Supplementary Online Material
*for*
**Indifferent or Impartial?**
**Actor-Observer Asymmetries in Expressing and Evaluating Socio-Political Neutrality**

**Table of Contents**



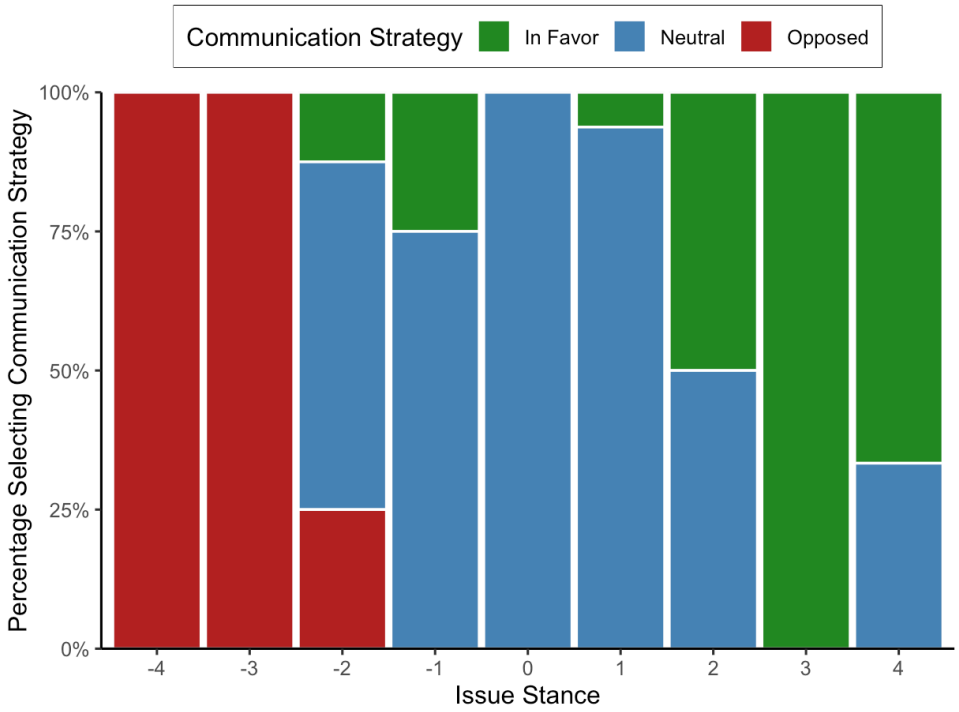
For the most current version of this document, including any post-publication updates, please visit this project's OSF page: https://osf.io/ye9mn/

## SECTION 1: SUPPLEMENTAL ANALYSES

### 1.1 Study 1A Supplemental Analyses

We also examined participants' communication strategy as a function of their position on the continuous attitude measure (9-point scale). There were few participants with strongly opposed responses, violating assumptions for a multinomial regression, so we conducted a chi-square analysis. This test revealed a significant effect of issue position on communication strategy selection, $\chi^2(16) = 134.04$, $p < .001$, $V = 0.82$ (see Figure S1).

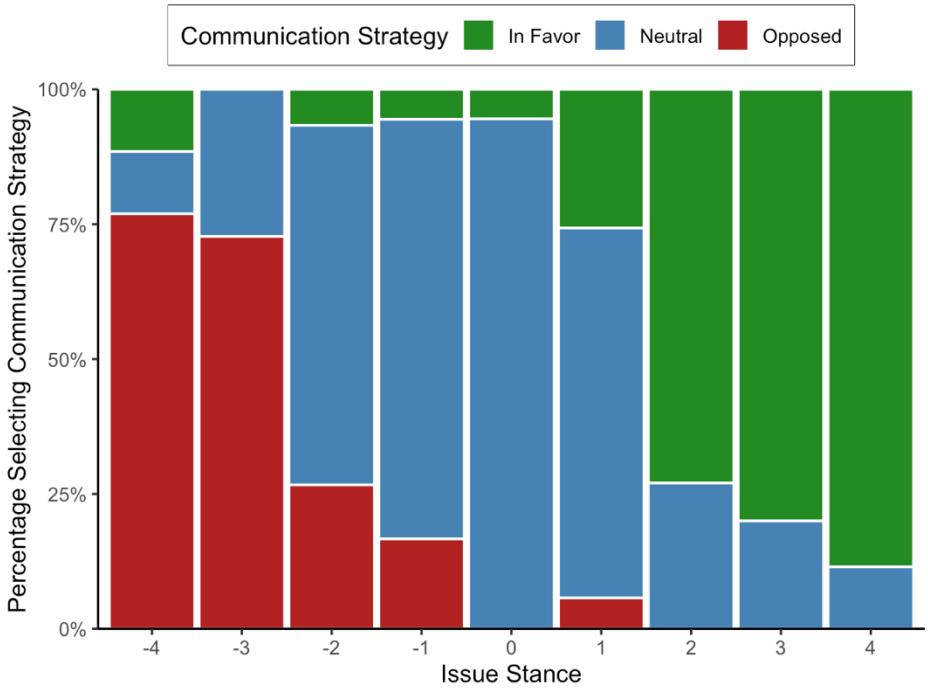
**Figure S1.** Prevalence of communication strategy by issue position in Study 1A.




### 1.2 Study 1B Supplemental Analyses

We again examined the results using the continuous measure of issue position. A multinomial regression revealed a significant effect of issue position on strategy selection (see Figure S2), such that greater private support for affirmative action decreased the likelihood of selecting "opposed" rather than "neutral," $B = -1.12$, $SE = 0.17$, $p < .001$, and increased the likelihood of selecting "in favor" rather than "neutral," $B = 0.91$, $SE = 0.11$, $p < .001$; see Figure S2).

**Figure S2.** Prevalence of communication strategy by issue position in Study 1B.




### 1.3 Study 1C Supplemental Analyses

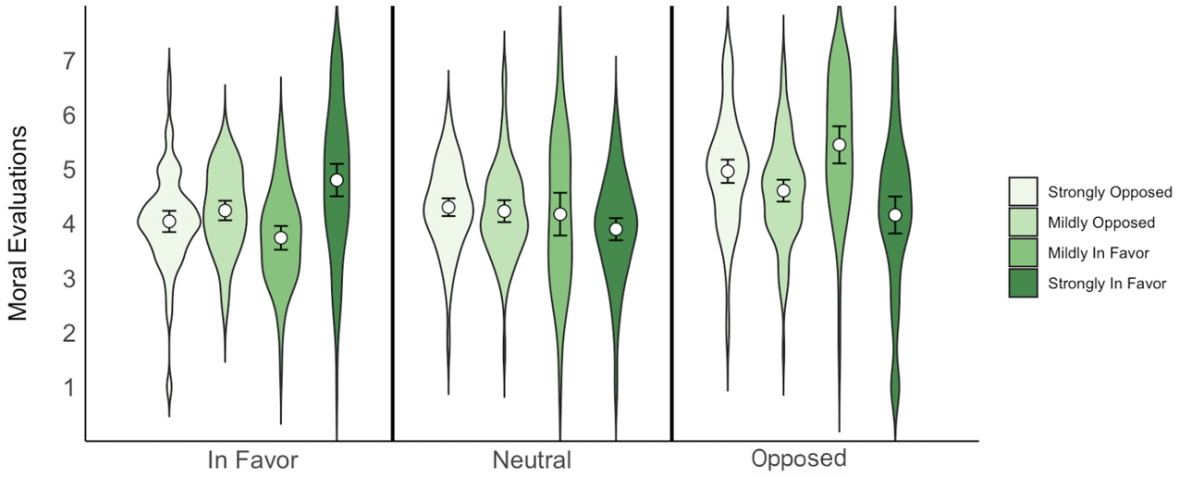
In addition to the measures described in the main text, participants also completed a four-item measure of authenticity (authentic; sincere; expresses their true values; tries to be something they're not (reverse coded); ω = .89). They also indicated the extent to which the target was competent, cowardly, self-interested, deceitful, and how much they thought the average Facebook user would like them, all on 7-points scales. Given the context, we also collected exploratory variables assessing how participants would respond to the post, including how they would respond in an open-ended comment, and use of Facebook's pre-existing responses (like, love, sad, angry, no reaction), and their openness to hearing more from the poster. To assess our preregistered prediction that neutrality might reduce moralization of the issue, participants completed a two-item measure of moralization: "To what extent is your position on the issue of affirmative action a reflection of your core moral beliefs and convictions?" and "To what extent are your feelings about the issue of affirmative action connected to your beliefs about 'right' and 'wrong'? (Skitka et al. 2005).

Participants indicated what they thought the social media poster's true private position on the issue was, as in previous studies on a scale from -4 (*Completely against affirmative action*) to 0 (*Neutral*) to +4 (*Completely in favor of affirmative action*), and completed the same item to assess their own stance on the issue.

**Moderation by Participants' Attitude Extremity.** To assess whether participants' own attitudes toward affirmative action affected how they responded to targets, we examined the interaction between stance condition (in favor vs. neutral vs. opposed) and attitude extremity (mildly in favor, i.e., those who selected +1 or +2, vs. strongly in favor, i.e., those who selected

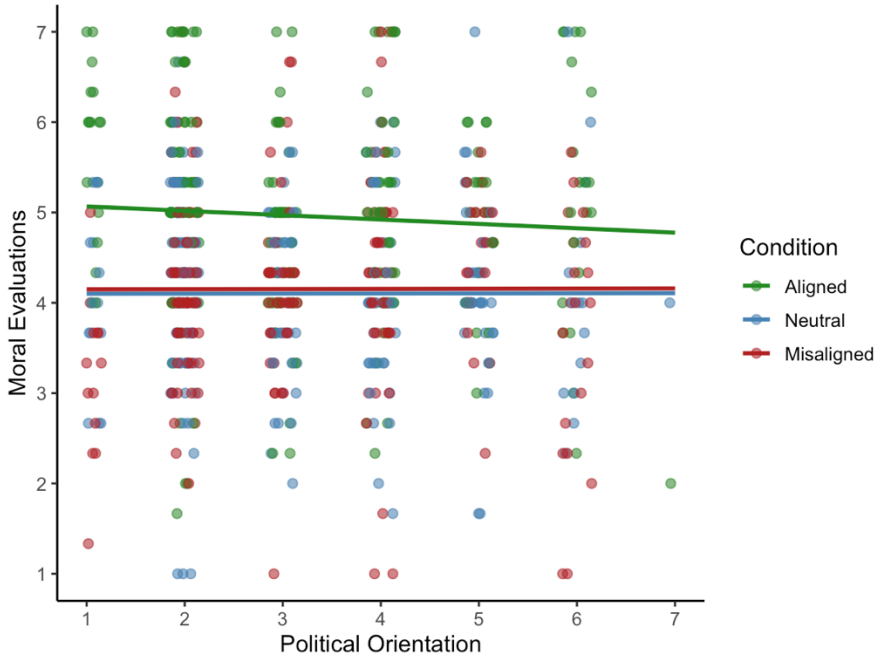
+3 or +4, vs. mildly opposed, i.e., those who selected -1 or -2, vs. strongly opposed, i.e., those who selected -3 or -4) on moral evaluations, and found a significant effect of position condition, $F(2, 900) = 33.31$, $p < .001$, partial-$\eta^2 = 0.069$, no main effect of attitude extremity, $F(3, 900) = 1.20$, $p = .309$, partial-$\eta^2 = 0.004$, and a significant interaction, $F(6, 900) = 16.03$, $p < .001$, partial-$\eta^2 = 0.097$, such that participants who were more strongly opposed or more strongly in favor of affirmative action responded more favorably to those who were aligned with their position than did those who held weaker positions on the issue (see Figure S3). The results are similar using the continuous measure of attitude extremity.

**Figure S3.** The effect of condition and issue position on moral evaluations in Study 1C.




**Moderation by Participants' Political Orientation.** We also examined the interaction between alignment condition (aligned, neutral, and misaligned) and participants' political orientation (1 = extremely liberal; 7 = extremely conservative) on moral evaluations using PROCESS Model 1. We found a significant main effect of condition, $b_{neutral} = -0.93$, $SE = .22$, $p < .001$, $b_{misaligned} = -0.75$, $SE = .21$, $p = .005$. There was no main effect of political orientation, $b = -.05$, $SE = .04$, $p = .210$. Additionally, the interaction between condition and political orientation was not significant, $b = -0.05$, $SE = .04$, $p = .493$ (see Figure S4).

**Figure S4.** The effect of alignment group and political orientation on moral evaluations in Study 1C.




## Additional Variables

**Perceived True Stance.** There was a significant main effect of target position condition (in favor vs. neutral vs. opposed) on perceptions of the poster's "true" (privately held) position on the issue, $F(2, 909^{1}) = 257.18$, $p < .001$, partial-$\eta^2 = .361$. Participants in the opposed condition thought the poster was more against the issue ($M = 3.52$, $SD = 2.50$, 95% CI [3.29, 3.75]) than those in the neutral condition ($M = 5.08$, $SD = 1.87$, 95% CI [4.85, 5.31]), or the in-favor condition ($M = 7.28$, $SD = 1.73$, 95% CI [7.05, 7.51]). All conditions significantly differed from each other, $t$s > 9.28, $p$s < .001, $d$s > 0.71. To examine whether this inference accounting for our results in the main text, we examined whether the results of alignment group (aligned versus neutral versus misaligned) on moral evaluations held while controlling for perceived true stance. We found support for this result, $F_{alignment}(2, 908) = 32.21$, $p < .001$, $\eta^2 = .070$, suggesting that our results are not driven by the perception that the neutral target holds a privately opposing position.

**Engagement with Post.** There was no significant effect of alignment group on participants' likelihood of commenting on the post, $F(2, 909) = 0.78$, $p = .461$, partial-$\eta^2 = 0.002$. Participants who observed a neutral target were no more likely to comment on the post ($M = 2.01$, $SD = 1.36$, 95% CI [1.93, 2.25]) than those who observed an aligned target ($M = 2.09$, $SD = 1.35$, 95% CI [1.93, 2.24]) or a misaligned target ($M = 2.15$, $SD = 1.53$, 95% CI [1.98, 2.32]), $p$s ≥ .201. The aligned and misaligned groups did not differ significantly ($p = .968$). Comment content from the neutral group included a question asking (e.g., "Can I ask why you're neutral?" as well as stating the commenters' own thoughts on the matter (e.g., "I am all for fairness be it male or female we are all equal and should be treated as such").

[1] Due to missing responses for some participants' dependent measures, the degrees of freedom vary between outcomes.

There was also a significant effect of condition on participants' emotional reactions to the post using Facebook's pre-existing responses, $\chi^2(8, 912) = 66.23$, $p < .001$, $\phi = .27$ (see Table S1). Of note, significantly more participants in the misaligned group chose the "angry" and "sad" reactions than did participants responding to a neutral or misaligned target; those responding to the neutral or aligned targets did not differ in their anger.

**Table S1**. Responding to the Facebook Post by Alignment Condition.

| | Condition | | | |
|---|---|---|---|---|
| | Misaligned | Neutral | Aligned | Omnibus Test df(2, 912) |
| Like | $16.23\%_a$ | $24.34\%_b$ | $32.33\%_c$ | $\chi^2 = 21.46$, $p < .001$, $\phi = .15$ |
| Love | $1.95\%_a$ | $1.64\%_a$ | $6.33\%_b$ | $\chi^2 = 13.06$, $p = .001$, $\phi = .12$ |
| Sad | $10.39\%_a$ | $4.93\%_a$ | $6.00\%_a$ | $\chi^2 = 7.74$, $p = .021$, $\phi = .09$ |
| Angry | $12.34\%_a$ | $2.63\%_b$ | $5.00\%_b$ | $\chi^2 = 25.14$, $p < .001$, $\phi = .17$ |
| No Reaction | $59.10\%_a$ | $66.45\%_b$ | $50.33\%_a$ | $\chi^2 = 16.20$, $p < .001$, $\phi = .13$ |

*Note*. Means containing different subscripts differ at the $p \leq .016667$ level (to adjust for multiple comparisons).

**Openness to Hearing More.** There was a significant effect of alignment on participants' interest in hearing more from the other side, $F(2, 908) = 4.05$, $p = .018$, $\eta^2 = .009$, such that participants in the aligned condition ($M = 4.37$, $SD = 1.95$, 95% CI [4.15, 4.60]) were significantly more interested in hearing more about the poster's perspective than were those in the misaligned group ($M = 3.92$, $SD = 2.04$, 95% CI [3.64, 4.15]), $t = 2.80$, $p = .014$, $d = 0.23$, 95% CI[0.07, 0.38]. Participants in the neutral condition ($M = 4.08$, $SD = 1.98$, 95% CI [3.86, 4.30]) did not differ significantly from either the aligned, $p = .165$, or opposed, $p = .581$, groups.

**Authenticity.** There was also a significant effect of alignment on judgments of the poster's authenticity, $F(2, 909) = 54.91$, $p < .001$, $\eta^2 = .109$. Participants in the neutral condition rated the poster as significantly less authentic ($M = 3.94$, $SD = 1.19$, 95% CI [3.81, 4.08]) than did those in the aligned group ($M = 4.91$, $SD = 1.13$, 95% CI [4.79, 5.04]), $t(602) = 10.46$, $p < .001$, $d = 0.84$, 95% CI [0.67, 1.00], or misaligned group ($M = 4.44$, $SD = 1.21$, 95% CI [4.35, 4.52]), $t(609) = 5.75$, $p < .001$, $d = 0.46$, 95% CI [0.30, 0.62]. Participants also thought that the poster was more authentic when they were aligned than when they were misaligned, $t(605) = 4.75$, $p < .001$, $d = 0.39$, 95% CI [0.23, 0.55].

**Moralization.** Counter to our preregistered prediction, condition did not affect participants' own moralization of affirmative action at Time 2, $F(2, 909) = 0.04$, $p = .958$ ($M_{aligned} = 4.83$, $SD = 1.52$; $M_{neutral} = 4.85$, $SD = 1.50$; $M_{misaligned} = 4.85$, $SD = 1.50$), suggesting that neutrality does not inhibit (nor boost) issue moralization.

Results for additional variables are reported in Table S2.

**Table S2. Exploratory Items by Alignment Condition.**

| | Alignment Condition | | |
|---|---|---|---|
| | Aligned | Neutral | Misaligned |
| Trustworthy* | 4.36 (1.19)$_a$ | 3.89 (1.10)$_b$ | 3.85 (1.14)$_b$ |
| Cowardly | 2.48 (1.43)$_a$ | 3.39 (1.62)$_b$ | 3.16 (1.60)$_b$ |
| Self-interested | 3.81 (1.60)$_a$ | 4.01 (1.49)$_a$ | 4.43 (1.65)$_b$ |
| Deceitful | 2.57 (1.33)$_a$ | 3.15 (1.40)$_b$ | 3.09 (1.41)$_b$ |
| Liking by average Facebook user | 3.73 (1.09)$_a$ | 3.96 (1.07)$_b$ | 4.04 (1.21)$_b$ |

*Note*. Cells containing different subscripts differ from one another at $p < .05$ based on the results of Tukey post hoc tests (to adjust for multiple comparisons).

**Additional Analyses for the Moral Evaluations Measure.** We conducted a factor analysis using principal axis factoring to examine whether the moral evaluations scale items represented a single factor. The analysis revealed a one-factor solution (eigenvalue = 2.18, explaining 72.74% of the total variance).

However, because honesty is only sometimes linked to overall moral perceptions (e.g., Levine & Schweitzer, 2014), and because the strategic attribution might be linked with dishonesty, we wanted to ensure that the results were robust to the exclusion of the item "honest." We describe the results at the item level below (as well as the results for other related items), as well as the results of the analyses when excluding "honest."

**Table S3. Results by Condition at the Item Level.**

| | Alignment Condition | | |
|---|---|---|---|
| | Aligned | Neutral | Misaligned |
| Moral | 4.71 (1.32)$_a$ | 4.20 (1.01)$_b$ | 3.97 (1.33)$_b$ |
| Ethical | 4.67 (1.36)$_a$ | 4.09 (1.11)$_b$ | 3.89 (1.40)$_b$ |
| Honest | 5.11 (1.37)$_a$ | 4.14 (1.39)$_b$ | 4.64 (1.33)$_c$ |

*Note*. Cells containing different subscripts differ from one another at $p < .05$ based on the results of Tukey post hoc tests (to adjust for multiple comparisons).

The results here are robust to the exclusion of "honesty," $F(2, 909) = 31.38$, $p < .001$, such that participants in the aligned group were rated as more moral ($M = 4.69$, $SD = 1.30$) than were those in the neutral ($M = 4.14$, $SD = 1.00$) and opposed conditions ($M = 3.93$, $SD = 1.32$), $p$s $< .001$. The neutral and opposed conditions did not significantly differ, $p = .074$.

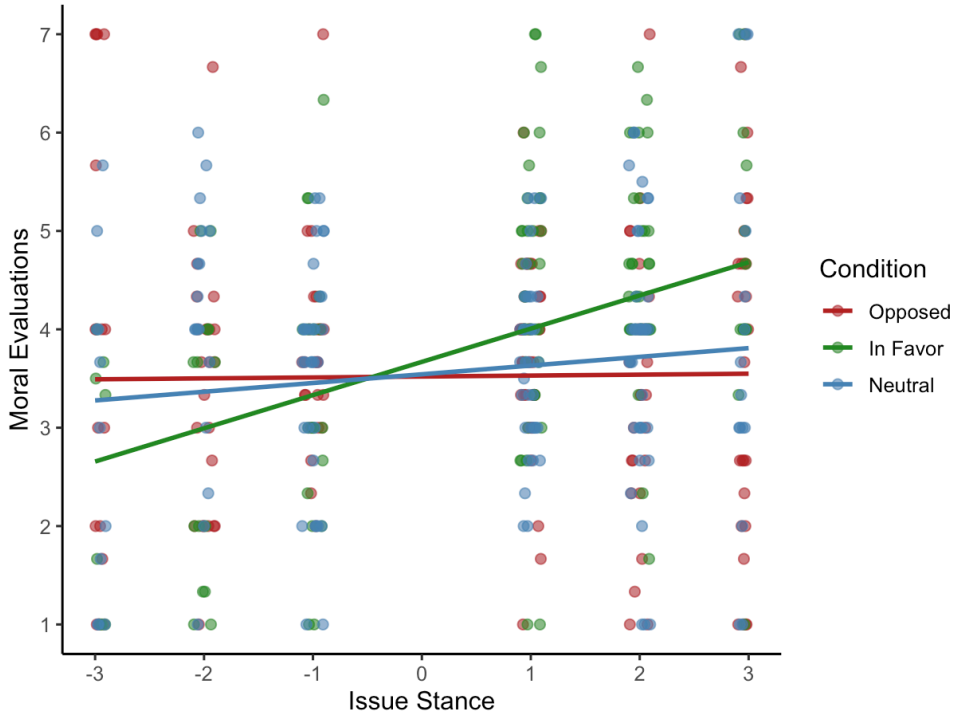
*A final measurement note is that we conducted our analyses in the main text with the preregistered three item scale described above. However, given related work (e.g., Zlatev, 2019), and high correlations in this study with trustworthy, in subsequent studies, we also include this item. The results are unchanged in this study when including this item.

### 1.4 Study 1D Supplemental Analyses

**Moderation by Participants' Attitude Extremity.** To assess whether participants' own attitudes toward affirmative action affected how they responded to targets, we examined the

interaction between stance condition (in favor vs. neutral vs. opposed) and participants' continuous attitude measure (7-point scale in Study 1D) on moral evaluations. There was again a significant interaction ($con_{opposed}$ × participant stance: $B = 0.33$, $p < .001$), such that participants who were more strongly opposed or more strongly in favor of affirmative action responded more favorably to targets who were aligned with their position than did those who held weaker positions on the issue. This interaction was not significant for neutral targets ($con_{neutral}$ × participant stance: $B = 0.08$, $p = .241$). As seen in Supplemental Figure 5 below, the influence of participants' own attitudes on moral evaluations was strongest among those who were more strongly in favor of affirmative action. We do not observe this finding in the other studies (i.e., we typically observe similar effects from those in favor or opposed to an issue), so we are reluctant to draw conclusions from this particular pattern of data.

**Figure S5.** The effect of stance condition and participants' own issue stance on moral evaluations in Study 1D.




**Neutral Participants.** There was a larger share of neutral participants in this study than in others, so we also examined the effect of target position condition on neutral participants, $F(2, 468) = 5.10$, $p = .006$, partial-$\eta^2 = 0.021$. Neutral participants judged the target opposed to affirmative action as less moral ($M = 3.56$, $SD = 1.27$) than the target in favor of affirmative action ($M = 3.99$, $SD = 1.22$), $p = .004$, but similar to the neutral target ($M = 3.76$, $SD = 1.11$), $p = .286$. The neutral target also did not differ from the in favor target, $p = .239$.

## Additional Measures

### Authenticity

There was no main effect on authenticity judgments, $F(2, 542) = 2.10$, $p = .124$, partial-$\eta^2 = 0.01$. Participants in the neutral group were rated as similarly authentic ($M = 3.74$, $SD = 1.44$, 95% CI [3.53, 3.96]) as those in the aligned group ($M = 4.06$, $SD = 1.56$, 95% CI [3.84, 4.27]),

$t(373) = 2.07$, $p = .109$, $d = 0.21$. Misaligned targets ($M = 3.85$, $SD = 1.53$, 95% CI [3.63, 4.08]) were rated as similarly authentic as were those in the other two groups, $p$s > .404.

**Openness to Target**

There was a significant effect of alignment on participants' interest in hearing more from the other side, $F(2, 552) = 6.30$, $p = .002$, partial-$\eta^2 = .022$, such that participants in the aligned group ($M = 3.26$, $SD = 1.93$, 95% CI [3.00, 3.52]) were significantly more interested in hearing more about the poster's perspective than were those in the misaligned group ($M = 2.58$, $SD = 1.81$, 95% CI [2.31, 2.86]), $t = 3.79$, $p < .001$, $d = 0.23$. Participants in the neutral condition ($M = 2.91$, $SD = 1.76$, 95% CI [2.65, 3.17]) were similarly interested in hearing about the poster's perspective as those in the misaligned group, $p = .210$. They also did not differ significantly from those in the aligned group, $p = .146$, suggesting that, although the neutral target was rated as less moral, participants indicated being just as willing to engage with this target as those who agreed with them.

**Moral Evaluations**

We conducted a factor analysis using principal axis factoring to examine whether the moral evaluations scale items represented a single factor. The analysis revealed a one-factor solution (eigenvalue = 2.47, explaining 82.47% of the total variance). We again provide the results at the item level for the moral evaluations measure below in Table S4.

**Table S4.** Means (and standard deviations) at the item level by whether the target was neutral or had a view that was aligned/misaligned with participants.

| | Condition | | |
|---|---|---|---|
| | Aligned | Neutral | Misaligned |
| Moral | 3.93 (1.63)$_a$ | 3.54 (1.46)$_b$ | 3.07 (1.53)$_c$ |
| Ethical | 3.99 (1.60)$_a$ | 3.53 (1.46)$_b$ | 3.12 (1.66)$_c$ |
| Honest | 4.08 (1.57)$_a$ | 3.66 (1.53)$_b$ | 3.75 (1.72)$_{ab}$ |

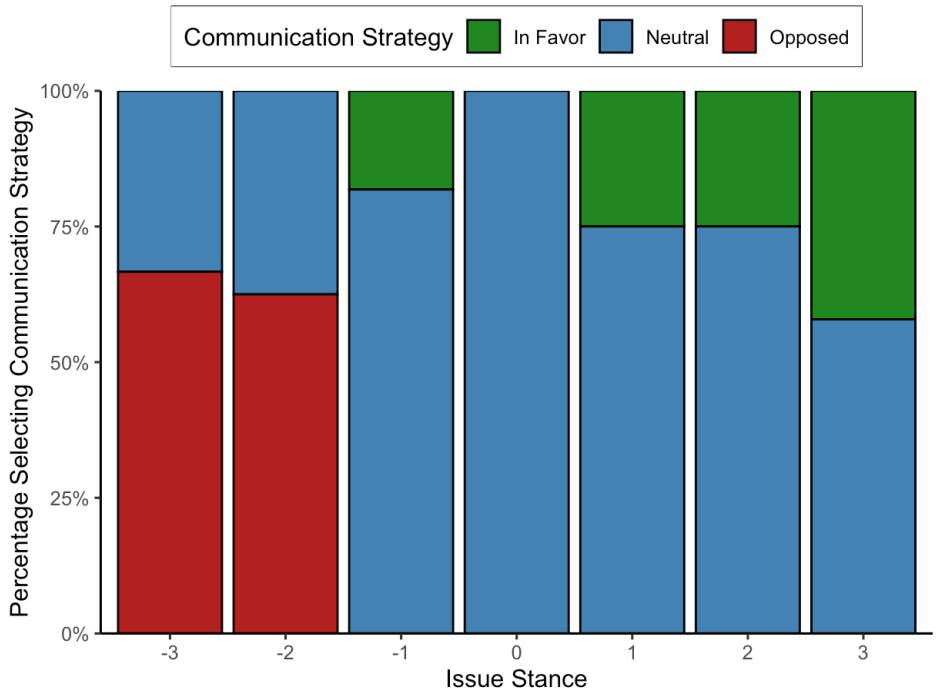
*Note*. Cells containing different subscripts differ from one another at $p < .05$ based on Tukey post hoc tests.

Examining the results of alignment condition on moral evaluations without the "honest" item, we again observe the main effect on moral evaluations, $F(2, 542) = 15.39$, $p < .001$, partial-$\eta^2 = 0.054$. In this study, aligned targets were rated as more moral ($M = 3.96$, $SD = 1.56$) than were neutral ($M = 3.55$, $SD = 1.41$) or misaligned ($M = 3.08$, $SD = 1.51$), $p$s < .008. Neutral targets were rated as more moral than misaligned targets when the honest item was excluded, $p = .010$.

## 1.5 Study 2A Supplemental Analyses

We also examined the recommended communication strategy as a function of participants' own continuous stance on the issue. Given the smaller sample size due to our sample of experts, the data violate the statistical assumptions necessary for chi-square analyses (some cells have fewer than 5 observations). Thus, we present the results below graphically in Figure S6 for descriptive interpretation.

**Figure S6.** Communication strategy by issue position in Study 2A.

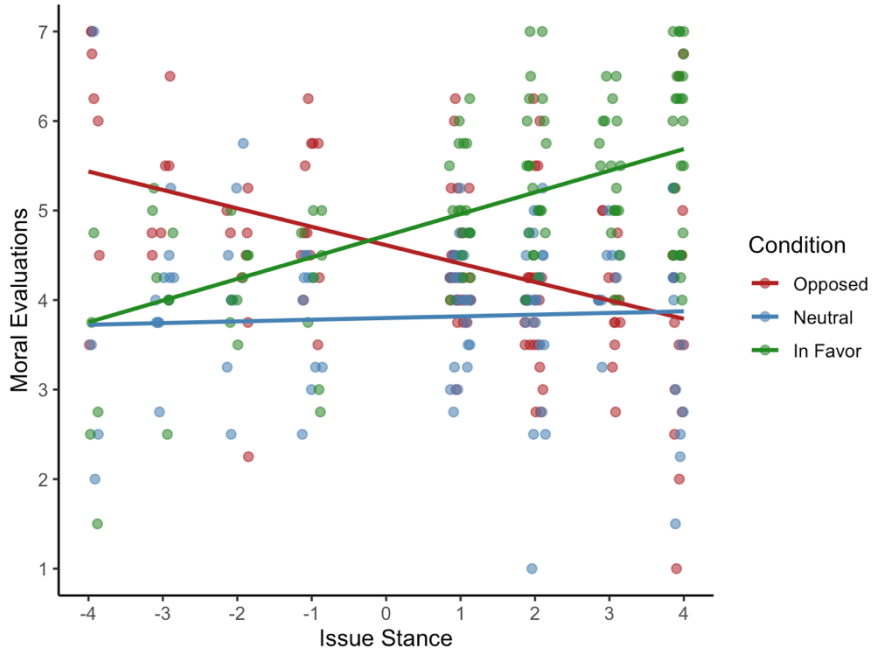



### 1.6 Study 2B Supplemental Analyses

We measured supplemental variables on scales ranging from 1 (not at all) to 7 (very much) and conducted omnibus tests (and post-hoc tests) of the effects of alignment condition on these variables. Means with different subscripts are significantly different from one another according to contrast tests at $p \leq .05$ (see Table S5 below). All other dependent measures are available on OSF.

**Moderation by Participants' Attitude Extremity.** To assess whether participants' own attitudes toward affirmative action affected how they responded to targets, we examined the interaction between stance condition (in favor vs. neutral vs. opposed) and attitude extremity (1-7 scale) on moral evaluations. The analysis revealed a significant main effect of stance condition, $B_{neutral} = -1.96$, $SE = 0.30$, $t = -6.52$, $p < .001$, $B_{misaligned} = 2.25$, $SE = .31$, $t = -7.22$, $p < .001$. There was also a main effect of attitude extremity, $B = -0.21$, $SE = 0.03$, $t = -6.13$, $p < .001$. However, a significant interaction emerged between stance condition and attitude extremity. Specifically, participants' attitude extremity interacted with the in-favor condition, $B = -0.22$, $SE = 0.05$, $t(391) = -4.93$, $p < .001$, such that more strongly opposed participants rated the in-favor target as less moral; similarly, they rated the opposed target as more moral, $B = 0.23$, $SE = 0.04$, $t(391) = 5.09$, $p < .001$ (see Figure S7 below).

**Figure S7.** The effect of stance condition and participants' attitudes on moral evaluations in Study 2B.




**Additional variables:**

**Table S5.** Results by alignment condition for Study 2B.

| | Condition | | | Omnibus test |
|---|---|---|---|---|
| Measures | Aligned | Neutral | Misaligned | |
| Authenticity (4 items, $\alpha$ = .86) | 5.13 (1.09)$_a$ | 3.52 (1.18)$_b$ | 4.55 (.95)$_c$ | $F(2, 394) = 76.05$, $p < .001$, partial-$\eta^2 = .279$ |
| Trustworthy* | 5.05 (1.17)$_a$ | 3.69 (1.14)$_b$ | 3.99 (1.15)$_b$ | $F(2, 394) = 50.40$, $p < .001$, partial-$\eta^2 = .204$ |
| Likeable | 5.05 (1.04)$_a$ | 3.73 (1.08)$_b$ | 3.41 (1.18)$_c$ | $F(2, 394) = 82.96$, $p < .001$, partial-$\eta^2 = .296$ |
| Competent | 4.99 (1.14)$_a$ | 3.66 (1.27)$_b$ | 3.82 (1.05)$_b$ | $F(2, 394) = 52.20$, $p < .001$, partial-$\eta^2 = .209$ |
| Cowardly | 2.20 (1.23)$_a$ | 4.30 (1.59)$_b$ | 3.42 (1.41)$_c$ | $F(2, 394) = 73.35$, $p < .001$, partial-$\eta^2 = .271$ |
| Self-interested | 3.20 (1.47)$_a$ | 4.44 (1.43)$_b$ | 4.29 (1.40)$_b$ | $F(2, 394) = 29.52$, $p < .001$, partial-$\eta^2 = .130$ |
| Deceitful | 2.56 (1.21)$_a$ | 3.64 (1.34)$_b$ | 3.09 (1.27)$_c$ | $F(2, 394) = 23.86$, $p < .001$, partial-$\eta^2 = .108$ |
| Liked by others | 3.33 (0.75)$_a$ | 2.77 (0.78)$_b$ | 3.02 (0.79)$_c$ | $F(2, 394) = 18.79$, $p < .001$, partial-$\eta^2 = .087$ |
| Issue moralizing at Time 2 | 5.48 (1.22)$_a$ | 5.07 (1.17)$_b$ | 5.31 (1.22)$_{ab}$ | $F(2, 394) = 3.93$, $p = .020$, partial-$\eta^2 = .020$ |

*Note*. Cells containing different subscripts differ at the $p < .05$ level based on Tukey post hoc tests (to adjust for multiple comparisons).

**Moralization.** We used the same two items to measure moralization as in Study 1C, in both a pretest, $r(397) = .74$, $p < .001$, and after the dependent measures, $r(397) = .68$, $p < .001$.

Our preregistered analysis was a test of moderation: whether the degree to which participants see the issue of affirmative action as moral moderates the effects of target alignment or misalignment vs. neutrality. The interaction between target alignment (1) vs. neutrality (0) and moralization was significant, $B = .29$, $SE = .09$, $p < .001$, $R^2 = .02$, but the interaction between target misalignment (1) vs. neutrality (0) and moralization was not, $B = -.09$, $SE = .08$, $p = .31$, $R^2 = .02$. Aligned targets were seen as more moral than neutral targets at low levels of moralization, $B = 1.10$, $SE = .17$, $p < .001$, and this effect was even more pronounced at high levels of moralization, $B = 1.83$, $SE = .17$, $p < .001$. The between cell differences for moralization are available in Table S5.

**Perceived True Stance.** There was a significant main effect of target position condition on perceptions of the poster's "true" (privately held) position on the issue, $F(2, 394) = 147.82$, $p < .001$, $\eta^2 = .429$. Participants in the opposing condition thought the politician was more against the issue ($M = 2.31$, $SD = 1.57$, 95% CI [2.07, 2.56]) than did those in the neutral condition ($M = 3.52$, $SD = 1.49$, 95% CI [3.27, 3.77]), or the in-favor condition ($M = 5.38$, $SD = 1.30$, 95% CI [5.13, 5.63]). All conditions significantly differed from each other, $p$s $< .001$. The results of alignment condition on moral evaluations hold when controlling for perceived true position, $F_{condition}(2, 541) = 12.70$, $p < .001$, $\eta^2 = .04$, suggesting that our results are not driven by the perception that the neutral target holds a privately opposing position.

**Attributions.** We asked participants in the neutral condition six items measuring what attribution they made for the target's neutrality: "they see the pros and cons to each side or option, and they balance each other out" (indecision); "they don't care about the issue" (apathy); "They had no opinion on the issue" (lacking information); "They didn't want to take sides between people" (conflict avoidance); "They feared the consequences of taking a stance" (strategic 1); and "They didn't want to tell others their true opinion" (strategic 2). Given that only participants in the neutral condition completed this measure, we do not have condition-level results, but the correlations with the morality measure are listed in the Table S6.

**Table S6.** Correlations between moral evaluations and the attributions.

| | 1. | 2. | 3. | 4. | 5. | 6. | 7. |
|---|---|---|---|---|---|---|---|
| 1. Moral | | .65** | -.30** | .24** | -.01 | -.46** | -.43** |
| 2. Indecision | | | -.23** | .30** | -.05 | -.41** | -.32** |
| 3. Apathy | | | | .32** | .13 | .28** | .23** |
| 4. Lacking information | | | | | .19* | -.05 | -.05 |
| 5. Conflict avoidance | | | | | | .40** | .53* |
| 6. Strategic 1 | | | | | | | .78** |
| 7. Strategic 2 | | | | | | | |

**$p < .01$

**Moral Evaluations**

We again provide the results at the item level for the moral evaluations measure below in Table S7.

**Table S7**. Means (and standard deviations) for Study 2B by whether the target was neutral or had a view that was aligned/misaligned with participants.

| | Condition | | |
|---|---|---|---|
| | Aligned | Neutral | Misaligned |
| Moral | 5.42 (1.13)$_a$ | 3.73 (1.15)$_b$ | 3.64 (1.36)$_b$ |
| Ethical | 5.39 (1.20)$_a$ | 3.73 (1.17)$_b$ | 3.44 (1.42)$_b$ |
| Honest | 5.50 (1.29)$_a$ | 4.10 (1.52)$_b$ | 5.22 (1.28)$_a$ |

*Note*. Cells containing different subscripts differ at the $p < .05$ level based on Tukey post hoc tests (to adjust for multiple comparisons).

These results are again robust to the exclusion of "honest," $F(2, 394) = 98.72$, $p < .001$, $\eta^2 = .334$. Examining the composite moral evaluations score without the "honest" item, participants in the aligned condition were rated as more moral ($M = 5.41$, $SD = 1.12$) than were those in the neutral ($M = 3.73$, $SD = 1.12$) or opposed ($M = 3.54$, $SD = 1.31$) conditions, $p$s < .001. The neutral and opposed conditions did not significantly differ, $p = .396$.

*A final measurement note is that we conducted our analyses in the main text with the preregistered three item scale described above. However, given related work (e.g., Zlatev, 2019), and high correlations in this study with trustworthy, in subsequent studies, we also include this item. Of note, when trustworthy is included in the moral evaluations composite, the neutral target is rated as less moral than the misaligned target, $t(394) = 2.12$, $p = .035$, $d = 0.26$.

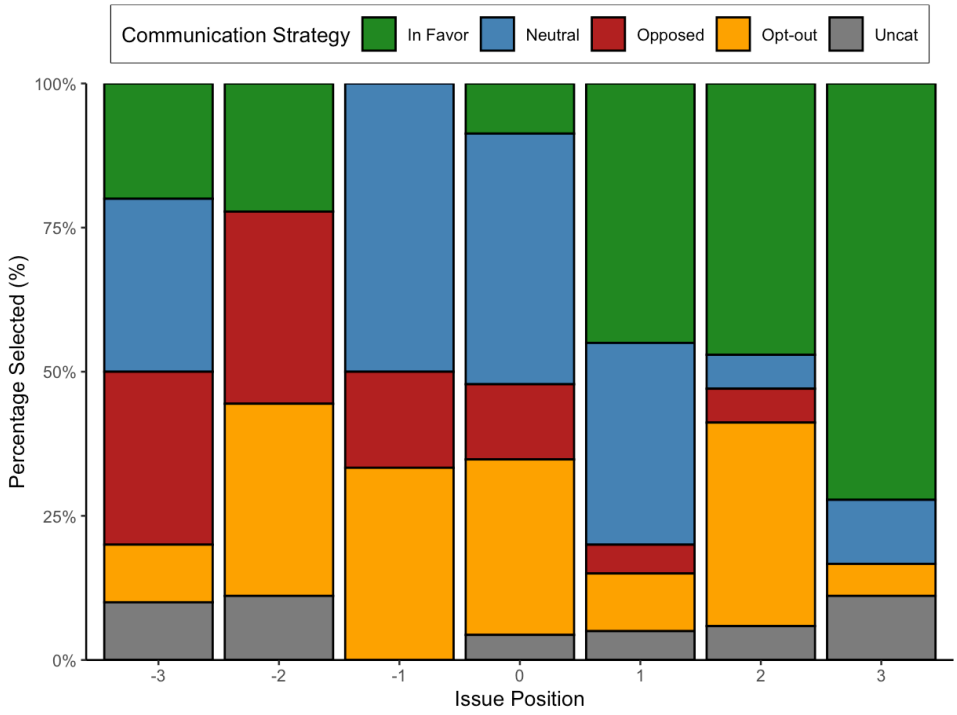
### 1.7 Study 3A Supplemental Analyses

Figure S8 displays the raw data for communication strategy by condition via a scatterplot.

**Figure S8.** Scatterplot of communication strategy by condition in Study 3A.


**Participants' Own Stance.** Among participants in the public condition, continuous issue position also significantly predicted participants' communication strategy (this was irrelevant in the private condition), $\chi^2(3) = 23.56$, $p < .001$ (see Figure S9).

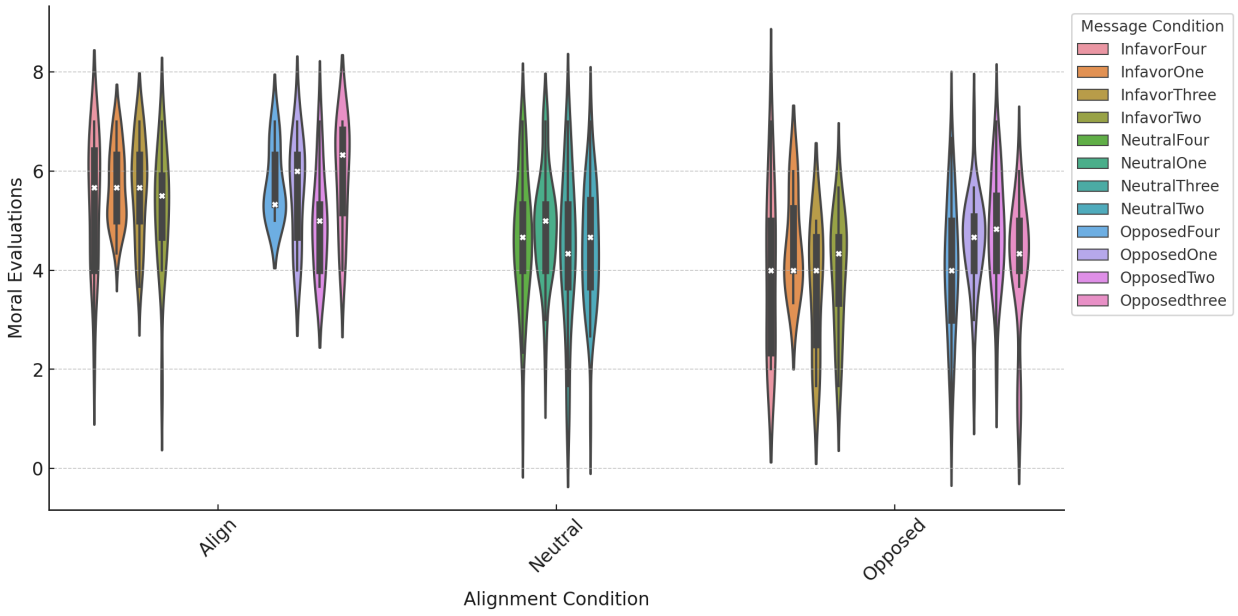
**Figure S9.** Selection of communication strategy by issue position among participants in the public condition in Study 3A.



## 1.8 Study 3B Supplemental Analyses

A mixed linear model was conducted to examine the effects of message type and alignment group on moral evaluations. The model included 12 messages and 3 conditions as fixed effects. There was a significant effect of alignment condition, $F(2, 482) = 77.45$, $p < .001$, no main effect of message type, $F(10, 482) = 1.17$, $p = .306$, and no interaction, $F(7, 482) = 1.97$, $p = .057$. Relative to the aligned group, the misaligned group led to significantly lower moral evaluations ($p < .001$), as did the neutral condition ($p < .001$). The misaligned and neutral groups did not differ, $p = .050$. Following the recommendations of Simonsohn et al. (2025), Figure S10 presents the results for moral evaluations by alignment group and the specific stimuli (message) to which participants were exposed.

**Figure S10.** Results of alignment group by stimulus in Study 3B.



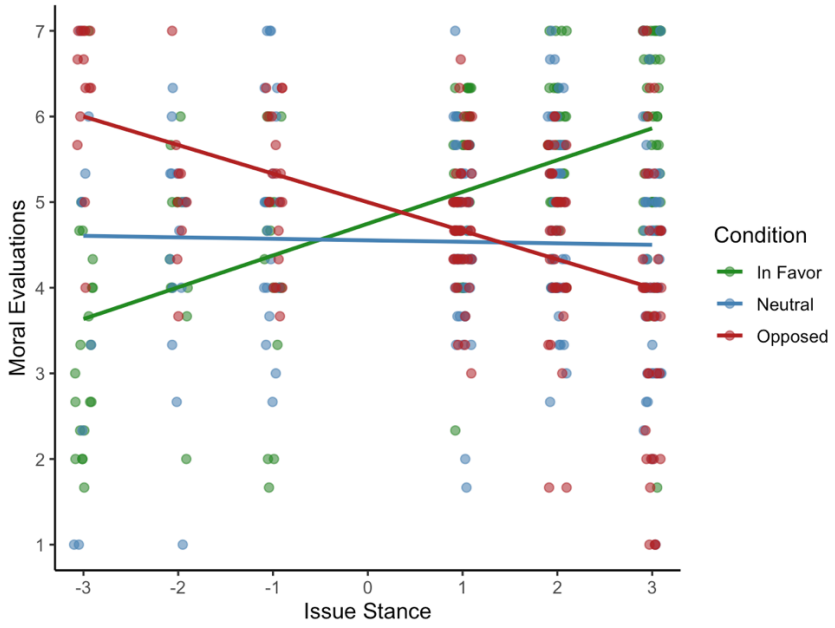

*Note.* The message labels correspond to the following:
In favor one: "Yes, I would say immigration is a good thing and I think it's how this country was built and what makes this country unique."
In favor two: "You know, I think this is an important issue, and personally I think we should increase immigration."
In favor three: "You know, all things considered, I personally think increasing immigration levels would be a good thing for the country."
In favor four: "Personally, I think we should increase immigration levels at this time."
Neutral one: "You know, I really don't have a stance on the issue."
Neutral two: "I don't really have a view on it."
Neutral three: "I don't have an opinion, I don't follow the stuff."
Neutral four: "You know, I think it can go either way really."
Opposed one: "No - I would say that immigration in this country is not well monitored, and I think the government could do a much better job than they are presently doing to make it work."
Opposed two: "You know, I think this is an important issue, and personally I think we should keep it at current levels."

Opposed three: “You know, all things considered, I personally don't think increasing immigration levels would be a good thing for the country.”
Opposed four: “Personally, I don't think we should increase immigration levels at this time.”

**Perceived True Stance.** In a mixed linear model to account for the different messages, we see that all target stance conditions significantly differed from each other (In favor = 5.95, *SE* = .31; Neutral = 3.69, *SE* = .31; Opposed = 2.69, *SE* = 0.31). In the model with message as a random effect, adding perceived true stance as a covariate, we again observed a significant main effect of condition, $F_{condition}(2, 49) = 33.97$, $p < .001$.

**Moderation by Participants' Attitude Extremity.** We again observed a significant interaction between participants' own attitude extremity and target stance conditions ($B_{opposedXstance} = -0.69$, $p < .001$; $B_{neutralXstance} = -0.41$, $p < .001$) (see Figure S11). Here, we collapsed across message condition for clarity.

**Figure S11.** The effect of stance condition and participants' issue stance on moral evaluations in Study 3B.




**Moral Evaluations.** In a linear mixed model, we also examined the results at the item level. Each model included 12 messages as a random effect, and 3 alignment groups as fixed effects. The results indicated that all overall models were significant, $F$s > 22.53, $p$s < .001. Relative to the aligned group, the misaligned and neutral groups led to significantly lower moral evaluations ($p$s < .001). Neutral targets were viewed as more ‘ethical’ than misaligned targets ($p$ = .028), but similarly ‘honest’ and ‘trustworthy’ ($p$s = .089 and 1.00). The results are robust to the exclusion of ‘honest,’ $F = 50.98$, $p < .001$. Relative to the aligned group, the misaligned group led to significantly lower moral evaluations, as did the neutral group, ($p$s < .001). The neutral group was more moral than the misaligned here, $p = .036$.

## 1.9 Study 5 Supplemental Analyses

**Perceived True Stance.** Participants were asked about the target's true (privately held) opinion on increasing immigration on a 7-point scale. There was a significant effect of target stance condition, $F(2, 874) = 45.83$, $p < .001$, $\eta^2 = .296$. Participants in the in-favor condition rated the target as more in favor of immigration ($M = 6.14$, $SD = 1.09$) than did participants in any other condition, $p$s < .001. Likewise, the opposed target was rated as significantly more opposed to immigration ($M = 3.05$, $SD = 1.64$) than those in the opt out ($M = 3.71$, $SD = 1.28$), undecided ($M = 3.81$, $SD = 1.58$), conflict avoidance ($M = 3.88$, $SD = 1.18$), lacking information ($M = 4.02$, $SD = 1.25$), and in-favor conditions, but were rated as similar to the target in the strategic ($M = 3.23$, $SD = 1.57$), apathy ($M = 3.53$, $SD = 1.17$), and neutral baseline ($M = 3.60$, $SD = 1.35$) conditions. As in the previous studies, the effects of the alignment group on moral evaluations again remained when controlling for perceptions of true position, $F_{condition}(8, 873) = 10.30$, $p < .001$, partial-$\eta^2 = .086$.

**Exploratory information sharing measure.** We also collected an exploratory measure assessing participants' willingness to disclose sensitive information to the target: "Imagine that you learned confidential information about upcoming layoffs at your workplace. This is useful information, but you only want to share with people that you believe will keep this information confidential. To what extent would you be inclined to share this information with the colleague you just read about?" on a scale from 1 (*Not at all*) to 7 (*Very much so*).

There was an effect of condition on willingness to share, $F(2, 874) = 4.07$, $p < .001$, partial-$\eta^2 = .036$. Participants were more willing to share with targets in the aligned, opt out, undecided, conflict avoidance, and lacking information conditions than with targets in the misaligned, strategic, apathetic, or neutral baseline conditions. Means, standard deviations, and the results of post hoc tests are included in Table S8 below.

**Table S8**. The effect of alignment group on the information sharing measure in Study 5.

| Condition | M (SD) |
|---|---|
| Aligned | $4.39\ (1.93)_{a}$ |
| Neutral – baseline | $3.67\ (1.88)_{abc}$ |
| Misaligned | $3.14\ (1.89)_{c}$ |
| N-Conflict avoid | $3.90\ (1.77)_{abc}$ |
| N-Apathy | $3.27\ (1.98)_{bc}$ |
| N-Both sides | $3.58\ (2.01)_{abc}$ |
| N-Lack info | $3.51\ (1.84)_{bc}$ |
| N-Strategic | $3.80\ (1.82)_{abc}$ |
| Opt out | $4.03\ (2.12)_{ab}$ |

*Note*. Cells containing different subscripts differ at the $p < .05$ level based on Tukey post hoc tests (to adjust for multiple comparisons).

**Moral Evaluations.** We again present the item level results for the moral evaluations measure in Table S9.

**Table S9**. Means (and standard deviations) for Study 5 by whether the target was neutral or had a view that was aligned/misaligned with participants.

| | Dependent Measure | | |
|---|---|---|---|
| Condition | Moral | Trustworthy | Honest |
| Aligned | 5.74 (1.15)$_a$ | 5.71 (1.19)$_a$ | 6.02 (1.07)$_a$ |
| Neutral – baseline | 4.44 (1.40)$_b$ | 4.38 (1.54)$_b$ | 4.58 (1.66)$_c$ |
| Misaligned | 4.57 (1.20)$_b$ | 4.50 (1.28)$_b$ | 5.24 (1.40)$_{bc}$ |
| N-Conflict avoid | 4.83 (1.30)$_b$ | 4.89 (1.38)$_{ab}$ | 5.12 (1.50)$_c$ |
| N-Apathy | 4.34 (1.23)$_b$ | 4.45 (1.34)$_b$ | 5.08 (1.55)$_c$ |
| N-Both sides | 4.84 (1.41)$_b$ | 4.80 (1.49)$_b$ | 4.86 (1.61)$_c$ |
| N-Lack info | 4.66 (1.30)$_b$ | 4.75 (1.49)$_b$ | 5.03 (1.57)$_c$ |
| N-Strategic | 4.55 (1.50)$_b$ | 4.35 (1.55)$_b$ | 5.02 (1.73)$_c$ |
| Opt out | 5.48 (1.27)$_a$ | 5.44 (1.26)$_a$ | 5.87 (1.25)$_{ba}$ |

*Note*. Cells containing different subscripts differ at the $p < .05$ level based on Tukey post hoc tests (to adjust for multiple comparisons).

The results are robust to the exclusion of "honest," $F(8, 874) = 14.04$, $p < .001$, partial-$\eta^2 = .114$. Participants were rated as more moral in the aligned ($M = 5.73$, $SD = 1.12$) and opt out ($M = 5.46$, $SD = 1.21$) conditions than in the other conditions ($p$s $< .001$), and did not differ from each other, $p = .877$. The other conditions did not significantly differ from each other.

### 1.10 Study 6 Supplemental Analyses

Participants also completed the authenticity measure as in previous studies. We also asked how likely the participant would be to participate in this conversation at the dinner table (from 1= *not at all* to 7 = *definitely*). Like the previous study, we then also asked, "What do you think the person's true (privately held) opinion is?" (-3 = *against the construction*; 0 = *neutral*; +3 = *in favor of the construction*). Participants also indicated the extent to which the target was people-pleasing, cowardly, strategic, evasive, likable, competent, warm, fair, objective, open-minded, agreeable, courageous, self-interested, deceitful, difficult, overly sensitive, and not confident, all on 7-point scales from 1 (*Not at all*) to 7 (*Very much*). We again asked participants about their own position on the issue, the strength of their position, and the degree to which they moralize this issue (using the same moralization scale used previously). To address the possible alternative explanation that the neutrality scenario was more disfluent than the other two scenarios (for or against), which could separately drive negative reactions (e.g., Reber et al., 1998), we also asked participants, "To what extent was the scenario you read easy to imagine happening in real life?" (from 1 = *not at all* to 5 = *very much so*).

**Authenticity.** Alignment group significantly predicted participants' judgments of authenticity, $F(2, 420) = 22.37$, $p < .001$, $\eta^2 = .096$. Participants perceived the aligned dinner guest to be more authentic ($M = 5.43$, $SD = 1.13$, 95% CI [5.24, 5.61]) than the misaligned dinner guest ($M = 5.00$, $SD = 1.11$, 95% CI [4.81, 5.19]), $t(420) = 3.03$, $p = .007$, $d = 0.36$, 95% CI[.13, .60], or the neutral dinner guest ($M = 3.52$, $SD = 1.18$, 95% CI [3.31, 3.72]), $t(420) = 6.68$, $p < .001$, $d = 0.79$, 95% CI [.55, 1.03]. The misaligned dinner guest was rated as more authentic than the neutral dinner guest, $t(420) = 7.14$, $p = .001$, $d = 0.43$, 95% CI [.19, .66].

**Perceived True Stance.** There was a significant effect of the target's stance condition on perceived true stance, $F(2, 599) = 161.85$, $p < .001$, partial-$\eta^2 = .351$. Participants in the in-favor condition thought the target was more in favor of the issue ($M = 5.37$, $SD = 1.59$) than the neutral target ($M = 3.79$, $SD = 1.59$) and the opposed target ($M = 2.52$, $SD = 1.59$), $t$s > 7.92, $p$s < .001. The effect of alignment condition held when controlling for perceived true position, $F_{condition}(3, 419) = 13.04$, p < .001, $\eta^2 = .059$.

**Willingness to Participate in the Conversation.** There was a significant effect of alignment group on participants' willingness to engage in the conversation at the dinner table, $F(2, 420) = 6.07$, $p = .003$, partial-$\eta^2 = .023$. Participants were significantly more likely to engage in conversation with an aligned target ($M = 4.79$, $SD = 1.85$) than a misaligned target ($M = 4.00$, $SD = 1.99$), $p = .002$, but were similarly likely to converse with a neutral target ($M = 4.47$, $SD = 1.87$), $p = .338$. The neutral and misaligned groups did not differ, $p = .095$.

**Ease of Imagining the Scenario.** There was no significant effect of condition on ease of imagining, $F(2, 420) = 1.18$, $p = .308$, partial-$\eta^2 = .006$, suggesting that the neutrality scenario was not more disfluent than the other two conditions.

**Moderation by Issue Moralization.** In examining potential moderation by participants' moralization using PROCESS Model 1, there was no significant interaction between alignment group and moralization on moral evaluations, $p = .270$. There was similarly no interaction observed for authenticity judgments, $p = .365$.

**Additional measures.** In a series of ANOVAs, we also analyzed the effect of alignment group on the person perception measures described above (see means, standard deviations, and the results of post hoc tests in Table S10 below). The data tended to follow the same pattern as the moral evaluations measure, with neutral targets being rated as similar to misaligned targets, and aligned targets being rated more favorably. There were a few exceptions: neutral targets were rated as more people pleasing, strategic, evasive, unconfident, and less courageous than both aligned and misaligned targets; neutral targets were rated as less open-minded than aligned targets, but more so than misaligned targets, and less difficult than misaligned targets. All targets were rated as similarly overly sensitive.

Given that we measured warmth and competence in this study, it also gave us an opportunity to examine whether our results hold even after accounting for the 'big two' of person perception, which we observed, $F(2, 418) = 8.31$, $p < .001$. The results also hold after accounting for the effects of the likability measure, $F(2, 419) = 13.12$, $p < .001$.

**Table S10.**

| Dependent Measure | Condition | | |
|---|---|---|---|
| | Aligned | Neutral | Misaligned |
| People Pleasing | 3.29 (1.62)$_a$ | 3.79 (1.60)$_b$ | 3.28 (1.59)$_a$ |
| Cowardly | 2.18 (1.44)$_a$ | 3.01 (1.63)$_b$ | 2.64 (1.48)$_b$ |
| Strategic | 3.13 (1.56)$_a$ | 3.78 (1.64)$_b$ | 3.34 (1.33)$_a$ |
| Evasive | 2.43 (1.39)$_a$ | 3.70 (1.76)$_b$ | 2.91 (1.60)$_c$ |
| Likable | 4.81 (1.50)$_a$ | 4.29 (1.47)$_b$ | 4.18 (1.55)$_b$ |
| Competent | 4.97 (1.36)$_a$ | 4.31 (1.43)$_b$ | 4.47 (1.37)$_b$ |
| Warm | 4.61 (1.42)$_a$ | 4.08 (1.48)$_b$ | 3.99 (1.46)$_b$ |
| Fair | 4.85 (1.40)$_a$ | 4.26 (1.49)$_b$ | 3.97 (1.30)$_b$ |

| | | | |
|---|---|---|---|
| Objective | 4.43 (1.47)$_a$ | 4.01 (1.42)$_b$ | 3.91 (1.33)$_b$ |
| Open-minded | 4.85 (1.58)$_a$ | 4.11 (1.59)$_b$ | 3.58 (1.61)$_c$ |
| Agreeable | 4.68 (1.36)$_a$ | 4.33 (1.37)$_a$ | 3.89 (1.44)$_b$ |
| Courageous | 4.54 (1.42)$_a$ | 3.55 (1.41)$_b$ | 4.10 (1.31)$_c$ |
| Self-Interested | 3.37 (1.95)$_a$ | 3.57 (1.75)$_{ab}$ | 3.91 (1.68)$_b$ |
| Deceitful | 2.46 (1.58)$_a$ | 2.68 (1.55)$_a$ | 2.51 (1.35)$_a$ |
| Difficult | 3.03 (1.72)$_a$ | 3.17 (1.77)$_a$ | 3.51 (1.57)$_b$ |
| Overly Sensitive | 2.69 (1.58)$_a$ | 2.81 (1.49)$_a$ | 3.01 (1.50)$_a$ |
| Not Confident | 2.45 (1.43)$_a$ | 3.26 (1.56)$_b$ | 2.74 (1.55)$_a$ |

**Additional Mediation Analyses.** In Table S11, we report the correlations between the attributions.

**Table S11.** Correlations between the attributions.

| | | 1. | 2. | 3. | 4. | 5. |
|---|---|---|---|---|---|---|
| 1. | Indecision | | .58** | .42** | .03 | -.09* |
| 2. | Apathy | | | .56** | .23** | .08 |
| 3. | Lacking information | | | | .29** | .17** |
| 4. | Conflict avoidance | | | | | .64** |
| 5. | Strategic | | | | | |

**$p < .01$

Because of the correlations between the attributions, we conducted separate mediation analyses using PROCESS Model 4. The results were largely similar to the simultaneous mediation model: There was a significant, positive indirect effect of conflict avoidance, indirect effect = .13, *SE* = .04, 95% CI [.05, .22], a negative indirect effect of apathy, effect = -.53, *SE* = .09, 95% CI [-.71, -.36], no effect of indecision, effect = -.11, *SE* = .06, 95% CI [-.23, .01], a negative effect of lacking information, effect = -.31, *SE* = .07, 95% CI [-.45, -.17] (perhaps because of its relationship to apathy), and no effect of the strategic attribution, effect = .01, *SE* = .06, 95% CI [-.14, .11]. Thus, the results are similar to the simultaneous mediation presented in the main text except for the strategic and lacking information attributions. Given the correlation between conflict avoidance, which has a positive relationship to the outcome measure, and strategic, which has a negative relationship to the outcome measure, there may be some suppression occurring.

**Moral evaluations.** We again examine the results at the item level for the moral evaluations measure, as presented in Table S12, and without the item "honest."

**Table S12**. Means (and standard deviations) for Study 6 by whether the target was neutral or had a view that was aligned/misaligned with participants.

| Dependent Measure | Condition |
|---|---|

| | Aligned | Neutral | Misaligned |
|---|---|---|---|
| Moral | 5.23 (1.39)$_a$ | 4.28 (1.43)$_b$ | 4.23 (1.42)$_b$ |
| Trustworthy | 5.06 (1.49)$_a$ | 4.32 (1.58)$_b$ | 4.61 (1.41)$_b$ |
| Ethical | 5.20 (1.41)$_a$ | 4.33 (1.43)$_b$ | 4.07 (1.46)$_b$ |
| Honest | 5.46 (1.41)$_a$ | 4.56 (1.62)$_b$ | 5.21 (1.32)$_a$ |

*Note*. Cells with different subscripts are significantly different from one another at $p < .05$ based on Tukey post hoc tests.

The results are again robust to the exclusion of "honest," $F(2, 420) = 19.23$, $p < .001$, $\eta^2 = .084$. Participants were rated as more moral in the aligned ($M = 5.16$, $SD = 1.35$) group than in the neutral ($M = 4.31$, $SD = 1.39$) and misaligned ($M = 4.30$, $SD = 1.27$) groups ($p$s $< .001$), and did not differ from each other, $p = .999$.

**11. Summary of Patterns Across Exploratory Measures**

| Measure | Studies Used In | Rationale for Inclusion | Key Findings | Reason for Supplemental Placement |
|---|---|---|---|---|
| **Authenticity** | 1C, 1D, 2B, 6, S4, S5, S6, S8 | To examine whether neutrality is perceived as less authentic than stance-taking | Neutral targets consistently rated as less authentic than aligned targets, and typically less authentic than misaligned targets (except in Study 1D and S6) | Some authenticity items (e.g., "expresses true values") could be tautological with certain attributions for neutrality |
| **Warmth & Likeability** | 2B, 6, S6 | To examine whether neutrality affects warmth and likability judgments | Generally followed similar pattern as moral evaluations; neutral targets rated similarly to misaligned targets | Less theoretically central than morality |
| **Competence** | 1C, 2B, 6, S4, S6 | To assess whether neutrality affects competence judgments, given warmth and competence are the 'big two' of person perception | Mixed results: neutral targets rated as similar to misaligned targets in Studies 2B, 6, and S6, but more competent than misaligned targets in Study 1C | Less theoretically central than morality |
| **Perceived Target Stance** | 1C, 2B, 5, 6, S4, S5, S6 | To examine whether observers infer "true" positions from neutral expressions | On average, neutral targets perceived as neutral, with variance; effects on moral evaluations robust when controlling for perceived true stance | Important robustness check showing that attributions, not inferred positions, drive effects |
| **Agreeableness/ People-pleasing** | 6 | To examine specific trait perceptions | Neutral targets rated as more agreeable/people- | Results consistent with conflict avoidance attribution, |

| Measure | Studies Used In | Rationale for Inclusion | Key Findings | Reason for Supplemental Placement |
|---|---|---|---|---|
| | | related to conflict avoidance | pleasing than misaligned targets | but less central than morality |
| **Open-mindedness** | 6 | To test whether neutrality signals intellectual receptivity | Neutral targets rated as more open-minded than misaligned targets but less than aligned targets | Demonstrates a potential benefit of neutrality, but less central than morality |
| **Cowardice/Courage** | 1C, 2B, 6, S6 | To examine whether neutrality is seen as lacking moral courage | Neutral targets rated as more cowardly/less courageous than aligned targets except in S6 | Helps explain moral devaluation, but outside of core theoretical focus |
| **Willingness to Engage** | 1C, 1D, 6, S6, S8 | To assess intentions to interact with targets | Mixed results: interest in hearing more from neutral targets similar to aligned targets in some studies, but not in others | Complex pattern suggests neutrality may not always lead to social exclusion despite moral penalties |
| **Issue Moralization** | 1C, 2B, 6, S4, S5, S6, S8 | To examine whether issue moralization moderates effects, or if it is affected by condition | No significant moderation by moralization; no effect of condition on moralization; effects stronger for moral vs. non-moral issues in S5 | Important boundary condition for supplemental study but not central to main findings |

## SECTION 2: SUPPLEMENTAL STUDIES

We report additional studies below. These studies were not included in the main text because they test slightly different hypotheses, dependent measures, or for reasons otherwise noted (see also Appendix in the main text).

### 2.1 Supplemental Study 1

While we assumed that actors would implicitly prioritize making positive moral impressions in political conversations (Goodwin, 2015), we ran a supplemental study to examine whether our results would replicate when participants were given the explicit goal to be viewed as moral.

#### Method

We recruited 100 English-speaking American participants recruited on Prolific. The paradigm was nearly identical to Study 1A in the main text with two key tweaks. First, we used the context of safe injection drug facilities in one's city. Second, we tweaked the prompt to include, "You're hoping to make a good impression, having them [colleagues at work] think you're a moral, good person." As in Study 1A, participants selected whether they would say they were in favor/neutral/opposed to the construction of safe injection sites in the area, and indicated their private opinions.

#### Results

There were significant differences between participants' private and public stances, $\chi^2(4, 100) = 97.01$, $p < .001$, $V = 0.70$ (consistent with Study 1 in the main text). In public, 33% said they would express being in favor, 31% would express neutrality, and 36% would express opposition. By contrast, in private, 39% were in favor, 20% were neutral, and 41% were opposed. Thus, 1.55 times as many participants would claim neutrality in public than were truly neutral.

### 2.2 Supplemental Studies S2A and 2B: Lay Theories of Neutrality

To deepen our understanding of lay theories about *why* individuals express neutrality, we conducted a series of studies investigating the attributions people make for their own and others' neutral statements. In Supplemental Studies 2A and 2B, we collected and coded qualitative data about the attributions participants made about their own or another's neutrality.

#### Supplemental Studies 2A and 2B Method

In Supplemental Study 2A and 2B, we asked participants to describe past instances in which they themselves (Study 2A) or another person (Study 2B) expressed neutrality on an issue. For Study 2A, we recruited 101 participants from MTurk (49 men, 51 women; $M_{age}$ = 38.48, $SD_{age}$ = 11.73). For Study 2B, we recruited 100 participants from MTurk (58 men, 40 women, 1 non-binary/third gender, 1 prefer not to say; $M_{age}$ = 39.56, $SD_{age}$ = 11.45).

We asked Supplemental Study 2A participants to "think about a time when you interacted with someone who said that they were neutral on an issue, such as a disagreement between friends, a choice between two options, or any other situation." Then, we asked them to describe the situation and "why you think the other person told you that they were neutral." Finally, we asked them to "please explain why you believed or did not believe that the other person was neutral" (open-ended).

We asked Supplemental Study 2B participants to "Think about a time when you were neutral on an issue, such as a disagreement between friends, a choice between two options, or any other situation." Then, we asked them to describe the situation and why they were neutral.

We asked two research assistants to read all narratives and generate reasons that could explain *why* participants thought the target was neutral (described below). The authors then discussed and generated five broader categories that captured these attributions. Next, two research assistants dummy-coded whether each of these categories was present in each narrative. Narratives could contain more than one reason and therefore receive codes in multiple categories. Across all categories, both raters agreed at an average rate of 81.3%. A third coder resolved codes where the two raters disagreed.

**Supplemental Studies 2A and 2B Results**

People generated five different types of attributions for their own or others' neutrality: trying to be conflict avoidance, strategic concerns, apathy (i.e., lack of issue concern), indecision, or lacking information. In Table S13, we report the category labels and examples of participants in each study who cited each category as a reason they were (or thought someone else was) neutral.

Researchers generated five categories of attributions for neutrality (Column 1). Examples of each attribution for neutrality are shown in Column 2, and the frequency of each category's appearance across all stories in Supplemental Studies 2A and 2B is shown in Column 3.

**Table S13.** Categories of neutrality cited by participants, Supplemental Studies 2A-2B.

| 1. Category | 2. Example | 3. Frequency |
|---|---|---|
| Conflict avoidance | Didn't want to "cause strain on the friendship" (1A)<br>"I did not want to get involved in the argument." (1B) | 1A: 39.6%<br>1B: 36.1% |
| Strategic | "So that they did not get judged" or "didn't want to expose themselves" (1A)<br>"Didn't want to be called a racist" (1B) | 1A: 39.6%<br>1B: 24.1% |
| Apathy | An issue "did not matter to him" (1A) or "I do not care at all what we [choose]"(1B) | 1A: 12.2%<br>1B: 36.1% |
| Indecision | They "were completely fine with either [choice]" (1A)<br>"I was neutral because I was fine with either choice." (1B ) | 1A: 10.8%<br>1B: 6.8% |
| Lacking Information | "Did not have enough information to say one way or the other" (1A)<br>"I did not fully know what happened" (1B) | 1A: 7.2%<br>1B: 27.1% |

## 2.3 Supplemental Study 3

In Supplemental Study 3, we examine whether the results from Supplemental Study 2 and Study 4 in the main text generalize to self-report measures capturing attributions for expressed neutrality.

**Method**

We recruited 400 participants (234 women, 148 men, 12 non-binary/other, 1 prefer not to say, 5 no response; $M_{age} = 35.75$; $SD_{age} = 12.31$) from Prolific Academic in a 2-cell (self vs.

other's neutrality) between-subjects design. We preregistered this study (https://aspredicted.org/GXY_M3P). Four participants did not provide an experience, and one copied and pasted a quote from elsewhere. The results presented below exclude these participants, but are robust to their inclusion.

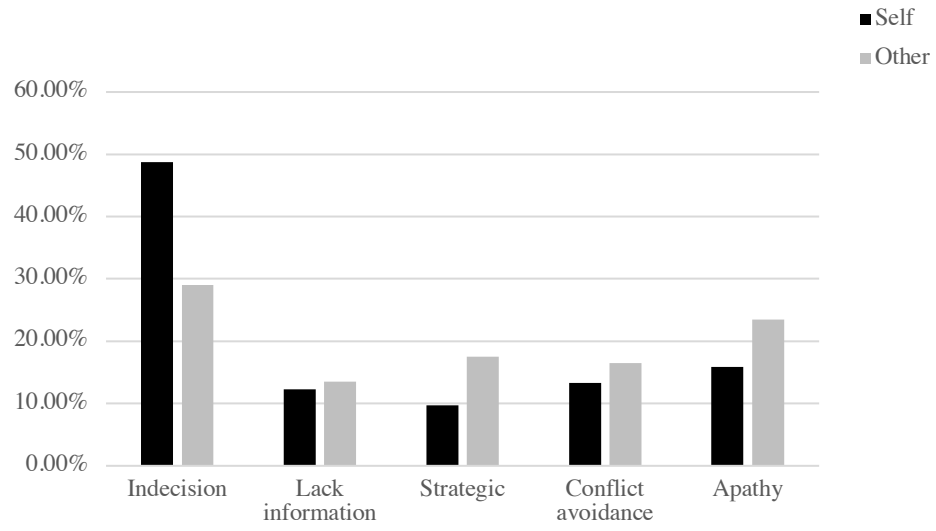
Participants were randomly assigned to think about a time either they, or someone else, claimed to be neutral, and then to write about the situation. Based on the categories identified in the pilot studies, participants were then asked, "If you had to best categorize the reason WHY you said they (they said you) were neutral, which option would you select?" They could select one of five attributions: apathy, indecision, lacking information, strategic, trying to avoid conflict (conflict avoidance). Participants then indicated the extent to which they felt truly neutral (or believed the other person felt neutral) on a scale from 1(*Not at all neutral*) to 7 (*Perfectly neutral*).

As an exploratory measure, participants in the self (actor) condition also indicated the extent to which they thought they had successfully convinced the other person they were neutral on a scale from 1 (*Not at all successful*) to 7 (*Very successful*).

**Results and Discussion**

**Attributions.** We observed a significant effect of condition (self vs. other) on the type of attribution, $\chi^2(4, 395) = 17.92$, $p = .001$, Cramer's $V = .21$. Compared to the other (observer) condition, participants in the self (actor) condition were significantly less likely to select strategic, $\chi^2(1, 395) = 5.03$, $p = .025$, Cramer's $V = .11$; marginally less likely to select apathy, $\chi^2(1, 395) = 3.60$, $p = .058$, Cramer's $V = .09$; and *more* likely to select indecision, $\chi^2(1, 395) = 16.18$, $p < .001$, Cramer's $V = .20$. No significant differences emerged in lacking information or conflict avoidance, $p$s > .377 (see Figure S12).

**Figure S12.** Percentage selecting attributions by condition in Supplemental Study 3.




**Felt Neutrality.** Participants in the self condition reported feeling more neutral ($M$ = 4.51, $SD$ = 1.85, 95% CI [4.25, 4.76]) than those in the other condition perceived them to be ($M$ = 4.05, $SD$ = 1.81, 95% CI [3.82, 4.92]), $t(398) = 2.47$, $p = .014$, $d = 0.25$.

**Convinced of Neutrality.** Compared to observers' beliefs about how much the target felt neutral ($M$ = 4.05, $SD$ = 1.81, 95% CI [3.82, 4.29]), participants in the self condition overestimated how convincing their neutrality would be to observers ($M$ = 5.02, $SD$ = 1.56, 95% CI [4.73, 5.32]), $t(331) = 5.02$, $p < .001$, $d = 0.56$.

Supplemental Study 3 revealed that people were more likely to attribute their own expressed neutrality to indecision, but were more likely to attribute others' expressed neutrality to being strategic and somewhat more likely to attribute their neutrality to apathy. People also overestimated how much others would believe they were truly neutral.

### 2.4 Supplemental Study 4

In Supplemental Study 4 we examined how these five categories of attributions for neutrality affect perceptions of a target's morality. We recruited 500 participants from Prolific (155 men, 337 women, 6 non-binary/third gender, 2 prefer not to say, $M_{age}$ = 35.52, $SD_{age}$ = 13.12). Our criteria were that participants be US residents and fluent in English. We preregistered this study at https://aspredicted.org/KXV_YDK.

**Method**

All participants read a scenario about sexual harassment in which their colleague remained neutral: "At your workplace, a woman has voiced complaints that a man in her same department has been making her uncomfortable with sexual jokes and unusual glances. The man says that he did not intend to make her uncomfortable and did not realize his jokes and glances were inappropriate. At a group lunch, someone asks another colleague what they think about this situation, and they tell the group that they are neutral on the issue. They explain that they are neutral because…" We manipulated the reason the colleague provided for their neutrality, between-participants, based on the attribution categories generated in Study 4, and Supplemental Studies 2A and 2B. These included apathy ("they don't care either way about who is right or wrong"); strategic ("they want to avoid getting judged for their true opinion on the issue"); lacking information ("they don't have enough information to make a decision"); conflict avoidance ("they want to avoid saying anything for fear of having to pick sides"); and indecision ("they can't decide which person is right or wrong").

Participants then rated the perceived morality of the target (moral, trustworthy, ethical, honest, $\alpha = .95$) as well as some supplemental measures, which are described below. We also asked them to explain their answer (open-ended). We further asked, "To what extent would you actually believe that the other colleague you talked to was neutral in this situation?" (1 = *did not believe them at all*; 7 = *completely believed them*).

**Results**

Descriptive statistics, omnibus tests, and pairwise comparisons for the key outcome variables are reported in Supplemental Table 14.

**Moral Evaluations.** There was a significant effect of attribution condition on perceived morality. The results indicated that the colleague lacking information was rated as the most moral, followed by the indecisive colleague, and then the apathetic, strategic, and the conflict avoidant colleague. The results are again robust to the exclusion of "honest" (see Table S14).

**Beliefs about Neutrality.** People were less likely to believe a target was truly neutral when they were being conflict avoidant or strategic, relative to targets who said they were apathetic, uninformed, or unable to decide.

**Table S14.** Means (and standard deviations and 95% CIs) for the main measures by condition, Supplemental Study 4.

| | Apathy | Strategic | Lacking Information | Conflict Avoidant | Indecision | Omnibus ANOVA |
|---|---|---|---|---|---|---|
| Moral (4 items) | 3.27 (1.32)$_{a}$ [3.01, 3.53] | 3.25 (1.45)$_{a}$ [2.96, 3.54] | 4.12 (1.50)$_{b}$ [3.82, 4.42] | 3.17 (1.49)$_{a}$ [2.87, 3.47] | 3.52 (1.53)$_{a}$ [3.21, 3.82] | $F(4, 496) = 7.08$, $p < .001$, $\eta^2 = .054$ |
| Moral (no "honest") | 3.30 (1.32)$_{a}$ [3.02, 3.59] | 3.22 (1.46)$_{a}$ [2.94, 3.51] | 4.11 (1.48)$_{b}$ [3.82, 4.39] | 3.18 (1.50)$_{a.}$ [2.89, 3.46] | 3.51 (1.52)$_{a}$ [3.23, 3.80] | $F(4, 495) = 6.81$, $p < .001$, $\eta^2 = .052$ |
| Believe truly neutral | 3.51 (1.70)$_{cd}$ [3.18, 3.84] | 2.45 (1.49)$_{a}$ [2.16, 2.75] | 4.11 (1.48)$_{d}$ [3.82, 4.40] | 2.89 (1.61)$_{ab}$ [2.57, 3.21] | 3.41 (1.51)$_{bc}$ [3.11, 3.70] | $F(4, 496) = 16.27$, $p < .001$, $\eta^2 = .116$ |

*Note.* Cells with different subscripts are significantly different from one another at $p \leq .05$ according to Tukey HSD post-hoc tests correcting for multiple comparisons.

**Supplemental Measures.**

**Perceived True Stance.** We asked, "What do you think the other person's actual position on this issue was?" (-3 = *completely sided with the woman*; 0 = *neutral*; +3 = *completely sided with the man*). All colleagues were perceived to have an actual stance in favor of the man (versus the woman) in the sexual harassment issue (one-sample *t*-tests comparing the value to the scale midpoint; $t \geq 2.85$, $p$s $\leq .005$, $d$s $\geq .28$). Further, participants' estimates of their neutral colleagues' actual position were shaped by the reason provided. Participants were particularly likely to think the target favored the man rather than the woman when the target said they were apathetic, strategic, or indecisive.

**Authenticity.** Participants completed a four-item measure of authenticity (authentic; sincere; expresses their true values; tries to be something they're not (reverse coded); $\omega = .89$). Participants' assessments of authenticity varied significantly across the different reasons the target provided for their neutrality. People thought the target was less authentic when the colleague said they wanted to be conflict avoidant or avoid being judged for their opinion, relative to a target who said they were apathetic, uninformed, or unable to decide. See Table S15 for descriptive statistics and the results of post-hoc tests.

**Competence.** We used a single-item measure to assess participants' perceptions of the target's competence (1 = *not at all*, 7 = *very much*). Participants rated the uninformed and indecisive targets as similarly competent, and the uninformed target was seen as more competent than the apathetic, conflict avoidant, and strategic targets.

**Own Position.** We asked, "How would you likely feel about the accusation you read about earlier in the study?" (-3 = *I would side completely with the woman in this situation*; 0 = *I am neutral*; +3 = *I would side completely with the man in this situation*). There was no main effect of condition, and the aggregate mean ($M = 2.48$, $SD = 1.24$, 95% CI (1.41, 1.63)) was significantly different from the neutral midpoint of the scale (one-sample $t(499) = 27.34$, $p < .001$, $d = 1.22$, indicating participants would be more likely to side with the woman than to be neutral.

**Moralization.** To confirm that participants did, on aggregate, view this issue as a moral one, and to test whether there might be differences by condition, participants responded to "To what extent was the workplace issue a moral issue (having to do with your beliefs about 'right'

or 'wrong')?" (1 = *not at all*; 7 = *very much so*). The main effect of condition of moralization was not significant. Further, moralization did not interact with the manipulated reason the colleague provided for their neutrality to predict perceptions of morality, $p = .810$.

**Table S15.** Descriptive statistics for the supplemental measures by condition.

| | Apathy | Strategic | Lacking Information | Conflict Avoidant | Indecision | Omnibus ANOVA |
|---|---|---|---|---|---|---|
| Target's true position (-3 to +3) | 1.60 (1.29)$_{ab}$ [4.34, 4.85] | 1.72 (1.55)$_{b}$ [4.41, 5.03] | 1.12 (1.08)$_{a}$ [3.91, 4.33] | 1.09 (1.51)$_{a}$ [3.79, 4.40] | 1.34 (1.19)$_{ab}$ [4.1, 4.57] | $F(4, 496) = 4.44$, $p = .002$, $\eta^2 = .035$ |
| Authenticity (4 items) | 3.56 (1.44)$_{ab}$ [3.28, 3.85] | 3.11 (1.37)$_{a}$ [2.83, 3.38] | 4.36 (1.31$_{c}$ [4.10, 4.62] | 3.12 (1.35)$_{a}$ [2.85, 3.39] | 3.73 (1.49)$_{b}$ [3.44, 4.03] | $F(4, 496) = 13.76$, $p < .001$, $\eta^2 = .100$ |
| Competence | 3.93 (1.31)$_{a}$ [3.67, 4.19] | (3.98 (1.34)$_{a}$ [3.71, 4.25] | 4.60 (1.33)$_{b}$ [4.34, 4.86] | 3.97 (1.44)$_{a}$ [3.68, 4.26] | 4.10 (1.45)$_{ab}$ [3.81, 4.38] | $F(4, 496) = 4.10$, $p = .003$, $\eta^2 = .032$ |
| How participant would side* | 2.42 (1.24)$_{a}$ [2.17, 2.66] | 2.46 (1.36)$_{a}$ [2.19, 2.74] | 2.67 (1.19)$_{a}$ [2.43, 2.91] | 2.25 (1.14)$_{a}$ [2.03, 2.48] | 2.60 (1.26)$_{a}$ [2.36, 2.85] | $F(4, 496) = 1.75$, $p = .138$, $\eta^2 = .014$ |
| Moralization of issue ^ | 5.15 (1.49)$_{a}$ [4.85, 5.44] | 4.76 (1.74)$_{a}$ [4.41, 5.10] | 5.19 (1.43)$_{a}$ [4.91, 5.47] | 4.85 (1.71)$_{a}$ [4.51, 5.19] | 4.87 (1.60)$_{a}$ [4.55, 5.19] | $F(4, 496) = 1.46$, $p = .212$, $\eta^2 = .012$ |

* (-3 = *completely with the woman*; 0 = *neutral*; +3 = *completely with the man)*
^ (1 = not at all, 7 = very much)

Consistent with the results of Study 5 in the main text, we found that these attributions differentially predict morality, such that the apathetic, and strategic targets were perceived to be relatively less moral than the indecisive target or the target who lacked information. The one difference here was that the conflict avoidant target was also rated as less moral. However, in this context, avoiding taking sides also involved a potential perpetrator, which would have different moral implications, and may explain these diverging results.

### 2.5 Supplemental Study 5 – Moral versus Non-Moral Issues

Supplemental Study 5 sought to test a potential boundary condition around the effects of claiming neutrality. Our central hypothesis is that claiming neutral will backfire for moral and political issues. People's positions on moral issues come to feel like self-evident, objective truths – like there is no need for deliberation as to whether abortion is right or wrong, for example (e.g., Skitka, 2010). This is not true for non-moral decisions, where contemplating pros and cons is common (e.g., selecting Apartment A or B) and involves matters of taste or preference (e.g., deciding whether Frank Sinatra is a better singer than Michael Bolton) (Turiel, 1983). Given the potential implausibility of true indecision along moral issues, claimed neutrality for moral issues may elicit attributions of strategic concealment as opposed to the more favorable attributions surrounding "true" neutrality. Testing this possibility, in Supplemental Study 5, we sought to explore our proposed boundary condition around the effects of claiming neutrality.

#### Method

In this study, participants were randomly assigned to 1 of 6 conditions in a 2(Issue type: moral vs. non-moral) x 3(Issue position: In favor vs. neutral vs. opposed) between-subjects design. All participants read about an acquaintance speaking about an issue in a group setting

and then completed the interpersonal judgment measures as in previous studies. Our design and analysis plan were preregistered (https://aspredicted.org/blind.php?x=9h35gy).

We recruited 700 Prolific participants who had previously completed a panel indicating their stance on a variety of issues. We sampled from a pool who were not neutral on the key issues for the study (affirmative action in the moral condition, online courses in the non-moral condition[2]). Of the 700 participants who enrolled, some participants provided incomplete data, and others incorrectly entered their Prolific IDs; as a result, we could not link participants' positions on the issues to their responses in the study. We were left with 605 participants (274 men, 410 women, 15 non-binary/third gender/prefer not to say; $M_{age}$ = 36.04, $SD_{age}$ = 12.81), which still gave us 95% power to detect the small-to-medium-sized effects observed in the previous studies.

Participants read a scenario describing a group conversation about either affirmative action (moral condition) or online courses (non-moral condition). In the in-favor condition, participants read that an acquaintance had said they were in favor of affirmative action (online courses), neutral in the neutral condition, and opposed in the opposed condition.

As in previous studies, we asked participants about their own positions on these issues to measure alignment with the target's beliefs about affirmative action/online courses. All participants then responded to the dependent measures.

**Results**

**Manipulation Checks.** Participants in the moral issue condition (affirmative action) rated the issue as more moral ($M$ = 5.09, $SD$ = 1.29, 95% CI [4.95, 5.22]) than did participants in the non-moral issue condition (online courses) ($M$ = 3.60, $SD$ = 1.59, 95% CI [3.44, 3.77]), $t(697) = 13.54$, $p < .001$, $d$ = 1.02, 95% CI [1.27, 1.70]. There was also a significant effect of stance condition on perceptions of the acquaintance's publicly stated position, $F(2, 694) = 670.38$, $p < .001$, $\eta^2 = 0.66$. Participants in the opposing condition thought the acquaintance was more against the issue ($M$ = 2.38, $SD$ = 1.94, 95% CI [2.13, 2.63]) than did those in the neutral condition ($M$ = 4.99, $SD$ = 0.87, 95% CI [4.88, 5.10]), or the in-favor condition ($M$ = 7.76, $SD$ = 1.73, 95% CI [7.53, 7.98]). All conditions significantly differed from each other, $t$s > 17.76, $p$s < .001, $d$s > 1.73.

**Authenticity.** Examining the effect of condition on authenticity judgments revealed a significant effect of alignment, $F(2, 599) = 50.54$, $p < .001$, $\eta^2 = 0.067$, a main effect of moral vs. non-moral issue condition, $F(1, 599) = 10.72$, $p = .003$, $\eta^2 = 0.015$, and no significant interaction, $F(2, 599) = 1.90$, $p = .150$, $\eta^2 = .006$. For both moral and non-moral issues, the neutral and opposing target was rated as less authentic than the aligned target ($t$s > 3.11, $p$s < .002, $d$s > 0.48). The neutral and opposing groups did not significantly differ in either condition ($p$s > .657).

**Moral Evaluations.** There was a significant effect of alignment on moral evaluations, $F(2, 599) = 73.09$, $p < .001$, $\eta^2 = 0.195$, a main effect of issue type condition, $F(1, 599) = 15.86$,

[2] Before selecting the moral and non-moral issues for Study 3, we first ran a pretest examining how 13 issues would be rated in terms of participants' preferences along these issues, as well as their attitude strength, and issue moralization. 300 participants were recruited via MTurk (159 women; $M_{age}$ = 38.98, $SD_{age}$ = 11.53) and randomly assigned to rate 3 of the 13 issues to keep the survey length manageable. These participants completed attitude position, strength, and moralization items as in Study 2. We selected affirmative action and taking online (versus in-person) courses as our issues, given that these issues varied significantly in terms of the degree to which they were moral in nature ($M_{AA}$ = 3.26, $SD$ = 1.13; $M_{Online}$ = 1.59, $SD$ = 1.16), *paired* $t$ = 3.70, $p$ = .002, $d$ = 0.90.

$p < .001$, $\eta^2 = 0.03$, and a significant interaction, $F(2, 599) = 20.10$, $p < .001$, $\eta^2 = 0.06$. Within the moral issue condition, both the neutral and misaligned targets were rated as significantly less moral than the aligned target, $t$s > 6.57, $d$s > 1.00. The misaligned target was also rated as less moral than the neutral target, $t(205) = 5.08$, $p < .001$, $d = 0.68$, 95% CI [0.40, 0.96]. However, within the non-moral issue condition, only the opposed target was penalized, being rated as less moral than the aligned target, $t(185) = 4.68$, $p < .001$, $d = 0.68$, 95% CI[0.39, 0.98], and the neutral target, $t(211) = -3.65$, $p = .001$, $d = 0.53$, 95% CI [0.26, 0.81]. The neutral target was rated similarly moral to the aligned target, $t(191) = 1.31$, $p = .392$, $d = 0.17$, 95% CI [-0.10, 0.45]. The results follow a similar pattern with or without the "honest" item.

**Attributions for Neutrality.** We again examined participants' attributions for neutrality within the neutral condition. Consistent with our theorizing, participants in the moral (versus non-moral issue) conditions generally made different attributions for the acquaintance's neutrality (see Table S16). Participants in the moral condition thought that the neutrality was driven less by assessing the pros and cons and less by having no opinion than did those in the non-moral condition. They thought that the neutrality was more likely to be driven by not wanting to take sides, not wanting to share their opinion, and fearing the consequences of speaking up. Participants in the moral and non-moral issue conditions did not vary in their attributions of lacking information or apathy. In other words, when it comes to moral issues, people are more likely to attribute claimed neutrality to those reasons less linked to believability.

**Table S16.** Attributions for Neutrality by Issue Type Condition.

| Attribution | Moral Issue | Non-Moral Issue | Test ($df = 229$) |
|---|---|---|---|
| | *M* (*SD*) [95% CI] | *M* (*SD*) [95% CI] | |
| Indecision | 4.91 (1.68) [4.60, 5.22] | 5.70 (1.18) [5.48, 5.92] | $t = -4.11$, $p < .001$, $d = -0.54$, 95% CI [-1.16, -.41] |
| Apathy | 4.33 (1.70) [4.02, 4.65] | 4.54 (1.53) [4.26, 4.82] | $t = -1.00$, $p = .319$, $d = -0.13$, 95% CI [-0.63, .21] |
| No Opinion | 4.17 (1.67) [3.86, 4.47] | 4.76 (1.61) [4.46, 5.06] | $t = -2.74$, $p = .007$, $d = -0.36$, 95% CI [-1.02, -0.17] |
| Lacks Knowledge | 4.65 (1.52) [4.37, 4.93] | 4.37 (1.55) [4.09, 4.66] | $t = 1.40$, $p = .164$, $d = 0.18$, 95% CI [-0.12, 0.70] |
| Conflict Avoidance | 5.32 (1.31) [5.08, 5.56] | 4.66 (1.43) [4.40, 4.93] | $t = 3.65$, $p < .001$, $d = 0.48$, 95% CI [0. 30, 1.01] |
| Doesn't Want to Share Opinion | 4.74 (1.65) [4.43, 5.04] | 4.02 (1.69) [3.71, 4.33] | $t = 3.28$, $p = .001$, $d = 0.43$, 95% CI [0. 30, 1.01] |
| Fears the Consequences | 4.97 (1.53) [4.68, 5.25] | 3.62 (1.81) [3.29, 3.95] | $t = 6.11$, $p < .001$, $d = 0.80$, 95% CI [0. 30, 1.01] |

Consistent with our other studies, perceiving a weighting of pros and cons was positively linked to authenticity and moral evaluations ($r$s $\geq .48$, $p$s $< .001$), whereas not wanting to take sides, not wanting to share their opinion, and fearing the consequences of speaking up were all negatively related to authenticity and moral evaluations ($r$s $\leq -.22$, $p$s $< .001$). Therefore, the very attributions invoked by neutrality along moral (versus non-moral issues) also shape perceptions of authenticity and morality. To further examine this, we also entered the moral issue condition (moral = 1; non-moral = 0), the attributions, and moral evaluations into a mediation analysis testing multiple possible mediators (PROCESS Model 6). The results revealed significant indirect effects via pros/cons, indirect effect = .21, 95% [.09, .36]. No other significant indirect effects emerged. In short, participants were more likely to believe that neutrality was driven by a weighting of pros and cons in the non-moral condition than in the moral condition, which drove more favorable moral evaluations.

**Addressing Potential Alternatives.** Given that moral issues differ from non-moral issues along a variety of dimensions (e.g., attitude strength, issue importance), we conducted analyses to ensure that these other differences between moral and non-moral issues were not driving our effects. To do so, we entered participants' issue strength ("How strongly do you feel about your position on the issue?") and importance ("How important is this issue to you?") in an analysis of covariance (ANCOVA).

Examining the effect of condition on authenticity judgments revealed a significant effect of alignment group, $F(2, 597) = 21.73$, $p < .001$, $\eta^2 = 0.07$, a main effect of issue type condition,

$F(1, 597) = 7.03$, $p = .008$, $\eta^2 = 0.01$, and no significant interaction, $F(2, 597) = 1.69$, $p = .185$, $\eta^2 = .006$. There was no effect of issue strength or importance ($ps \geq .266$).

Examining the effect of condition on moral evaluations revealed a significant effect of alignment group, $F(2, 597) = 72.35$, $p < .001$, $\eta^2 = 0.19$, a main effect of issue type condition, $F(1, 597) = 14.14$, $p < .001$, $\eta^2 = 0.02$, and a significant interaction, $F(2, 597) = 19.65$, $p < .001$, $\eta^2 = .06$.There was again no effect of issue strength or importance ($ps \geq .581$). Therefore, results replicate when controlling for issue strength and importance.

These results show that although people who claimed neutrality were seen as similarly low in authenticity regardless of whether the issue was moral, neutral targets were only seen as less moral when the issue was moral. Further, this decrease in perceptions of morality appears to be related to the different attributions that people form for neutrality.

### 2.6 Supplemental Study 6 – The Job Candidate

The goal of Supplemental Study 6 was to examine whether neutrality version would generalize to a new issue and a new context. Specifically, we examined how participants responded to a job candidate who indicated being in favor of, neutral, or against the COVID-19 vaccine. After reading about the candidate, participants indicated their willingness to interview the candidate and completed the authenticity and morality measures as in previous studies. The design and analysis plan were preregistered (https://aspredicted.org/blind.php?x=gk2jf6).

**Method**

We recruited 407 participants from Prolific who indicated that they had hiring experience via a panel (data collected by them independently of our study). As pre-registered, after screening for possible bots ($n = 0$) and including only those who indicated that they were either pro ($n = 334$) or anti ($n = 26$) the COVID-19 vaccine, our final sample included 360 participants (145 men, 211 women, 3 non-binary/third gender/prefer not to say, 1 unreported; $M_{age} = 46.68$, $SD_{age} = 12.09$).[3] We ran this study in early April 2021, when the COVID-19 vaccine was available to essential workers (including grocery store employees), but not yet available to the general population in most states.

Participants were asked to imagine that, as a manager of a branch of the Crunchy grocery store chain, they are responsible for hiring a cashier. We displayed a fictitious, handwritten job application from James Smith that included an employment history. Embedded in the application was the following question: "How do you feel about getting the COVID-19 vaccine?" (1 = *not at all comfortable*; 2= *somewhat uncomfortable*; 3 = *neutral*; 4= *somewhat comfortable*; 5 = *completely comfortable*). We manipulated the candidate's response—he either wrote down a 2, 3, or 4 in response to this question. Thus, participants were randomly assigned to one of these three conditions (anti, neutral, or pro). We used these values rather than the extreme values (1 and 5) to reduce the likelihood of potential demand effects from extra attention given to such positions.

As a manipulation check, we asked participants how the applicant felt about the vaccine (which varied by condition); as an attention check, we asked participants how the applicant felt

[3] In this study, we collected participants' views of the issue in the study itself rather than recruiting them from a panel where we knew their attitudes independently. We did this after extensive pretesting showed that our manipulations never affected individuals' own attitudes towards the issue (see OSF). As in those studies, here we do not see that there were differences in participants' own attitudes as a result of the manipulation ($F$ (2, 399) = 1.47, $p$ = .232), nor did the strength of their position vary by condition ($F$ (2, 399) = .33, $p$ =.717).

about lifting heavy objects (which did not vary; correct answer was, "4 – somewhat comfortable").

Our primary dependent measures of interest were whether the participant thought that the job candidate should be interviewed for a job (on a scale from 1 = *definitely should not be interviewed* to 9 = *definitely should be interviewed*) and perceptions of the job candidate's morality (measured with the same three items as previous studies). We also measured whether participants felt that he is qualified, on a scale (from 1 = *not at all* to 9 = *very*), since it could be that people felt that the vaccine was a requirement for working at a grocery store during a pandemic. Participants completed additional exploratory perception measures (cowardly, self-interested, competent, deceitful, liking). They were also asked to explain how they came up with their assessment of the candidate (open-ended question). Participants in the neutral condition were asked about the reasons behind the candidate's response (e.g., "he sees both the pros and cons of getting the vaccine"). We also asked, "Are there any additional reasons not listed above why this person might have responded the way they did to the COVID vaccine question?"

In addition, we asked participants, "What do you think the candidate's true (privately held) opinion of the COVID vaccine is?" (-3 = *against the vaccine*; 0 = *neutral*; +3 = *in favor of the vaccine*), and whether participants had ever worked in a grocery store (no/yes (when?)) and how readable the job application was (1= *not at all readable* to 7 = *very readable*). Finally, we asked 2 items to measure the extent to which the COVID-19 vaccine is a moral issue to participants as in the previous studies, and attitude strength with a single item.

**Results**

Means and omnibus tests are presented in Table S17.

**Attention and Manipulation Checks**. Most participants correctly responded to the question about how the candidate felt about lifting heavy objects (99.97%) and about the COVID vaccine (99.99%). We retained all participants in the analyses in accordance with our pre-registration.

**Perceived True Stance**. Looking at individuals' beliefs in what they thought the applicants real (privately held) stance on the vaccine, we found significant differences across all conditions (using the randomly assigned condition variable, not the alignment group variable), $F(2, 399) = 258.16$, $p < .001$, $\eta^2 = .56$, 95% CI [.503, .613]. People believed that the candidate was more in favor of the vaccine when they read that he was comfortable with the vaccine ($M_{supportive} = 5.05$, $SD_{supportive} = 1.21$, 95% CI [4.84, 5.25] than when they read that he was neutral ($M_{neutral} = 3.47$, $SD_{neutral} = 1.14$, 95% CI [3.28, 3.66]; $t = 11.84$, $p < .001$) or uncomfortable with the vaccine ($M_{opposed} = 2.03$, $SD_{opposed} = .89$, 95% CI [1.88, 2.18], $t = 22.67$, $p < .001$. In addition, participants rated the applicant as more in favor of the vaccine in the neutral than opposed conditions ($t = 10.92$, $p < .001$). However, the means indicated that people believed the neutral candidate was slightly opposed to the vaccine ($M_{neutral} = 3.47$, $SD_{neutral} = 1.14$), which differed significantly from the scale midpoint of 4, one-sample $t(133) = -5.40$, $p < .001$, $d = -0.47$.

For the remainder of the analyses below, we used alignment group as the independent variable rather than condition.

**Interviewing the Candidate.** An omnibus ANOVA showed that alignment group significantly affected interest in interviewing the candidate, $F(2, 357) = 9.28$, $p < .001$, $\eta^2 = .05$, 95% CI [.013, .096]. For our key dependent variable for this study, Tukey tests showed that participants were more likely to want to interview the candidate when he was aligned with them on the issue ($M_{\text{aligned}} = 7.04$, $SD_{\text{aligned}} = 1.74$, 95% CI [6.73, 7.34]) than when he held the opposite

position from them ($M_{misaligned}$ = 6.03, $SD_{misaligned}$ = 2.05, 95% CI [5.64, 6.41]), $t(238) = 4.04$, $p < .001$, $d = 0.53$, 95% CI [0.27, 0.79]). Participants were less likely to want to interview the neutral candidate ($M_{neutral}$ = 6.26, $SD_{neutral}$ = 1.99, 95% CI [5.90, 6.62]) than candidates who held participants' same vaccine position, $t(245) = 3.12$, $p = .004$, $d = 0.42$, 95% CI [0.16, 0.67], but not candidates who held an opposite one, $p = .629$. Importantly there were no significant differences across groups in perceptions of qualifications for the job, $F(2, 357) = 1.81$, $p = .165$; competence, $F(2, 357) = 1.27$, $p = .282$; or readability of the application, $F(2, 357) = 1.79$, $p = .168$, and therefore these were not viable alternative explanations.

**Moral Evaluations.** An ANOVA indicated that the conditions differed significantly in their moral evaluations of the candidate, $F(2, 357) = 6.16$, $p = .002$, $\eta^2 = .033$, 95% CI [.005, .074]. Post-hoc Tukey tests on evaluations of morality of the candidate showed that those who held the same position saw the candidate as more moral ($M_{aligned}$ = 5.04, $SD$ = 0.91, 95 % CI [4.89, 5.21]) than those who held the opposite position ($M_{misaligned}$ = 4.66, $SD$ = .86, 95% CI [4.50, 4.82]), $t(238) = 3.45$, $p = .002$, $d = 0.43$, 95% CI [0.17, 0.68]. The neutral candidate ($M_{neutral}$ = 4.79, $SD$ = 0.81, 95% CI [4.65, 4.94]) was seen as marginally less moral than the candidate who held the same position, $t(245) = 2.27$, $p = .055$, $d = 0.29$, 95% CI [0.04, 0.54]. The neutral candidate was rated as similarly moral as the candidate who held the misaligned position, $p = .494$. These results are again robust to the exclusion of "honest," $F(2, 357) = 8.49$, $p < .001$, $\eta^2 = .045$, such that aligned participants were rated as more moral ($M$ = 4.31, $SD$ = 0.95) than neutral participants ($M$ = 4.65, $SD$ = 0.85) or misaligned participants ($M$ = 4.43, $SD$ = 0.97), $p$s < .023, and the neutral and misaligned participants were rated similarly, $p = .068$.

**Authenticity.** Unexpectedly, there were no significant differences in evaluations of the candidate's authenticity, $F(2, 357) = .658$, $p = .518$.

**Mediation.** To examine whether changes in the moral evaluations of the candidate drove the effect of alignment group on willingness to interview, we conducted a mediation analysis using Model 4 in PROCESS. There was a significant indirect effect of condition on interviewing likelihood through moral evaluations of the candidate, *Effect* = .31, *Boot SE* = .10; 95% CI (5,000) [.13, .52].

**Cowardice.** An ANOVA indicated that participants across conditions differed in their evaluations of the candidate's cowardice, $F(2, 357) = 4.04$, $p = .018$, $\eta^2 = .022$, 95% CI [.00, .057]. Post-hoc Tukey tests demonstrated that the candidate opposed to participants' position ($M_{opposed}$ = 2.88, $SD_{opposed}$ = 1.44, 95% CI [2.62, 3.15]) were perceived as more cowardly than those who held the same position ($M_{same}$ = 2.41, $SD_{same}$ = 1.21, 95% CI [2.20, 2.62]; $t = 2.82$, $p = .016$, but the neutral candidate ($M_{neutral}$ = 2.72, $SD_{neutral}$ = 1.31, 95% CI [2.48, 2.95]) was not seen as significantly more or less cowardly than either ($p$s = .595 and .162, respectively).

**Liking.** There were no significant differences across conditions in evaluating the job candidate's likability, $F(2, 357) = 2.14$, $p = .120$.

**Trustworthiness.** There were significant differences across conditions of evaluations of the candidate's trustworthiness, $F(2, 357) = 3.76$, $p = .024$, $\eta^2 = .021$, 95% CI [.0001, .0548]. Those who evaluated a job candidate with the opposing position on vaccines ($M_{opposed}$ = 4.73, $SD_{opposed}$ = .98, 95% CI [4.54, 4.91] were seen as less trustworthy than an aligned candidate ($M_{same}$ = 5.07, $SD_{same}$ = 1.04, 95% CI [4.89, 5.25], $t$ = - 2.69, $p$ = .022). There were no other significant differences across conditions.

**Competence.** There were no significant differences across conditions in the evaluations of the applicant's competence, $F(2, 357) = 1.27$, $p$ =. 282.

**Self-Interest.** There were significant differences across conditions in evaluations of the job candidate's self-interestedness, $F(2, 357) = 3.70$, $p = .026$, $\eta^2 = .020$, 95% CI [.000, .054]. Those who held the position opposing the candidate felt that he was more self-interested ($M_{opposed} = 4.29$, $SD_{opposed} = 1.29$, 95% CI [4.05, 4.53]) than those aligned in their position ($M_{same} = 3.82$, $SD_{same} = 1.42$, 95% CI [3.57, 4.07]; $t = 2.61$, $p = .022$). There were no other significant differences across conditions.

**Deceit.** There were no significant differences across conditions on assessments of the candidate's deceitfulness, $F(2, 357) = 1.17$, $p = .312$.

**Moralization.** There were no significant differences across conditions in their moralization of the issue, $F(2, 357) = .56$, $p = .571$.

**Issue Strength.** There were no differences across conditions in the strength of how people felt towards the issue, $F(2, 357) = .03$, $p = .970$.

**Perceived True Stance.** There were was a significant effect of stance condition on perception of true stance, $F(2, 399) = 258.16$, $p < .001$. Participants in the opposed condition rated the target as less in favor of the vaccine ($M = 2.03$, 95% CI [1.88, 2.18], than did those in the neutral condition ($M = 3.47$, 95% CI [3.28, 3.66]), or those in the in-favor condition ($M = 5.05$, 95% CI [4.84, 5.25). All conditions significantly differed from each other, $p$s $< .001$.

When a job applicant claimed neutrality, they were rated just as immoral, and were just as unlikely to be interviewed as an applicant holding an opposing stance on vaccination. Unexpectedly, we observed no differences in perceived authenticity in this study. It is perhaps because at the time of data collection (April, 2021), there was a high amount of normative and institutional pressure to receive the vaccine, and thus even claiming neutrality as opposed to high positivity may have been seen as counter-normative and thus boosted authenticity. Any decrements in authenticity resulting from perceived concealment may be offset by a boost in claiming a socially undesirable response, resulting in a null effect.

**Table S17**. The effect of alignment group on the supplemental variables in Study S6.

| Measure | Alignment Condition Mean and 95% CI | | | Omnibus test |
|---|---|---|---|---|
| | Aligned | Neutral | Misaligned | |
| Interview | $7.23_{ab}$ [6.99, 7.47] | $6.70_{a}$ [6.45, 6.94] | $6.58_{b}$ [6.33, 6.84] | $F(2, 357) = 7.79$, $p < .001$ |
| Moral Evaluations | $5.05_{a}$ [4.89, 5.21] | 4.79 [4.65, 4.94] | $4.66_{a}$ [4.50, 4.82] | $F(2, 357) = 6.16$, $p = .002$ |
| Qualified | 7.42 [7.16, 7.67] | 7.13 [6.90, 7.36] | 7.14 [6.91, 7.37] | $F(2, 357) = 1.81$, $p = .165$ |
| Cowardly | $2.41_{a}$ [2.20, 2.62] | 2.72 [2.48, 2.95] | $2.88_{a}$ [2.62, 3.15] | $F(2, 357) = 4.04$, $p = .018$ |
| Authenticity | 5.33 [5.18, 5.49] | 5.26 [5.12, 5.41] | 5.21 [5.06, 5.36] | $F(2, 357) = .66$, $p = .518$ |
| Likeable | 4.73 [4.55, 4.92] | 4.54 [4.38, 4.71] | 4.49 [4.31, 4.66] | $F(2, 357) = 2.14$, $p = .120$ |
| Trustworthy | $5.07_{a}$ [4.89, 5.25] | 4.83 [4.65, 5.01] | $4.73_{a}$ [4.54, 4.91] | $F(2, 357) = 3.76$, $p = .024$ |
| Competent | 5.36 [5.19, 5.53] | 5.19 [4.99, 5.39] | 5.16 [4.95, 5.37] | $F(2, 357) = 1.27$, $p = .282$ |
| Self-interested | $3.82_{a}$ [3.57, 4.07] | 3.96 [3.71, 4.21] | $4.29_{a}$ [4.05, 4.53] | $F(2, 357) = 3.70$, $p = .026$ |
| Deceitful | 2.13 [1.94, 2.33] | 2.20 [1.99, 2.41] | 2.35 [2.14, 2.57] | $F(2, 357) = 1.17$, $p = .312$ |
| Issue moralizing | 3.53 [3.31, 3.74] | 3.68 [3.49, 3.88] | 3.59 [3.38, 3.81] | $F(2, 357) = .56$, $p = .571$ |
| Issue strength | 4.33 [4.19, 4.47] | 4.35 [4.20, 4.50] | 4.35 [4.21, 4.50] | $F(2, 357) = .03$, $p = .970$ |

*Note.* Means that share a subscript are significantly different at $p < .05$.

## 2.7 Supplemental Study 7 – Strategic Attributions

In Supplemental Study 7, we test whether neutrality explains others' judgments above and beyond information about other strategic impression management attempts.

### Method

We recruited 1500 participants from Prolific (954 women; $M_{age}$ = 33.77, $SD_{age}$ = 12.51). The design and analysis plan were preregistered (https://aspredicted.org/3L7_RN8). Participants were randomly assigned to 1 of 12 conditions in a 2(issue: affirmative action or vaccine mandates) x 3(issue position: in favor vs. neutral vs. opposed) x 2(strategic attribution: strategic vs. control) between-subjects design.

All participants imagined being at a work get-together when a topic was brought up. To ensure the results were not unique to a single issue, we used stimulus sampling across two different salient moral issues at the time (affirmative action, vaccine mandates), but planned to collapse across these issues if no significant interactions by condition emerged, which was the case (*p*s > .434). Participants were randomly assigned to read about a colleague who publicly indicated being in favor of, opposed to, or neutral about the issue.

We constructed the position alignment variable using the same method as the previous studies. After reading about this interaction, participants in the strategic attribution condition were informed that they find the colleague "is motivated to make a good impression on their colleagues and boss." In the control condition, participants were given no additional information.

Participants completed the moral evaluations measure. Participants responded to five items measuring why they thought the colleague responded the way they did (*wanting to be agreeable*; *wanting to be liked by their colleagues*; *their level of concern about the issue; their decisiveness about the issue; the amount of information they have about the issue*) on 7-point scales. We also asked, "What do you think the person's true (privately held) opinion is?" (-3 = *against the construction*; 0 = *neutral*; +3 = *in favor of the construction*). Participants completed a manipulation check, "in general, how much is the colleague motivated to make a good impression on their colleagues and boss?" on a scale (from 1 = *not at all* to 7 = *very much so*). Finally, participants completed an attention check asking them to indicate which issue was brought up for discussion at the get-together.[4]

### Results

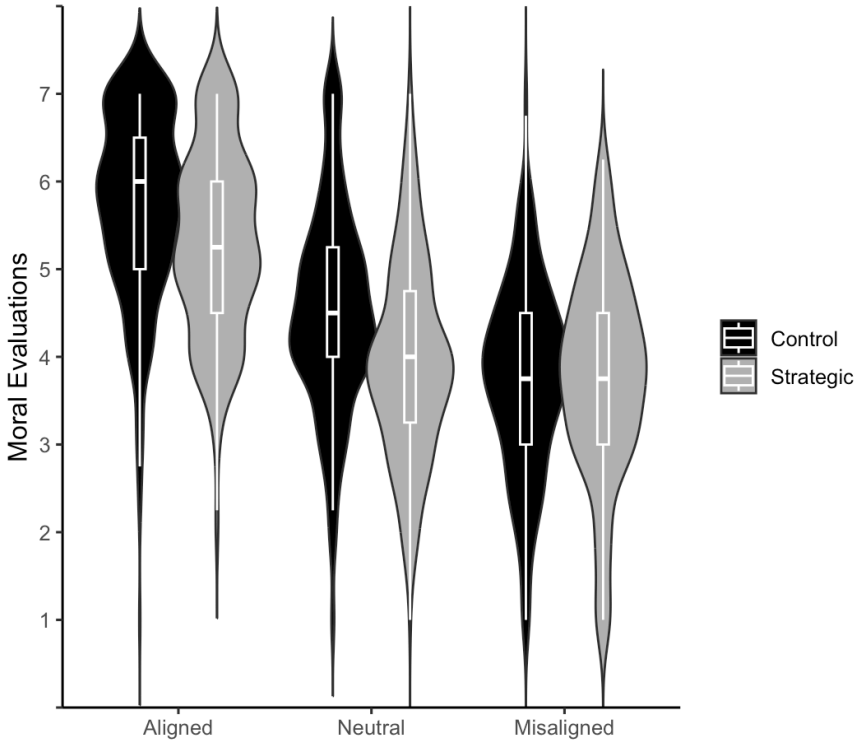
**Strategic Attributions Manipulation Check.** There was a significant effect of the target's position condition, $F(2, 1491) = 98.01$, $p < .001$, partial-$\eta^2 = .116$, a significant effect of strategic attribution condition, $F(1, 1491) = 397.91$, $p < .001$, partial-$\eta^2 = .211$, and a significant interaction between issue position and strategic attribution condition predicting perceptions of being strategic, $F(2, 1491) = 8.26$, $p < .001$, partial-$\eta^2 = .011$. Importantly, participants thought that the strategic target ($M = 5.32$, $SD = 1.46$, 95% CI [5.23, 5.43]) was more strategic than the control target ($M = 3.88$, $SD = 1.51$, 95% CI [3.78, 3.99]). The neutral target ($M = 5.18$, $SD = 1.39$, 95% CI [5.01, 5.26]) was seen as more strategic than the in-favor target ($M = 4.69$, $SD = 1.53$ 95% CI [4.63, 4.88]), $p < .001$, $d = .34$, who was in turn seen as more strategic than the opposed target ($M = 3.95$, $SD = 1.77$, 95% CI [3.81, 4.05]), $p < .001$, $d = .45$.

[4] Consistent with our preregistration document, we retain all participants, but also note the results are robust to the exclusion of the participants who did not pass the check.

**Moral Evaluations.** When analyzing how much participants perceived their colleague to be moral, the results revealed a significant main effect of alignment, $F(2, 1315) = 286.82$, $p < .001$, $\eta^2 = .302$, a significant main effect of strategic attribution condition, $F(1, 1315) = 26.79$, $p < .001$, $\eta^2 = .020$, and the predicted significant interaction, $F(2, 1315) = 7.78$, $p < .001$, $\eta^2 = .012$. Within the strategic attribution condition, the target whose views aligned with participants was rated as significantly more moral than the neutral and misaligned targets, both $p$s $< .001$, $d$s $\geq 1.24$, and the neutral and misaligned targets did not significantly differ in terms of moral evaluations, $p = .095$ (see Figure S13). In the control/not strategic condition, the aligned target was rated as more moral than the neutral target, $p < .001$, $d = 1.07$, who was rated as more moral than the misaligned target, $p < .001$, $d = .66$.

The results are again robust to the exclusion of "honesty," whereby there is a significant main effect of stance, $F(2, 1315) = 345.58$, $p < .001$, $\eta^2 = .345$, a main effect of strategic attribution condition, $F(1, 1315) = 12.87$, $p = .002$, $\eta^2 = .007$, and a significant interaction, $F(2, 1315) = 6.73$, $p = .001$, $\eta^2 = .012$.

**Figure S13.** Moral evaluations by whether the target's views were neutral or aligned versus misaligned with the participants' views.




In both the strategic and control attribution conditions, the aligned target was rated as more moral than the neutral target, both $p$s $< .001$, $d$s $\geq 1.07$. Even when participants thought a target was aligned with them for strategic reasons ($M = 5.29$, $SD = 1.06$, 95% CI [5.14, 5.44]), they rated the target was as more moral than a non-strategic neutral target, ($M = 4.50$, $SD = 1.17$, 95% CI [4.35, 4.65]), $t(410) = 7.17$, $p < .001$, $d = 0.71$. This suggests that a target's position on an issue—including expressed neutrality— had a stronger influence on moral evaluations than did information about whether the target was behaving strategically.

Supplemental Study 7 suggests that neutral expressions explain variance in negative moral perceptions above and beyond perceptions of the target as strategic.

### 2.8 Supplemental Study 8

The primary goal of Supplemental Study 8 was to directly test the discrepancy between actors and observers within a single experiment. Specifically, we held the social issue constant and randomized participants to be message senders (actors) or recipients (observers). We predicted that actors would overestimate how moral observers perceive them to be, and that these actor-observer gaps in moral perceptions would be connected to the differential attributions formed for neutrality.

#### Method

**Participants.** We recruited 600 participants from Prolific Academic in a preregistered design (https://aspredicted.org/8R6_T32). We recruited 600 participants and ended up with 602 participants with complete data (271 men, 318 women, 8 non-binary, 5 prefer not to disclose; $M_{age}$ = 36.38; $SD_{age}$ = 12.74).

**Procedure.** Participants were randomly assigned to one of two conditions (actor or observer) in a between-subjects design. Those in the recipient condition were further randomly assigned to one of three conditions: in favor vs. neutral vs. opposed to affirmative action. All participants were informed that they were taking part in a study on first impressions in which they would either send a message to or receive a message from another Prolific participant. In the actor condition, participants were asked to respond to a question from their partner: "hey, what do you think about affirmative action?" They could choose to send a message that they were either in favor of, neutral on, or opposed to affirmative action. In the observer condition, participants were told their partner had been asked this same question, and had responded that they were in favor of, neutral on, or opposed to affirmative action.

In the actor condition, participants predicted how moral they would be seen by their partner using the same items as in Study 1C ($\alpha$ = .92); observers rated their ostensible partners' morality on the same scale ($\alpha$ = .89). Participants also completed items assessing attributions for issue position, drawing on the categories we discovered in Study 1A. Specifically, participants in the observer condition indicated the extent to which they thought the target's position was driven by five possible attributions (participants in the sender condition indicated the extent to which their partner would think their position was driven by these attributions) on a 7-point scale (*consideration for their feelings (conflict avoidance); consideration for the consequences of sharing their opinion for themselves (strategic); their level of concern about the issue (apathy); their level of indecision about the issue (indecision); the amount of information they have about the issue (information)*. Finally, all participants were asked about their true (privately held) stances on affirmative action (-4 = completely opposed; 0 = neutral; +4 = completely in favor).

#### Results

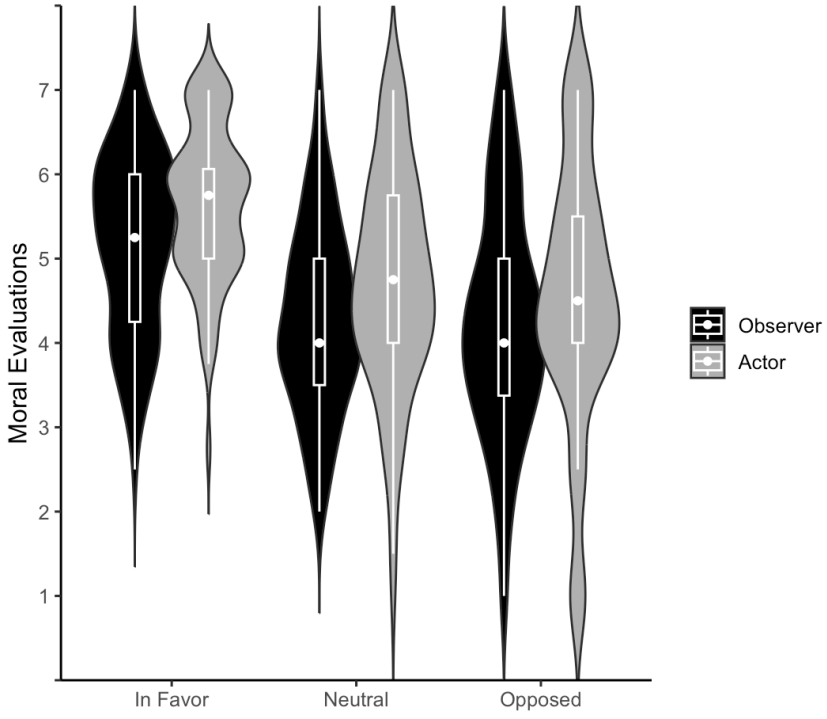
**Prevalence of Neutrality Among Senders**. Among actors, in private, 10.7% of participants claimed they were neutral on affirmative action, whereas 61.4% were in favor, and 27.9% were opposed. Yet, when sending messages to a partner, 33.6% claimed neutrality, 52.3% claimed being in favor, and 14.1% claimed being opposed, indicating significant differences between participants' private and public positions, $\chi^2(4, 298) = 185.83$, $p < .001$, Cramer's $V$ =

.56. Indeed, people were three times more likely to send a neutral message than to actually be neutral, a significant difference, $\chi^2(1, 298) = 34.82$, $p < .001$, Cramer's $V = .34$.

**Moral Evaluations**. When analyzing how much participants perceived their colleague to be moral, the results revealed a significant main effect of actor condition, $F(1, 560) = 14.54$, $p < .001$, partial-$\eta^2 = .025$. Actors predicted they would be rated as more moral than they were seen by observers. There was also a significant main effect of issue position condition, $F(2, 560) = 45.78$, $p < .001$, partial-$\eta^2 = .141$. Neutral and opposed-to-affirmative action participants were rated as significantly less moral than those in favor of affirmative action ($p$s < .001), but did not differ significantly from each other, $p = .431$. There was no actor/observer role by issue position interaction, $F(2, 560) = 0.16$, $p = .853$, $\eta^2 = .001$ (see Figure S14). These results are again robust to the exclusion of the "honest" item.

Our focal preregistered analyses involved the comparison of the actors and observers in the neutral condition. Participants who sent the neutral message anticipated that they would be rated as more moral than they were rated by observers, $t(186) = 2.55$, $p = .012$, $d = 0.37$. Among those in favor, participants in the actor condition again overestimated how moral they would be rated by observers, $t(247) = 3.53$, $p < .001$, $d = 0.46$. For those with stances opposed to affirmative action, actors and observers offered similar morality ratings, $p = .546$. Interestingly, while opposed- and neutral targets were rated as similarly (im)moral, only actors in the opposed condition correctly anticipated this.

**Figure S14.** Predicted vs. actual moral evaluations (1 = *not at all*; 7 = *very much*) by role (actor/observer) and issue position (in favor of, neutral, or opposed to affirmative action).




**Attributions**. We also examined whether the attributions for the target's position varied by actor/observer and position conditions. In examining the attributions, there were significant effects of issue position across measures, $F$s ≥ 7.56, $p$s < .001, $\eta^2 = .091$, and a significant effect

of actor/observer condition on attributions of conflict avoidance, $F(1, 560) = 55.81$, $p < .001$, $\eta^2 = .091$ (all other $p$s $\geq .096$). Because we were specifically interested in attributions for the neutral self versus other, we again examined the effects of message sender condition for neutral targets. Participants in the actor condition thought that receivers would render higher attributions of conflict avoidance ($M = 4.10$, $SD = 1.61$, 95% CI [3.78, 4.43]) than they did ($M = 3.14$, $SD = 1.67$, 95% CI [2.79, 3.48]), $t(201) = 4.22$, $p < .001$, $d = 0.58$, and thought that observers would rate them as being more concerned about the issue ($M = 4.36$, $SD = 1.45$) than they did ($M = 3.78$, $SD = 1.58$), $t(201) = 2.73$, $p = .007$, $d = 0.37$. No other significant differences emerged (strategic: $M_{self} = 4.28$, $SD = 1.81$; $M_{other} = 4.55$, $SD = 1.72$; indecision: $M_{self} = 4.60$, $SD = 1.78$; $M_{other} = 4.66$, $SD = 1.61$; lacking information: $M_{self} = 4.23$, $SD = 1.42$; $M_{other} = 4.40$, $SD = 1.41$), $p$s $> .305$.

Entering actor/observer condition (actor vs. observer) and position condition (in favor vs. neutral vs. opposed), the attributions, and moral evaluations into a moderated mediation model revealed that, relative to observers, actors' increased attributions of conflict avoidance, 95% CI [.04, .30], and issue concern, 95% CI [.05, .33] drove the tendency for senders to overestimate how well their neutrality would be received.

**Other variables.**

**Authenticity.** We also asked participants in the actor condition how much their partner would think they were authentic using the same four items from previous studies; participants in the observer condition were asked to rate their partner's authenticity with the same four items ($\alpha = .83$).

There was a main effect of actor/observer condition, $F(1, 596) = 9.58$, $p = .002$, partial-$\eta^2 = .016$, a main effect of stance, $F(2, 596) = 32.34$, $p < .001$, $\eta^2 = .098$, but no significant interaction, $F(2, 596) = .94$, $p = .392$. Participants in the actor condition ($M = 5.34$, $SD = 1.12$, 95% CI (5.21, 5.51)) overestimated how authentic others would perceive them to be ($M = 5.06$, $SD = 1.19$, 95% CI (4.94, 5.19)). Anticipated and perceived authenticity ratings were lowest among participants in the neutral condition ($M = 4.69$, $SD = 1.22$, 95% CI (4.54, 4.84)) but higher in the in-favor ($M = 5.44$, $SD = 1.04$, 95% CI (5.27, 5.53)) and opposed conditions ($M = 5.50$, $SD = 1.03$, 95% CI (5.35, 5.76), both $p$s $< .001$, $d$s $\geq .68$.

**Attributions.** In the main text, we present the attribution results for neutral targets (of central concern). We also examined the results for pro- and anti-affirmative action targets. For both groups, message senders thought that receivers would render higher attributions of conflict avoidance than they did, $p$s $< .001$, $d$s $> 0.58$, and higher attributions of issue concern than they did, $p$s $< .029$, $d$s $> 0.25$ (see Table S18).

**Table S18**. Descriptive statistics for the attributions by condition.

| Measure | Condition | | | | | |
|---|---|---|---|---|---|---|
| | Pro- Affirmative Action | | Neutral | | Anti- Affirmative Action | |
| | Actor | Observer | Actor | Observer | Actor | Observer |
| Conflict avoidance | 4.26 (1.80) | 2.84 (1.68) | 4.10 (1.61) | 3.13 (1.67) | 2.79 (1.87) | 1.99 (1.27) |
| Strategic | 3.63 (1.82) | 3.50 (1.72) | 4.28 (1.81) | 4.35 (1.77) | 2.98 (1.87) | 2.59 (1.46) |
| Issue concern | 5.11 (1.37) | 4.78 (1.45) | 4.36 (1.45) | 3.78 (1.58) | 3.98 (1.97) | 4.26 (1.63) |
| Indecision | 2.90 (1.64) | 3.01 (1.55) | 4.60 (1.77) | 4.61 (1.65) | 2.10 (1.56) | 2.32 (1.37) |

| | | | | | | |
|---|---|---|---|---|---|---|
| Amount of Information | 4.57 (1.44) | 4.69 (1.25) | 4.23 (1.42) | 4.29 (1.42) | 4.02 (1.96) | 4.01 (1.63) |

**Future Interactions.** We asked actors, "Do you predict that your partner would want to interact with you further?" Observers were asked, "Would you want to interact with your partner further?" Both questions were asked on a binary (yes/no) scale.

Most actors (yes: $n$ = 204/298, or 68.5%) and observers (yes: $n$ = 204/304 or 67.1%,) predicted that their partner would interact with them again and most observers said they would want to interact with their partner again. The difference in rates of predicted interaction between condition was not significant, $\chi^2(1, N = 602) = .13$, $p = .72$. Furthermore, the only significant self-other difference in how much the observer would want to interact with the sender was in the opposed-to-affirmative action condition, $\chi^2(1, N = 142) = 5.56$, $p = .018$.

**Moralization.** We used the same two items to measure moralization as in previous studies, $r(602) = .77$, $p < .001$. The three-way interaction term between moralization, actor/observer role, and stance did not significantly predict moral evaluations, $B = .12$, $SE = .10$, $p = .211$.

**Attitude Strength.** We asked two items to measure participants' attitudes toward affirmative action: "How strongly do you feel about your position on this issue?" and "How important is this issue?" (1 = not at all; 7 = very much), $r(602) = .60$, $p < .001$. The three-way interaction term between moralization, sender/receiver role, and stance did not significantly predict perceptions of morality, $B = .13$, $SE = .10$, $p = .172$.

Taken together, people who communicate a neutral stance overestimate the extent to which a communication partner assumes their neutrality is driven by taking the partner's feelings into account, as well as their issue concern, which leads them to overestimate how positively they will be morally evaluated by message receivers. One limitation of this study is that message selection is endogenous, and thus it makes interpreting the actors' predictions relative the observers' difficult to interpret. Thus, we present these data in the Supplement.

## 2.9 Supplemental Study 9

Thus far, we have documented that while people frequently rely on neutrality as a conversational strategy, observers rate neutral targets as poorly as those who oppose them. One limitation of the current studies is a reliance on self-report measures. In Supplemental Study 9, we examine the behavioral consequences of expressed neutrality in a field context. Specifically, we investigate how people interact with flyers posted around campus in which live discussion forum speakers claim neutrality or other stances on a political issue.

### Method

We conducted an observational field experiment on the campus of a large public East Coast university in the United States. We preregistered this study at https://aspredicted.org/z8ix7.pdf. We created three versions of a poster advertising a fictitious panel forum discussion about the recent legalization of marijuana in the state. Each included a quote from one of the three discussion panelists indicating that they were either in favor, opposed, or neutral on this topic. Because the state recently passed legislation supporting legalization, and the generally progressive community politics, we presumed that most observers of the posters would be "in favor" of legalization. The names of all three panelists were listed on

each poster, and we varied which panelist was quoted. Posters all included ten tabs with a fictitious registration web address that could be torn off.

We designed the field study to maximize power with at least 50 posters per condition, but with sensitivity to the amount of available space, possible cross-contamination, and access to campus buildings. We instructed 11 research assistants, blind to the hypotheses, to hang the flyers ($n_{posters}$ = 169; 55 opposed, 49 in favor, and 65 neutral) around the university grounds while maintaining a reasonable distance between flyers. Importantly, while we separated the locations to minimize cross-contamination across conditions, it is possible that people on campus saw posters from different conditions. However, the posters were not in conflict with one another, because the manipulation simply included a different panelist's quote rather than a different purpose of the forum itself. Therefore, this would not likely arouse suspicion. The research assistants varied in their level of access to different campus buildings and housing due to COVID-19 restrictions. Therefore, we divided the campus into 11 different areas, which we then randomized to one of the three conditions using a random number generator. Each RA was instructed to put the posters up on communal posting spaces (e.g., bulletin boards) within their location, and we gave each RA approximately 15 flyers each (one RA had a smaller location and was given 10 flyers, while another two RAs had larger areas and received 20 and 19 flyers each to post (one was misplaced); total *N* posters = 169; 55 opposed, 49 in favor, and 65 neutral).

Research assistants placed the posters in their assigned areas on Thursday, April 29, 2021, between noon and 5 PM, and went to retrieve them on Monday, May 3, 2021, before noon (exact timing for placement and retrieval were recorded). Research assistants took photographs of each poster when it was posted, and again when they went to retrieve it. When they retrieved each poster, they indicated whether it had been taken down (1 = *yes*, 0 = *no*), and if it was still up, they indicated whether it had been covered, torn, or written on (for each of these: 1 = *yes*, 0 = *no*), and the amount of weather damage it incurred if any (from 1 = *none* to 3 = *fully damaged*). Research assistants also counted the number of tabs taken, if any (out of ten tabs).

**Pretest.** We pretested the posters to ensure that they varied in terms of the panelists' perceived position on the issue, but not in terms of how interesting, attention-grabbing, surprising, or curiosity-evoking they were. We recruited 152 American participants on Prolific (50% identified as men and 50% identified as women; $M_{age}$ = 39.50, *SD* =12.43) and randomly assigned them to 1 of 3 conditions (neutral vs. opposed vs. in-favor of marijuana legalization). Since we would not have access to individuals' attitudes in the field, we did not examine the mismatches or matches between participants' own views and the randomized condition, as we do in the other experiments. We asked participants to imagine that they saw the flyer on a walk in their neighborhood and asked them about their perceptions of the flyer. As a manipulation check, participants indicated what the person's publicly stated position on legalization on a 5-point scale (where -2 = opposed to the issue, 0 = neutral, and +2 = in favor of the issue). They also indicated what they assumed was the person's true position on the issue was, using the same scale. Participants also indicated how interesting and attention-grabbing the flyer was on scales from 1 (*Not at all*) to 5 (*Extremely*), as well as their emotional reactions to the poster (angry; disgusted; calm; contemptuous; sad; curious; surprised; all on 7-point scales). At the end of the survey, we asked participants what the speaker had said on the poster (to examine recall accuracy).

The results of the manipulation check, which asked for the person's publicly stated position (from 1 = *opposed* to 5 = *in favor*), showed that all three conditions significantly differed from each other, $F(2, 149) = 425.10$, $p < .001$, $\eta^2 = 0.85$. The opposed poster was seen

as more opposed (i.e., lower values) to marijuana legalization ($M_{opposed}$ = 1.18, $SD_{opposed}$ = 0.72, 95% CI [.98, 1.38]) than was the neutral ($M_{neutral}$ = 3.12, $SD_{neutral}$ = 0.52; 95 % CI [2.97, 3.26]) and in favor posters ($M_{supportive}$= 4.84, $SD_{supportive}$ = 0.64, 95% CI [4.66, 5.02]), $t$s $\geq$ 13.80, $p$s <.001, ds > 2.73. Results showed that 96.71% of the sample accurately recalled the poster's stance and accuracy did not vary across condition, $F(2, 149) = .19$, $p$ =.826.

Ratings of the quoted person's true (privately held) position also differed by condition, $F(2, 149) = 136.29$, $p < .001$, $\eta^2 = .65$. The opposed poster ($M_{\text{opposed}}$ = 1.76, $SD_{\text{opposed}}$ = 1.04, 95% CI [1.46, 2.06]) was rated as less in favor than the neutral ($M_{\text{neutral}}$ = 3.47, $SD_{\text{neutral}}$ = .95, 95% CI [3.20, 3.74]) and in favor poster conditions ($M_{\text{supportive}}$ = 4.72, $SD_{\text{supportive}}$ = .67, 95% CI [4.53, 4.91]), $t$s $\geq$ 6.98, $p$s <.001, $d$s > 1.39.

Importantly, the posters did not differ in terms of their interestingness or attention-grabbing nature, $F(2, 149) = 0.34$, $p$ = .711 and $F(2, 149) = 0.46$, $p$ = .633 respectively. The posters also did not vary in how surprising they were, or the degree to which they evoked curiosity in participants, $F(2, 149) = 1.70$, $p$ = .187 and $F(2, 149) = 0.66$, $p$ = .520, addressing the possibility that more people attend to neutral posters because they are more counterintuitive in nature. We also found that the negative moral emotions of anger and contempt significantly predicted intentions for taking the poster down, $b$s > .22, $p$s < .001, $R^2$s > .33, while interest and curiosity positively predicted keeping the poster up, $b$s > .09, $p$s < .024, $R^2$s > .02. Thus, the behavior of taking down the poster appears related to the emotions linked to moral condemnation.

We explored participants' reactions to the poster, operationalized as their self-reported likelihood of taking the poster down, taking a tab from the poster, visiting the URL, and defacing and damaging the poster, all measured on a 1(*Not at all*) to 5(*Extremely*) likelihood scale. Participants were more likely to take a tab or visit the URL from the neutral poster than the opposed or in favor posters, but there were no condition differences in damaging or writing on the poster. Importantly, while the ANOVA did not reach significance, $F(2, 149) = .62$, $p$ =.541, the pattern of means qualitatively suggested that the neutral poster was most likely to be taken down ($M_{\text{neutral}}$= 1.41, $SD_{\text{neutral}}$ = 1.02, 95% CI [1.12, 1.70]), followed by the opposed poster ($M_{\text{opposed}}$ = 1.28, $SD_{\text{opposed}}$ = .86, 95 % CI [1.04, 1.52]) and then the in favor ($M_{\text{supportive}}$ = 1.22, $SD_{\text{supportive}}$ = .83, 95% CI [.98, 1.45]).

**Results**

We predicted that more neutral posters would be removed than in favor posters. We examined which flyers were removed over approximately four days as the dependent variable. A Pearson chi-square test indicated that the flyers differed significantly in terms of the frequency at which they were removed, $\chi^2(2, 167) = 10.66$, $p$ = .005) (see Table S19). More neutral flyers were removed (63.08%) than the "n favor flyers (32.65%), $\chi^2(1, 112) = 10.26$, $p$ = .001), or opposed flyers (45.45%), $\chi^2(1, 118) = 3.71$, $p$ = .054). Rates of removal of the in favor and opposed flyers did not significantly differ ($\chi^2(1, 102) = 1.75$, $p$ = .186).

Because posters were randomized by location, in analyzing the data, it occurred to us that the effects might driven by specific locations (e.g., being taken down in a building by custodial staff). Thus, we conducted exploratory robustness checks using a Lasso (L1) regularized logistic regression model to prevent overfitting. Specifically, we regressed condition (comparing neutral = 1 to in favor and opposed = 0), and the dummy coded location control onto the flyer removal outcome measure. The results revealed that, while there were some effects of poster location (see SOM), the significant effect of neutral condition on flyer condition remained when controlling

for location, beta = 0.49, $p = .026$, as well as significant location effects, such that posters at location 2, 3, 7, and 8 were significantly more likely to be taken down.

**Table S19.** Number of flyers removed in each condition. Cells refer to the number of flyers removed or not, based on whether the poster espoused a stance opposed to, neutral toward, or in favor of legalizing marijuana.

| | | Condition | | | |
|---|---|---|---|---|---|
| | | Opposed to Legalization | Neutral | In Favor of Legalization | Total |
| Taken Down | 0 (no) | 30 | 24 | 33 | 87 |
| | 1 (yes) | 25 | 41 | 16 | 82 |
| Total | | 55 | 65 | 49 | 169 |

Examining the other collected measures, posters did not differ significantly by condition in terms of the damage incurred due to weather, $F(2, 86) = 2.86$, $p = .063$, $\eta^2 = .062$. None of the posters had been covered and only one was written on (an opposed poster). However, there were significant differences between conditions in terms of the average time between drop off and attempted pick up, $F(2, 166) = 29.93$, $p < .001$, partial-$\eta^2 = .27$). Importantly, this would not explain the higher rate of neutral posters being taken down because the in-favor posters had been up for significantly longer ($M_{\text{in-favor}} = 97.99$ hours, $SD_{\text{in-favor}} = 3.95$, 95% CI [96.86, 99.12]) than the opposed ($M_{\text{opposed}} = 93.85$ hours, $SD_{\text{opposed}} = 2.59$; 95% CI [93.14, 94.55]; $t = 4.15$, $p < .001$, $d = 1.36$) and neutral posters ($M_{\text{neutral}} = 94.10$ hours, $SD_{\text{neutral}} = 2.63$; 95% CI [93.45, 94.76]); $t = 6.72$, $p < .001$, $d = 1.27$. The neutral and opposed posters did not differ significantly in time posted ($t = 0.46$, $p = .647$).

We also found that posters did not vary across conditions in terms of the number of tabs removed, $F(2, 86) = 1.93$, $p = .152$, $\eta^2 = .043$. This might be because tabs removed could indicate interest, support for, or opposition to legalization.

There were differences in how often the remaining 91 posters were torn, $\chi^2(2, 89) = 12.68$, $p = .002$. Posters from the in-favor-legalization condition were torn more often (36.11%) than posters from the opposed legalization (3.23%; $\chi^2(1, 60) = 10.73$, $p = .001$) or neutral conditions (12.50%; $\chi^2(1, 55) = 4.04$, $p = .045$). The photographs showed this appeared related to the tabs being torn off, rather than defacing the poster.

**Discussion**

Posters that expressed a neutral viewpoint on marijuana legalization were removed at a higher rate than posters expressing an in favor or opposed-legalization stance. However, as mentioned above, there were significant location effects, with 100% of the posters being taken down at location 7 (which was a neutral condition), leading us to believe this was an issue with this location. Thus, we include this study in the Supplement for transparency, but encourage caution when interpreting these results.

## 2.10 Supplemental Study 10

In Supplemental Study 10, we examined whether an informational intervention may attenuate the actor-observer gap.

### Method

We recruited 800 Americans employed full-time via Prolific. They were randomly assigned to 1 of 8 conditions in a 4(actor vs. in favor observer vs. neutral observer vs. opposed observer) x 2(debias vs. control) between-subjects design. We used the safe injection sites paradigm described previously.

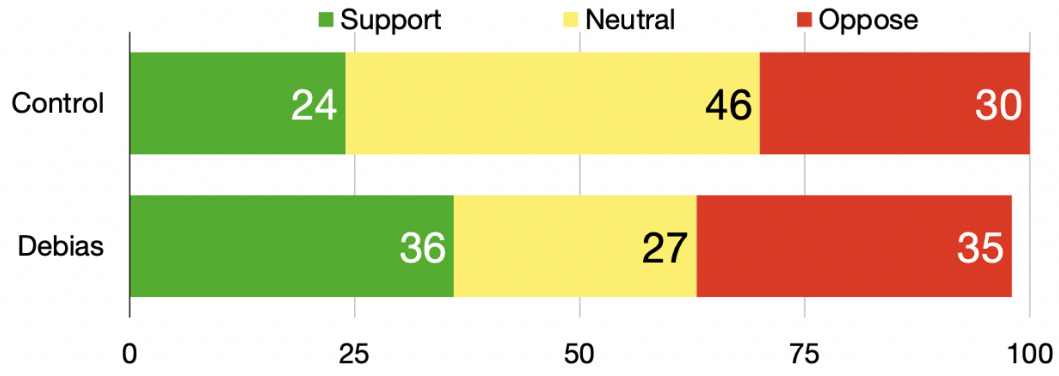
In the informational intervention condition, participants first read, "Previous research has found that stating neutrality is often misperceived: People who say they're neutral often do so because they're undecided or would like additional information. However, when other people hear them express neutrality, they think the person is trying to be strategic, manipulative, or doesn't care enough about the issue." Participants in the control condition proceeded to the study.

In the actor condition indicated how they would respond to a workplace conversation about safe injection sites in the community (in favor/neutral/opposed), whereas participants in the observer condition provided moral evaluations of a colleague who indicated being in favor, neutral, or opposed.

### Results

Among actor participants, whereas 46% of participants selected neutrality in the control condition, only 27% did so in the intervention condition, a significant difference, $\chi^2 = 7.71$, $p = .021$, Cramer's V = .20 (see Figure S15).

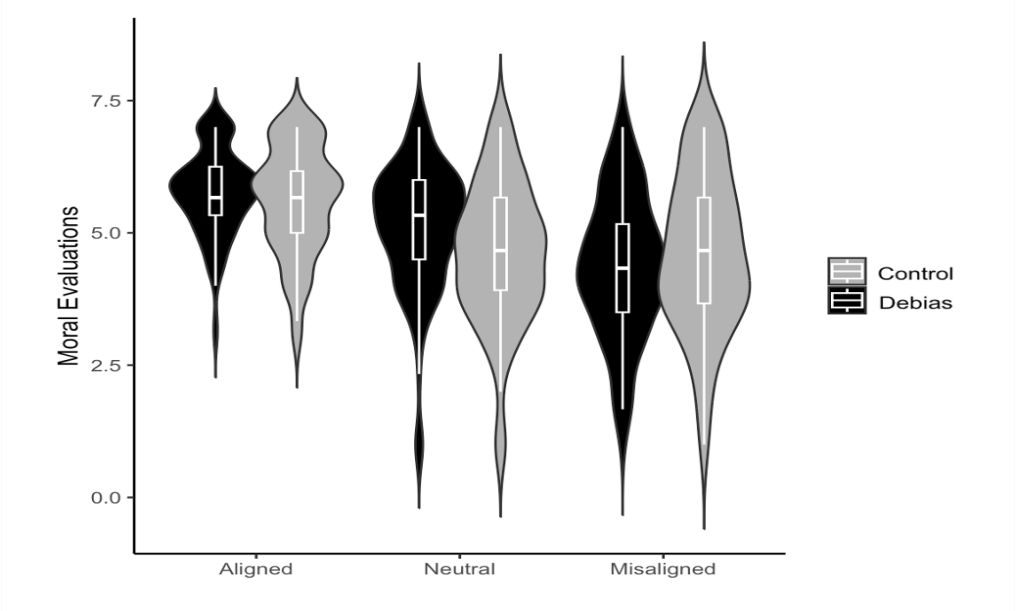
**Figure S15**. The effect of intervention condition on message selection.




Among observer participants, there was a significant main effect of alignment group, $F(2, 597) = 50.76$, $p < .001$, $\eta^2 = .145$, no effect of the intervention condition, $F(1, 597) = 1.67$, $p = .197$, $\eta^2 = .003$, and a significant interaction, $F(2, 597) = 6.40$, $p = .014$, $\eta^2 = .014$.

Decomposing this interaction, within the control condition, we replicated the effect observed throughout the paper: the neutral target was rated as less moral than aligned targets ($p < .001$), but similarly moral to the misaligned target ($p = .975$). However, within the intervention condition, the neutral target was rated as less moral than the aligned target ($p < .001$), but more moral than the misaligned target ($p < .001$) (see Figure S16).

**Figure S16**. The effect of condition on moral evaluations.




Therefore, the informational intervention reduced the rate at which actors used neutrality as a strategy, and also reduced the extent to which observers rated neutral targets as immoral. However, these effect sizes were small, especially in light of the magnitude of the intervention (explicitly describing the effect before the study).

## References

Goodwin, G. P. (2015). Moral character in person perception. *Current Directions in Psychological Science, 24*(1), 38-44.

Kreps, T. A., Laurin, K., & Merritt, A. C. (2017). Hypocritical flip-flop, or courageous evolution? When leaders change their moral minds. *Journal of Personality and Social Psychology, 113*(5), 730–752.

Simonsohn, U., Montealegre, A., & Evangelidis, I. (2025). Stimulus sampling reimagined: Designing experiments with mix-and-match, analyzing results with stimulus plots. *Journal of Personality and Social Psychology*.

Skitka, L. J., Bauman, C. W., & Sargis, E. G. (2005). Moral conviction: Another contributor to attitude strength or something more? *Journal of Personality and Social Psychology, 88*(6), 895-917.

Skitka, L. J. (2010). The psychology of moral conviction. *Social and Personality Psychology Compass, 4*(4), 267-281.

Turiel, E. (1983). The development of social knowledge: morality and convention. United Kingdom: Cambridge University Press.