.68, p < .001).

Among Republicans (n = 600), our between-subjects analyses reveal no significant differences between evaluations of punishers who did (vs. did not) look with respect to overall positivity ratings (b = 2.08 [-1.67, 5.84], t = 1.09, p = .276), money shared (b = 2.49 [-1.82, 6.80], t = 1.14, p = .257) or ratings of competence (b = 1.76 [-1.82, 5.33], t = 0.96, p = .335). Punishers who looked were, however, rated as significantly fairer (b = 4.32 [7.0, 7.94], t = 2.34, p = .020) but also significantly less loyal supporters of Blue Lives Matter (b = -10.83 [-15.66, -5.99], t = -4.40, p < .001).

1.4 Analyses of evaluations of non-punishers who did vs. did not look

Our main text focuses on how Evaluators in Studies 1a-b evaluated punishers who did vs. did not look. However, as described in the main text, the designs of Studies 1a-b also allow us to investigate evaluations of non-punishers who did vs. did not look. Here, we report analyses of these evaluations.

Among Democrats (n = 629), we find that Evaluators formed more positive overall impressions of non-punishers who did look (b = 13.31 [11.65, 14.97], t = 15.78, p < .001), shared more money with them (b = 7.74 [6.36, 9.12], t = 11.02, p < .001), and rated them as significantly fairer (b = 17.66 [15.78, 19.53], t = 18.52, p < .001), more competent (b = 14.38 [12.63, 16.14], t = 16.13, p < .001), and more loyal (b = 7.19 [5.75, 8.63], t = 9.81, p < .001). Similarly, Republicans (n = 600) rated non-punishers who looked as more positive overall (b = 9.22 [7.30, 11.14], t = 9.43, p < .001), shared more money with them (b = 4.82 [3.34, 6.29], t = 6.40, p < .001), and rated them as well as fairer (b = 12.59 [10.48, 14.69], t = 11.74, p < .001), more competent (b = 10.14 [8.23, 12.05], t = 10.44, p < .001), and marginally more loyal (b = 1.95 [-1.14, 4.04], t = 1.83, p = .068).

These analyses demonstrate that Evaluators created reputational incentives for Actors to consider opposing perspectives, even if they did not ultimately choose to punish. In this way, they serve to bolster our conclusion that looking can have positive reputational consequences.

1.5 Analyses of perfect comprehenders

As reported in the main text, for all studies in this paper, our main text analyses include all subjects, regardless of performance on comprehension questions. However, per our pre-registration, here we report secondary analyses that restrict to subjects who correctly answered all comprehension questions (and produce very similar results).
1.5.1 Evaluations of punishers vs. non punishers

We begin by reporting analyses of how Evaluators who showed perfect comprehension reacted to Actors who did vs. did not punish, when given no information about whether the Actor chose to look.

Among Democrats \((n = 533)\), Evaluators formed more positive overall impressions of punishers than non-punishers \((b = 23.97 \ [21.25, 26.69], t = 17.31, p < .001)\) and also shared significantly more money with punishers \((b = 15.42 \ [13.25, 17.59], t = 13.95, p < .001)\). Furthermore, subjects rated punishers as more loyal \((b = 37.80 \ [35.40, 40.19], t = 30.96, p < .001)\), fair \((b = 18.92 \ [16.55, 21.29], t = 15.69, p < .001)\), and competent \((b = 17.32 \ [15.10, 19.54], t = 15.34, p < .001)\).

1.5.2 Evaluations of punishers who did vs. did not look

Next, we report analyses of how Evaluators who showed perfect comprehension reacted to Actors who chose to punish, with vs. without first considering opposing perspectives.

Among Democrats \((n = 533)\), we find that Evaluators formed more positive overall impressions of punishers who did \((vs. \ did \ not)\) look \((b = 10.75 \ [8.75, 12.75], t = 10.55, p < .001)\). They also shared more money with punishers who looked \((b = 4.60 \ [3.11, 6.09], t = 6.06, p < .001)\), and rated them as fairer \((b = 16.71 \ [14.48, 18.95], t = 14.69, p < .001)\) and more competent \((b = 11.59 \ [9.72, 13.46], t = 12.16, p < .001)\), but less loyal \((b = -6.15 \ [-7.89, -4.42], t = -6.98, p < .001)\). Similarly, Republicans \((n = 494)\) rated punishers who looked more positively overall \((b = 4.34 \ [2.34, 6.32], t = 4.30, p < .001)\), shared more money with them \((b = 1.21 \ [-.39, 2.82], t = 1.48, p = .138)\) and rated them as fairer \((b = 9.04 \ [7.08, 10.99], t = 9.10, p < .001)\) and more competent \((b = 4.93 \ [3.12, 6.75], t = 5.33, p < .001)\), but less loyal \((b = -7.47 \ [-9.72, -5.21], t = -6.51, p < .001)\).

2. Supplemental Analyses of Studies 2-3

2.1 Analyses of Study 3a by batch, and corrections for “peeking”

As described in the main text, for Study 3a, we pre-registered and recruited an initial target \(n = 1200\) subjects. Upon reaching this initial target, all of our key results were significant. However, while finishing data collection for Study 3b (which recruited Republicans, who are under-represented on MTurk), we decided to direct Democrats to Study 3a, increasing its sample size. After making this decision, we registered an amendment to our Study 3a pre-registration, increasing our target to \(n = 2000\) subjects. In the main text, we thus report analyses of Study 3a that combine data from both batches of data collection. However, in Table S3, we report our Study 3a results within each individual batch (and also re-report the overall results from both batches combined, to facilitate comparison).

We also pre-registered a plan to correct, in our combined analyses of both batches, for the fact that we “peeked” at first batch of data before deciding to collect a second batch (using the approach of Sagarin, Ambler, & Lee, 2014, Perspectives on Psychological Science). In particular, we planned, for each key result that is significant in our final sample, to report
whether the result continues to be significant when accounting for peeking by computing an adjusted alpha threshold that allows us to maintain an actual type-I error rate of .05.

More precisely, instead of computing a single adjusted alpha threshold, we planned to report an adjusted alpha range. This reflects that the required alpha threshold depends on the maximum p-value observed in the initial batch of data collection for which we would have collected more data rather than declaring the initial results non-significant; this could range from a “best-case scenario” of the p-value observed after the initial data collection to a “worst-case scenario” of 1.

Thus, for each result in Table S3, we report a best- and worst-case scenario for the adjusted alpha threshold. Importantly, for each of the three results that are significant in the overall sample, we find that the p-values are smaller than even the worst-case scenario adjusted alpha threshold. Thus, we find that all significant Study 3a results remain significant, even after accounting for peeking.

<p>| Effect of making punishment observable on... |  | Effect of making looking observable on... |  |</p>
<table>
<thead>
<tr>
<th>Punishment overall</th>
<th>Punishment without looking</th>
<th>Looking overall</th>
<th>Punishment without looking</th>
</tr>
</thead>
<tbody>
<tr>
<td>Batch 1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(n = 1222)</td>
<td>b = .09 [.03, .15]</td>
<td>b = .05 [.002, .09]</td>
<td>b = .12 [.06, .19]</td>
</tr>
<tr>
<td>(n = 752)</td>
<td>t = 3.05, p &lt; .000, n = 615</td>
<td>t = 2.05, p = .040, n = 615</td>
<td>t = 3.56, p &lt; .001, n = 797</td>
</tr>
<tr>
<td>Overall</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(n = 1974)</td>
<td>b = .16 [.09, .24]</td>
<td>b = .07 [.01, .13]</td>
<td>b = .01 [.00, .08]</td>
</tr>
<tr>
<td>Adjusted Alpha Threshold:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Best case</td>
<td>0.049998</td>
<td>0.031093</td>
<td>0.031099</td>
</tr>
<tr>
<td>Worst case</td>
<td>0.353, n = 487</td>
<td>0.279, n = 415</td>
<td>0.686, n = 1222</td>
</tr>
</tbody>
</table>

Table S3. Analyses of Study 3a by batch, and adjusted Alpha Thresholds that account for “peeking” at data between batches.

### 2.2 Analyses of how making punishment observable influences looking

In our main text analyses of Studies 2-3, when considering the effects of making punishment observable (i.e., when comparing our “Nothing Observable” and “Punishment Observable” conditions), we focus on rates of punishment overall, and punishment without looking, as dependent variables. However, our design also allows us to ask how making punishment observable influences overall rates of looking. Indeed, we pre-registered analyses of this question, but do not report them in the main text because they are less relevant to our key theoretical questions. Here, however, we report these analyses.

In Study 2, among Democrats, rates of looking were similar in Punishment Observable (35%) and Nothing Observable (34%), b = .01 [-.04, .06], t = .40, p = .686, n = 1222. Republicans were also similarly likely to look in Punishment Observable (24%) and Nothing Observable (28%), b = -.04 [-.09, .01], t = -1.45, p = .148, n = 1214.

In Study 3, among Democrats, rates of looking were significantly higher in Punishment Observable (41%) than in Nothing Observable (34%), b = .07 [.02, .12], t = 2.73, p = .006, n = 1319. In contrast, Republicans were similarly likely to look in Punishment Observable (15%) and Nothing Observable (18%), b = -.03 [-.08, .02], t = -1.08, p = .279, n = 763.

Thus, we mostly find that making punishment observable has no reliable effect on rates of looking. However, in one case (Democrats in Study 3), making punishment observable increased looking.

In the main text, we note that the effect of making punishment observable on looking may help shed light on the question of whether potential punishers sometimes decline to look because they (i) anticipate a reputational upside to punishing and (ii) therefore wish to avoid...
being persuaded against punishing by opposing perspectives. At face value, the above-reported analyses are inconsistent with this suggestion: creating reputational incentives to punish (by making punishment observable) did not make subjects in our studies significantly less likely to engage with opposing perspectives.

That being said, the above-described logic assumes that potential punishers, before engaging with opposing perspectives, expect such engagement to, with some meaningful probability, decrease their propensity to punish (i.e., they expect the opposing perspectives to be potentially persuasive). If potential punishers do not hold this expectation, a motivation to support punishment should not deter the consideration of opposing perspectives. And if potential punishers hold the opposite expectation (i.e., they expect the opposing perspectives to be unpersuasive), a motivation to support punishment could actually encourage looking.

In our studies, before deciding whether or not to look (i.e., click the links to read the opposing perspective articles), subjects saw headlines for these articles. And because we did not collect any data about subjects’ expectations, on the basis of these headlines, about the persuasiveness of the opposing perspective articles, it is unclear what expectations subjects held. Thus, while our data do not support the proposal that reputational incentives for punishment can deter the consideration of opposing perspectives, we do not see them as providing a definitive refutation of this hypothesis. It is possible that making punishment observable would have decreased looking if subjects had expected the opposing perspective articles to be more persuasive.

2.3 Analyses of how making looking observable influences punishment

In our main text analyses of Studies 2-3, when considering the effects of making looking observable (i.e., when comparing our “Punishment Observable” and “Both Observable” conditions), we focus on rates of looking overall, and punishment without looking, as dependent measures. However, our design also allows us to ask how making looking observable influences overall rates of punishment. We likewise pre-registered analyses of this question, but do not report them in the main text because they are less relevant to our key theoretical questions. Here, however, we report these analyses.

In Study 2, among Democrats, rates of punishment were similar in Both Observable (24%) and Punishment Observable (23%), b = .01 [-.04, .05], t = .27, p = .786, n = 1206. Republicans were also similarly likely to punish in Both Observable (28%) and Punishment Observable (27%), b = .02 [-.03, .07], t = .72, p = .474, n = 1197.

In Study 3, among Democrats, rates of punishment were similar in Both Observable (28%) and Punishment Observable (30%), b = -.02 [-.07, .03], t = -.77, p = .439, n = 1284. Republicans were also similarly likely to punish in Both Observable (25%) and Punishment Observable (25%), b = -.003 [-.06, .06], t = -.10, p = .921, n = 779.

Thus, making looking observable did not influence rates of punishment in our studies.

2.4 Analyses of perfect comprehenders

As reported in the main text, for all studies in this paper, our main text analyses include all subjects, regardless of performance on comprehension questions. However, per our pre-registration, below we report secondary analyses that restrict to subjects who correctly answered our comprehension questions on their first try (and produce very similar results).

We note that in Studies 2-3, subjects answered two sets of comprehension questions. The first three questions tested comprehension of the Dictator Game structure, and the second three
questions tested comprehension of the extent to which subjects’ punishment and looking
decisions were observable to the Evaluator. Importantly, in the below analyses, rather than
restricting to subjects who answered all questions correctly on their first try, we merely restrict to
subjects who correctly answered the Dictator Game questions on the first try. This decision
reflects that in both Study 2 and Study 3, initial performance on each of the three observability
questions differed significantly across observability conditions; thus, restricting to subjects who
correctly answered all observability questions on their first try would undermine random
assignment and thus causal inference.

In particular, in Study 2, a series of chi squared tests reveals significant condition effects
on binary variables indicating whether the first (Democrats: \(X^2 = 30.37, p < .001\); Republicans:
\(X^2 = 34.10, p < .001\)), second (Democrats: \(X^2 = 174.58, p < .001\); Republicans: \(X^2 = 178.69, p < .001\)), and third (Democrats: \(X^2 = 6.81, p = .033\); Republicans: \(X^2 = 13.68, p = .001\))
observability questions were initially answered correctly. Similarly, in Study 3, condition
significantly predicted initial performance on the first (Democrats: \(X^2 = 92.94, p < .001\);
Republicans: \(X^2 = 38.76, p < .001\)), second (Democrats: \(X^2 = 204.96, p < .001\); Republicans: \(X^2 = 146.42, p < .001\)), and third (Democrats: \(X^2 = 13.54, p = .001\); Republicans: \(X^2 = 6.83, p = .033\))
observability questions.

### 2.4.1 Effects of making punishment observable

First, we report effects of making punishment observable (i.e., comparisons between our
Nothing Observable and Punishment Observable conditions) among perfect comprehenders.

In Study 2, among Democrats, overall rates of punishment were significantly higher in
Punishment Observable (24%) than Nothing Observable (18%), \(b = .07 [.02, .12], t = 2.83, p = .005, n = 1080\). However, rates of punishment without looking did not significantly differ across
the Punishment Observable (13%) and Nothing Observable (10%) conditions, \(b = .03 [-.01, .07],
t = 1.53, p = .127, n = 1080\). Among Republicans, overall rates of punishment were marginally
higher in Punishment Observable (28%) than Nothing Observable (22%), \(b = .05 [-.01, .11], t = 1.91, p = .056, n = 1022\). Furthermore, Republicans were significantly more likely to punish
without looking in Punishment Observable (20%) than in Nothing Observable (14%), \(b = .06 [.02, .11], t = 2.65, p = .008, n = 1022\).

In Study 3, among Democrats, overall rates of punishment were significantly higher in
Punishment Observable (31%) than Nothing Observable (18%), \(b = .13 [.08, .18], t = 5.22, p < .001, n = 1172\). Moreover, we find significantly higher rates of punishment without looking in
Punishment Observable (16%) than Nothing Observable (9%), \(b = .07 [.03, .11], t = 3.50, p < .001, n = 1172\). Similarly, Republicans were more likely to punish in Punishment Observable
(26%) than Nothing Observable (16%), \(b = .10 [.03, .16], t = 3.02, p = .003, n = 659\). And they
were also more likely to punish without looking in Punishment Observable (20%) than in
Nothing Observable (13%), \(b = .08 [.02, .13], t = 2.59, p = .010, n = 659\).

### 2.4.2 Effects of making looking observable

Next, report effects of making looking observable (i.e., comparisons between our
Punishment Observable and Both Observable conditions) among perfect comprehenders.

In Study 2, among Democrats, overall rates of looking were significantly higher in Both
Observable (52%) than Punishment Observable (36%), \(b = .16 [.10, .22], t = 5.41, p < .001, n = 1056\). Furthermore, rates of punishment without looking were significantly lower in Both
Observable (8%) than Punishment Observable (13%), \(b = -.05 [-.09, -.01], t = -2.64, p = .009, n
= 1056. Among Republicans, we likewise observed higher rates of looking in Both Observable (39%) than Punishment Observable (24%), \( b = .15 \ [.09, .21], t = 5.13, p < .001, n = 986. \)
However, we observed no significant difference between rates of punishment without looking in Both Observable (16%) and Punishment Observable (20%), \( b = -.04 \ [-.09, .01], t = -1.57, p = .116, n = 986. \)

In Study 3, among Democrats, overall rates of looking were significantly higher in Both Observable (50%) than in Punishment Observable (42%), \( b = .07 \ [.02, .13], t = 2.52, p = .012, n = 1137. \)
However, we found no significant difference in rates of punishment without looking between Both Observable (13%) versus Punishment Observable (16%), \( b = -.03 \ [-.07, .01], t = -1.47, p = .116, n = 1137. \)

2.5 Analyses of alternative specifications of looking

As described in the main text, per our pre-registration, our primary analyses define “looking” in Studies 2-3 as clicking at least one link to an opposing perspective article. However, for Study 3 (but not Study 2), we also preregistered secondary analyses that define looking as either (i) time spent on the “looking” screen (in which subjects were presented with links to opposing perspective articles, and invited to search the Internet for other opposing perspectives) or (ii) the continuous number of opposing perspective article links that subjects clicked. We report these alternative analyses, which produce very similar results, below.

2.5.1 Looking time (natural-log transformed seconds)

We begin by defining looking in terms of time spent on the “looking” screen. Per our pre-registration, we specifically consider the number of natural-log transformed seconds spent on this screen.

First, we use this new definition of looking to analyze overall rates of looking. In particular, we report the effect of making looking observable on overall rates of looking. In Study 2, among Democrats, we find that subjects spent more time looking in Both Observable (\( M = 3.66 \)) than Punishment Observable (\( M = 3.16 \)), \( b = .50 \ [.35, .65], t = 6.39, p < .001. \)
Similarly, Republicans subjects spent more time looking in Both Observable (\( M = 3.36 \)) than Punishment Observable (\( M = 3.03 \)), \( b = .34 \ [.20, .47], t = 4.83, p < .001. \)
In Study 3, among Democrats, subjects spent more time looking in Both Observable (\( M = 3.41 \)) than Punishment Observable (\( M = 3.20 \)), \( b = .21 \ [.06, .36], t = 2.72, p = .007. \)
Similarly, Republicans subjects spent more time looking in Both Observable (\( M = 3.11 \)) than Punishment Observable (\( M = 2.71 \)), \( b = .39 \ [.24, .54], t = 5.02, p < .001. \)

Next, we use our new definition of looking to redefine punishment without looking as punishing after looking for a below-median amount of time. In doing so, we compute the median (and thus define subjects as being above vs. below the median) separately for Democrats and Republicans. Then, we report the effects of making (i) punishment observable and (ii) looking observable, on punishment without looking.

In Study 2, among Democrats, subjects punished without looking at comparable rates in Punishment Observable (.10) and Nothing Observable (.10), \( b = .002 \ [-.03, .04], t = .12, p = .907, \) and comparable rates in Both Observable (.08) and Punishment Observable (.10), \( b = -.02 \ [-.09, .01], t = -1.57, p = .116, n = 986. \)

In Study 3, among Republicans, we likewise saw higher rates of looking in Both Observable (30%) than Punishment Observable (16%), \( b = .14 \ [.08, .20], t = 4.36, p < .001, n = 676. \)
However, we found no significant difference in rates of punishment without looking between Both Observable (20%) and Punishment Observable (20%), \( b = -.03 \ [-.06, .01], t = -1.47, p = .116, n = 676. \)
10

[-.06, .01], \( t = -1.42, p = .157 \). Among Republicans, subjects punished without looking at higher rates in Punishment Observable (.14) than Nothing Observable (.10), \( b = .05 [.01, .08], t = 2.56, p = .011 \), and lower rates in Both Observable (.10) than Punishment Observable (.14), \( b = -.04 [-.08, -.01], t = -2.28, p = .023 \).

In Study 3, among Democrats, subjects punished without looking at higher rates in Punishment Observable (.14) than Nothing Observable (.08), \( b = .06 [.02, .09], t = 3.26, p = .001 \), and comparable rates in Both Observable (.11) and Punishment Observable (.14), \( b = -.03 [-.06, -.01], t = -1.37, p = .170 \). Similarly, Republican subjects punished without looking at higher rates in Punishment Observable (.13) than Nothing Observable (.06), \( b = .07 [.03, .11], t = 3.38, p < .001 \), and comparable rates in Both Observable (.12) and Punishment Observable (.13), \( b = -.01 [-.06, .04], t = -.45, p = .651 \).

2.5.2 Number of articles

Next, we define looking as the continuous number of opposing perspective articles that subjects clicked the link for. When defining looking this way, in Study 2, we find that Democrats read a larger number of articles in Both Observable (\( M = .67 \)) than Punishment Observable (\( M = .44 \)), \( b = .24 [.16, .31], t = 5.96, p < .001 \). Similarly, Republicans read a larger number of articles in Both Observable (\( M = .61 \)) than Punishment Observable (\( M = .38 \)), \( b = .23 [.15, .32], t = 5.23, p < .001 \). In Study 3, among Democrats, subjects read a larger number of articles in Both Observable (\( M = .67 \)) than Punishment Observable (\( M = .54 \)), \( b = .13 [.05, .21], t = 3.12, p = .002 \). Similarly, Republicans subjects read a larger number of articles in Both Observable (\( M = .44 \)) than Punishment Observable (\( M = .20 \)), \( b = .24 [.15, .33], t = 5.12, p < .001 \).

We do not use this definition of looking to redefine punishment without looking. This reflects that, in all studies, less than half of subjects read even one article; thus, if we were to mirror our above approach and redefine punishment without looking as punishing after looking at a below-median number of articles, we would wind up with the same definition of punishment without looking that we already used in our primary analyses.

3. Discussion of preregistration deviations

All studies were pre-registered (Study 1a: https://aspredicted.org/blind.php?x=RB5_VT1; Study 1b: https://aspredicted.org/522_B9Z; Study 2a: https://aspredicted.org/blind.php?x=WM4_C5K; Study 2b: https://aspredicted.org/XX6_24X; Study 3a: https://aspredicted.org/2HX_QQI; Study 3b: https://aspredicted.org/XRC_RY3). As described in the main text, our analyses closely to our preregistered analysis plans with some minor deviations. Here, we describe these deviations.

3.1 Studies 1a-b

We begin by describing deviations in our analyses of Studies 1a-b. First, our Study 1a-b preregistrations describe our between-subject analyses (of evaluations of punishers who did vs. did not look) as primary analyses. However, for reasons of brevity and because these analyses produce results that are similar to our within-subject analyses but afford less statistical power, we report them only in the SI and not in the main text.

Second, our preregistrations describe our analyses of money sent in the Dictator Game as primary. In contrast, analyses of all other dependent variables (i.e., ratings of overall positivity, fairness, competence, and loyalty) are described in our pre-registrations as secondary analyses. Yet our main text does not preferentially focus on money sent. Instead, we treat money sent and
also ratings of overall positivity, fairness, competence, and loyalty as primary dependent variables.

To explain this decision, we begin by noting that we saw money sent in the Dictator Game and ratings of overall positivity and as “global” evaluation variables (i.e., as variables that reflected Evaluators’ overall impressions of Actors). In contrast, we saw ratings of fairness, competence, and loyalty as more specific evaluation variables. Originally, we planned to privilege money sent as our “primary” global evaluation variable. However, upon seeing our data, we noticed that many Evaluators chose either to share none or to exactly half of their endowment, and almost no Evaluators shared more than half of their endowment. This distributional feature likely reflects that money shared in the Dictator Game tapped both (i) the extent to which the Evaluator had a positive impression of the Actor and (ii) the extent to which the Evaluator was selfish vs. generous, or committed to avoiding inequality, etc. For example, Evaluators who were motivated to maximize their own payoffs may have never shared any money with any Actors, regardless of how positively they regarded them. And Evaluators who were committed to fairness may have always shared half of their money with all Actors, regardless of how positively they regarded them.

In contrast, ratings of overall positivity were more continuously distributed, perhaps reflecting that this global evaluation variable was a “purer” reflection of positive regard for the Actor. Thus, we ultimately felt that analyses of this overall positivity were at least as informative, if not more informative, than analyses of money sent. Moreover, we also ultimately felt that analyses of perceived fairness, competence, and loyalty provided critical insight into why Evaluators formed the global evaluations of Actors that they did. For these reasons, we prominently feature analyses of all DVs in our main text.

Third, our pre-registrations describe our analyses of Evaluations of punishers who do vs. do not look as primary, but our analyses of Evaluations of punishers vs. non-punishers (in the absence of looking information) as secondary. However, our main text prominently features both sets of analyses. This decision reflects that, while our primary theoretical focus was on evaluations of punishers who did vs. did not look, we felt that understanding whether punishment was evaluated positively in general was critical for interpreting Evaluators’ preferences for punishment with vs. without looking.

Fourth, we report a few analyses that were not pre-registered. In particular, we did not pre-register any of our exploratory analyses investigating individual preferences for punishment with vs. without looking, or our mediation analyses. While exploratory, we ultimately felt that these analyses were informative additions to our paper.

Finally, we also note that for these non-registered analyses, we used overall positivity ratings as our DV, rather than our other global evaluation variable, money shared in the Dictator Game. We made this choice despite the fact that money shared was pre-registered as our primary DV, for the reasons described above. Furthermore, for the same reasons, when we visualize our key Study 1 results (in Figure 1) and explore how ideology moderates those results (in Table 1), we also use overall positivity ratings as our DV.

3.2 Studies 2-3

Next, we describe deviations in our analyses of Studies 2-3. First, our Study 2-3 preregistrations describe as primary two analyses that we ultimately report only in the SI (and not in our main text): (i) analyses of the effect of making punishment observable on looking, and (ii) analyses of the effect of making looking observable on punishment. As noted in the main
text, we ultimately saw these analyses as less central to our theoretical questions; we therefore felt that they would distract from the flow of the main text and obfuscate the relationship between our analyses and the hypotheses we sought to test.

Second, for Study 3 but not Study 2, we preregistered a plan to report secondary analyses that used alternative definitions of looking. However, for completeness, in this SI we report these secondary analyses for both Study 3 and Study 2.

4. Appendix: Experimental stimuli

Here, we describe and illustrate with screenshots our experimental stimuli for all studies. Additionally, on OSF ([https://osf.io/3es2k/?view_only=e272c31aa7304eca841f084d582187fb](https://osf.io/3es2k/?view_only=e272c31aa7304eca841f084d582187fb)), we provide .qsf and PDF export files of the Qualtrics surveys used for each study, as well as PDFs containing the full texts of all opposing perspective articles (that subjects in Studies 2-3 viewed if they clicked the relevant links).

4.1 Studies 1a-b

After reporting their Prolific IDs and providing informed consent, subjects began by reporting their age, gender, and political party affiliation, and answering two attention check questions:

Age:

Which gender do you identify with more closely?

- Male
- Female
- Non-binary / other identity

Do you think of yourself as more of a Democrat or a Republican?

If you instead consider yourself an Independent, please indicate whether, as of today, you lean more Democratic or Republican.

Democrat

Republican

Thank you. Please carefully read the following story about a woman named Sarah.

Sarah works at a local grocery store. At the store, Sarah's job is to serve as the cashier. Normally, Sarah works Monday-Friday but does not work weekends. However, last week Sarah's coworker Ben asked her to cover his Saturday shift. So this Saturday, Sarah has to work a 7-hour shift.
What is Sarah’s job at the grocery store?

- Manager
- Cashier
- Stocker
- Customer service
- It was not specified in the story

The meal pictured above consists of multiple food items. From the options below, please select the food that is probably not one of these food items.

- ketchup
- onion
- mustard
- bread
- olives
- relish
- hotdog

Next, we introduced the Dictator Game (described as the “Sharing Game”), informed subjects that they would be interacting with another participant (described as “Player 2”) who shares their political party, and presented a set of comprehension questions about the game:
In this study, you will also participate in an interactive game with another participant.

The game is called the Sharing Game, and it has two players: Player 1 and Player 2. You will be Player 1.

Another participant will be Player 2. Like you, this participant ALSO identifies as a Democrat. And, just like we just told you that Player 2 is a Democrat, we will also tell Player 2 that you are a Democrat.

In the Sharing Game, you (Player 1) start with 50 cents. You then choose how many cents, if any, to share with Player 2.

Please answer the following questions, to make sure you understand the Sharing Game.

Imagine that you are deciding how much to share with Player 2.

Which decision will result in you earning the most money?

- You deciding to share 0 cents
- You deciding to share 25 cents
- You deciding to share 50 cents

Imagine that you are deciding how much to share with Player 2.

Which decision will result in Player 2 earning the most money?

- You deciding to share 0 cents
- You deciding to share 25 cents
- You deciding to share 50 cents

What political affiliation is Player 2?

- Player 2 is a Republican
- Player 2 is a Democrat

Note: the above screenshot shows the screen for Democrats in Study 1a. For Republicans in Study 1b, references to Democrats were replaced with references to Republicans.

Next, we explained that Player 2 signed a petition, and described the petition. Below we show how this looked for each of the three petitions. Recall that in Study 1a, Democrats were randomly assigned to either the Moore or Negy petition; in Study 1b, Republicans were always assigned to the Amazon petition.
Democrats, Moore Petition:

On this page, we would like you to tell you a bit more about Player 2.

We already recruited Player 2 to complete a study, which had an additional component.

In addition to participating in the Sharing Game, Player 2 also had the opportunity to sign a petition, hosted on Change.org.

The petition calls for the LAPD to fire Chief Michael Moore. It argues that Chief Moore should be fired following comments he made that blamed protestors for George Floyd's death.

On the next page, we will provide you with the full text of the petition and ask you to read it.

We also showed this full text to Player 2, who then decided whether or not to help get Chief Moore fired by signing the petition.

Democrats, Negy Petition:

On this page, we would like you to tell you a bit more about Player 2.

We already recruited Player 2 to complete a study, which had an additional component.

In addition to participating in the Sharing Game, Player 2 also had the opportunity to sign a petition, hosted on Change.org.

The petition calls for the University of Central Florida to fire professor Charles Negy. The petition alleges that professor Negy should be fired on the basis of abhorrent racist comments he has made on his personal Twitter account.

On the next page, we will provide you with the full text of the petition and ask you to read it.

We also showed this full text to Player 2, who then decided whether or not to help get professor Negy fired by signing the petition.

Republicans, Amazon Petition:

On this page, we would like you to tell you a bit more about Player 2.

We already recruited Player 2 to complete a study, which had an additional component.

In addition to participating in the Sharing Game, Player 2 also had the opportunity to sign a petition, hosted on Change.org.

The petition calls for the remove of “Blue Lives Murder” merchandise from Amazon. It alleges that the merchandise is hateful and must be immediately removed.

On the next page, we will provide you with the full text of the petition and ask you to read it.

We also showed this full text to Player 2, who then decided whether or not to help remove “Blue Lives Murder” merchandise from Amazon by signing the petition.
Next, we presented a screenshot of the petition.

Democrats, Moore Petition:

Below is a screenshot of the petition that we showed to Player 2. Over 300,000 people have already signed the petition, as highlighted in the screenshot below.

Again, the petition calls for the LAPD to fire Chief Michael Moore. It argues that Chief Moore should be fired following comments he made that blamed protesters for George Floyd’s death.

On this screen, please read through the petition.

LAPD is a corrupt department who claims the peaceful protests all around Los Angeles are civil unrest and criminal acts. This man should be fired.

On June 1st, Moore stated that George Floyd’s death is on protesters and losers hands. He is belittling the struggles and ways of the Black Lives Matter movement. This is only the shadow of an underlying racist and discriminating department.

Let LAPD Police Chief Michael Moore know how you feel about his blaming protesters for George Floyd’s death.
Democrats, Negy Petition:

Below is a screenshot of the petition that we showed to Player 2. Over 30,000 people have already signed the petition, as highlighted in the screenshot below.

Again, the petition calls for the University of Central Florida to fire professor Charles Negy. It alleges that professor Negy should be fired on the basis of abhorrent racist comments he has made on his personal Twitter account.

On this screen, please read through the petition.

We are calling on the University of Central Florida to dismiss psychology professor Charles Negy due to abhorrent racist comments he has made and continue to make on his personal Twitter account. In addition to racism, Negy has engaged in perverse transphobia and sexism on his account, which is just as reprehensible. While he has a right to free speech, he does not have a right to dehumanize students of color and other minority groups, which is a regular occurrence in his classroom, by allowing him to continue in his position, UCF would simply be empowering another cog in the machine of systemic racism.

UCF is a diverse, welcoming campus of 60,000 students from all across the country and the world. Our diversity is our strength, and we should have faculty who understand and promote that - not the opposite. College is a place to learn and be exposed to new ideas, but for so many Black students, racism is not just an idea, but something they have experienced for their entire life, and the same goes for the LGBT community with trans and homophobia. They deserve better, and so does the entire campus community.

I once again want to reiterate my love for this university. We are a community which always has each other’s back. As a student coming from a place of privilege, this is me having the backs of my fellow Knights.

Please dismiss Professor Negy for the good of the entire campus community.
Republicans, Amazon Petition:

Below is a screenshot of the petition that we showed to Player 2. Over 115,000 people have already signed the petition, as highlighted in the screenshot below.

Again, the petition calls for the removal of "Blue Lives Murder" merchandise from Amazon. It argues that the merchandise is hateful and must be immediately removed.

On this screen, please read through the petition.
Next, we explained to subjects that Player 2 had the chance to “look” at opposing perspectives, and show an example headline from an opposing perspective article.

Democrats, Moore petition:

After viewing this petition, Player 2 decided whether or not to sign it.

Additionally, we gave Player 2 an opportunity to take some time—if Player 2 wanted it—to consider OPPOSING perspectives before deciding whether to sign. It was completely up to Player 2 whether to use this time, and how.

For Player 2's convenience, we provided links to some specific articles that may provide opposing perspectives pertaining to the petition. For example, we provided a link to the below article published by the LA Times describing how elected politicians in Los Angeles, including the mayor, continue to support Police Chief Moore.

Player 2 also had the option of taking time to search the Internet for other opposing perspectives.

And, importantly, Player 2 also had the option NOT to read any opposing perspectives before deciding whether or not to sign the petition.

Democrats, Negy petition:

After viewing this petition, Player 2 decided whether or not to sign it.

Additionally, we gave Player 2 an opportunity to take some time—if Player 2 wanted it—to consider OPPOSING perspectives before deciding whether to sign. It was completely up to Player 2 whether to use this time, and how.

For Player 2's convenience, we provided links to some specific articles that may provide opposing perspectives pertaining to the petition. For example, we provided a link to the below article, published by the Orlando Sentinel, describing professor Negy's claim that he is the subject of a "witch hunt".

Player 2 also had the option of taking time to search the Internet for other opposing perspectives.

And, importantly, Player 2 also had the option NOT to read any opposing perspectives before deciding whether or not to sign the petition.
Republicans, Amazon Petition:

After viewing this petition, Player 2 decided whether or not to sign it.

Additionally, we gave Player 2 an opportunity to take some time—If Player 2 wanted it—to consider OPPOSING perspectives before deciding whether to sign. It was completely up to Player 2 whether to use this time, and how.

For Player 2’s convenience, we provided links to some specific articles that may provide opposing perspectives pertaining to the petition. For example, we provided a link to the below article, published by a local Houston news outlet, reporting on the perspective of an activist who does not think “Blue Lives Murder” merchandise constitutes hate speech.

Player 2 also had the option of taking time to search the Internet for other opposing perspectives.

And, importantly, Player 2 also had the option NOT to read any opposing perspectives before deciding whether or not to sign the petition.

Next, we explained that Player 2s varied with respect to the amount of time they spent looking:

Importantly, participants assigned to the role of Player 2 varied in the amount of time they spent considering opposing perspectives before deciding whether or not to sign the petition.

Some Player 2s spent no or very little time considering opposing perspectives before deciding whether or not to sign the petition.

In contrast, other Player 2s spent a lot of time considering opposing perspectives before deciding whether or not to sign the petition.

Next, we collected our dependent variables, for a set of different Player 2s with different behavioral profiles (i.e., a set of Player 2s about whom we provided different information about their punishment and looking behavior).

The first two Player 2s that subjects evaluated were punishers who did vs. did not look. We manipulated, between subjects, the order in which these Player 2s were presented. After the first but before the second was presented, subjects saw a transition screen.
This was the screen subjects used to evaluate a punisher who did *not* look:

Now, we would like you to make a Sharing Game decision.

Another participant, in the role of Player 2, spent a BELOW-AVERAGE amount of time considering opposing perspectives, and then chose TO sign the petition.

After this study is completed, we will match Player 1s with Player 2s and compute bonuses for all players. You may be matched with this Player 2.

If you are paired with this Player 2, how many cents, if any, would you like to share?

<table>
<thead>
<tr>
<th>0</th>
<th>5</th>
<th>10</th>
<th>15</th>
<th>20</th>
<th>25</th>
<th>30</th>
<th>35</th>
<th>40</th>
<th>45</th>
<th>50</th>
</tr>
</thead>
</table>

Note: this Player 2 is a REAL participant who really did make the decisions described above. If you are paired with this Player 2, your decision really will be used to compute your bonus and their bonus in this study.

Please also answer the following questions about your impression of this Player 2.

How **positive** is your evaluation of this Player 2?

<table>
<thead>
<tr>
<th>Very negative</th>
<th>Neutral</th>
<th>Very positive</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>25</td>
<td>50</td>
</tr>
<tr>
<td></td>
<td>75</td>
<td>100</td>
</tr>
</tbody>
</table>

To what extent do you think this Player 2 is a **loyal supporter** of Black Lives Matter?

<table>
<thead>
<tr>
<th>Not a supporter at all</th>
<th>A moderate supporter</th>
<th>A very loyal supporter</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>25</td>
<td>50</td>
</tr>
<tr>
<td></td>
<td>75</td>
<td>100</td>
</tr>
</tbody>
</table>

How **competent** do you think this Player 2 is?

<table>
<thead>
<tr>
<th>Very incompetent</th>
<th>Neutral</th>
<th>Very competent</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>25</td>
<td>50</td>
</tr>
<tr>
<td></td>
<td>75</td>
<td>100</td>
</tr>
</tbody>
</table>

How **fair** do you think this Player 2 is?

<table>
<thead>
<tr>
<th>Very unfair</th>
<th>Neutral</th>
<th>Very fair</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>25</td>
<td>50</td>
</tr>
<tr>
<td></td>
<td>75</td>
<td>100</td>
</tr>
</tbody>
</table>

Note: the above screenshot shows the screen for Democrats in Study 1a, where we measured loyalty towards Black Lives Matter; for Republicans in Study 1b, we instead measured loyalty towards Blue Lives Matter. We also randomized between-subjects the order in which we presented the questions about how competent, fair, and loyal Player 2 was; subjects were randomly assigned to one order and then that order was consistently applied across all Player 2s that they evaluated.

The screen that subjects used to evaluate a punisher who *did* look was identical, except that we replaced “below-average” with “above-average” in the second sentence on the screen.
This was the transition screen that subjects saw (between evaluating a punisher who did vs. did not look):

Thank you for making that decision.

On the subsequent screens, you will make a series of decisions about how much to share with OTHER potential Player 2s that you could be paired with.

These Player 2s are all REAL participants, who, like you, identify as Democrats. They all had the opportunity to sign the petition about Professor Charles Negy. And they all chose how long to spend considering opposing perspectives before deciding whether or not to sign the petition. However, these Player 2s varied in terms of the decisions that they made in the study.

On the subsequent screens, please decide how much you would like to share with each of them, if you are ultimately paired with them.

Note: the above screenshot shows screen for Democrats who were assigned to the Negy petition; the text was adapted to reference the correct party and petition for all subjects.

The next two Player 2s that subjects evaluated were non-punishers who did vs. did not look. Subjects who evaluated a punisher who did look before a punisher who did not look correspondingly evaluated a non-punisher who did look before evaluating a non-punisher who did not look, and vice versa. The screens that subjects used to evaluate non-punishers were identical to the screens they used to evaluate punishers, except that we replaced “chose TO sign the petition” with “chose NOT to sign the petition”.

The next two Player 2s that subjects evaluated were punishers and non-punishers, with no information provided about looking. Before presenting these two Player 2 profiles, we showed subjects this transition screen:

Thank you.

On the next screens, you will make two more sharing game decisions about Player 2s that you might be matched with.

For these decisions, we will tell you whether or not Player 2 decided to sign the petition.

However, we will NOT tell you anything about how long Player 2 spent considering opposing perspectives before making this decision.

The screens that subjects used to evaluate Player 2s without looking information were identical to the screens they used to evaluate Player 2s with looking information, except that we did not provide information about looking. So, the non-punisher screen began like this:

Now, we would like you to make another Sharing Game decision.

Another participant, in the role of Player 2, chose NOT to sign the petition.

And punisher screen began identically, except that we replaced “chose NOT to” with “chose TO”.

Next, subjects evaluated a Player 2 about whom they had no information:

Thank you for making these decisions.

Finally, how many cents would you like to share with Player 2, if given no information about Player 2's decisions?

<table>
<thead>
<tr>
<th></th>
<th>0</th>
<th>5</th>
<th>10</th>
<th>15</th>
<th>20</th>
<th>25</th>
<th>30</th>
<th>35</th>
<th>40</th>
<th>45</th>
<th>50</th>
</tr>
</thead>
<tbody>
<tr>
<td>Options</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td></td>
</tr>
</tbody>
</table>

Note: this decision may really be used to compute your bonus and the bonus of a Player 2 in this study.

Given no information about Player 2's decisions, how positive is your impression of Player 2?

<table>
<thead>
<tr>
<th></th>
<th>Very negative</th>
<th>Neutral</th>
<th>Very positive</th>
</tr>
</thead>
<tbody>
<tr>
<td>Options</td>
<td>0</td>
<td>25</td>
<td>50</td>
</tr>
<tr>
<td></td>
<td>75</td>
<td>100</td>
<td></td>
</tr>
</tbody>
</table>

Given no information about Player 2's decisions, to what extent do you think Player 2 is a loyal supporter of Black Lives Matter?

<table>
<thead>
<tr>
<th></th>
<th>Not a supporter at all</th>
<th>A moderate supporter</th>
<th>A very loyal supporter</th>
</tr>
</thead>
<tbody>
<tr>
<td>Options</td>
<td>0</td>
<td>25</td>
<td>50</td>
</tr>
<tr>
<td></td>
<td>75</td>
<td>100</td>
<td></td>
</tr>
</tbody>
</table>

Given no information about Player 2's decisions, how competent do you think Player 2 is?

<table>
<thead>
<tr>
<th></th>
<th>Very incompetent</th>
<th>Neutral</th>
<th>Very competent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Options</td>
<td>0</td>
<td>25</td>
<td>50</td>
</tr>
<tr>
<td></td>
<td>75</td>
<td>100</td>
<td></td>
</tr>
</tbody>
</table>

Given no information about Player 2's decisions, how fair do you think Player 2 is?

<table>
<thead>
<tr>
<th></th>
<th>Very unfair</th>
<th>Neutral</th>
<th>Very fair</th>
</tr>
</thead>
<tbody>
<tr>
<td>Options</td>
<td>0</td>
<td>25</td>
<td>50</td>
</tr>
<tr>
<td></td>
<td>75</td>
<td>100</td>
<td></td>
</tr>
</tbody>
</table>

Finally, subjects completed the following post-experimental survey:

Thank you for making these decisions.

Now we would like you to provide your own opinion about the petition. We have included a screenshot of the petition below for your reference.

Please rate your agreement with the following statement: "I strongly support the petition and the underlying cause behind it."

<table>
<thead>
<tr>
<th></th>
<th>1 - Strongly disagree</th>
<th>2</th>
<th>3 - Somewhat disagree</th>
<th>4</th>
<th>5 - Neither agree nor disagree</th>
<th>6</th>
<th>7 - Somewhat agree</th>
<th>8</th>
<th>9 - Strongly agree</th>
</tr>
</thead>
<tbody>
<tr>
<td>Options</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
</tr>
</tbody>
</table>

[Below this question, we re-printed the screenshot of the petition]

[Page break]

Please describe how you made your choices in this study.

[Page break]

Is there anything you'd like to share with us about your impression of this study?

[Page break]
To what extent have you previously participated in other studies like to this one?

<table>
<thead>
<tr>
<th>1 - Nothing like this scenario</th>
<th>2</th>
<th>3 - Something like this scenario</th>
<th>4</th>
<th>5 - Exactly this scenario</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

To what extent do you believe that the petition is real?

<table>
<thead>
<tr>
<th>1 - Very skeptical that it is real</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7 - Very confident that it is real</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

To what extent do you believe that you really will be matched with a real Player 2?

<table>
<thead>
<tr>
<th>1 - Very skeptical that I will be matched with a real Player 2</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7 - Very confident that I will be matched with a real Player 2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

[Page break]
Note: the above screenshot shows the screen for Democrats in Study 1a, where we measured support for Black Lives Matter; for Republicans in Study 1b, we referenced Republicans (rather than Democrats) and measured support for Blue Lives Matter ("To what extent do you support "Blue Lives Matter" (a countermovement, started in response to Black Lives Matter, advocating that those who are prosecuted and convicted of killing law enforcement officers should be sentenced under hate crime statutes)?").
4.2 Studies 2-3

After reporting their Prolific IDs and providing informed consent, subjects began by reporting their age, gender, and political party affiliation, and answering two attention check questions; these measures were all identical to those presented at the beginning of Study 1. So, see Study 1 in this document for screenshots.

Next, we introduced the Dictator Game (described as the “Sharing Game”), informed subjects that they would be interacting with another participant (described as “Player 1”) who shares their political party, and presented a set of comprehension questions about the game.

For Democrats in Study 2a, Player 1 was described as relatively less ideological:

In this HIT, you will participate in an interactive game with another MTurk worker.

The game is called the Sharing Game, and it has two players: Player 1 and Player 2. You will be Player 2.

Another MTurk worker will be Player 1. Like you, this worker ALSO indicated to us in a survey that they identify as a Democrat.

More specifically, Player 1 indicated that they identify as a weak Democrat, who only leans towards the party.

And, just like we just told you that Player 1 is a Democrat, we will also tell Player 1 that you are a Democrat.

In the Sharing Game, Player 1 starts with 50 cents. Player 1 then chooses how many cents, if any, to share with you (Player 2).

Please answer the following questions, to make sure you understand the Sharing Game.

Imagine that Player 1 is deciding how much to share with you.

Which decision will result in Player 1 earning the highest payoff?

- Player 1 deciding to share 0 cents
- Player 1 deciding to share 25 cents
- Player 1 deciding to share 50 cents

Imagine that Player 1 is deciding how much to share with you.

Which decision will result in you earning the highest payoff?

- Player 1 deciding to share 0 cents
- Player 1 deciding to share 25 cents
- Player 1 deciding to share 50 cents

What information do you know about Player 1?

- Player 1 identifies as a Republican
- Nothing
- Player 1 identifies as a weak Democrat, who only leans towards the party
For Democrats in Study 3a, Player 1 was described as relatively more ideological. Thus, the screen instead began as follows:

In this HIT, you will also participate in an interactive game with another MTurk worker.

The game is called the Sharing Game, and it has two players: Player 1 and Player 2. **You will be Player 2.**

Another MTurk worker will be Player 1. **Like you, this worker ALSO indicated to us in a survey that they identify as a Democrat.**

More specifically, Player 1 indicated that they identify as a strong Democrat who strongly supports Black Lives Matter.

And, just like we just told you that Player 1 is a Democrat, we will also tell Player 1 that you are a Democrat.

In the Sharing Game, Player 1 starts with 50 cents. Player 1 then chooses how many cents, if any, to share with you (Player 2).

The comprehension questions that followed were identical to those presented in Study 2a, except that the final answer chance for last question instead read “Player 1 identifies as a strong Democrat who strongly supports Black Lives Matter”.

For Republicans in Study 2b, the screen was identical to what we showed in Study 2a, except that we referenced Republicans instead of Democrats.

For Republicans in Study 3b, the screen was identical to what we showed in Study 3a, except that we (i) referenced Republicans instead of Democrats and (ii) the sentence starting with “more specifically” instead read “More specifically, Player 1 indicated that they identify as a strong Republican who strongly supports “Blue Lives Matter” (a countermovement, started in response to Black Lives Matter, advocating that those who are prosecuted and convicted of killing law enforcement officers should be sentenced under hate crime statutes).” Furthermore, the final answer choice for last question read “Player 1 identifies as a strong Republican who strongly supports Blue Lives Matter”.

If subjects answered any comprehension questions incorrectly, they were presented with the screen involving the questions a second time, along with the text: “On the previous page, you answered one or more questions incorrectly. You MUST answer ALL questions correctly to continue on with the study. Please carefully re-read the below instructions and answer the questions again”. Here, they were not allowed to precede to the next screen until they answered all questions correctly.

Next, we told subjects that they had the opportunity to sign a petition. Below we show the text we presented for each of the three petitions. Below this text, we displayed a screenshot of the petition. These screenshots were identical to those presented in Study 1; so, see Study 1 in this document for screenshots.
Democrats, Moore petition (Study 2a):

This HIT will also have another component.

In addition to participating in the Sharing Game, you will also have the opportunity to sign a petition, hosted on Change.org. Below, we have provided a screenshot of the petition. Later in this HIT, we will provide a link to the petition. Over 300,000 people have already signed the petition, as highlighted in the screenshot below.

The petition calls for the LAPD to fire Chief Michael Moore. It argues that Chief Moore should be fired following comments he made that blamed protestors for George Floyd's death.

On this screen, please read through the petition. Then, later in this HIT, you will have the chance to decide whether you’d like to help get Chief Moore fired by signing the petition.

You can choose to sign or not to sign. One thing you should keep in mind is that we will tell Player 1 whether or not you choose to sign.

Democrats, Negy petition (Study 3a):

This HIT will also have another component.

In addition to participating in the Sharing Game, you will also have the opportunity to sign a petition, hosted on Change.org. Below, we have provided a screenshot of the petition. Later in this HIT, we will provide a link to the petition. Over 30,000 people have already signed the petition, as highlighted in the screenshot below.

The petition calls for the University of Central Florida to fire professor Charles Negy. It argues that professor Negy should be fired on the basis of abhorrent racist comments he has made on his personal Twitter account.

On this screen, please read through the petition. Then, later in this HIT, you will have the chance to decide whether you’d like to help get professor Negy fired by signing the petition.

You can choose to sign or not to sign. One thing you should keep in mind is that we will tell Player 1 whether or not you choose to sign.

Note: the above two screenshots show the screens from the “Punishment Observable” and “Both Observable” conditions, where punishment was observable.

Republicans, Amazon petition (Studies 2b and 3b):

This HIT will also have another component.

In addition to participating in the Sharing Game, you will also have the opportunity to sign a petition, hosted on Change.org. Below, we have provided a screenshot of the petition. Later in this HIT, we will provide a link to the petition. Over 115,000 people have already signed the petition, as highlighted in the screenshot below.

The petition calls for the removal of “Blue Lives Murder” merchandise from Amazon. It argues that the merchandise is hateful and must be immediately removed.

On this screen, please read through the petition. Then, later in this HIT, you will have the chance to decide whether you’d like to help remove “Blue Lives Murder” merchandise from Amazon by signing the petition.

You can choose to sign or not to sign. One thing you should keep in mind is that your decision will be completely private: we will NOT tell Player 1 whether or not you choose to sign.

Note: the above screenshot shows the screen from the “Nothing Observable” condition, where punishment was not observable.
Next, we told subjects that they would have a chance to look at opposing perspective articles, and showed an example headline from an opposing perspective article.

Democrats, Negy Petition (Study 3a):

Additionally, we will give you an opportunity to take some time—if you would like—to consider OPPOSING perspectives before deciding whether to sign. It is completely up to you whether to use this time, and how.

For your convenience, we will provide you with links to some specific articles that may provide opposing perspectives pertaining to the petition. For example, we will provide a link to the below article, published by the Orlando Sentinel, describing professor Negy’s claim that he is the subject of a “witch hunt.” You can also choose to use the time to search the Internet for other opposing perspectives.

Importantly, you can also choose NOT to read any opposing perspectives and instead move forward with the survey.

You can take as much or as little time as you would like to consider opposing perspectives. One thing you should keep in mind is that we will tell Player 1 how long you spend considering opposing perspectives before deciding whether to sign.

Note: the above screenshot shows the screen from the “Both Observable” condition, where looking was observable.

Democrats, Moore petition (Study 2a):

Additionally, we will give you an opportunity to take some time—if you would like—to consider OPPOSING perspectives before deciding whether to sign. It is completely up to you whether to use this time, and how.

For your convenience, we will provide you with links to some specific articles that may provide opposing perspectives pertaining to the petition. For example, we will provide a link to the below article published by the LA Times describing how elected politicians in Los Angeles, including the mayor, continue to support Police Chief Moore. You can also choose to use the time to search the Internet for other opposing perspectives.

Importantly, you can also choose NOT to read any opposing perspectives and instead move forward with the survey.

You can take as much or as little time as you would like to consider opposing perspectives. One thing you should keep in mind is that your decision will be completely private: we will NOT tell Player 1 how long you spend considering opposing perspectives before deciding whether to sign.
Republicans, Amazon Petition (Studies 2b and 3b):

Additionally, we will give you an opportunity to take some time—if you would like—to consider OPPOSING perspectives before deciding whether to sign. It is completely up to you whether to use this time, and how.

For your convenience, we will provide you with links to some specific articles that may provide opposing perspectives pertaining to the petition. For example, we will provide a link to the below article, published by a local Houston news outlet, reporting on the perspective of an activist who does not think “Blue Lives Murder” merchandise constitutes hate speech. You can also choose to use the time to search the Internet for other opposing perspectives.

Importantly, you can also choose NOT to read any opposing perspectives and instead move forward with the survey.

You can take as much or as little time as you would like to consider opposing perspectives. One thing you should keep in mind is that your decision will be completely private: we will NOT tell Player 1 how long you spend considering opposing perspectives before deciding whether to sign.

Note: the above two screenshots show the screens from the “Nothing Observable” and “Punishment Observable” conditions, where looking was not observable.
Next, we summarized for subjects the relevant information about their observability condition, and presented a second set of comprehension questions about what was observable to Player 1. This screen looked different across conditions.

Nothing Observable:

Thank you. On this page, we’d like to review how the petition and Sharing Game components of this HIT relate to each other.

In the Sharing Game, before deciding how much to share with you, Player 1 will see the full text of the petition and learn that you had the opportunity to sign it. Furthermore, Player 1 will learn that you were given time to consider opposing perspectives before deciding whether to sign the petition.

Importantly, however, we will NOT tell Player 1 how long you spend considering opposing perspectives, OR whether or not you ultimately decide to sign the petition.

In other words, your signing decision—and the time you spend considering opposing perspectives—will both remain completely private.

Please answer the following questions, to make sure you understand.

Before Player 1 decides how much money to share with you…

Will Player 1 learn that you have the opportunity to sign the petition?

☐ No
☐ Yes, but Player 1 will not get to read the full text of the petition
☐ Yes, and Player 1 will also get to read the full text of the petition

Will Player 1 learn that you were given time to consider opposing perspectives before committing to signing?

☐ No
☐ Yes

What will Player 1 learn about your behavior?

Nothing. Player 1 will not find out how long I spend considering opposing perspectives, or whether I ultimately sign the petition

Player 1 will only find out whether I ultimately sign the petition (but not how long I spend considering opposing perspectives)

Player 1 will both find out how long I spend considering opposing perspectives, and whether I ultimately sign the petition
In Punishment Observable, we presented the same questions, but introduced them as follows:

Thank you. On this page, we’d like to review how the petition and Sharing Game components of this HIT relate to each other.

In the Sharing Game, before deciding how much to share with you, Player 1 will see the full text of the petition and learn that you had the opportunity to sign it. Furthermore, Player 1 will learn that you were given time to consider opposing perspectives before deciding whether to sign the petition.

Importantly, we will NOT tell Player 1 how long you spend considering opposing perspectives. We WILL, however, tell Player 1 whether or not you ultimately decide to sign the petition.

In other words, while the time you spend considering opposing perspectives will remain completely private, your signing decision will be shared with Player 1.

In Both Observable, we likewise presented the same questions, but introduced them as follows:

Thank you. On this page, we’d like to review how the petition and Sharing Game components of this HIT relate to each other.

In the Sharing Game, before deciding how much to share with you, Player 1 will see the full text of the petition and learn that you had the opportunity to sign it. Furthermore, Player 1 will learn that you were given time to consider opposing perspectives before deciding whether to sign the petition.

And, importantly, we will tell Player 1 how long you spend considering opposing perspectives, AND whether or not you ultimately decide to sign the petition.

In other words, your signing decision—and the time you spend considering opposing perspectives—will both be shared with Player 1.

Like with the first set of comprehension questions, if subjects answered any questions on this screen incorrectly, they were presented with the screen a second time, along with the text: “On the previous page, you answered one or more questions incorrectly. You MUST answer ALL questions correctly to continue on with the study. Please carefully re-read the below instructions and answer the questions again”. Here, they were not allowed to precede to the next screen until they answered all questions correctly.

Next, we informed subjects that they were about to precede to the page where they could look at opposing perspectives:

Thank you.

On the next screen, we will give you the opportunity to take some time— if you would like—to consider OPPOSING perspectives before deciding whether to sign the petition.

As a reminder, it is completely up to you whether to use this time, and how. We will provide you with links to some specific articles that may provide opposing perspectives pertaining to the petition. You can also choose to use the time to search the Internet for other opposing perspectives.

Importantly, you can also choose NOT to read any opposing perspectives and instead move forward with the survey.

Remember, we will tell Player 1 how long you choose to spend considering opposing perspectives before deciding whether to sign the petition. So Player 1 will learn whether you make a quick decision about whether to sign the petition, or choose to extensively consider opposing perspectives first.

Note: the above screenshot shows the screen from “Both Observable”. In the other two conditions, the last paragraph instead read: “Remember, we will NOT tell Player 1 how long you choose to spend considering opposing perspectives before deciding whether to sign the petition. So Player 1 will NOT learn whether you make a quick decision about whether to sign the petition, or choose to extensively consider opposing perspectives first”.
Next, we measured looking by presenting the following screens, and measuring which, if any, links subjects clicked on. Subjects were not informed that their link-clicking would be tracked.

Democrats, Negy petition (Study 3a):

Please take as much or as little time as you would like to consider opposing perspectives.

Remember, we will tell Player 1 how long you spend on this page considering opposing perspectives.

When you have decided whether to sign the petition, please advance the screen.

Previously, we showed you the headline of the below article, published by the Orlando Sentinel, describing professor Negy's claim that he is the subject of a "witch hunt". The link to view this article is here.

![UCF professor behind tweets deemed racist says he is subject of ‘witch hunt’](image)

Below is the headline of another article you may consider, published by the National Association of Scholars, arguing that firing professor Negy would be a violation of his rights. The link to view this article is here.

![Let the Tweeter Beware](image)

Note: the above screenshot shows the screen from the “Both Observable” condition, where looking was observable.
Democrats, Moore petition (Study 2a):

Please take as much or as little time as you would like to consider opposing perspectives.

Remember, we will NOT tell Player 1 how long you spend on this page considering opposing perspectives.

When you have decided whether to sign the petition, please advance the screen.

Previously, we showed you the headline of the below article, published by the LA Times, describing the support Chief Moore has maintained in Los Angeles political circles since the controversy erupted. The link to view this article is [here].

![Los Angeles Times](https://via.placeholder.com/150)

Despite criticism, LAPD Chief Michel Moore maintains support in political circles

Below is the headline of another article you may consider, published by FOX 11 Los Angeles, reporting on Los Angeles Mayor Eric Garcetti's statement of support for Chief Moore. The link to view this article is [here].

Garcetti says he has confidence in LAPD Chief Moore after he 'misspoke'

Published June 3 | Death of George Floyd | City News Service
Republicans, Amazon Petition (Studies 2b and 3b):

Please take as much or as little time as you would like to consider opposing perspectives.

Remember, we will NOT tell Player 1 how long you spend on this page considering opposing perspectives.

When you have decided whether to sign the petition, please advance the screen.

Previously, we showed you the headline of the below article, published by Click2Houston, reporting on the perspective of an activist who does not think “Blue Lives Murder” merchandise constitutes hate speech. The link to view this article is here.

Below is the headline of another article you may consider, published by FIRE, about what constitutes hate speech and the ways that it is protected by the first amendment. The link to view this article is here.

Note: the above two screenshots show the screens from the “Nothing Observable” and “Punishment Observable” conditions, where looking was not observable.
Next, we measured punishment by presenting the following screens, and measuring whether subjects clicked the link to the petition. Subjects were not informed that their link-clicking would be tracked.

On the next page, you will decide whether or not to sign the petition.

If you choose to sign, we will NOT collect your identifying information. Instead, we will ask you to show us that you really did sign another way.

Specifically, immediately after you sign, you will be redirected to a new screen. At the top of this new screen, you will see the 1-2-3 image below. However, instead of seeing the black box, you will see some text. Specifically, you will see a simple phrase, written at the top of the page in relatively small font.

When you see the 1-2-3 image, you will be DONE signing the petition. You do NOT need to take any further action. Instead, please STAY ON THIS PAGE and pay attention to the phrase that is written above the 1-2-3 image (in place of the black box).

If you sign the petition, we will ask you to report that phrase back to us, to show that you really did sign.

[Page break]

Thank you. Now, you will decide whether or not to sign the petition.

Remember, Player 1 will learn about the petition (and will get to read the full text of the petition). And we will tell Player 1 whether or not you ultimately chose to sign the petition.

If you would like to sign, please do so now by clicking here.

Remember, if you sign, please stay on the subsequent screen and pay attention to the phrase written above the 1-2-3 image!

Did you choose to sign the petition?

☐ Yes
☐ No

If you DID sign the petition, you should be able to see the 1-2-3 image, pictured above, on the next screen. To show us that you signed, please type the phrase written above the 1-2-3 image (instead of the black box).

If you did NOT sign the petition, please instead just type "no".

Note: if you DID sign the petition, but are confused by the above instructions, please just type any phrase that you saw after signing the petition.

Please also rate your agreement with the following statement:

“I am strongly committed to supporting the petition and the underlying cause behind it.”

Note: the above screenshot shows the screens from the “Punishment Observable” and “Both Observable” conditions, where punishment was observable. In “Nothing Observable”, we replaced the sentence that read “And we will tell Player 1 whether or not you ultimate choose to sign the petition” with “But we will NOT tell Player 1 whether or not you sign the petition.”
Finally, subjects completed the following post-experimental survey:

Thank you for your decision. Now, we would like you to answer a few questions about the petition. (Unlike your signing decision, your responses to these questions will NOT be shown to Player 1).

<table>
<thead>
<tr>
<th></th>
<th>Not at all</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
</tr>
</thead>
<tbody>
<tr>
<td>How moral do you think the petition is?</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td></td>
</tr>
<tr>
<td>To what extent do you agree with the petition?</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td></td>
</tr>
<tr>
<td>How comfortable are you with the petition's approach?</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td></td>
</tr>
<tr>
<td>To what extent do you think the petition's approach is proportionate and appropriate?</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td></td>
</tr>
</tbody>
</table>

[Note: the above screenshot shows the screens from the “Punishment Observable” and “Both Observable” conditions, where punishment was observable. In “Nothing Observable”, the parenthetical instead read: (As was the case for your signing decision, your responses to these questions will NOT be shown to Player 1).]

[Page break]

Thank you. For our records, on the screen where you had the chance to consider opposing perspectives, did you do any of the things listed below?

If so, please check all that apply. If not, that's totally fine; please leave this question blank and advance to the next page.

**Note**: your response to this question will NOT be shown to Player 1. It is only for our records.

- [ ] I read the Herald Mail Media article (headline: UCF professor behind tweets deemed racist says he is subject of "witch hunt")
- [ ] I read the National Association of Scholars article (headline: Let the tweeter beware)
- [ ] I searched the Internet for other perspectives that would oppose the petition
- [ ] I searched the Internet for other perspectives that would support the petition

[Note: the above screenshot shows the screen from Study 3a, where Democrats saw the Negy petition. The text was adapted to reference the correct party and petition for all subjects.]

[Page break]

Thank you.

Before completing this HIT, how much did you know about the events described in the petition?

<table>
<thead>
<tr>
<th></th>
<th>1 - I had never heard of them</th>
<th>2</th>
<th>3 - I had heard of them, but knew only some details</th>
<th>4</th>
<th>5 - I had been closely following the situation and know many details</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>○</td>
<td>2</td>
<td>○</td>
<td>4</td>
<td>○</td>
</tr>
</tbody>
</table>

[Page break]

Please describe how you made your choices in this HIT.

[Page break]
Now, we would like to ask you a few questions regarding the amount of time you decided to spend considering opposing perspectives.

To what extent did you make your decision (regarding how long to spend considering opposing perspectives) because...

<table>
<thead>
<tr>
<th>1 - Not at all</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7 - Entirely</th>
</tr>
</thead>
<tbody>
<tr>
<td>... you personally felt that it was truly the right decision?</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
</tr>
<tr>
<td>... you wanted to see yourself as a good person?</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
</tr>
<tr>
<td>... you wanted others to see you as a good person?</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
</tr>
<tr>
<td>... you wanted Player 1 to see you as a good person?</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
</tr>
</tbody>
</table>

Now, we would like to ask you a few questions about your decision regarding whether to sign the petition.

To what extent did you make your decision (regarding whether to sign the petition) because...

<table>
<thead>
<tr>
<th>1 - Not at all</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7 - Entirely</th>
</tr>
</thead>
<tbody>
<tr>
<td>... you personally felt that it was truly the right decision?</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
</tr>
<tr>
<td>... you wanted to see yourself as a good person?</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
</tr>
<tr>
<td>... you wanted others to see you as a good person?</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
</tr>
<tr>
<td>... you wanted Player 1 to see you as a good person?</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
</tr>
</tbody>
</table>

[Page break]

To what extent do you believe that the petition and articles providing opposing perspectives are real?

<table>
<thead>
<tr>
<th>1 - Very skeptical that they are real</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7 - Very confident that they are real</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
</tr>
</tbody>
</table>

To what extent do you believe that Player 1 and the sharing game are real?

<table>
<thead>
<tr>
<th>1 - Very skeptical that they are real</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7 - Very confident that they are real</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
</tr>
</tbody>
</table>

[Page break]
Note: the above screenshot shows the screen for Democrats in Study 1a, where we measured support for Black Lives Matter; for Republicans in Study 1b, we referenced the Republican party (rather than the Democrat party) and measured support for Blue Lives Matter (“To what extent do you support "Blue Lives Matter" (a countermovement, started in response to Black Lives Matter, advocating that those who are prosecuted and convicted of killing law enforcement officers should be sentenced under hate crime statutes)?”).