# Immoral stereotype, or acceptable generalization? Beliefs about stereotypes shape their moral evaluation and use

Thalia H. Vrantsidis,[1] William A. Cunningam[2]

[1] Princeton University, Program in Cognitive Science and University Center for Human Values.
[2] University of Toronto, Psychology Department.

Preprint of manuscript submitted to Science Advances

## One sentence teaser

Problematic stereotype use may stem in part from people thinking they are merely using acceptable group-based generalizations.

## Abstract

People sometimes use and express stereotypes, even when others consider this morally problematic. We examined an under-explored explanation for this: people's beliefs about a stereotype (e.g., how overgeneralized or likely to incite discrimination it is) may lead them to see it as merely an acceptable group-based generalization, and thus freely use it, while others with different beliefs consider this immoral stereotyping. Five studies supported this explanation, using belief differences across political groups as a case study. Results showed that, despite largely using the same general moral principles, liberals and conservatives differed in their beliefs about, and moral evaluations of, specific group-based generalizations, with these beliefs affecting willingness to use those generalizations. Furthermore, a predicted consequence of these differences was observed: people viewed problematic stereotyping as primarily done by others, especially political outgroups, rather than themselves. This work provides novel insights into why people use problematic stereotypes, and informs stereotype-reduction strategies.

*Keywords*: stereotypes; stereotyping; morality; acceptability; social categories; beliefs

## Introduction

*Far more serious in the modern world than any difference of moral code is the difference in the assumptions about facts to which the code is applied.*

-Walter Lippmann, 1922

Despite apparently widespread moral norms against stereotyping (*1–4*), people continue to use and express stereotypes. For example, gender stereotypes often influence hiring decisions, so that women are less likely to be hired for male-dominated jobs (e.g., jobs in STEM fields: science, technology, engineering and math (*5*, *6*)). Furthermore, people often express explicit gender stereotypes in line with these biases (e.g., reporting that men tend to have more natural scientific aptitude compared to women (*7*)) The use of these stereotypes occurs despite the fact that it may be met with criticisms of being offensive, sexist, discriminatory, or even illegal (*2*, *8*).

What leads people to use and express stereotypes like these, in spite of the moral criticisms this may illicit? Past research has largely explained such stereotype use in two ways. First, people may use stereotypes unintentionally, such as when stereotypes subtly influence reaction times, non-verbal behavior, or complex tasks like hiring decisions (*9–13*). Second, people may use stereotypes due to competing, and often unsavory, intentions (e.g., justifying inequalities, boosting one's self-esteem, cognitive laziness, etc. (*6, 14–18*)). Adding to these explanations, here we highlight another important reason that people may use stereotypes, even when others find this morally problematic. This may happen precisely because people do not think they are using morally problematic stereotypes, but instead think they are merely using acceptable group-based generalizations (e.g., akin to "German people speak German"). Thus, people may freely use a generalization – perhaps even intentionally and without conflicting intentions – and think there is nothing wrong with this, and perhaps that are not even 'stereotyping'. Yet, at the same time, other people might disagree, and see this person as using an immoral stereotype.

While these types of moral disagreements can have many sources (*6*), the current work focuses on one source in particular: differences in people's other beliefs about a generalization (e.g., how overgeneralized, harmful, or widespread it is), which may be used to decide whether that generalization counts as immoral stereotyping. When people differ in these beliefs, it means that two people could be following the same general moral principles yet disagree on the "facts" about a generalization to which their moral principles are applied, and thus disagree on whether that specific generalization counts as immoral.

The current work investigates this proposed process: where different beliefs may lead people to different moral judgments of a generalization, and thus lead people to freely use what they see as merely acceptable generalizations, even while others see this as immoral stereotyping – perhaps all while using the same general moral principles. Investigating this process can help more fully understand the reasons people use problematic stereotypes, despite the moral and social issues associated with them. This can shed light on cases of stereotyping not captured by previous theories – e.g., cases when people intentionally use problematic stereotypes, despite having no ill-intentions – as well as provide a fuller understanding of the factors that promote and reduce stereotype use in cases where unintentional biases or ill-intentions are also involved.

**Causes and consequences of disagreements over whether a generalization counts as immoral stereotyping**

At the core of the proposed process is the idea that people may disagree over whether a generalization counts as immoral stereotyping. To understand how this may occur, it is important to understand how people make these decisions about whether a generalization counts as immoral stereotyping or not.

One possibility is that people may think that all group-based generalizations count as immoral stereotyping. This view relies on defining 'stereotypes' simply as beliefs about social groups, and 'stereotyping' as making generalizations based on these beliefs (as in some social psychological definitions (*9, 19, 20*). Furthermore, this view involves seeing all such group-based generalizations as immoral – e.g., due to violating fundamental principles of justice that require treating people as individuals (*1, 3, 21*). Under such a view, there is little room for reasonable disagreement over whether a group-based generalization counts as immoral stereotyping, since all of them should count. It therefore follows that any use of stereotypes can be explained by either unintentional stereotype use, or competing ill-intentions. Indeed, previous

research may largely have assumed this view, as past work has largely focused on these two types of explanations for problematic stereotype use.

However, we suggest that this view does not capture the more nuanced way people may typically decide if generalizations count as immoral, and as 'stereotyping.' In particular, we suggest people do not typically see all group-based generalizations as immoral. Instead, some generalizations may be seen as morally neutral (e.g., thinking that someone who lives in Germany likely speaks German), while others may be seen as morally beneficial, if they lead to prosocial responses (e.g., recognizing that a group has been unfairly disadvantaged, if it motivates people to help that group). Furthermore, we suggest that lay people may tend to restrict the term 'stereotype' to refer to more problematic cases (*22*). This use of the term would be in line with other definitions within social psychology, which characterize stereotypes specifically as certain, more problematic, beliefs about groups: e.g., those that are overgeneralized, exaggerated, rigid, culturally widespread, and/or used to justify prejudice (*23–25*; see *19*, *26*) (To avoid confusion, herein we will use the terms 'stereotype' and 'stereotyping' in this narrower, more negative, sense, and 'group-based generalization' as a broader, more neutral, term.)

Importantly, if people endorse this alternative view, then deciding whether a generalization counts as immoral stereotyping is a more nuanced decision process, with greater potential for disagreements. To make these decisions, people may need to rely on their various beliefs about a generalization, including its content, context, and use. For example, people may consider beliefs about the various factors identified in the definitions above (e.g., how overgeneralized, exaggerated, etc., a generalization is), and perhaps even beliefs about a range of other factors (e.g., how disadvantaged the target group is, the chance of inciting discrimination towards that group, how much harm the generalization is likely to cause). Thus, whenever these types of beliefs differ between people, people might come to disagree over whether a generalization counts as immoral stereotyping or merely an acceptable group-based generalization.

These disagreements have several important consequences. As suggested above, even though people may typically avoid using generalizations they see as problematic, they might freely use other generalizations that they see as acceptable – perhaps even using these intentionally and without competing ill-intentions. Thus, when disagreements occur, people might end up freely using generalizations that others see as immoral stereotyping. Furthermore, these disagreements may lead to a form of "bias blind-spot" (*27–29*), where people tend not to think they use immoral stereotypes, but instead see this as something done mainly by other people. Thus, even those who might be criticized for using problematic stereotypes may not think they are doing this at all, and, perhaps, even think it is their critics who use problematic stereotypes, not themselves.

**Political differences as an initial case study**

As an initial case study, we focus on how the process proposed here may occur across political divides, particularly in the United States. Initial evidence suggests that political differences may be associated with differences in the beliefs about, evaluations of, and use of group-based generalizations. For example, liberals and conservatives tend to differ in how acceptable they think it is to use racial stereotypes, as well as their endorsement of (i.e., beliefs about/use of) certain racial and gender stereotypes (*7*). Liberals and conservatives have also been shown to hold divergent beliefs on politically relevant topics, such as the causes for economic inequalities (*30*), which might contribute to different evaluations of certain stereotypes. These

types of diverging beliefs could be created or reinforced by differences in the information liberals and conservatives tend to be exposed to (e.g., through media, social media, and in-person interactions (*31*–*34*)), or differences in the motivations and values that tend to be held by liberals and conservatives (*35*), which could affect beliefs through motivated reasoning processes.

While the current work does not examine the sources of these belief differences, it instead focuses on testing more broadly whether such belief differences exist, and then on examining the consequences of these differences. We extend previous research in several ways. First, we greatly expand the range of generalizations and beliefs studied. Second, we link these beliefs directly to their consequences for people's evaluations and use of group-based generalizations. Third, we also test for another consequence of these belief differences: the presence of a novel form of 'bias blind-spot', where people primarily think that stereotyping is done by other people – perhaps especially political outgroups – and not themselves. A series of five studies examined different aspects of this process.

Though the processes discussed here apply well beyond the political domain, they may be especially important to understand in this domain, given increasingly polarized and antagonistic political climates (*36*, *37*). When criticisms of stereotyping occur across political lines, this may further increase antagonism between political groups, and create backlash that exacerbates existing political conflicts (*38*). One possible way to address these issues is to recognize when the use of problematic stereotypes stems from differences in people's beliefs, rather than differences in their broader moral values, or even a complete disregard for moral values. In such cases, one might be able to effectively confront stereotyping by addressing these belief differences, while promoting mutual respect by acknowledging shared moral values that bridge across political divides. The current work therefore also examined if there might indeed be widely shared moral principles used to evaluate stereotyping, that are used across both liberals and conservatives. We return to this point at the end of the five studies.

## Results

### People think problematic stereotyping primarily occurs in other people, especially in those with different political views, and not themselves (Study 1)

As discussed, the current perspective predicts that people will show a novel form of bias blind-spot, where they think that stereotyping (or at least problematic forms of stereotyping – e.g., immoral or inaccurate stereotyping) tends to occur primarily in others, not themselves. This prediction stems from two ideas. The first idea is that people generally try to avoid using generalizations they deem problematic, which should lead people to think that they largely do *not* use problematic stereotypes. The second idea is that people will continue to use other generalizations that they deem acceptable, which other people, especially those who hold differing beliefs about the generalization, will sometimes view as problematic. This should lead people to think that stereotyping continues to exist, but primarily in other people – and perhaps especially in members of opposing political groups, who may tend to differ more in their beliefs, and thus disagree more about which generalizations count as problematic.

Study 1 tested this prediction, by examining how often people thought that they intentionally used problematic generalizations (i.e., 'stereotypes', 'inaccurate stereotypes', or 'immoral stereotypes'), and compared this to how much they thought others with the same or with opposing political views did this. The hypotheses, design, sample size/stopping criteria, exclusion criteria and analysis plan for this study were all preregistered.

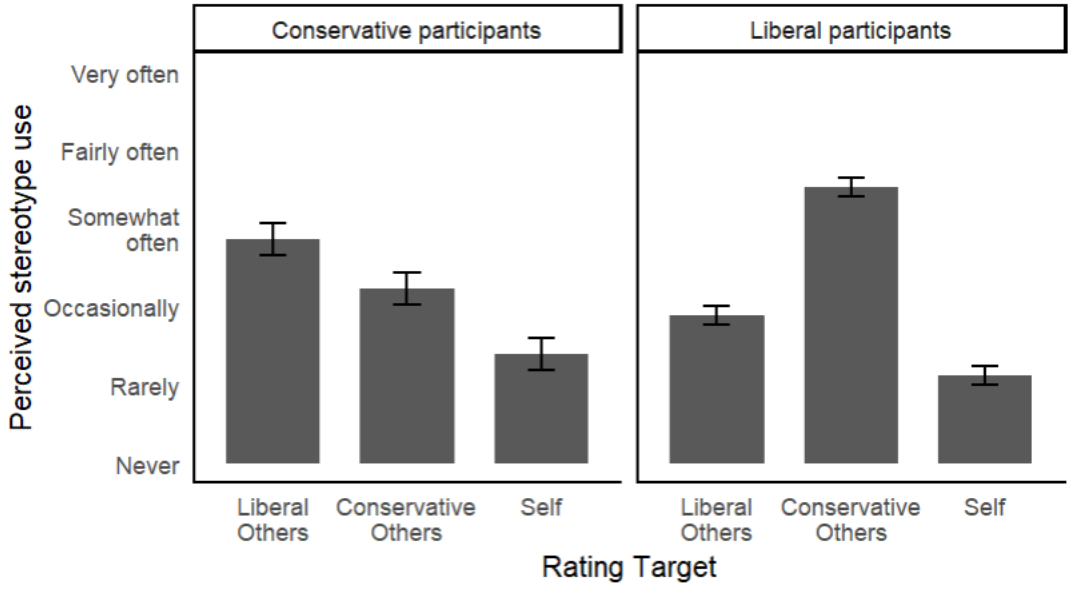
Results were similar across wording conditions (i.e., describing the generalizations as "stereotypes", "inaccurate stereotypes" or "immoral stereotypes"). Results below are therefore

collapsed across conditions (see supplementary materials for full results). In line with predictions, on average, participants reported that they personally used stereotypes only infrequently – between 'rarely' and 'occasionally' ($M$ = 1.20, $SD$ = 0.86) – while they viewed other people as using stereotypes more often ($M$ = 2.67, $SD$ = 1.28; $r$ = 0.51, $p$ < .001). Furthermore, perceptions depended on the political orientation of the participant and of the other people being rated (interaction: $F(2, 806) = 157.96$, $p < .001$). As show in Figure 1, both liberal and conservative participants tended to view others with opposing political views as using stereotypes most, more than others with the same political views as the participant, who in turn were thought to do so more than oneself.

These perceived differences between self and others' stereotype use are broadly consistent with previous findings on "bias blind spots," including findings that people tend to think of themselves as less biased, more objective, morally better, and less likely to use specific stereotypes, compared to other people (*27*, *29*, *39–41*). Furthermore, the perceived differences in political in- and out-group members' stereotype use are broadly consistent with findings that Americans often have quite negative attitudes about those with opposing political views (*36*, *37*, *42*). The current results extend this previous work by showing that these self/other and political in/outgroup asymmetries apply to judgments of how often people use stereotypes in general – which, importantly, can depend crucially on which generalizations people view as problematic stereotyping. Thus, while multiple factors could have led to the asymmetries observed here (e.g., a desire to view oneself or one's ingroup positively (*43–45*)), the subsequent studies test whether these findings might be explained by the processes proposed here: specifically, differences in what generalizations people consider to be problematic stereotyping.

**Figure 1**

*Perceptions of How Often Oneself and Other People use Stereotypes*




*Note.* The stereotype use ratings reflect averages across wording conditions ('stereotypes', 'immoral stereotypes', 'inaccurate stereotypes'). Error bars reflect 95% confidence intervals.

**Not all group-based generalizations are seen as immoral or as stereotypes (Study 2)**

The current explanation for Study 1's results relies on the assumption that people do not think all group-based generalizations are immoral, but instead distinguish between morally acceptable and morally problematic cases, and perhaps largely consider the problematic cases to be 'stereotyping'. This would allow people to disagree on which generalizations count as immoral/stereotyping, and to freely use some group-based generalizations that they view as acceptable.

Study 2 therefore tested this assumption, by asking participants how often they think it is ok to make generalizations based on groups, and by comparing this to the equivalent question about generalizations based on *stereotypes*. Two versions of each question were asked, involving either generalizations about individuals or about groups as a whole. (An additional preregistered study, Study S1, asked about how often it is *morally* ok to use generalizations or stereotypes, and replicated the current results. See supplementary materials for details.)

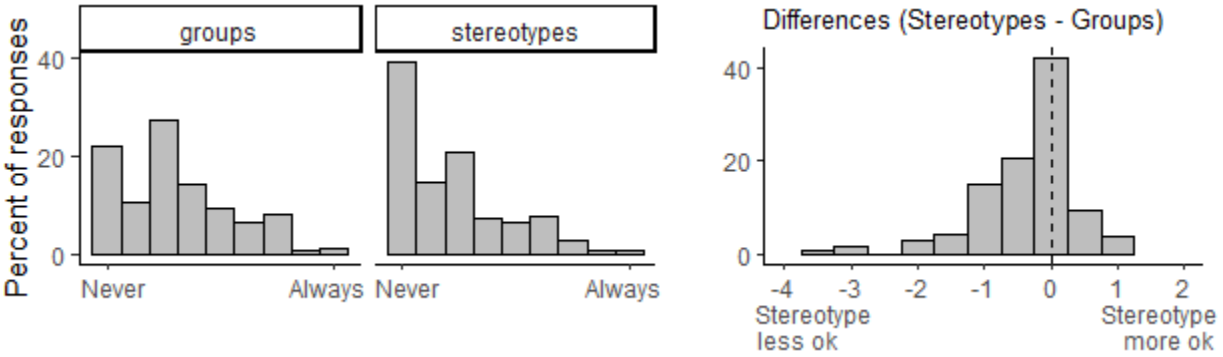
In line with the assumption that people distinguish between acceptable and unacceptable group-based generalizations, most participants thought that generalizations based on people's groups were sometimes ok, rather than never ok, or always ok (76.92% of people responded to at least one version of this question with an answer other than 'never' or 'always'; Figure 2). Furthermore, many (though not all) participants appeared to distinguish between generalizations based on 'stereotypes', and based on 'groups', with those based on 'stereotypes' seen as more problematic. This can be seen in the fact that, on average, stereotype-based generalizations were rated as less often ok than group-based generalizations ($M_{\text{stereotype - group}} = -0.36$; $SD_{\text{stereotype - group}} = 0.91$, $d = -.40$, $p < .001$; with 45% of participants showing average differences in this direction). Study S1 replicated the results of Study 1 for both liberal and conservative participants, and showed that, despite some small differences, both liberal and conservative participants gave largely similar ratings (see supplementary materials).

The fact that, in these studies, participants tended to think that only some group-based generalizations were unacceptable, immoral, or 'stereotypes' implies that people need to decide where to draw the line between different types of generalizations. This allows for the possibility that people might differ in how they evaluate a given case, and thus freely use what they see as merely acceptable group-based generalizations, even while other people see those as immoral stereotypes. The following studies examined if and how this may occur.

Note that the following studies focus primarily on differences in moral judgments of group-based generalizations, rather than differences in how much they are considered 'stereotypes', since the results of Study 2 suggest that there may be variability in how people use the term 'stereotype', while moral judgments may be more unambiguous.

**Figure 2**

*Perceptions of How Often it is Ok to Make Generalizations Based on People's Groups, or Based on Stereotypes*




*Note.* Left: Histogram of responses. Right: Histogram of differences between ratings of generalizations based on stereotypes, and those based on groups. Responses in all plots reflect individual participants' responses averaged over the two versions of each question (about individuals, or groups as a whole).

**Differences in people's beliefs can lead to differences in their moral judgments of a given group-based generalization (Study 3)**

The fact that people distinguish between acceptable and unacceptable group-based generalizations raises the question of if and how people might come to differ in their evaluations of a given case. We propose that these types of differences do in fact occur (focusing on differences in moral judgments here), and, furthermore, that these differences can be caused by differences in people's other beliefs about that generalization that might influence these moral judgments. Study 3 tested this by examining differences associated with political orientation. Specifically, it tested whether liberals and conservatives would differ in their moral evaluations of, and beliefs about, a given generalization, and whether these moral differences could be explained by differences in these other beliefs.

Participants imagined that someone made various statements containing group-based generalizations (e.g., "Artists are creative", "Black people are good at sports", or "Muslims are terrorists"; see Figure 3), and then reported their perceptions of how immoral each statement was, as well as various other beliefs about each statement that might influence their moral judgments (see Table 1). The beliefs reported were: two aspects of how overgeneralized the statements are (the percent of the group with the trait, and the percent of the group the speaker thought had the trait), how much the statement was based on a stereotype, how likely the statement was to incite discrimination against the group, how much the statement reflected widespread cultural or media representations of the group, how disadvantaged the group was, and the size of the group (i.e., what percent of society was part of the group).

These beliefs were selected because they were expected to affect judgments of how immoral each generalization was. For example, previous work suggests that more overgeneralized statements can be seen as more problematic (*8*). As per study 2, being a 'stereotype' may also imply that a generalization has various problematic qualities. Other beliefs here were expected to affect the perceived negative impact of the statement on the target group. Specifically, statements seen as more likely to incite discrimination should be seen as having more negative impact on the group; statements reflecting widespread cultural/media representations might be seen as more able to convince others to believe and act on the statement, thus have more potential negative impacts on the group; and disadvantaged or

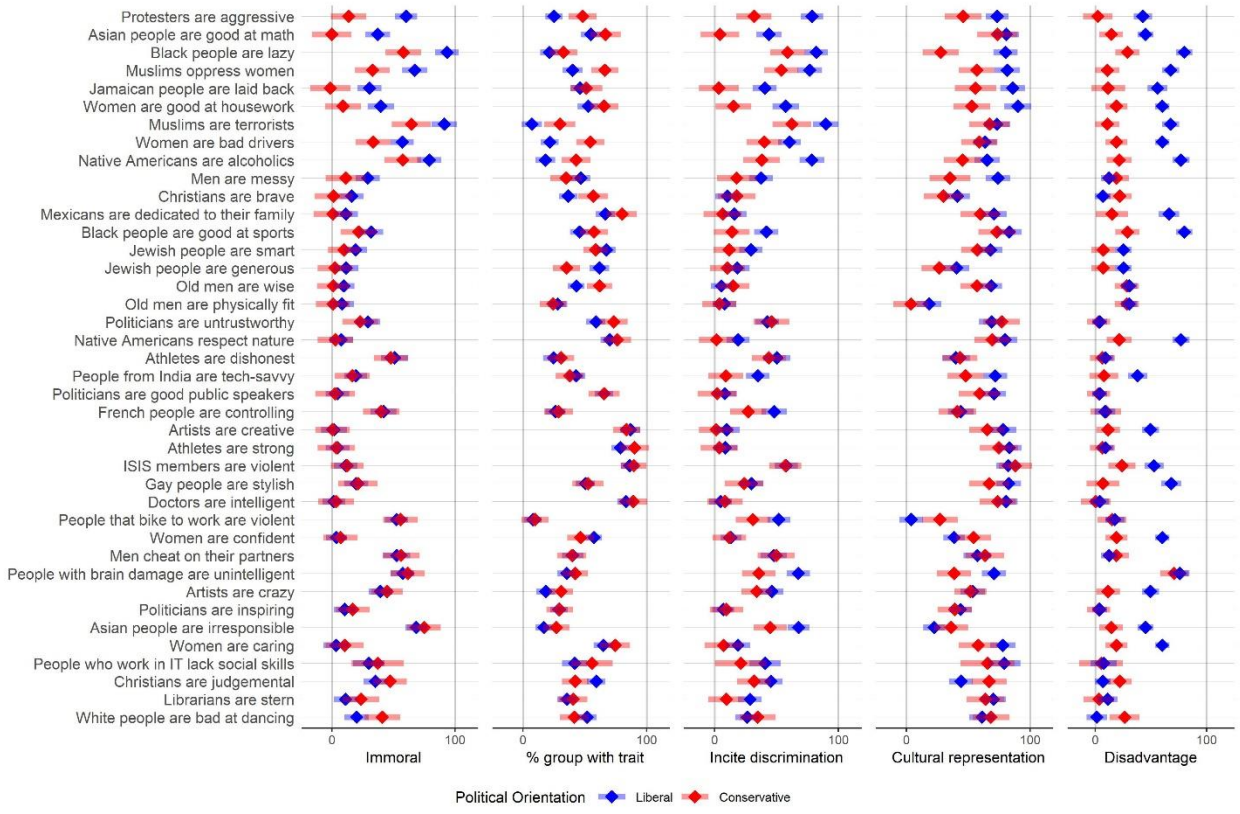
minority groups might be seen as more vulnerable to any negative consequences of the statement (*2*, *46*). At the end of the study, participants reported their political orientation.

***Political differences in moral judgments of, and beliefs about, group-based generalizations***

We first examined whether liberal and conservative participants would differ systematically in how immoral a given statement seemed. Political orientation interacted with the specific statement to predict moral judgments, $F(39, 2562) = 2.37$, $p < .001$ (with no main effect of political orientation, $F(1, 139) = 3.24$, $p = 0.07$), implying that liberals and conservatives tended to differ in how immoral they thought the statements were, in a way that depended on the specific statement. As seen in Figure 3, liberals rated several of the statements as more immoral than conservatives, such as "Protestors are aggressive", "Asian people are good at math" and "Black people are lazy". On the other hand, conservatives also rated some statements as more immoral than liberals, such as "White people are bad at dancing", "Librarians are stern", and "Christians are judgmental", though differences in this direction were smaller and less frequent (perhaps due to the specific statements used in this study).

**Figure 3**

*Differences between Liberals and Conservatives in Moral Evaluations and Beliefs about the Group-based Generalizations used in Study 3*




*Note.* Plots display ratings and 95% CIs, estimated for the most extremely liberal and extremely conservative participants in order to make differences are more easily visible. Statements are ordered based on the size of the difference in moral evaluations (liberal minus conservative). "Immoral" = ratings of how immoral the statement is. "% group with trait" = ratings of the percent of the group with the trait. "Incite discrimination" = ratings of how likely the statement is to incite discrimination. "Cultural representation" = ratings of how much the statement reflects

widespread cultural or media representations of a group. "Disadvantage" = ratings of how disadvantaged the group is in society.

We then examined whether liberals and conservatives might show differences in their other beliefs about these statements, beliefs which may be used to form these moral judgments. In many cases, there were significant differences in liberals' and conservatives' beliefs about these statements, again in a way that depended on the specific statement (see Figure 3). Significant interactions were observed for beliefs about: the percent of the group with the trait ($F(39, 2578) = 2.39$, $p < .001$), the statement's likelihood of inciting discrimination ($F(39, 2560) = 2.27$, $p < .001$), how much it reflected a widespread cultural or media representation ($F(39, 2567) = 2.51$, $p < .001$), and how disadvantaged the group was ($F(24, 2018) = 9.13$, $p < .001$); though not for beliefs about how stereotype-based a statement was or the size of the group.

Examining the pattern of results in Figure 3 provides some initial indication that these belief differences may help account for the observed differences in liberals' and conservatives' moral judgments. For example, for the statement "Protesters are aggressive", more liberal participants tended to find this statement more immoral, and they also tended to have beliefs that could supporting this: tending to think that fewer protestors were in fact aggressive, that this statement was more likely to incite discrimination, that protestors are more disadvantaged in society, and that this statement is more reflective of widespread cultural or media representations of protesters.

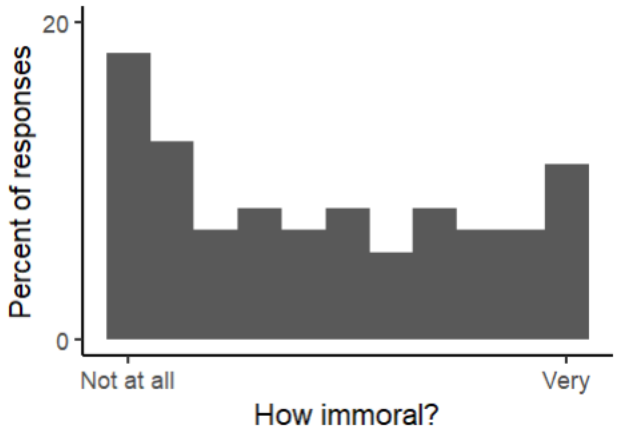
This relationship was tested more formally by examining whether observed differences in moral judgments (specifically, the interaction effect of statement and political orientation on moral judgments) was statistically mediated by these beliefs. The results indicated partial mediation: the indirect effect through these beliefs (combined across all beliefs and statements) was significant, $X^2(39) = 157.96$, $p < .001$, though the direct effect of the interaction on moral judgments continued to be significant after accounting for these beliefs, $X^2(39) = 120.33$, $p < .001$. This indicates that the measured beliefs statistically accounted for part of the observed differences in liberals' and conservatives' moral judgments of a given statement. This result is consistent with the idea that these belief differences may have causally contributed to differences in these moral judgments.

***Idiosyncratic differences in moral judgments of, and beliefs about, group-based generalizations***

Political orientation is not the only reason people may differ in their beliefs about, and thus moral judgments of, a given generalization. Indeed, people's beliefs may vary for all sorts of reasons (e.g., different experiences with a group, exposure to different media). We therefore also examined differences in these beliefs and moral judgments, regardless of political orientation. Looking at individuals' moral judgments of a given statement highlighted again how widely these judgments can differ. For example, as seen in Figure 4, the same statement could be seen as not at all immoral by some participants, and very immoral by others (see supplementary materials for all statements; average within-statement range of the middle 95% of responses = 77.59 on a 100-point scale).

**Figure 4**

*Example Distribution of how Immoral Participants thought a Given Statement was.*




*Note.* Histogram of all participants' immorality ratings for the statement "Protesters are aggressive".

We also examined whether more idiosyncratic variation in beliefs about a given generalization was related to, and thus might contribute to, variation in how immoral that generalization seemed. To isolate more idiosyncratic sources of variation, this analysis controlled for political orientation, the specific statement, and the interaction of these. Results showed that these idiosyncratic differences in beliefs about a given statement were in associated with differences in moral judgments of that statement, with most of the measured beliefs independently contributing to predicting these moral judgements (see Table 1). Specifically, the same statement was seen as more immoral when participants thought it was: true of a smaller percent of the group, seen by the speaker as true of a larger percent of the group, more stereotype-based, more likely to incite discrimination, more reflective of cultural or media representations, and about a more disadvantaged group. (Group size was not significant.) These results therefore further show that people meaningfully differ in their moral evaluations of specific group-based generalizations and are consistent with the idea that differences in people's other beliefs might contribute to these types of moral disagreements.

**Table 1**

*Idiosyncratic Differences in Beliefs about a Group-Based Generalization Predict Differences in how Immoral the Generalization seems*

| Belief | *β* | P |
|---|---|---|
| What percent of the group the trait applies to | -.22 | **< .001** |
| What percent of group the speaker thinks the trait applies to | .03 | **.04** |
| How stereotype-based the statement is | .06 | **< .001** |
| How likely to incite discrimination the statement is | .36 | **< .001** |
| How reflective of widespread cultural/media representations the statement is | .03 | **.04** |
| How disadvantaged the group is | .04 | **.004** |
| What percent of society is in the group | .02 | .40 |

*Note.* Positive coefficients indicate that higher values of the measured belief are associated with judging the generalization as more immoral, after controlling for all other measured beliefs. Significant p-values are marked in bold.

***Summary of Study 3***

The results of Study 3 showed that, in many cases, how immoral a generalization seems is not widely accepted, instead differing across individuals, and between liberals and conservatives. These differences in moral judgements were in turn related to differences in people's other beliefs about a given generalization – beliefs that can differ idiosyncratically between individuals, and systematically across political orientations. These results are thus consistent with the idea that differences in beliefs can lead one person to see a generalization as morally acceptable (and thus perhaps freely use it), even while other people who hold different beliefs might see it as immoral. Furthermore, the systematic political differences observed in the current study suggest that this is more likely to occur when people hold different (vs. similar) political orientations, consistent with our explanation of the Study 1 finding, where political outgroup members are seen using stereotypes more often than political ingroup members. These results thus extend previous findings of differences in liberals and conservatives' beliefs about and evaluations of group-based generalization (*7*, *30*), by showing that this holds for a broader array of beliefs and a broader array of generalizations, as well as by providing initial evidence that these belief differences may contribute these differences in moral judgments.

The current results also highlight the multifaceted, nuanced nature of moral judgments in this domain, and thus how easy it may be for these moral disagreements to arise, even among people who may largely see eye to eye. In particular, the current results identified a wide range of beliefs that all appeared to contribute to these moral judgments (by explaining independent variance in these judgments; see Table 1). Therefore, for example, two people might agree on the moral norms that should be used to evaluate group-based generalizations, and even agree on most of their beliefs about a specific generalization, yet if they disagree on one of these beliefs (say, about how disadvantaged the group is), they may still differ in their moral judgments. These results thus help understand the some of the many ways people might come to disagree on the morality of a group-based generalization, both across individuals and across political divides.

**Establishing causal direction (Study 4)**

The results of study 3 suggest that differences in people's beliefs about a generalization may lead them to differ in whether they think that generalization is morally acceptable or not. Yet the previous study was correlational, and thus cannot establish whether these belief differences actually caused people's moral judgments to differ. In contrast, it is possible that people's moral judgments are driven purely by other factors, and that these other factors, or the moral judgments themselves, drive people to hold corresponding beliefs. Effects of this type could occur through a variety of processes, including rationalizing one's past behaviors, justifying one's intuitive gut feelings, or internalizing social norms (*47–49*). Without casting doubt on the importance of these other processes, Study 4 aimed to more directly test whether people's beliefs about a given generalization can causally affect how immoral people think it is. This was tested by using scenarios about artificial groups, which allow these beliefs to be easily manipulated, and then examining how this affected moral judgments. (Additional results from Study 4 addressing different questions are presented in a later section.)

Participants read scenarios where someone makes a group-based generalization (e.g., saying, "Oh, that guy is a Lupite. He is probably short-tempered"). Participants then evaluated how morally good or bad this statement was. The scenarios were manipulated to vary five factors which might affect these moral judgments: two aspects of how overgeneralized the statement was – 1) the percent of the group members that actually had the trait, and 2) how strongly the

statement was worded (roughly akin to the percent of group members that the speaker thinks had the trait, as asked in Study 3) – as well as 3) if the trait was positive or negative, 4) if the group was disadvantaged or advantaged, and 5) the size of the group (if it was a minority or majority of the society).

Results indicated that people's manipulated beliefs about a generalization could causally influence their moral judgments. Statements were seen as more immoral when the statement was more overgeneralized, in the sense of a smaller percent of the group having the trait ($\beta = 0.06$, $p = .002$), as well as when the trait was negative rather than positive ($\beta = 0.47$, $p < .001$). The overgeneralization effect here provides particularly clear support for the idea that people's beliefs about a generalization influence their moral judgment, since the actual statement was identical in both cases, and it was only people's background knowledge about the group that varied. The other three manipulations did not have significant effects (strength of statement wording: $\beta = 0.02$, $p = .34$; disadvantage: $\beta = 0.03$, $p = .16$; group size: $\beta = 0.02$, $p = .44$).

The specific pattern of results here is fairly consistent with Study 3, in that both studies showed larger effects for the same aspect of overgeneralization (i.e., the percent of the group with the trait), smaller effects for some of the other beliefs (e.g., disadvantage, strength of statement working/the speaker's view of the percent of the group with the trait), and non-significant effects for group size. However, the current study had generally smaller/less significant effects, which may stem from limitations on people's attention due to manipulating all five factors at once. Indeed, previous work shows that type of setup can lead people to attend only to the most influential factor(s) (*50*) – which in this case, appeared to be the valence of the trait. Despite this limitation, the current results support the idea that people's beliefs about a statement, including about how overgeneralized and how negative it is, can causally influence their moral judgments. This supports the current view that differences in these types of beliefs can cause one person to think a generalization is acceptable, while other people with different beliefs view it as immoral.

**Differences in beliefs can lead people to use group-based generalizations that are deemed immoral stereotypes by others (Study 5)**

The current view predicts not only that belief differences will lead to differences in moral judgments, but also that these should lead to corresponding differences in people's willingness to use a generalization. That is, when people's beliefs lead them to see a generalization as morally acceptable, they should also be more likely to freely use the generalization, while people with different beliefs may see it as a case of immoral stereotyping and avoid using it. Study 5 tested this prediction by manipulating beliefs about how overgeneralized a generalization is and examining people's willingness to explicitly make that generalization themselves. These beliefs were manipulated through presenting summaries of actual scientific research suggesting that there either were, or were not, real group differences in line with a particular cultural stereotype (e.g., that boys do better than girls at math). We then examined whether people would use this cultural stereotype themselves (e.g., inferring that a particular boy is more likely to be doing better at math than a particular girl), rather than avoiding using it (e.g., inferring equal likelihood for the boy and girl) or perhaps using the counter-stereotype (e.g., inferring the girl is more likely to be doing better).

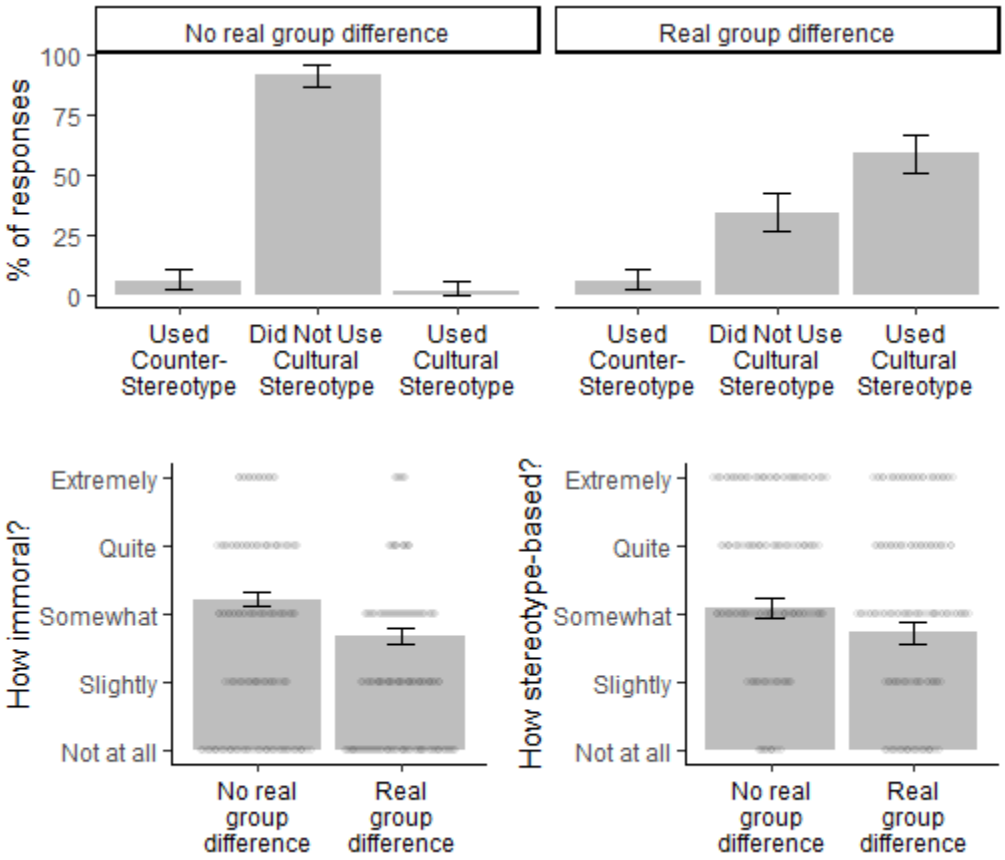
In addition to examining willingness to use a generalization, Study 5 also tested whether this belief manipulation would affect how immoral and stereotype-based people think this generalization is. This would extend Study 4 by replicating the effect on moral judgments using real social groups, and by also investigating perceptions of 'stereotype' use (in line with the idea

that morally worse generalizations should also tend to be seen as more based on stereotypes). This study was fully preregistered, including the hypotheses, design, sample size, exclusion criteria, and analysis plan.

We examined if manipulating people's beliefs about whether there were real group differences affected their use of generalizations based on those differences, as well as judgments of how immoral and stereotype-based this generalization would be. As hypothesized, participants who were told there were (vs. were not) real group differences in line with a cultural stereotype were more likely to use the cultural stereotype themselves ($B$ = -1.21, $p$ < .001; see Figure 5). Furthermore, being told that there were (vs. were not) real group differences led people to see generalizations based on these differences as less immoral ($B$ = -0.27, $p$ < .001), and less based on a stereotype ($B$ = -0.18, $p$ < .001). An exploratory mediation analysis showed that moral judgments partially mediated the observed changes in use of the generalization (proportion mediated = 0.04, $p$ = .002), and ratings of how stereotype-based it was (proportion mediated = 0.57, $p$ = .002). An additional study replicated most of the results of Study 5 using a variation on these methods, further supporting the robustness of these conclusions (see Supplementary Materials, Study S2).

**Figure 5**

*Effects of Manipulating Beliefs about whether there were Real Group Differences on People's Use of, and Perceptions, of Cultural-Stereotype-Based Generalizations*




*Note.* Effects of belief manipulation (saying that there was vs. was not a real group difference) on: (top) use of the corresponding cultural stereotype, (bottom left) perceptions of how immoral a generalization based on the cultural stereotype would be, and (bottom right) perceptions of how

stereotype-based this generalization would be. All plots display post-manipulation responses with error bars indicating 95% CIs.

These results therefore imply that changing people's beliefs about a generalization (i.e., about the extent to which it is overgeneralized, vs. reflects real group differences) can causally affect their willingness to use that generalization, along with their evaluations of how immoral and stereotype-based it is. These results also provide evidence for the full process proposed in the current paper: where differences in people's beliefs led to some people freely using a generalization, without seeing this as especially immoral or stereotype-based, even while other people with different beliefs tended to avoid using it and were more likely to see this as immoral and stereotype-based.

**Estimates of undeserved harm underlie moral judgments of group-based generalizations (Additional Study 4 results)**

The previous studies highlight an important challenge in fighting problematic cases of stereotyping: the fact that people may not necessarily agree on what they should be fighting, that is, on which generalizations are problematic. One could take this to mean that these disputes are fundamentally unresolvable, and that there is no basis for saying that one person's view is right and one person's is wrong. Indeed, in the political realm, some previous work has suggested that liberals and conservatives may hold fundamentally different moral values (*51*, *52*). This could imply that when these disagreements occur across political lines, there may be no way for people on both sides to agree on whether a generalization is morally problematic or not. However, this conclusion risks undermining attempts to reduce problematic cases of stereotype use, by justifying things like racist, sexist or prejudiced views as simply different but equally valid opinions.

One way to avoid this conclusion, and provide a basis for resolving these moral disagreements, would be to identify widely shared moral principles used to distinguish immoral from acceptable generalizations. If such principles exist, it provides hope that, through discussion, people may be able to come to a consensus on whether something in fact violates these shared moral principles. We therefore examined whether, despite disagreeing on the morality of specific cases, people might typically evaluate group-based generalizations based on widely shared moral principles, principles which are held across people of different political orientations.

The specific principle that we propose fills this role is the *undeserved harm principle*. Specifically, this principle says that one should evaluate how immoral a group-based generalization is based on the extent to which it is expected to cause undeserved or unjustified harm. This principle is in part an application of harm-based theories of morality (*53*), which have been argued more generally to underlie apparent differences in liberals' and conservatives' moral values (*53*, *54*). Here, harm is defined broadly to include things like physical, economic and emotional harms, as well as lack of, or loss of, benefits. Based on other work (e.g., *60*, *61*), the current principle also includes the idea that some harms may be seen as deserved or justified, and thus morally ok (e.g., people who work less deserve less pay, criminals deserve to be punished).

To estimate undeserved harm for a given generalization, we suggest that people often rely on their various beliefs about that generalization, such as those identified in the previous studies. For example, generalizations that are seen as negative or inciting discrimination may tend to be seen as especially harmful, while those that are seen as more overgeneralized may tend to be seen as having especially undeserved consequences (e.g., assuming that all Muslims are terrorists

will likely lead to treating many innocent Muslims in ways they do not deserve). Other factors like seeing a group as disadvantaged might be thought to make a group more vulnerable, thus amplifying any potential harms of the generalization. If people use these various beliefs to estimate undeserved harm, then it follows that differences in these beliefs could lead people to have different estimates of undeserved harm, and thus different moral judgments of a given generalization – even while using the same moral principles.

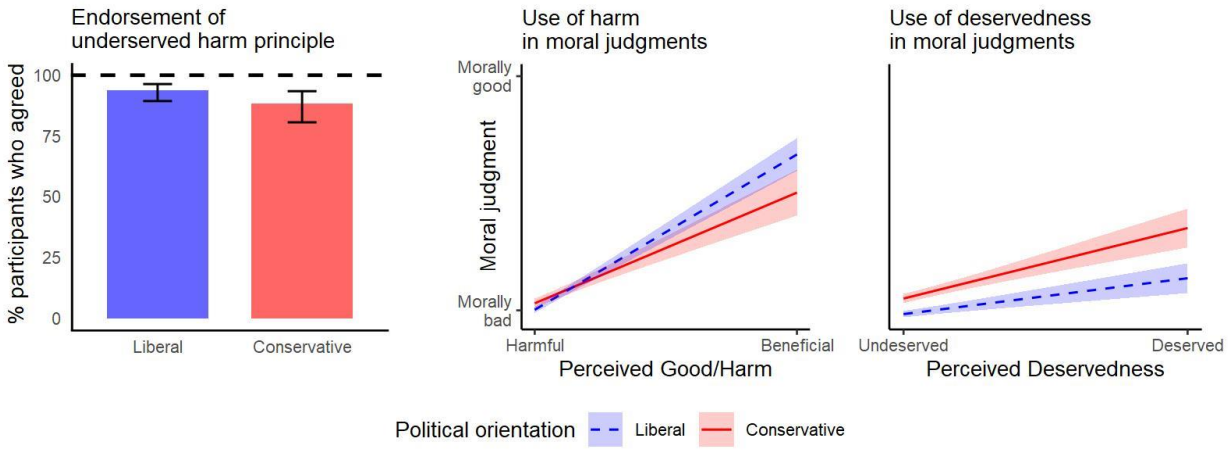
We tested whether the undeserved harm principle might be widely endorsed and applied as a way to judge the morality of group-based generalizations, including across people of differing political orientations. We also tested whether other beliefs found to affect these moral judgments do so through affecting perceptions of undeserved harm. If these two results are observed, it would help understand how people might end up with different moral judgments about a given generalization even while using the same moral principles, and also highlight moral common ground that may help resolve these differences.

These questions were examined using additional aspects of the data from Study 4. For the scenarios in Study 4, in addition to making morality ratings about each group-based generalization, participants also rated their perceptions of undeserved harm: specifically, how much good or harm the generalization was likely to cause, and how much anyone affected by the generalization deserves what happens because of it. A portion of participants also reported their agreement with using the undeserved harm principle to evaluate group-based generalizations and reported their political orientation.

Results provided support for the undeserved harm principle as being widely used to evaluate group-based generalizations. There was widespread endorsement of this principle, which held across political orientations, with 94% of liberal participants and 88% of conservative participants agreeing that expectations of underserved harm should affect how morally bad a group-based inference is (Figure 6, right). This principle also appeared to be applied to evaluate the specific group-based generalizations in the scenarios. Specifically, statements were seen as more immoral when they were seen as causing more harm ($\beta = 0.66$, $p < .001$) and more undeserved consequences ($\beta = 0.21$, $p < .001$). This was true for both liberal and conservative participants (Figure 6 center and left; liberals: good/harm: $\beta = 0.67$, $p < .001$; deservedness: $\beta = 0.18$, $p < .001$; conservatives: good/harm: $\beta = 0.48$, $p < .001$; deservedness: $\beta = 0.35$, $p < .001$). This indicates that the undeserved harm principle was applied across political orientations. However, there were some differences in how much harm vs. deservedness drove different people's moral judgments: liberals based judgments slightly more on harm compared to conservatives (interaction: $\beta = -0.12$, $p < .001$), and conservatives based judgments slightly more on deservedness compared to liberals (interaction: $\beta = 0.11$, $p < .001$).

**Figure 6**

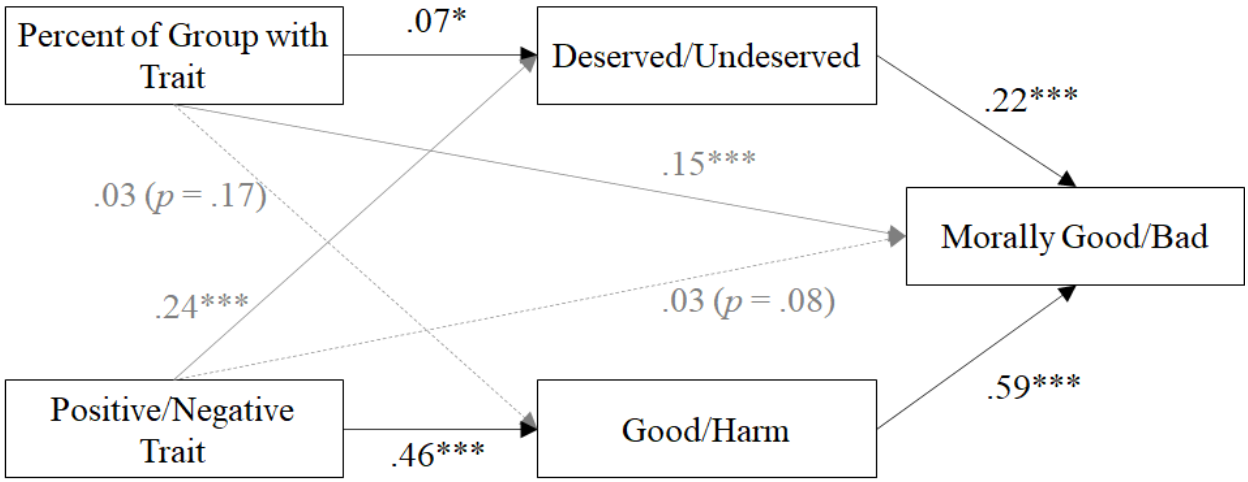
Endorsement and Use of the Undeserved Harm Principle in Moral Evaluations of Group-based Generalizations by Liberal and Conservative Participants




*Note.* 95% CIs shown.

Results also suggested that changes in people's other beliefs affect moral judgments through the undeserved harm principle. Specifically, the two manipulations that influenced people's moral judgments in the scenarios appeared to do so through altering perceptions of undeserved harm (see Figure 7). More overgeneralized statements (i.e., statements true of a smaller percent of the group) were seen as more undeserved in their consequences ($\beta = 0.07$, $p = .01$), and statements with negative rather than positive traits were seen as more harmful ($\beta = 0.46$, $p < .001$), and, unexpectedly, less deserved ($\beta = 0.24$, $p < .001$) perhaps because people think others generally deserve to be viewed positively. These effects on perceived harm and deservedness fully mediated the effects of the trait's negativity on morality ratings, and partially mediated the effect of overgeneralization on morality ratings (see Figure 7, negativity indirect effect: $\beta = 0.32$, $p < .001$; overgeneralization indirect effect: $\beta = 0.04$, $p = .03$). These results underscore how people may be using the same moral principle, yet still disagree in their moral judgments of a particular generalization: because different beliefs can lead people to have different estimates of the undeserved harm it is likely to cause.

**Figure 7**
The Effects of Overgeneralization (Percent of the Group with the Trait) and Trait Positivity/Negativity on Moral Judgements are Mediated by Perceptions of Harm and Deservedness




*Note:* Standardized regression coefficients shown. *** indicates $p < .001$, * indicates $p < .05$. Solid (vs. dotted) lines indicate significant (vs. non-significant) paths. Black (vs. grey) paths indicate paths of primary theoretical interest.

Data from an additional study replicated most of these results, further supporting the role of undeserved harm these types of moral judgments (see Supplementary Materials, Study S2).

Together, these results show that people may often be using the same moral principle – the undeserved harm principle – yet, due to differences in their beliefs, still disagree on whether a generalization is acceptable to use. More optimistically, identifying undeserved harm as widely shared moral principle highlights a basis upon which these moral disagreements may begin to be addressed. This is because, ideally, through open dialogue and discussion, people might be able to come to a consensus about the undeserved harm that is likely to be caused by a given generalization.

## General discussion

The current research examined one reason why people might continue to use stereotypes and group-based generalizations, even when this might be viewed as racist, sexist, or otherwise morally problematic by other people. Previous work has largely explained this as due to people's unawareness of the subtle, unintentional influences of stereotypes on their actions(*9–11*), or due to competing ill-intentions that may overwhelm any desire not to stereotype (*6*, *14*, *15*). Without denying the importance of these explanations, the current work highlights another, less explored, possibility. Specifically, people may freely make group-based generalizations because they see those generalizations as acceptable – rather than as cases of immoral stereotyping – as determined based on their various beliefs about the generalization. Yet differences in people's beliefs can mean that other people might disagree and see this generalization instead as a case of immoral stereotyping. Thus, people may freely use generalizations that others deem immoral, simply because they do not think they are doing anything wrong. This process may also lead to a form of 'bias blind-spot', where people see problematic stereotyping as occurring primarily in other people, especially in those whose beliefs differ more strongly from one's own (e.g., members of political outgroups), rather than in themselves.

A series of five studies and two supplementary studies provided evidence for the current view. As predicted, Study 1 showed that people think that stereotypes (or immoral or inaccurate stereotypes) are used primarily by others, rather than themselves – and especially by members of

political outgroups. Study 2 confirmed that most people do indeed distinguish between immoral and acceptable group-based generalizations, with 'stereotype' tending to refer to the more problematic kind. However, as indicated by Study 3, people often disagree on how morally acceptable a given generalizations is, and these disagreements are associated with differences in people's other beliefs about the generalization (e.g., about how overgeneralized, or likely to incite discrimination it is). These types of differences were observed as systematic differences between liberals and conservatives, as well idiosyncratic differences between individuals. Study 4 and 5 confirmed that these types of belief differences can in fact cause people to differ in how immoral and stereotype-based they think a generalization is, as well as in their willingness to use a generalization themselves. Data from Study 4 further highlighted how these moral disagreements can arise even while people are following the same widely shared moral principles. Specifically, it showed that people – including both liberals and conservatives – appear to widely endorse and apply the undeserved harm principle when making moral judgments of group-based generalizations. However, differences in people's other beliefs could still lead them to have different estimates of the undeserved harm likely to be caused by a generalization, and thus different views of how immoral that generalization is. Overall, these studies highlight an important yet under-explored reason that people might end up using problematic stereotypes, or at least, what others view as problematic stereotypes: because, due to their other beliefs, people think they are merely using acceptable group-based generalizations.

**Relationship to previous theories of stereotyping and prejudice**

The current work complements existing explanations for why people use stereotypes (i.e., problematic group-based generalizations), by highlighting another factor that can influence people's stereotype use – how morally acceptable people think the generalization is – and by identifying some of the beliefs that play into this judgment. The current work can thus help to understand cases of stereotyping not accounted for by most previous views. For example, existing accounts often explain stereotyping as due on unintentional use of stereotypes or competing ill-intentions (*6*, *9–11*, *14*, *15*). However, the current view can help explain cases where people have no competing ill-intentions, but still intentionally use stereotypes. In these cases, people may simply not see anything wrong with using the stereotype, thus have no reason to avoid using it. For example, if asked about their views on the topic, someone might express the view that men have more natural scientific aptitude than women, simply because they do not see anything wrong with this view: e.g., they do not think it is overgeneralized, likely to incite discrimination, or otherwise likely to cause undeserved harm.

The current work also deepens our understanding of the factors leading to stereotype use in cases where ill-intentions or unintentional biases are involved. Indeed, whether and how these other factors affect stereotype use may depend crucially on people's moral evaluations of their stereotypes, and the beliefs that feed into these evaluations. For example, stereotype use might sometimes show up in the form of unconscious, unintentional biases, but reflect stereotypes that the person would in fact deem appropriate, and continue to use, if they were made aware of this bias. Thus, bringing stereotype use into conscious, intentional control might only reduce stereotyping if people in fact judge it as unacceptable. Furthermore, ill-intentions might primarily affect stereotype use by biasing one's beliefs, and thus making one's stereotypes seem more acceptable (e.g., through motivated reasoning processes (*57*, *58*)). For instance, members of white supremacist groups might be motivated to look for "evidence" that non-white people are inferior, so that any consequences of their generalizations are seen as more deserved (*59*). However, in cases where people are unable to alter the relevant beliefs, and thus unable to

convince themselves that their stereotypes are acceptable, people largely avoid stereotyping regardless of any ill-intentions they hold.

In highlighting the ways people might try to justify their stereotype use, the current work parallels and complements related work on how people justify their prejudices, such as the justification-suppression model (JSM) of prejudice (*60*). In particular, the JSM highlights how various beliefs, values, and ideologies can affect people's willingness to express dislike towards a group, by either justifying their pre-existing dislike, or motivating people to suppress it. The current work extends this logic by showing how various beliefs may similarly affect people's willingness to use stereotypes. Moreover, the current work provides a more general framework that can apply to both stereotyping and prejudice, while highlighting unique implications for the case of stereotyping. Specifically, the current work suggests that, in general, people's moral judgments may depend on the expected undeserved or unjustified harm something is likely to cause. When applied to prejudice, negative attitudes towards a group may always be seen as likely to cause harm. Therefore, the main influence on how acceptable this attitude seems should be how deserved or justified this harm seems. In line with this, many of the justifiers and suppressors identified in the JSM appear to primarily affect perceived deservedness (e.g., 'just world beliefs', that is, the belief that people generally get what they deserve (*61*, *62*)). For this reason, many of these factors may similarly affect stereotype use. However, applying the undeserved harm principle to stereotype use highlights a host of other factors that can affect stereotype use (and perhaps to some extent, prejudice expression): any belief that affects perceptions of how much harm is likely to be caused. Thus, the current perspective can integrate our understanding of stereotyping and prejudice expression within a broader theoretical framework and enrich our understanding of how people sometimes come to see these as acceptable.

Returning to explanations for stereotype use, the current view also differs from, and complements, another less common explanation for stereotype use: the idea that people may use stereotypes in order to try make more accurate inferences about other people (*63–67*). This explanation is related to the one proposed here (that people may use stereotypes because they see them as merely acceptable generalizations), in that both explanations could account for cases where people use stereotypes intentionally and without ill-intentions. Yet, in the current view, a stereotype's perceived potential to increase accuracy is just one factor that can affect how acceptable it seems (akin to seeing it as less overgeneralized), by affecting how deserved its consequences seem. Thus, there might be cases when people see a generalization as largely accuracy-promoting, yet avoid using it, because other factors mean that they still see it as likely to cause undeserved harm. For example, a teacher who noticed gender differences in their students' math performance might try to avoid expressing this view, or letting it influence their interactions with students, for fear of discouraging students of the worse-performing gender.

**Implications for understanding moral evaluations of stereotyping**

This research also provides important insights into how people think about the morality of stereotyping and group-based generalizations. While norms against stereotyping are sometimes taken to mean that all group-based generalizations are seen as immoral (*1*, *3*, *21*), the current studies showed that this view is not widely endorsed or applied. In contrast, most people appear to distinguish between immoral and acceptable group-based generalizations. The fact that people readily make this distinction suggests that the problems with stereotyping (i.e., the bad kind of generalization) cannot be explained as simply due to inherent issues with group-based generalizations. Instead, the current work suggests that we need a more nuanced approach to

understand why stereotyping is often considered immoral. In particular, we suggested that these nuances moral judgments may largely stem from people's expectations of the undeserved harm likely to be caused by a generalization.

Identifying the undeserved harm principle as a potential basis for these judgments not only helps understand how people distinguish the good from the bad generalizations, but also why people might so often disagree on this. In particular, estimating undeserved harm is an inherently complex task, and any number or combination of factors could be relevant to this estimate for a given generalization, including the specifics of the content, how it is applied, who it is about, and the broader social and historical context. Furthermore, people's beliefs about these various factors may often be inaccurate, due to limitations and biases in people's knowledge and experience. Thus, it is no surprise that people might frequently disagree on their estimates of undeserved harm, and thus perhaps disagree on whether something counts as a case of immoral stereotyping, or simply an acceptable group-based generalization.

Future research can further explore the implications of the undeserved harm principle for understanding stereotyping and other related phenomenon. As discussed, many factors related to stereotyping, prejudice, and discrimination might be productively viewed as affecting perceptions of harm and/or deservedness. For instance, dehumanizing other groups might make them seem less affected by more "human" forms of harm, while portraying groups as cruel or malicious might make harms towards them seem more deserved. By testing these types of predictions, this principle may help clarify the mechanisms by which these various factors can affect stereotyping, prejudice expression, or discrimination.

The undeserved harm principle may also be useful for clarifying the relationship between epistemic values (e.g. accuracy goals) and moral values in cases of stereotyping (*68–70*). The current results suggest that epistemic and moral values may often align in these cases: e.g., with more accurate (less overgeneralized) generalizations typically be viewed as less immoral. However, there might also be cases where these values conflict (*68–70*) – as in the case where a teacher's use of gender stereotypes might increase predictive accuracy, yet have the negative moral consequence of discouraging students. Future work can further examine how people think about these types of moral-epistemic conflicts (as in *75*). It can also examine potential ways to avoid these conflicts, and how people view such resolutions. For example, maybe people think it is ok to discuss gender differences in math if the appropriate caveats are included in the discussion (e.g., that differences may be socialized, rather than innate). How people approach these conflicts has important implications for cases where real group differences exist, but where believing in or communicating about these differences might lead to problematic consequences.

**Implications for reducing problematic stereotype use**

The current research also has important practical implications for how people try to reduce problematic cases of stereotyping. One implication is that attempts to reduce stereotyping in general – e.g., warning people not to let stereotypes influence their hiring decisions, without mentioning specific stereotypes – might often be ineffective (as shown in previous work (*71*)). The current work highlights another explanation for this finding: people may think they are following the advice to avoid stereotyping, because they do not see their generalizations as the 'stereotyping', or problematic kind of generalization that they are being told to avoid. When this occurs, addressing specific stereotypes and the moral norms around using them may be a more effective strategy.

This work also has implications for how stereotype use might be confronted in specific cases. In an ideal world, people confronted about their problematic stereotype use would respond

with open-mindedness and a willingness to learn and improve. However, in reality, such confrontations often lead to backlash for the confronter (*72*) or more broadly – e.g. fueling political conflict and polarization (*38*). Confronting may also create defensiveness by threatening people's sense of being a morally good person (*73*) This might lead people to make more superficial changes, rather than deeper moral improvements (e.g., to stop expressing the stereotype, but not actually change their underlying beliefs or attitudes).

In cases where confronters want to avoid these negative reactions, the current work highlights one way they might do this. Specifically, rather than directly addressing the immorality of the person or their actions (e.g., calling the person racist, or saying that their comment is not ok (*74*, *75*)), one could instead address errors in the person's *beliefs*, which may have led the person to think the stereotype was ok to use in the first place. For example, one might point out factual errors in the person's beliefs about a group, unnoticed forms of harm that the stereotype may cause. Focusing on belief errors in this way, rather than moral errors, may reduce defensiveness, as this should be less threatening to people's sense of being a morally good person (*73*). In addition, defensiveness may be even further reduced if this is combined with a re-affirmation of shared moral values (e.g., acknowledging and reminding the person of their desire to avoid causing undeserved harm). The current work provides guidance for this approach by identifying various beliefs that may be addressed in this way, as well as widely shared moral values that may be re-affirmed.

This approach to confrontations may be especially valuable when confrontations occur across political lines, as it may be well suited to promoting open dialogue and mutual respect within an increasingly antagonistic political landscape (*36*, *37*). And, indeed, in cases where the stereotype use is truly not caused by bad intentions, but instead by differences in the information liberals and conservatives are exposed to, using a belief-focused strategy may be most appropriate. However, even if ill-intentions are involved, this strategy may still be helpful, since reminding people of their moral values may increase their motivation to act in line with these values, rather than any competing goals.

This work also serves as an important reminder for people who are confronted about their own stereotype use to remain open to feedback about their moral errors, and to use it as an opportunity to learn and improve. The difficulty of estimating undeserved harm highlights how even the most well-intentioned person might still use problematic stereotypes, without realizing they are doing anything wrong. Thus, feedback from other people can ideally be treated not as a threat, but as an important source of information to help better act in line with one's moral values. Indeed, even if someone is not confronted about a generalization, and even if that generalization is seen as acceptable by those around them, it might still be causing underserved harm. This highlights the importance of not only learning from confrontations, but also proactively trying to understand and avoid any negative impacts of one's generalizations, especially by seeking out the perspectives of the people in the targeted groups. This potential to be unaware of these negative impacts may perhaps be best appreciated by considering historical changes in the acceptability of different stereotypes (*76*): just as many stereotypes that were widely considered acceptable in the past can seem offensive to modern sensibilities, people in the future may look back at us today and wonder how we did not see the problems with many of our widely used generalizations.

**Limitations**

The current studies were designed to each test different aspects of the process proposed here: i.e., where belief differences CAN lead people to have different moral judgements of A

group-based generalization (even while using the same general moral principles), in turn lead people to sometimes freely use generalizations that others view as immoral stereotypes. Though each of the current studies has its own limitations, as a whole, the studies complement each other to offset these limitations and provide stronger evidence for the proposed process. For example, the correlational data in Study 3 highlighted a range of beliefs that may contribute to differences in moral judgments; however, this study could not show the causal role of these beliefs. This limitation was addressed by using experimental manipulations of these beliefs in Study 4 and 5. Similarly, Study 4 only used artificial scenarios to show the causal role of beliefs, but Study 5 showed that these effects generalize using real social groups. As another example, one possible concern with Study 5 is that the manipulation of perceived group differences may have affected evaluations and use of the generalizations through experimenter demand, rather than though genuinely altering people's beliefs. However, reducing such concerns, the link between beliefs and moral judgments also held in Study 3, which did not include any manipulations, and in Study 4, which used more complex scenarios about novel groups, likely reducing the perception that a certain answer was desired by the experimenter. (See Supplementary Materials for further discussion of possible demand effects.) The current work thus provides an important step towards establishing the role of the processes discussed here. Future research can continue to develop our understanding of these processes, for example, by more fully examining the role of other beliefs in people's evaluations and use of stereotypes.

## Conclusion

People often continue to use stereotypes, even in cases where this might be criticized for being racist, sexist, or otherwise immoral. The current work investigated an under-explored reason that this can occur: because differences in people's beliefs may lead them to think they are merely using acceptable group-based generalizations, even while other people see this as morally problematic stereotyping. The current work therefore provides a richer understanding of the factors that lead people to use problematic stereotypes, and thus how this might be reduced.

## Open Practices

Raw data, study materials, analysis scripts, and pre-registrations (where applicable) for the studies in this article are available online through the Open Science Framework [link to be inserted upon publication or upon request].

## Materials and methods

Sample sizes reported for all studies reflect the number of included participants (i.e., who completed the task, completed all attention checks correctly, and, upon debriefing, agreed that their data could be used in the study). In any case where this led to a participant being included more than once (i.e., if they completed the task twice), only the first dataset was included. Except where otherwise noted, sample sizes for the number of collected participants were determined a priori and independent of the data. For all studies, participants were considered liberal/left-wing if they rated themselves below the scale midpoint, conservative/right-wing if above the scale midpoint, and neither if at the scale midpoint. (See Study 1 methods for question format.)

### Study 1

#### *Participants*

Study 1 included 405 participants recruited through the Prolific research platform from the United States (Gender: 228 female, 171 male, 6 other/prefer not to answer; Mean age (SD) = 36.69 (15.83); political orientation: 61% liberal/left-wing; 26% conservative/right-wing; 11% neither).

#### *Materials and Procedure*

All participants rated how often they thought that themselves, and other people who were either politically liberal or conservative, used stereotypes/immoral stereotypes/inaccurate stereotypes. For self-ratings, participants responded to the following question: "How often do you intentionally use [stereotypes/inaccurate stereotypes/immoral stereotypes]?" (wording manipulated between participants). They also answered similar questions about political liberals, and political conservatives: e.g., "On average, how often do people who are politically liberal intentionally use [stereotypes/inaccurate stereotypes/immoral stereotypes]?" Responses were made on a six-point scale (0 = Never, 1 = Rarely, 2 = Occasionally, 3 = Somewhat often, 4 = Fairly often, 5 = Very often). At the end of the study, participants reported their own political orientation on a 100-point scale (0 = "Extremely liberal/left-wing", 100 = "Extremely conservative/right-wing").

***Analytic Approach***

To compare perceptions of self and others' stereotype use, multilevel modelling was used to predict participants' responses (perceptions of the use of stereotypes/immoral stereotypes/inaccurate stereotypes) from whether the question was about oneself or others, with random intercepts included for each participant. To examine how this depended on the political orientation of the other people being rated, as well as of participants, a second multilevel model predicted responses from the question target (self, liberals, conservatives), interacted with participants' political orientation, with random intercepts included for each participant.

## Study 2

***Participants***

Study 2 included 169 participants recruited through MTurk and the Cloud Research platform (*77*) from Canada and the United States (Gender: 71 female, 98 male; Mean age (SD) = 36.28 (11.63)).

***Materials and Procedure***

Participants were asked how often they thought it was ok to make generalizations based on groups, and based on stereotypes. Two versions of each question were asked, involving either generalizations about individuals or about groups as a whole. The specific questions were: "How often is it ok to draw conclusions or make judgments [about individual people based on the groups they are part of?/about groups as a whole?]" and "How often is it ok to use stereotypes to draw conclusions or make judgments [about individual people?/about groups as a whole?]" All participants responded to all four questions. The stereotype questions were asked first, to avoid the other questions affecting how participants interpret the term "stereotype". Responses were provided on five-point Likert scales (1 = Never, 2 = Sometimes, 3 = About half the time, 4 = Most of the time, 5 = Always).

***Analytic Approach***

To test whether stereotypes were seen as different than, and worse than, beliefs about groups, a one-sample t-test based was computed on differences between the 'stereotype' and 'group' version of each question (stereotype – group).

## Study 3

***Participants***

Study 3 included 140 participants recruited through MTurk and the Cloud Research platform (*77*) from the United States (Gender: 63 female, 75 male, 2 other/prefer not to answer; Mean age (SD) = 38.79 (11.88); political orientation: 63% liberal/left-wing; 26% conservative/right-wing; 11% neither).

***Materials and Procedure***

Participants read various statements containing group-based generalizations, and then reported their perceptions of how immoral each statement was, as well as various other beliefs about each statement that might influence their moral judgments. For each statement, participants were asked to imagine that someone made a generic statement where a group was described as having a trait (e.g., "Artists are creative"); see Figure 3 for all statements. Participants saw a random 20 out of 40 possible statements, where half of the possible statements involved positive traits and half involved negative traits. For each statement, participants responded to the following questions on 100-point sliding scales. (The statement "Artists are creative" is used throughout as an example.)

- "How much do you think this statement is based on a stereotype?" (0 = Not at all, 100 = Very much)
- How **immoral** do you think this statement is? (0 = Not at all immoral, 100 = Very immoral)
- Of [**artists**], how many of them do you think [**are creative**]? (0 = None, 50 = Half, 100 = All)
- How do you think **the person making the statement** would respond to the question: Of [**artists**], how many of them do you think [**are creative**]? (0 = None, 50 = Half, 100 = All)
- If this statement was heard by others, how likely would it be to incite them to act in **discriminatory** or otherwise problematic ways towards [artists]? (0 = Not at all likely, 100 = Extremely likely)
- How much does this statement reflect **cultural or media representations** of [artists]? (0 = Not at all, 100 = Extremely)

In the first section, participants answered the stereotype question for all statements (to avoid the other questions affecting how they interpret the term "stereotype"). In a second section, they answered the remaining questions for each statement. Finally, since some groups were used in multiple statements, in a third section, participants responded to the following two questions for each unique group in the previously seen statements:

- Of people in America, what percent of them are [artists]? (0 = 0%, 100 = 100%)
- In general, how disadvantaged do you think [artists] are in society? (0 = Not at all, 100 = Extremely)

At the end of the study, participants reported their political orientation on a 100-point scale.

***Analytic approach***

Multilevel models were used to test for differences in moral judgments and beliefs between liberals and conservatives. Each model involved predicting one type of question response from participants' political orientation, interacted with the specific statement, with random intercepts for each participant. For predicting group size and group disadvantage, the interaction was with the group, rather than the statement, since each group was rated only once but could occur in multiple statements. (In Figure 3, these results are repeated for each statement involving a given group, for ease of visual comparison.)

A mediation model was used to test whether the observed moral differences were mediated by the measured beliefs. Specifically, this model tested whether the interaction effect of statement and political orientation on moral judgments was statistically mediated by the measured beliefs (including all beliefs in the model simultaneously), while controlling for the main effects of statement and of political orientation at each step of the model. Standard errors

were clustered by participant to account for repeated measures. Wald's chi-squared tests were used to examine whether the indirect and direct effects were significant when considering all statements simultaneously. The estimated indirect effect summed over the indirect effects through each of the measured beliefs.

To test whether idiosyncratic variations in people's beliefs about a generalization were related to variation in their moral judgments of that generalization, moral judgments were predicted from all belief ratings simultaneously (all variables standardized), controlling for political orientation, the specific statement, and the interaction of these two variables. Random intercepts were included for each participant. Controlling for these variables ensured that the results were not driven by average differences in how liberals and conservatives viewed these statements (i.e., the political differences identified in the previous analyses), nor average differences between statements (e.g., if, on average, participants viewed some statements as more immoral than others). Additionally, modelling random intercepts for each participant ensured that these results do not just reflect other forms of consistent individual differences (e.g., if some participants on average viewed all statements as more immoral than other participants). This analysis therefore examined whether a person's idiosyncratic beliefs about specific generalization were related to seeing that specific generalization as more or less immoral. Significant results here indicate both that there is meaningful idiosyncratic variation in these moral judgments for a given statement (i.e., not merely due to measurement noise, or other factors like political orientation), and that this may be accounted for by variation in people's beliefs.

## Study 4

### *Participants*

Study 4 included 533 participants recruited through MTurk and the Cloud Research platform (*77*) from the United States (Gender: 205 female, 327 male, 1 other/prefer not to answer; Mean age (SD) = 35.98 (12.32)). This study was collected as two separate sub-studies (ns = 172 and 361, respectively), where the second sub-study included additional questions at the end of the study. The data from the first sub-study was analyzed before collecting the second sub-study, thus the final sample size when combined is not fully independent of the data. However, the almost all of the key results reported here replicate even when only considering data from the first sub-study. (The one effect that not reach significance when analyzing data from the first sub-study was the effect of the percent of the group with the trait on deservedness judgments, $\beta = 0.06$, $p = .21$, see Figure 7 for comparison, likely due to the small size of the effect.)

### *Materials and Procedure*

Participants read two scenarios about fake social groups (e.g., the Lupites). In each scenario, someone makes a group-based generalization, saying, for example, "Oh, that guy is a Lupite. He is probably short-tempered." After reading a scenario, participants made moral judgments of this statement, and evaluated the expected harm and deservedness of it on 100-point sliding scales. Specifically, they were asked:

- How MORALLY GOOD or BAD is it for this person to say or think this? (0=Very morally bad, 50=Morally neutral, 100=Very morally good)
- How much GOOD or HARM is this statement or belief likely to cause? (0=A lot of harm, 50=No good or harm, or equal good and harm, 100=A lot of good)
- How much do you think anyone affected by this statement or belief DESERVES what happens because of it? (0=Does not deserve it at all, 100=Completely deserves it)

Note that in the second sub-study the wording of the deservedness question was changed to say "How much do you think anyone else affected…" (i.e., adding the word "else") to clarify that it refers to consequences for people other than the speaker of the statement, since consequences for other people may be more morally relevant.

The scenarios were manipulated to vary five factors which might affect these moral judgments:

1) the percent of the group members that actually had the trait (either 90%, or 10% above the stated baserate in that society),
2) how strongly the statement was worded (i.e., if it said the person "might be", "is probably", or "must be" …),
3) if the trait was positive (i.e., "very intelligent", "generous", or "forgiving"), or negative (i.e., "dishonest", "short-tempered", or "lazy"),
4) if the group was described as disadvantaged or advantaged, and
5) if the group was described as a minority (20%) or majority (80%) of the society.

These variables were fully randomized for the first scenario, and the second scenario was set so that all five variables differed from the first scenario. An example scenario is presented below (bolding in original):

> Imagine you are in a world much like our own, and you are about to visit a country you have never been to before.
> In this country, one of the groups of people you may encounter are Lupites.
> Lupites form a **minority** of their society, comprising about **20%** of the country's population.
> Lupites tend to be relatively **disadvantaged**, and are generally not very well-off in their society. For example, due to historical circumstances, they tend to be poorer, have worse education, and are less likely to hold government offices compared to other groups.
> If you meet Lupites while travelling, you may notice certain things about them. For example, they tend to be **short-tempered**: **90%** of Lupites are **short-tempered**, compared to only **20%** of the general population.
> Now imagine you have arrived in this new country. As you are walking down the street, you overhear someone talking on the phone. You don't hear much of their conversation, but you hear them say:
> "Oh, that guy is a Lupite. He is probably short-tempered."

The stated baserates of each trait in the general population were depended on the specific trait, and were intended to approximate realistic baserates for that trait.

For participants in the second sub-study, a question at the end of the study asked how much they endorsed the undeserved harm principle as a way to evaluate group-based inferences. Specifically, participants were asked how much they agreed with the following statement: "Inferences about people based on their groups are more morally bad when… the inference is likely to lead to other people being treated in harmful ways, especially if they have done nothing to deserve it." Responses were provided on a 7-point Likert scale (1 = Strongly disagree, 4 = Neither agree nor disagree, 7 = Strongly agree). All responses of 5 or above were considered agreement with this principle. Participants in the second-sub study reported their political orientation at the end of the experiment on a 100-point scale.

***Analytic approach***

To test whether manipulating people's beliefs about a generalization affected their moral judgments, morality ratings (standardized) were predicted simultaneously from all five manipulations. Aside from the strength of the statement wording, all predictors were effect coded, with higher values indicating a larger percent of the group with the trait, positive statements, advantaged groups, and majority groups. How strongly the statement was worded was coded as -1, 0 and 1 and then standardized, with higher values indicating stronger wording. Random intercepts were included for each participant.

To examine the role of perceptions of undeserved harm in moral judgments, moral judgments were predicted simultaneously from good/harm and deservedness ratings in a multilevel model with random intercepts for each participant, with all variables standardized. To examine how this depended on political orientation, this model was run separately for liberal and conservative participants (i.e., political orientation responses below and above the scale midpoint, respectively). An additional model also included interactions of both harm and deservedness with political orientation (also standardized).

To test whether perceptions of harm and deservedness mediated the effect of the manipulated factors on moral judgments, the following path model was run. This analysis focused only on the two manipulations which were found to influence moral judgments: one aspect of overgeneralization (i.e., the percent of the group with the trait), and negativity (i.e., whether the trait was positive or negative). For the path model, good/harm was predicted from overgeneralization and negativity, deservedness was predicted from overgeneralization and negativity, and then moral judgments were predicted from harm, deservedness, overgeneralization, and negativity (variables coded as in the analyses above). Indirect effects were computed through the product of coefficients, with a single indirect effect computed for each manipulation which sums the indirect effects through harm and deservedness. Standard errors were clustered by participant.

## Study 5

### *Participants*

Study 5 included 300 participants recruited through the Prolific platform from the United States (Gender: 190 female, 97 male, 10 other/prefer not to answer; Mean age (SD) = 32.37 (12.41)).

### *Materials and Procedure*

This study involved measuring participants' use of group-based generalizations, as well their perceptions of how immoral and stereotype-based this generalization was. This was measured after manipulating their beliefs about the actual group differences, and thus beliefs about whether the generalization should count as an overgeneralization. (These same measures were also included before the manipulation, to help increase power by controlling for initial differences in participant's responses.)

Participants were randomly assigned to do this task for either a gender- or race-related cultural stereotype: boys doing better at math than girls, or black people tipping less than white people (*78, 79*). Using the gender case as an example, to measure participants' use of group-based generalizations, participants responded to the following scenario:

> Imagine there is an elementary school in a wealthy US neighbourhood. At this school, there are two students in the same math class. One of these students is a boy, one is a girl. One of these students is doing better than the other in the math class. Which student do you think is more likely to doing better in the class?"

There were three possible responses to this question: "the boy" (i.e., using the cultural stereotype), "the girl" (i.e., using the counter-stereotype), or "these are both equally likely" (i.e., not using the cultural stereotype). The key focus here was on whether participants would or would not use the cultural stereotype, however, the third counter-stereotypical response was included for completeness. Participants were then asked to imagine that someone had made the stereotypical response to that scenario (e.g., "Suppose someone had inferred that the boy was more likely to be doing better in the math class than the girl"). Participants rated how immoral and how stereotype-based they thought this inference was on a 5-point Likert scale (ranging from "Not at all" to "Extremely").

To manipulate beliefs about whether or not there were real group differences in this domain, and gave participants a summary of a real scientific article that either suggested a) that there were *no* group differences in this domain (e.g., that boys and girls did equally well in math classes), or b) that there *were* group differences in this domain that were consistent with the cultural stereotype (e.g., that boys did better than girls at math). These summaries built on stimuli developed by Corey Cusimano and Tania Lombrozo (to be published in forthcoming work). See Supplementary Materials for details of manipulation.

Participants then responded to the three key measures both before and after the manipulation. For the judgments of how immoral and stereotype-based the inference was, participants were reminded of their previous response to this question, and it was noted that they could change their responses if they wished. Reminding participants of their previous responses helped ensure that any changes in their responses reflected genuine changes, rather than random variability in their scale use, and thus may increase power by reducing measurement noise. See supplementary materials for further methods details.

***Analytic approach***

This study aimed to examine whether manipulating people's beliefs about the real group differences affected their use of group-based generalizations, as well as judgments of how immoral and stereotype-based this generalization would be. To examine the manipulation's effect on participants' use of the generalization, an ordered probit regression was used to predict participants' post-manipulation scenario responses (responses ordered as follows: using the cultural stereotype, not using cultural stereotype, and using the counter-stereotype). Thresholds between each of these three response types were allowed to vary freely with all predictors, though we focus on changes in the threshold between using and not using the cultural stereotype. Responses were predicted from the manipulation condition (1 = group differences, -1 = no group differences), while controlling for the scenario topic (race or gender, effect coded), and pre-manipulation responses to the same question (coded as 1, 0 or -1, ordered as above).

To examine whether this belief manipulation also affected how immoral and stereotype-based this generalization seemed, post-manipulation responses to these questions were predicted in a linear regression from the manipulation condition, while controlling for the scenario topic and pre-manipulation responses to the same question (variables coded as above).

For mediation analyses, the same models were used as in the analyses above, except that instead of using an ordinal regression to predict use of the generalization, a linear model was used, and data was subset to only include two outcomes: using the cultural stereotype, or not using it. This enabled simplifying the model, since the manipulation did not significantly affect the third response option (using the counter-stereotype).

## References

1. V. Yzerbyt, B. Dardenne, J.-P. Leyens, "Social judgeability concerns in impression formation" in *Metacognition: Cognitive and social dimensions* (Sage Publications, Inc, Thousand Oaks, CA, US, 1998), pp. 126–156.
2. A. Timmer, Judging Stereotypes: What the European Court of Human Rights Can Borrow from American and Canadian Equal Protection Law. *Am. J. Comp. Law*. **63**, 239–284 (2015).
3. M. Banaji, R. Bhaskar, "Implicit stereotypes and memory: The bounded rationality of social beliefs" in *Memory, brain, and belief*, Schacter, D. L., Scarry, E., Eds. (Harvard University Press, Cambridge, Massachusetts, 2000), pp. 139–175.
4. J. D. Carter, J. A. Hall, D. R. Carney, J. C. Rosip, Individual differences in the acceptance of stereotyping. *J. Res. Personal.* **40**, 1103–1118 (2006).
5. D. Kübler, J. Schmid, R. Stüber, Gender discrimination in hiring across occupations: a nationally-representative vignette study. *Labour Econ.* **55**, 215–229 (2018).
6. E. L. Uhlmann, V. L. Brescoll, E. Machery, The Motives Underlying Stereotype-Based Discrimination Against Members of Stigmatized Groups. *Soc. Justice Res.* **23**, 1–16 (2010).
7. C. Stern, J. Axt, Ideological differences in race and gender stereotyping. *Soc. Cogn.* **39**, 259–294 (2021).
8. J. Cao, M. Kleiman-Weiner, M. R. Banaji, People Make the Same Bayesian Judgment They Criticize in Others. *Psychol. Sci.* **30**, 20–31 (2019).
9. M. Banaji, A. G. Greenwald, "Implicit stereotyping and prejudice" in *The psychology of prejudice: The Ontario symposium, Vol. 7.* (Lawrence Erlbaum Associates, Inc, Hillsdale, NJ, US, 1994), *Ontario symposium on personality and social psychology, Vol. 7.*, pp. 55–76.
10. I. V. Blair, "Implicit stereotypes and prejudice" in *Cognitive social psychology: On the tenure and future of social cognition.*, G. Moskowitz, Ed. (Lawrence Erlbaum, Mahwah, NJ, 2001), pp. 359–374.
11. A. G. Greenwald, L. H. Krieger, Implicit Bias: Scientific Foundations. *Calif. Law Rev.* **94**, 945–967 (2006).
12. J. F. Dovidio, K. Kawakami, S. L. Gaertner, Implicit and explicit prejudice and interracial interaction. *J. Pers. Soc. Psychol.* **82**, 62 (2002).
13. I. Régner, C. Thinus-Blanc, A. Netter, T. Schmader, P. Huguet, Committees with implicit biases promote fewer women when they do not believe gender bias exists. *Nat. Hum. Behav.* **3**, 1171–1179 (2019).
14. C. S. Crandall, A. J. Bahns, R. Warner, M. Schaller, Stereotypes as Justifications of Prejudice. *Pers. Soc. Psychol. Bull.* **37**, 1488–1498 (2011).
15. C. N. Macrae, M. Hewstone, R. J. Griffiths, Processing load and memory for stereotype-based information. *Eur. J. Soc. Psychol.* **23**, 77–87 (1993).
16. Z. Kunda, S. J. Spencer, When do stereotypes come to mind and when do they color judgment? A goal-based theoretical framework for stereotype activation and application. *Psychol. Bull.* **129**, 522 (2003).
17. J. T. Jost, M. R. Banaji, B. A. Nosek, A Decade of System Justification Theory: Accumulated Evidence of Conscious and Unconscious Bolstering of the Status Quo. *Political Psychology*. **25**, 881–919 (2004).
18. J. T. Jost, "Outgroup favoritism and the theory of system justification: A paradigm for investigating the effects of socioeconomic success on stereotype content" in *Cognitive*

*social psychology: The Princeton symposium on the legacy and future of social cognition* (2001), pp. 89–102.
19. R. D. Ashmore, F. K. Del Boca, Conceptual approaches to stereotypes and stereotyping. *Cogn. Process. Stereotyping Intergroup Behav.* **1**, 35 (1981).
20. M. Snyder, "On the self-perpetuating nature of social stereotypes" in *Cognitive processes in stereotyping and intergroup behavior*, D. L. Hamilton, Ed. (Lawrence Erlbaum Associates, Inc., 1981).
21. B. K. Payne, C. D. Cameron, Divided minds, divided morals: How implicit social cognition underpins and undermines our sense of social justice. *Handb. Implicit Soc. Cogn. Meas. Theory Appl.*, 445–460 (2010).
22. L. Blum, Stereotypes And Stereotyping: A Moral Analysis. *Philos. Pap.* **33**, 251–289 (2004).
23. G. W. Allport, *The nature of prejudice.* (Addison-Wesley, Cambridge, MA, 1954).
24. J. A. Bargh, T. L. Chartrand, The unbearable automaticity of being. *Am. Psychol.* **54**, 462 (1999).
25. J. Sidanius, F. Pratto, Social dominance theory. *Handb. Theor. Soc. Psychol.* **2**, 418–438 (2011).
26. L. Jussim, *Social perception and social reality: Why accuracy dominates bias and self-fulfilling prophecy* (OUP USA, 2012).
27. E. Pronin, T. Gilovich, L. Ross, Objectivity in the eye of the beholder: divergent perceptions of bias in self versus others. *Psychol. Rev.* **111**, 781 (2004).
28. J. Ehrlinger, T. Gilovich, L. Ross, Peering into the bias blind spot: People's assessments of bias in themselves and others. *Pers. Soc. Psychol. Bull.* **31**, 680–692 (2005).
29. N. Epley, D. Dunning, Feeling "holier than thou": are self-serving assessments produced by errors in self-or social prediction? *J. Pers. Soc. Psychol.* **79**, 861 (2000).
30. E. Suhay, M. Tenenbaum, A. Bartola, Explanations for Inequality and Partisan Polarization in the U.S., 1980–2020. *Forum (Genova).* **20**, 5–36 (2022).
31. T. A. DiPrete, A. Gelman, T. McCormick, J. Teitler, T. Zheng, Segregation in social networks based on acquaintanceship and trust. *Am. J. Sociol.* **116**, 1234–83 (2011).
32. K. E. Walker, Political Segregation of the Metropolis: Spatial Sorting by Partisan Voting in Metropolitan Minneapolis–St Paul. *City Community.* **12**, 35–55 (2013).
33. A. Boutyline, R. Willer, The social structure of political echo chambers: Variation in ideological homophily in online networks. *Polit. Psychol.* **38**, 551–569 (2017).
34. D. Muise, H. Hosseinmardi, B. Howland, M. Mobius, D. Rothschild, D. J. Watts, Quantifying partisan news diets in Web and TV audiences. *Sci. Adv.* **8**, eabn0083 (2022).
35. R. J. Webster, M. D. Burns, M. Pickering, D. A. Saucier, The suppression and justification of prejudice as a function of political orientation. *Eur. J. Personal.* **28**, 44–59 (2014).
36. S. Iyengar, S. J. Westwood, Fear and loathing across party lines: New evidence on group polarization. *Am. J. Polit. Sci.* **59**, 690–707 (2015).
37. L. Mason, The rise of uncivil agreement: Issue versus behavioral polarization in the American electorate. *Am. Behav. Sci.* **57**, 140–159 (2013).
38. K. Durrheim, M. Okuyan, M. S. Twali, E. García-Sánchez, A. Pereira, J. S. Portice, T. Gur, O. Wiener-Blotner, T. F. Keil, How racism discourse can mobilize right-wing populism: The construction of identity and alliance in reactions to UKIP's Brexit "Breaking Point" campaign. *J. Community Appl. Soc. Psychol.* **28**, 385–405 (2018).

39. N. Klein, N. Epley, Maybe holier, but definitely less evil, than you: Bounded self-righteousness in social judgment. *J. Pers. Soc. Psychol.* **110**, 660 (2016).
40. Q. Wang, H. J. Jeon, Bias in bias recognition: People view others but not themselves as biased by preexisting beliefs and social stigmas. *PLOS ONE*. **15**, e0240232 (2020).
41. E. Pronin, D. Y. Lin, L. Ross, The bias blind spot: Perceptions of bias in self versus others. *Pers. Soc. Psychol. Bull.* **28**, 369–381 (2002).
42. J. E. Rothschild, A. J. Howat, R. M. Shafranek, E. C. Busby, Pigeonholing partisans: Stereotypes of party supporters and partisan polarization. *Polit. Behav.* **41**, 423–443 (2019).
43. D. Dunning, J. A. Meyerowitz, A. D. Holzberg, Ambiguity and self-evaluation: The role of idiosyncratic trait definitions in self-serving assessments of ability. *J. Pers. Soc. Psychol.* **57**, 1082–1090 (1989).
44. M. B. Brewer, In-group bias in the minimal intergroup situation: A cognitive-motivational analysis. *Psychol. Bull.* **86**, 307 (1979).
45. H. Tajfel, J. C. Turner, Social psychology of intergroup relations. *Annu. Rev. Psychol.* **33**, 1–39 (1982).
46. M. J. Rodin, J. M. Price, J. B. Bryson, F. J. Sanchez, Asymmetry in prejudice attribution. *J. Exp. Soc. Psychol.* **26**, 481–504 (1990).
47. C. S. Crandall, A. Eshleman, L. O'brien, Social norms and the expression and suppression of prejudice: the struggle for internalization. *J. Pers. Soc. Psychol.* **82**, 359 (2002).
48. F. Cushman, Rationalization is rational. *Behav. Brain Sci.* **43** (2020).
49. J. Haidt, The emotional dog and its rational tail: a social intuitionist approach to moral judgment. *Psychol. Rev.* **108**, 814 (2001).
50. M. F. Jonker, B. Donkers, E. W. de Bekker-Grob, E. A. Stolk, Effect of Level Overlap and Color Coding on Attribute Non-Attendance in Discrete Choice Experiments. *Value Health*. **21**, 767–771 (2018).
51. J. Graham, J. Haidt, B. A. Nosek, Liberals and conservatives rely on different sets of moral foundations. *J. Pers. Soc. Psychol.* **96**, 1029 (2009).
52. P. K. Hatemi, C. Crabtree, K. B. Smith, Ideology justifies morality: Political beliefs predict moral foundations. *Am. J. Polit. Sci.* **63**, 788–806 (2019).
53. C. Schein, K. Gray, The theory of dyadic morality: Reinventing moral judgment by redefining harm. *Personal. Soc. Psychol. Rev.* **22**, 32–70 (2018).
54. C. Schein, K. Gray, The unifying moral dyad: Liberals and conservatives share the same harm-based moral template. *Pers. Soc. Psychol. Bull.* **41**, 1147–1163 (2015).
55. J. A. Solecki, thesis, Rice University (2002).
56. F. Tan, E. Xiao, Third-party punishment: Retribution or deterrence? *J. Econ. Psychol.* **67**, 34–46 (2018).
57. Z. Kunda, The case for motivated reasoning. *Psychol. Bull.* **108**, 480 (1990).
58. N. Epley, T. Gilovich, The Mechanics of Motivated Reasoning. *J. Econ. Perspect.* **30**, 133–140 (2016).
59. M. Billig, The notion of 'prejudice': Some rhetorical and ideological aspects. *Text-Interdiscip. J. Study Discourse*. **8**, 91–110 (1988).
60. C. S. Crandall, A. Eshleman, A justification-suppression model of the expression and experience of prejudice. *Psychol. Bull.* **129**, 414 (2003).
61. E. Staub, Cultural-societal roots of violence: The examples of genocidal violence and of contemporary youth violence in the United States. *Am. Psychol.* **51**, 117 (1996).

62. A. Furnham, Belief in a just world: Research progress over the past decade. *Personal. Individ. Differ.* **34**, 795–817 (2003).
63. D. D'souza, *The end of racism: Principles for a multiracial society* (Free Press New York, 1995).
64. M. Levin, Responses to race differences in crime. *J. Soc. Philos.* **23**, 5–29 (1992).
65. L. Jussim, Précis of Social Perception and Social Reality: Why accuracy dominates bias and self-fulfilling prophecy. *Behav. Brain Sci.* **40** (2017), doi:10.1017/S0140525X1500062X.
66. S. Madon, M. Guyll, S. J. Hilbert, E. Kyriakatos, D. L. Vogel, Stereotyping the Stereotypic: When Individuals Match Social Stereotypes. *J. Appl. Soc. Psychol.* **36**, 178–205 (2006).
67. H. R. Arkes, P. E. Tetlock, Attributions of implicit prejudice, or "would Jesse Jackson 'fail' the Implicit Association Test?" *Psychol. Inq.* **15**, 257–278 (2004).
68. E. Beeghly, thesis, UC Berkeley (2014).
69. C. Cusimano, T. Lombrozo, Morality justifies motivated reasoning in the folk ethics of belief. *Cognition.* **209**, 104513 (2021).
70. G. Gardiner, "Evidentialism and moral encroachment" in *Believing in Accordance with the Evidence*, McCain, Kevin, Ed. (Springer, 2018), pp. 169–195.
71. J. R. Axt, C. K. Lai, Reducing discrimination: A bias versus noise perspective. *J. Pers. Soc. Psychol.* **117**, 26 (2019).
72. A. M. Czopp, "The consequences of confronting prejudice" in *Confronting Prejudice and Discrimination*, R. K. Mallett, M. J. Monteith, Eds. (Academic Press, 2019; https://www.sciencedirect.com/science/article/pii/B9780128147153000059), pp. 201–221.
73. I. K. Rösler, F. van Nunspeet, N. Ellemers, Don't tell me about my moral failures but motivate me to improve: Increasing effectiveness of outgroup criticism by criticizing one's competence. *Eur. J. Soc. Psychol.* **51**, 597–609 (2021).
74. A. M. Czopp, M. J. Monteith, A. Y. Mark, Standing up for a change: Reducing bias through interpersonal confrontation. *J. Pers. Soc. Psychol.* **90**, 784 (2006).
75. M. J. Monteith, M. D. Burns, L. K. Hildebrand, "Navigating successful confrontations: What should I say and how should I say it?" in *Confronting prejudice and discrimination*, Robyn K. Mallett, Margo J. Monteith, Eds. (Academic Press, 2019), pp. 225–248.
76. P. G. Devine, A. J. Elliot, Are racial stereotypes really fading? The Princeton trilogy revisited. *Pers. Soc. Psychol. Bull.* **21**, 1139–1150 (1995).
77. L. Litman, J. Robinson, T. Abberbock, TurkPrime.com: A versatile crowdsourcing data acquisition platform for the behavioral sciences. *Behav. Res. Methods*. **49**, 433–442 (2017).
78. Z. W. Brewster, G. R. Nowak, Racial Prejudices, Racialized Workplaces, and Restaurant Servers' Hyperbolic Perceptions of Black–White Tipping Differences. *Cornell Hosp. Q.* **60**, 159–173 (2019).
79. B. A. Nosek, M. R. Banaji, A. G. Greenwald, Math= male, me= female, therefore math≠ me. *J. Pers. Soc. Psychol.* **83**, 44 (2002).