Cognition 180 (2018) 165–181

Contents lists available at ScienceDirect

Cognition

journal homepage: www.elsevier.com/locate/cognit

Original Articles

When do circumstances excuse? Moral prejudices and beliefs about the true self drive preferences for agency-minimizing explanations

Simon Cullen

Princeton University, United States

ARTICLE INFO

Keywords:
True self
Action
Explanation
Person/situation distinction
Prejudice
Responsibility

ABSTRACT

When explaining human actions, people usually focus on a small subset of potential causes. What leads us to prefer certain explanations for valenced actions over others? The present studies indicate that our moral attitudes often predict our explanatory preferences far better than our beliefs about how causally sensitive actions are to features of the actor's environment. **Study 1** found that high-prejudice participants were much more likely to endorse non-agential explanations of an erotic same-sex encounter, such as that one of the men endured a stressful event earlier that day. **Study 2** manipulated participants' beliefs about how the agent's behavior depended on features of his environment, finding that such beliefs played no clear role in modeling participants' explanatory preferences. This result emerged both with low- and high-prejudice, US and Indian participants, suggesting that these findings probably reflect a species-typical feature of human psychology. **Study 3** found that moral attitudes also predicted explanations for a woman's decision to abort her pregnancy (3a) and a person's decision to convert to Islam (3b). **Study 4** found that luck in an action's etiology tends to undermine perceptions of blame more readily than perceptions of praise. Finally, **Study 5** found that when explaining support for a rival ideology, both Liberals and Conservatives downplay agential causes while emphasizing environmental ones. Taken together, these studies indicate that our explanatory preferences often reflect a powerful tendency to represent agents as possessing virtuous true selves. Consequently, situation-focused explanations often appear salient because people resist attributing negatively valenced actions to the true self. There is a person/situation distinction, but it is normative.

The concept of the *true self* plays a central role in folk psychology (Strohminger, Knobe, & Newman, 2017). Beliefs about the true self predict people's intuitions about personal identity (De Freitas, Cikara, Grossmann, & Schlegel, 2018; Prinz & Nichols, 2016, chap. 26; Strohminger & Nichols, 2014, 2015), what a person values (Newman, Bloom, & Knobe, 2013), whether a person is happy (Newman, De Freitas, & Knobe, 2014; Phillips, Misenheimer, & Knobe, 2011), weak-willed (Newman et al., 2014), morally responsible (Newman et al., 2014), and leading a meaningful life (Schlegel, Hicks, Arndt, & King, 2009; Schlegel, Hicks, King, & Arndt, 2011). Moreover, beliefs about the true self appear to moderate intergroup bias (De Freitas & Cikara, 2018) and decision satisfaction (Kim, Christy, Hicks, & Schlegel, 2017). Collectively, these studies reveal a powerful tendency for people to attribute characteristics they perceive as virtuous to the true self; immoral characteristics tend to be represented as more superficial aspects of the self (De Freitas, Cikara, Grossmann, & Schlegel, 2017; De Freitas, Tobia, Newman, & Knobe, 2016). For example, when participants consider an evangelical Christian man who believes homosexuality to be immoral while also finding himself sexually attracted to men, prejudiced participants are less likely to represent the agent's sexual orientation as part of his true self (Newman et al., 2013).

This paper explores the role that beliefs about the true self play in what may seem an unrelated area of psychology—the study of the cognitive processes that incline people to explain behavior in more or less situational terms. The distinction is a familiar one. Both commonsense and scientific psychology distinguish actions that arise from within an agent from those that are attributable to the circumstances in which the agent acts (e.g., Frankfurt, 1971; Heider, 1983/1958; Jones & Davis, 1965; Kelley, 1973; Smith, 2005; Watson, 1996). To help make the distinction more concrete, consider Darley and Batson's classic (1973) finding: seminary students could be made six times less likely to help an apparently injured person simply by being placed in circumstances where they felt they had to hurry to give a sermon. When we consider one of the hurried seminarians rushing off to give his sermon, ignoring the injured man, we tend to see his callousness as caused by his randomization into the Hurried experimental condition (Darley & Batson, 1973). To borrow a common metaphor, the experimental manipulation may seem to 'externally determine' the hurried seminarians' antisocial behavior (Batson, Darley, & Coke, 1978).

E-mail address: scullen@princeton.edu.

https://doi.org/10.1016/j.cognition.2018.06.021
Received 18 November 2017; Received in revised form 26 June 2018; Accepted 26 June 2018

Intuitions like this one appear to be widely shared (Kunda & Nisbett, 1986; Ross, 1977); however, the cognitive processes that underlie such intuitions remain unclear (Sabini, Siepmann, & Stein, 2001). How do people classify actions along the 'person/situation' dichotomy? A major theoretical tradition in social psychology holds that people locate the causes of actions and events in much the same 'commonsense' way that scientists do—namely, by assessing whether they occur only in the presence of an external pressure, or whether they also occur in the absence of that pressure (Kelley, 1967, 1973). Applied to our previous example, such accounts hold that we judge the seminarian's callous behavior to result from 'the situation' because we believe he would have acted benevolently in sufficiently many other sufficiently similar circumstances (Hewstone & Jaspars, 1987; for philosophical insights see, e.g., Lewis, 1986; Woodward, 2006).

Theorists have developed this basic picture in many ways, but they have tended to agree that laypeople, like scientists, aim to rely on causal-statistical ('covariation') information when explaining morally valenced human actions. However, recent research on the concept of the true self suggests that people may rely on strikingly unscientific considerations for this purpose. In particular, the degree to which an action appears to arise from features of the agent's circumstances may depend on whether the action appears to express the agent's true self. If we represent agents as fundamentally virtuous, our *explanatory preferences*—i.e., whether we tend to emphasize more agent- or more situation-focused factors when explaining an action—may in turn depend on our moral attitudes towards the action. That is, we may prefer situation-focused explanations to the extent that we perceive a mismatch in the moral valences of the agent's action and true self.

To illustrate this idea, consider again one of the hurried seminarians. On the hypothesis to be explored here—*the mismatch hypothesis*—people tend to explain his callousness in terms of the experimental condition into which he was randomized, to the extent that they believe (a) his action was immoral, and (b) his true self is virtuous. On this view, our beliefs about how valenced actions covary with features of the situation should have a small impact on our explanatory preferences relative to the impact of our beliefs about whether actions are *essence-disclosing*. (Psychologists often use 'self-disclosing' to refer to any behavior that expresses something about an agent. In philosophical action theory, the term is used more narrowly to refer only to actions that express something about an agent's *true self*. To avoid confusion, this paper uses the unfamiliar term 'essence-disclosing' in this narrower, action-theoretic sense.)

While the mismatch hypothesis has not been explicitly discussed or explored in previous research, several independent lines of evidence suggest that it warrants investigation. Jones and Nisbett (1972) famously hypothesized that we prefer to explain *our own* actions in terms of features of the situations in which we act, while we prefer to attribute *other agents'* actions to their 'internal' dispositions. The mismatch hypothesis predicts this asymmetry in the case of immoral behaviors. For, researchers have consistently found that we tend to regard ourselves as morally better than average (Epley & Dunning, 2000; Klein & Epley, 2016), which suggests that the valence of any given *immoral* behavior is somewhat more likely to conflict with our assessments of our own true selves than with our assessments of other agents' true selves. Thus, the mismatch hypothesis predicts the traditional actor-observer asymmetry when the target action is immoral. However, parallel reasoning suggests that the mismatch hypothesis predicts the opposite asymmetry for virtuous behavior—since good actions are *less* likely to conflict with our assessments of our own true selves than with our assessments of other actors'. Consistent with this prediction, an authoritative meta-analysis found no evidence for a morally neutral actor-observer asymmetry (Malle, 2006). Rather, the classic asymmetry appeared in studies where participants explained negative events, but reversed in studies where they explained positive events, as the mismatch hypothesis predicts.

The same reasoning appears to apply to intergroup explanatory preferences. If in-group members tend to think of themselves as having morally better true selves than out-group members, the mismatch hypothesis predicts that they will be more likely, compared to base rates, to produce agent-focused explanations for their own members' praiseworthy acts and situation-focused explanations for their blameworthy acts. Members of the out-group will get the opposite treatment. Social psychologists have coined the phrase 'ultimate attribution error' to describe this very patterning (Pettigrew, 1979). Taylor and Jaggi (1974) first investigated intergroup attribution in southern India, against the backdrop of Hindu-Muslim conflict. They asked Hindu participants to imagine themselves in various situations with either a Hindu or a Muslim interlocutor. In all scenarios, Hindus were more likely to give agent-focused explanations for the virtuous behavior of another Hindu agent. The study was replicated in Malaysia with Malay and Chinese subjects (Hewstone & Ward, 1985). If the tendency of in-group members to regard themselves as, on average, morally better than out-group members (Ellemers, Pagliaro, Barreto, & Leach, 2008; Leach, Ellemers, & Barreto, 2007; Levine & Campbell, 1972; although, cf., De Freitas & Cikara, 2018) extends to assessments of their *true* selves, the mismatch hypothesis appears to predict the patterning of intergroup explanatory preferences. (Note that the model does not assume all agents are represented as maximally or equally virtuous.)

The present studies

The patterning of laypeople's explanatory preferences suggests that the mismatch hypothesis is a promising initial account of the conditions that incline people to emphasize more agent- or situation-focused explanations. However, previous research has not investigated the influence of people's beliefs about the true self on their explanatory preferences. The present studies begin exploring this question.

Studies 1–3 found that participants' moral attitudes towards an action predict their explanatory preferences far better than their beliefs about how causally sensitive the action is to features of the agent's circumstances. This is true both for Western (North American) and non-Western (Indian) participants. Studies 4 and 5 supported the hypothesis that these surprising patterns reflect a more general feature of folk psychology identified in recent research, namely, a bias to represent agents as possessing morally virtuous true selves. The results indicate that people often prefer situation-focused explanations because they resist attributing negatively valenced actions to the true self. Study 4 tested this hypothesis by examining the conditions under which moral luck undermines the perception that an agent is fully responsible for his actions. Study 5 tested the hypothesis in the context of partisans' explanations of in-group and out-group political identities.

1. Study 1: Explaining gay sex

Consider the following vignette, adapted from Newman et al. (2013):

> Mark was born into a Christian family that eventually deteriorated, leading his parents to divorce. After being pushed out of home early, Mark met a new group of friends, some of whom were in same-sex relationships. Mark believed that homosexuality is morally wrong, and he encouraged his new friends to resist their attractions to people of the same sex. However, Mark himself was attracted to other men. He openly acknowledged this to his friends and discussed it as part of his own personal struggle. Mark believed that it was his duty to resist his feelings for other men, and he vowed to live a morally decent life the only way he could—by remaining celibate.
> But Mark sometimes failed to live up to his values. For example, one day, after a bad fight with his father, Mark went to see his friend Bill. They shared a bottle of wine and talked for hours.
> That night, Mark hit on Bill and they ended up having sex.

Many explanations for the agent's action are possible. On the one hand, his encounter with Bill plausibly depended to some degree on

features of his situation: for example, the influence of his new friends and the fight he had with his father. On the other hand, facts about the agent himself and his sexual dispositions also seem important. What inclines people to prefer more agent- or situation-focusedexplanations?

Using a similar vignette, Newman et al. (2013) found that both Liberals and Conservatives represent the agent's true self as virtuous. Thus, given the strong association of attitudes to homosexuality and political identification (Inbar, Pizarro, & Bloom, 2009; Inbar, Pizarro, Knobe, & Bloom, 2009), the mismatch hypothesis predicts that participants with relatively positive attitudes towards homosexuality should be more inclined to endorse agent-focused explanations and less inclined to endorse situation-focused explanations. Study 1 tested these predictions.

1.1. Methods

1.1.1. Participants

Participants were 403 adults ranging in age from 19 to 75 years (M = 36, 55% female). In all studies, participants were recruited using mTurk, provided informed consent, and were paid $0.20–$0.30.

1.1.2. Procedure

Participants read a vignette like the one described above. (All vignettes are reproduced in Appendix A.) The vignette made available various explanations for the agent's action, which participants rated (counterbalanced for order) on scales ranging from 1 ('completely disagree') to 9 ('completely agree'). Four explanations cited more agent-focused factors—e.g., the agent had sex with another man because he is gay—and four cited more situation-focused factors—e.g., the agent had sex with another man because of the influence of his new friends (Table 1). To report the degree to which his action seemed essence-disclosing—i.e., expressive of his true self—participants rated the statement "By having sex with Bill, Mark showed who he most truly is, deep down" on a scale from 1 ('completely disagree') to 9 ('completely agree').

The short form of Herek's (1998) *Attitudes to Gay Men* (ATLG) scale served as a measure of the degree to which participants perceived the agent's action as immoral. Participants rated the following statements (counterbalanced for order) on scales ranging from 1 ('completely disagree') to 5 ('completely agree'): *I think male homosexuals are disgusting; *Male homosexuality is a perversion; Male homosexuality is a natural expression of sexuality in men; *Sex between two men is just plain wrong; Male homosexuality is merely a different kind of lifestyle that should *not* be condemned (*reverse-coded). As an attention check, one item instructed participants to "select 'agree' for this question."

Participants responded to two questions intended to test their comprehension of the vignette: "How will Mark feel the next day when he reflects on his action?" (proud/ashamed), and "What does Mark believe?" (that homosexuality is moral/that homosexuality is immoral).

Table 1
Factor loadings from exploratory principal component analysis on ratings of the eight action-explanations (Study 1). Items are Likert-ratings of potential explanations of the agent's homosexual encounter.

	Component	
Explanation of agent's action	1	2
"…because he comes from a broken family"	**.88**	−.10
"…because his mother hurt him deeply …"	**.89**	−.13
"…because of the fight with his father"	**.69**	−.13
"…because of the influence of his new friends"	**.77**	−.07
"…because he is gay"	−.20	**.75**
"…because he is attracted to men"	−.21	**.84**
"…because he wanted to have sex with Bill"	−.18	**.85**
"…because he values intimacy …"	.07	**.61**

Factor loadings with values > .5 are bolded.

On the final page of the survey, participants rated themselves on a scale from 1 ('extremely conservative') to 7 ('extremely liberal') and indicated their age and sex.

1.2. Results

Three hundred and sixty-four participants correctly responded to the comprehension checks. Responses from participants who failed a check were excluded, but this had no meaningful effect.

1.2.1. Factor analysis

Tests of factorability indicated that participants' ratings of the eight explanations were suitable for factor analysis. For every item, there was at least one other with which it was correlated at $r \geqslant .5$. The Kaiser-Meyer-Olkin measure of sampling adequacy was .78, above the recommended threshold, and Bartlett's test of sphericity was significant, $\chi^2(28) = 1190$, $p < .001$. Principal component analysis with varimax rotation was therefore used to extract factor scores. Two components with eigenvalues greater than 1 emerged that together explain 65% of the variance in participants' ratings of the eight explanations (Table 1). The four intuitively situation-focused explanations loaded strongly onto the first component, which modeled 42% of the variance in responses, and the four intuitively agent-focused explanations loaded strongly onto the second component, which modeled the remaining 23%, indicating that the explanations were appropriately grouped into dichotomous categories.

1.2.2. Attribution and moral attitudes

The attitudes to homosexuality scale was highly reliable (Cronbach's α = .94). Correlation coefficients were therefore calculated for ATLG scores and the disaggregated person and situation components extracted via factor analysis. ATLG scores predicted participants' preferences for both agent-focused explanations, $r(362) = .32$, 95% CI: [.23, .42], $p < .001$, and situation-focused explanations, $r(362) = -.45$, 95% CI: [−.54, −.36], $p < .001$, indicating that prejudice may lead people to emphasize situation-focused rather than agent-focused explanations for homosexual behavior.

Bootstrap mediation analysis (Hayes, 2013) was used to test the hypothesis that beliefs about whether the agent's action expresses his true self ('essence-disclosure') mediate the effect of attitudes to homosexuality on explanatory preferences. ATLG scores were set as the independent variable with essence-disclosure ratings as mediator and person scores as DV (Fig. 1). Consistent with the mediation hypothesis, the analysis revealed a significant indirect effect: ab = .10, 95% CI (bootstrapped): [.05, .17]. (All bias-corrected bootstrap confidence intervals were calculated using 5000 bootstrap samples.)

1.3. Discussion

This study found that participants' moral attitudes towards homosexuality powerfully predicted the degree to which they favored more agent-focused or situation-focused explanations for an erotic encounter between two men. Moreover, the results were consistent with a model according to which the effect of people's moral attitudes on their explanatory preferences is partially mediated by their beliefs about

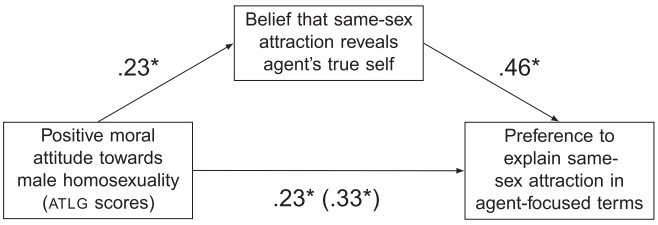

Fig. 1. Mediation (Study 1). *$p < .001$.

whether the encounter expressed the agent's true self. Since people tend to believe that the agent's true self is virtuous regardless of their attitudes towards homosexuality (Newman et al., 2013), these results appear to support the mismatch hypothesis. Participants who perceived a mismatch in the valences of the agent's action and true self rated the situation attributions more highly and saw the action as less expressive of the agent's true self than did those who perceived no such mismatch. The agent-focused explanations followed the reverse pattern.

The associations between participants' ratings of the eight explanations are also noteworthy. The fact that two-thirds of the variance in participants' explanatory preferences can be modeled by two orthogonal factors suggests that participants were responding to an underlying dichotomy when they assessed the explanations for the agent's action. The fact that the explanations intuitively classified as agent-focused and those intuitively classified as situation-focused neatly sorted onto these two factors (Table 1) suggests that the relevant dichotomy is the intuitive person/situation distinction. This result helps to allay concerns about whether the dichotomy reflects an important feature of how human beings actually explain intentional behavior (e.g., Malle, 2011; Malle, Knobe, O'Laughlin, Pearce, & Nelson, 2000).

2. Study 2: The role of causal information

It might be objected that whereas participants with negative attitudes towards homosexuality may be more inclined to think that the agent's sexual orientation is fundamentally something he has chosen, participants with more positive attitudes may think that he was born gay and there's not much use in fighting it (Haider-Markel & Joslyn, 2008; Jayaratne et al., 2006; Lewis, 2009; Suhay & Jayaratne, 2012). Thus, because people tend to believe that actions arising from innate traits are more stable across time and less sensitive to environmental variation than those acquired as result of choice (Gelman, 2003; Gelman, Heyman, & Legare, 2007; Haslam, Bastian, & Bissett, 2004), the results of Study 1 might be explained on the hypothesis that our attitudes to homosexuality influence the degree to which we perceive same-sex attraction as causally sensitive to variations in the agent's situation. In particular, high-prejudice participants may be more likely than low-prejudice participants to represent the agent's action as causally sensitive to specific features of his situation. This difference may explain why high levels of anti-gay prejudice predict a strong preference for situation-focused explanations of the agent's erotic same-sex encounter. Study 2 tested this rival explanation against the mismatch hypothesis by directly manipulating the degree to which participants viewed the agent's action as causally sensitive to his circumstances.

North American mTurkers tend to have highly positive attitudes towards homosexuality—the median ATLG score among North American participants is 21 out of 25 (Fig. 3). Study 2 also sought to replicate the basic valence-explanation asymmetry with participants who hold more negative attitudes towards homosexuality. Indian populations tend to have more negative attitudes towards sexuality and sexual orientation than North American populations (Asthana & Oostvogels, 2001; Patel, Mayer, & Makadon, 2012; Tahmindjis, 2014). Moreover, researchers have found evidence for belief in the virtuous true self when studying participants in the United States, Russia, Singapore, and Colombia (De Freitas et al., 2018), suggesting that this belief reflects a species-typical feature of human psychology. Thus, relative to North Americans, we should expect Indian participants to be significantly:

1. *more* inclined towards situation-focused explanations,
2. *less* inclined towards agent-focused explanations, and
3. *less* inclined to view the action as essence-disclosing.

Study 2 tested these predictions with a sample of English-speaking Indian participants recruited via mTurk.

2.1. Methods

2.1.1. Participants

A new group of 238 North American participants ranging in age from 18 to 74 years (M = 35.8, 48% female) were recruited using mTurk. Additionally, 252 Indian participants ranging in age from 21 to 68 years (M = 32.6, 22% female) were recruited using mTurk. TurkPrime (Litman, Robinson, & Abberbock, 2017) was used to verify that these participants were located in India.

2.1.2. Procedure

Participants were randomized into one of three conditions. The vignette for the Baseline condition was drawn from Study 1 without modification. The two other conditions were formed by appending one or the other of the following texts to the end of the original vignette:

> Person condition: Most people don't find Bill attractive, but Mark has often been sexually attracted to Bill. In fact, Mark often experiences attraction to other men, too.
> Situation condition: Most people don't find Bill attractive, and in the past Mark himself has rarely felt sexually attracted to Bill. In fact, Mark rarely experiences attraction to other men, either.

Participants who read that Mark rarely finds Bill or other men attractive should come to represent Mark's behavior as highly causally sensitive to his situation (i.e., it exhibited low "consistency" and high "distinctiveness"). For, if the traits which led to his encounter with Bill were causally insensitive to relevant changes in his circumstances, the agent would commonly experience same-sex attraction. Thus, because participants in the Situation condition read that he is *not* commonly attracted to other men, they should be more likely than participants in the Person condition to judge that his actions were highly causally sensitive to the details of the situation.

To test this hypothesis, participants rated the following (counterbalanced) statements:

> Weak robustness: If he were in *the very same circumstances* in the future, how probable do you think it is that Mark would have sex with Bill again? (1: Not at all probable – 7: Extremely probable.)
> Strong robustness: Imagine that a week goes by until Mark next sees Bill. Now Mark is feeling much better about the fight with his father and is generally back to his usual self. How probable do you think it is that Mark will have sex with Bill on this occasion? (1: Not at all probable – 7: Extremely probable.)

All other methods were drawn directly from Study 1.

2.2. Results (North American participants)

Two hundred and six participants passed the comprehension checks. Responses from participants who failed at least one check were excluded, but this did not meaningfully affect the pattern of results.

2.2.1. Manipulation checks

To test whether the experimental manipulation affected participants' beliefs about how causally sensitive the agent's action is to features of the situation, two-way ANOVA tests were conducted using experimental condition as the independent variable and weak or strong robustness as the dependent variable. The experimental manipulation had a significant effect on participants' ratings of both weak robustness, $F(2,203) = 9.2$, $p < .001$, and strong robustness, $F(2,203) = 7.3$, $p = .001$. Post-hoc tests revealed that participants in the Situation condition, who read that the agent has rarely experienced same-sex attraction, gave significantly lower ratings of weak robustness, $d = -0.63$, 95% CI: $[-0.92, -0.33]$, and strong robustness, $d = -0.56$, 95% CI: $[-0.85, -0.27]$, than did participants in the Person and Baseline conditions. However, the Person and Baseline conditions did not differ meaningfully on either measure ($ps > .5$).

Table 2
Summary of regression models from Study 2 (North American sample). Notes: baseline condition omitted; to compare the effect sizes of categorical and continuous predictors, this table provides η_p^2 values.

DV	Predictor	η_p^2	β	95% CI		t	p
Situation attribution[a]	Moral attitudes	.19	−.44	−.56,	−.31	−6.9	< .001
	Person condition	.00	.03	−.28,	.34	0.3	> .5
	Situation condition	.00	.13	−.17,	.43	0.9	.385
Agent attribution[b]	Moral attitudes	.11	.33	.20,	.46	5.1	< .001
	Person condition	.02	.35	.04,	.67	2.2	.037
	Situation condition	.00	−.15	−.46,	.15	−1.0	.325
Essence-disclosure[c]	Moral attitudes	.10	.32	.19,	.45	4.8	< .001
	Person condition	.00	.14	−.19,	.46	0.8	.404
	Situation condition	.01	−.22	−.53,	.10	−1.4	.185

[a] $R^2 = .20$, $F(202,3) = 16.8$, $p < .001$. [b] $R^2 = .15$, $F(202,3) = 12.2$. $p < .001$. [c] $R^2 = .13$, $F(202,3) = 9.7$, $p < .001$.

These results indicate that the experimental manipulation succeeded in making the agent's behavior seem more causally sensitive to environmental factors in the Situation condition. Notably, in this condition, participants were significantly less likely to predict that the agent would repeat his actions in similar circumstances in the future.

2.2.2. Factor analysis

Participants' ratings of the eight explanations were suitable for principal component analysis (Kaiser-Meyer-Olkin = .74; Bartlett's test of sphericity: $\chi^2(28) = 616$, $p < .001$). Replicating Study 1, the four intuitively situation-focused explanations loaded strongly onto a single component, which explained 40% of the variance, and the four intuitively agent-focused items loaded strongly onto a second component, which explained the remaining 22%.

2.2.3. Attribution and moral attitudes

Herek's attitudes to homosexuality scale was again highly reliable ($\alpha = .95$). ATLG scores and dummy variables representing the experimental conditions were regressed against situation scores, person scores, and essence-disclosure ratings. ATLG scores significantly predicted all three dependent variables (Table 2). Additionally, the experimental manipulation had a small, borderline-significant effect on participants' ratings of the agent-focused explanations but did not predict any of the other measures. Repeating the analysis using ratings of the individual explanations as DVs revealed that the Situation condition differed significantly only on two of the eight explanations: compared to the Person and Baseline conditions, participants in the Situation condition gave lower ratings to the statements "Mark had sex with Bill because he is gay," $t(204) = 2.4$, $p = .017$, $d = 0.35$, 95% CI: [0.06, 0.64], and "Mark had sex with Bill because he is attracted to men," $t(204) = 3.5$, $p < .001$, $d = 0.51$, 95% CI: [0.22, 0.84].

These data raise the possibility that the experimental manipulation did not affect participants' preferences for the agent-focused explanations directly; rather, learning that Mark reliably finds neither Bill nor other men attractive may cause participants to represent his encounter with Bill as less revealing of his true self, which may in turn make the agent-focused explanations seem less appropriate. To test this hypothesis, a bootstrap mediation analysis (Hayes, 2013) was conducted using an independent variable that was coded '1' if a participant was randomized into the Situation condition and '0' otherwise, essence-disclosure ratings as the mediator, and participants' mean ratings of the agent-focused explanations as the dependent variable (Fig. 2). The analysis indicated that the degree to which the agent's action appears to reflect his true self significantly mediated the effect of the Situation condition on participants' ratings of the agent-focused explanations: $ab = -.20$, 95% CI: [−.41, −.04].

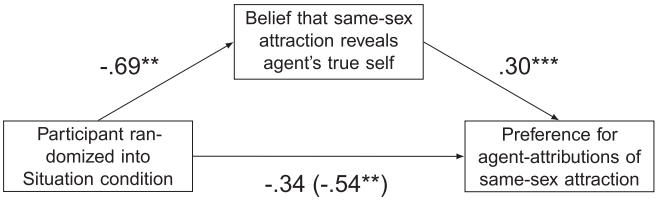

Fig. 2. Mediation (Study 2). $^*p < .05$; $^{***}p < .001$.

2.3. Replication with Indian participants

Two hundred and two Indian participants correctly responded to the attention and comprehension checks. Data from other Indian participants were excluded. The sample skewed strongly male, however, no significant differences emerged between male ($N = 159$) and female ($N = 43$) participants on any measure ($ps > .5$).

While the results found with North Americans replicated with Indian participants (Table 3), there were large differences between how the two groups explained the agent's action (Fig. 3). Compared to North Americans, Indian participants were significantly:

1. *more* inclined to favor situation explanations, $t(406) = 11.0$, $p < .001$, $d = 1.1$, 95% CI: [0.88, 1.30],
2. *less* inclined to favor person explanations, $t(406) = -7.1$, $p < .001$, $d = -0.70$, 95% CI: [−0.90, −0.50], and
3. *less* inclined to view the action as essence-disclosing, $t(406) = 2.4$, $p = .016$, $d = 0.24$, 95% CI: [0.04, 0.43].

To test whether moral attitudes towards homosexuality have a similar influence on Indian and North American participants' explanatory preferences, ATLG scores and dummy variables representing nationality and experimental condition were regressed against agent scores, situation scores, and essence-disclosure ratings in fully crossed models. Indian nationality did not have any effect ($ps > .5$). Similarly, there were no significant interactions between IVs ($ps > .5$). However, moral attitudes to homosexuality (as measured by ATLG scores) continued to predict participants' ratings of the agent-focused explanations, $\beta = .32$, 95% CI: [.23, .41], $p < .001$, the situation-focused explanations, $\beta = -.37$, 95% CI: [−.46, −.28], $p < .001$, and the essence-disclosure measure, $\beta = .27$, 95% CI: [.17, .36], $p < .001$, confirming all predictions derived from the mismatch hypothesis. This is noteworthy as Indian participants held strikingly more negative attitudes towards homosexuality than did North Americans, $d = -0.90$, 95% CI: [−1.10, −0.69], $p < .001$.

2.4. Discussion

This study found that both North American and Indian participants' moral attitudes towards homosexuality are highly predictive of their preferences for agent-focused vs. situation-focused explanations of an erotic encounter between two men. By contrast, beliefs about how the action covaried with changes in the actor's circumstances did not reliably predict explanatory preferences.

Perhaps unsurprisingly, people are less inclined to explain same-sex attraction in terms of the agent's sexual orientation when they are told that he does not reliably find other men attractive. However, it *is* surprising that this effect is so much smaller than the effect of people's moral attitudes towards homosexuality. *For predicting how someone will explain a same-sex encounter, it is far more useful to know about their prejudices than it is to know the degree to which they represent the agent's erotic feelings as causally covarying with his situation.* Moreover, data from Indian participants indicates that this finding is unlikely to reflect a uniquely WEIRD concept of human agency (Henrich, Heine, & Norenzayan, 2010). Instead, consistent with previous research (De Freitas et al., 2018), belief in the virtuous true self appears to reflect a species-typical feature of human psychology.

Table 3
Linear models from Indian sample (Study 2). Baseline omitted.

DV	Predictor	η_p^2	β	95% CI		t	p
Situation attribution[a]	Moral attitudes	.09	−.30	−.43,	−.16	−4.4	< .001
	Person condition	.00	.01	−.32,	.33	0.0	> .5
	Situation condition	.00	.11	−.22,	.45	0.7	> .5
Agent attribution[b]	Moral attitudes	.09	.30	.17,	.43	4.4	< .001
	Person condition	.02	.32	.01,	.64	1.9	.064
	Situation condition	.00	−.09	−.42,	.24	−0.5	> .5
Essence disclosure[c]	Moral attitudes	.04	.21	.19,	.45	4.8	.003
	Person condition	.00	.17	−.17,	.50	1.0	.335
	Situation condition	.00	−.09	−.43,	.25	−0.6	> .5

[a] $R^2 = .09$, $F(198,3) = 6.5$, $p < .001$. [b] $R^2 = .11$, $F(198,3) = 8.0$, $p < .001$. [c] $R^2 = .05$, $F(198,3) = 3.4$, $p = .02$.

3. Study 3: Generalizing

The case for the mismatch hypothesis would be considerably strengthened if the effect of moral attitudes on our explanatory preferences could be measured using a variety of target actions that, unlike sex between two men, do not reflect dispositions that we believe to be innate or chosen depending on our moral attitudes towards the actions (Haider-Markel & Joslyn, 2008). Studies 3a and 3b therefore attempted to replicate the basic mismatch effect using varied stimuli unrelated to sexual orientation.

3.1. Study 3a: Abortion

3.1.1. Participants

Participants were 204 North Americans ranging in age from 18 to 69 years ($M = 33$, 49% female).

3.1.2. Procedure

All participants read a vignette about a college senior, Kate, who discovers she is pregnant and later decides to have an abortion (see Appendix A for the complete vignette). The vignette made available various explanations for her decision (Table 4). Next, participants read that "Some of our actions reflect who we are deep down; they reveal our true selves," and rated the statement "Kate's decision to have an abortion reflects what she wants deep down." Participants then responded to two items aimed at measuring their beliefs about the causal sensitivity of Kate's decision to her circumstances:

> Weak robustness: If she were in the very same circumstances again, how probable is it that Kate would have another abortion?
> Strong robustness: Imagine that two years later Kate has another unwanted pregnancy. Now she lives in a different part of the country and is working as an intern at a hospital. How probable is it that she will also abort this second pregnancy?

Finally, participants completed a novel Moral Attitudes to Abortion (MATA) scale consisting of the following items (counterbalanced for order): Having an abortion is a disgusting thing to do; Abortion is just plain wrong; *Access to abortions should be free and easy; Abortion is only ever justified when a pregnancy seriously threatens the mother's life; Doctors who perform abortions are evil. All other materials were drawn without modification from previous experiments.

Table 4
Factor loadings from exploratory principal component analysis (Study 3a). Items are Likert-ratings of potential explanations for the agent's decision to abort her pregnancy.

Attribution:	Component 1	Component 2
"…because she believed it was not the right time for her to start a family"	**.88**	−.17
"…because she wanted to begin her career without having to care for a young child"	**.90**	.13
"…because her parents and friends peer pressured her into having an abortion"	−.20	**.82**
"…because her boyfriend broke up with her"	.16	**.83**

3.1.3. Results

One hundred and eighty-seven participants correctly responded to the comprehension checks. Responses from participants who failed more than one check were excluded, but this had no meaningful effect.

Factor analysis. The Kaiser-Meyer-Olkin measure of sampling adequacy was .50, slightly below the recommended threshold. However, the mean inter-item correlation was high, $r(185) = .77$, and Bartlett's test of sphericity was significant, $\chi^2(6) = 129$, $p < .001$. PCA with varimax rotation was therefore used to extract factor scores. Two components emerged that together explained 76% of the variance in participants' ratings of the four explanations (Table 4). Consistent with Studies 1 and 2, the two intuitively agent-focused explanations loaded strongly onto the first component which modeled 41% of the variance, and the two intuitively situation-focused items loaded strongly onto the second component which modeled the remaining 35%, indicating that participants drew the intuitive distinction, as intended.

Attribution and moral attitudes. The Moral Attitudes to Abortion (MATA) scale was highly reliable ($\alpha = .94$), so the items were summed to create an overall MATA score. MATA scores and robustness ratings were regressed against three dependent variables: situation scores, person scores, and essence-disclosure ratings (Table 5). In each model, moral attitudes significantly predicted the dependent variable. Weak

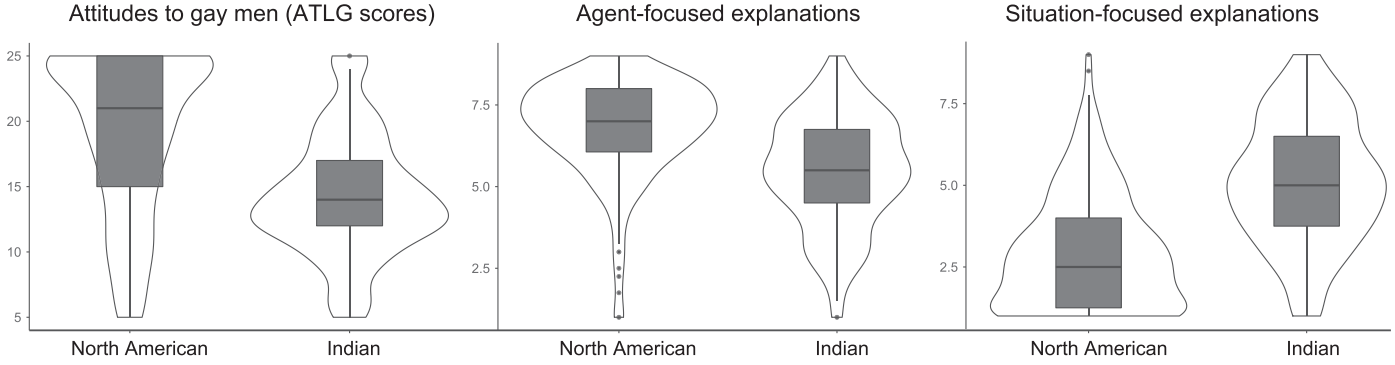

Fig. 3. North American and Indian participants (Study 2). Note: box plot boundaries are Q1, Q2, Q3, Q3 + 1.5 · IQR (Tukey-style).

robustness significantly predicted essence-disclosure ratings; however, strong robustness was not a significant predictor in any model.

3.2. Study 3b: Islamic conversion

3.2.1. Participants

Participants were 201 North Americans ranging in age from 18 to 55 years (M = 36, 45% female).

3.2.2. Procedure

Participants read a vignette about Jane, a girl born into a Christian family that eventually deteriorates, leading her parents to divorce. Jane's mother pushes her out of home, leading her to meet new friends at the local mosque: "Jane valued what she saw as their moral uprightness, and she perceived a kind of moral clarity in Islamic texts which she found reassuring. She came to believe that Islam is the one true path to God. Eventually, Jane decided she would convert to Islam." (See Appendix A for the complete vignette.) Participants then rated potential explanations for the agent's decision (Table 6). Next, participants rated the statement "Jane's decision to convert to Islam reflects her true self—who she is at the deepest level." Because religious conversion is an event that is unlikely to reoccur within a single person's life (Smith & Cooperman, 2015), Study 3b did not attempt to measure participants' beliefs about how the agent's decision would covary with features of her situation.

Participants then completed a novel Moral Attitudes to Muslims (MATM) scale consisting of the following items (counterbalanced for order): *Muslims are less intelligent; *Muslims are dirty; *Muslims are more likely to commit crimes; Most Muslims do NOT support violence against innocent people; I would feel comfortable being in close personal contact with a Muslim; Muslims are peaceful people.

All other methods were equivalent to Study 1.

3.2.3. Results

One hundred and eighty-three participants passed the comprehension checks. Responses from participants who failed at least one check were excluded, but this had no meaningful effect.

Factor analysis. Participants' ratings of the eight explanations were again suitable for PCA (Kaiser-Meyer-Olkin measure: .69; Bartlett's test of sphericity: $\chi^2(15) = 396$, $p < .001$). Consistent with previous experiments, factor loadings expressed the intuitive person/situation dichotomy (Table 6). The three intuitively situation-focused explanations loaded strongly onto the first component which modeled 46% of the variance, and the three intuitively agent-focused items loaded strongly onto the second component which modeled 24% of the variance, indicating that participants distinguished between the explanations in the expected way.

Table 5
Summary of regression models (Study 3a).

DV	Predictor	β	95% CI		t	p
Situation attribution[a]	Moral attitudes	−.44	−.57,	−.31	−6.6	< .001
	Weak robustness	−.14	−.27,	.00	−1.9	.065
	Strong robustness	.08	−.06,	.22	1.2	.245
Person attribution[b]	Moral attitudes	.30	.16,	.42	4.3	< .001
	Weak robustness	.23	.09,	.37	3.2	.002
	Strong robustness	.12	−.02,	.26	1.7	.104
Essence-disclosure[c]	Moral attitudes	.18	.05,	.31	2.7	.008
	Weak robustness	.39	.25,	.53	5.6	< .001
	Strong robustness	.06	−.08,	.19	0.8	.434

[a] $R^2 = .09$, $F(198,3) = 6.5$, $p < .001$. [b] $R^2 = .11$, $F(198,3) = 8.0$, $p < .001$. [c] $R^2 = .05$, $F(198,3) = 3.4$, $p = .026$.

Table 6
Factor loadings from exploratory principal component analysis (Study 3b). Items are Likert-ratings of potential explanations for an agent's decision to convert to Islam.

	Component	
Attribution:	1	2
"…because she admires Islam"	.06	**.70**
"…because she is a spiritual person"	−.11	**.84**
"…because of her deepest values"	−.43	**.78**
"…because her mother hurt her"	**.87**	−.10
"…because her peers pressured her"	**.75**	−.20
"…because she comes from a broken home"	**.90**	.01

Factor loadings with values > .5 are bolded.

Attribution and moral attitudes. The Moral Attitudes to Muslims (MATM) scale proved highly reliable ($\alpha = .93$), so responses to the individual items were summed to form overall MATM scores. Pearson correlations were calculated for MATM scores and explanatory preferences. Participants' MATM scores predicted both person ratings, $r = .33$, 95% CI: [.19, .47], $t(181) = 4.6$, $p < .001$, and situation ratings, $r = -.48$, 95% CI: [−.60, .−35], $t(181) = -7.6$, $p < .001$.

Replicating previous results, MATM scores strongly predicted the degree to which participants viewed the agent's decision as expressing her true self, $r = .40$, 95% CI: [.27, .54], $t(181) = 5.8$, $p < .001$, suggesting that beliefs about the true self may mediate the effect of moral attitudes on explanatory preferences. To test this hypothesis, a bootstrap mediation analysis (Hayes, 2013) was performed using MATM scores as the independent variable, essence-disclosure ratings as the mediator, and person scores as the dependent variable (Fig. 4). Consistent with the mediation hypothesis, the analysis revealed a significant indirect effect: $ab = .21$, 95% CI: [.12, .33]. Notably, the direct effect of moral attitudes on person attribution was relatively small when true-self ratings were included in the regression model, $\beta = .12$, 95% CI: [−.01, .25], $t(181) = 1.8$, $p = .081$.

3.2.4. Discussion

Consistent with the mismatch hypothesis, Studies 3a and 3b revealed powerful associations between participants' moral attitudes and their explanatory preferences. Thus, because these studies employed stimuli unrelated to sexual orientation, the results of Studies 1 and 2 are unlikely to reflect a distinctive bias associated with anti-gay prejudice. In particular, they are unlikely to reflect any systematic relationship between moral attitudes towards homosexuality and beliefs about whether same-sex attraction arises from choice or innate disposition. Rather, the association of moral attitudes both with explanatory style and with the degree to which actions appear expressive of actors' true selves appears to be surprisingly general.

4. Study 4: Good deeds of passion

Studies 1–3 indicate that people's explanatory preferences and beliefs about essence-disclosure are related. One hypothesis consistent with these results is that beliefs about whether an action expresses the agent's true self mediate the effect of participants' moral attitudes on their explanatory preferences. That is, people may treat a perceived mismatch in the moral valences of the action and the agent's true self as evidence that the action did not express the true self, and this may in turn incline them towards explanations of the action which do not implicate the true self. Such explanations tend to highlight features of the agent's environment or upbringing.

Because people tend to represent the true self as virtuous (Newman et al., 2013), this hypothesis predicts that people may favor a situation-focused explanation when explaining a bad deed but ignore that same explanation when explaining a virtuous one (cf., Newman et al., 2014; Pizarro, Uhlmann, & Salovey, 2003). We can return to Darley and

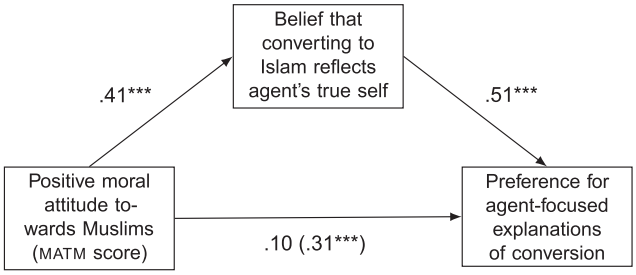

Fig. 4. Mediation (Study 3b). Note: only the indirect effect was significant. ***p < .001.

Batson's (1973) study to illustrate the asymmetry. Recall that Darley and Batson found that unhurried seminarians were six times more likely than hurried seminarians to help a stranger slumped in a doorway. As noted, when people focus on hurried seminarians who ignore the victim, they judge that the callous behavior is explained by the experimental manipulation. However, this situation-focused explanation can seem much less attractive when we consider the helpful participants in the Relaxed condition; it may seem that whereas the Hurried condition masked participants' true selves, the Relaxed condition allowed them to shine through.

So here we seem to have two agents that are causally sensitive to the environment in precisely the same way, yet a situation-focused explanation ("he acted because of the experimental condition into which he was randomized") does not seem appropriate for the unhurried seminarian who helped the victim. If our beliefs about whether actions are essence-disclosing mediate the effect of moral valence on our preferences for agent- or situation-focused explanations, this pattern is exactly what we should expect to find. Because we tend to represent agents as 'deep down' normatively aligned with ourselves, when the hurried seminarians ignore the victim, their actions seem less essence-disclosing than when the relaxed seminarians help the victim. This may be what leads us to favor situation-focused explanations in one case but not the other. Study 4 explored this hypothesis by investigating the conditions that incline people to attribute an action to luck.

4.1. Methods

4.1.1. Participants

Participants were 502 North Americans ranging in age from 18 to 70 years (M = 32, 44% female).

4.1.2. Procedure

The experiment employed a 2 (valence: good vs. bad) × 2 (moral luck: emphasized vs. not emphasized) × 2 (DV: responsibility vs. essence-disclosure) between-subjects design. Participants read one of four vignettes about a person born in the United States in the early 1800s who does something immoral (owns and mistreats slaves) or virtuous (helps slaves escape). In two conditions, the vignettes emphasized the role of chance in the agent's life:

> Chance played a big role in Tom's life. He was born in the Northern [Southern] United States in the early 1800s, but as a baby he was adopted by Southern plantation owners [Northern abolitionists] who raised him in the South [North]. If he had been raised by his biological parents, he would have grown up in the North [South], and he would have led a morally better [worse] life. But as a matter of fact, Tom himself went on to own slaves [work on the Underground Railroad to help people escape from slavery], and in this way he hurt [helped] many people over the course of his life.

The control conditions simply omitted the features expected to focus participants on the role of luck in the agents' lives:

> Tom was born in the Southern [Northern] United States in the early 1800s. Tom owned many slaves [worked on the Underground Railroad to help people escape from slavery] and in this way he hurt [helped] many people over the course of his life.

To ensure that any similarities in participants' responsibility and essence-disclosure ratings would not be an artifact of the survey design, participants were randomly assigned to rate either the agent's degree of responsibility or the degree to which his action was essence-disclosing:

> Responsibility: "How negatively [positively] does Tom deserve to be judged?" (1 = Not at all negatively [positively], 7 = Extremely negatively [positively].) "How much blame [praise] does Tom deserve?" (1 = No blame [praise] at all, 7 = Extreme blame [praise]).
> Essence-disclosure: "Helping [harming] people did not reflect Tom's true self—the person he really is deep down" (1 = strongly disagree, 9 = strongly agree).

The two responsibility measures were counterbalanced for order and the essence-disclosure measure was reverse-coded. All other methods were consistent with previous experiments.

This experiment required a measure of the degree to which participants attribute an agent's actions to luck that did not itself make the role luck plays in all people's lives salient. For example, from one's own perspective, the place of one's birth is entirely a matter of luck. Thus, if participants had been asked "Did the agent own and mistreat slaves [work on the Underground Railroad] *because* he was raised in the South [North]?" this would effectively have made the situation-focused explanation salient in *both* the Luck and Control conditions. The connection between luck and moral responsibility suggests that we can use participants' responsibility attributions to gauge the degree to which they explain an action in terms of situational luck (Levy, 2011; Nagel, 1979; Williams, 1981). This approach seems unlikely to make luck-based explanations salient to participants in the Control conditions.

4.2. Results

Four hundred and seventy-six participants correctly responded to the attention check; participants who failed were excluded.

4.2.1. Essence-disclosure

Essence-disclosure ratings were regressed against luck and valence in a fully crossed design. The resulting model, presented in Table 7a, was significant, R^2 = .20, $F(3,230)$ = 20, p < .001. Main effects emerged for both luck and valence (Fig. 5) such that participants rated working on the Underground Railroad as much more essence-disclosing than owning and mistreating slaves, d = 0.85, 95% CI: [0.58, 1.11]. Critically, a significant luck × valence interaction emerged. When the agent was described as owning and mistreating slaves, participants who read that he had been adopted by Southerners gave significantly lower essence-disclosure ratings than participants who did not read this information. Thus, emphasizing the role of luck in the agent's life reduced the degree to which participants rated his highly immoral behavior as expressive of his true self, d = −0.73, 95% CI: [−1.11, −0.34]. However, when the agent was described as helping people to escape from slavery, emphasizing the role that luck played in his life did not lead participants to rate his action as less essence-disclosing, d = 0.08.

4.2.2. Responsibility

Responses to the two responsibility items were highly consistent (α = .91), so they were averaged to create a composite responsibility score. The analysis was repeated with responsibility scores as the dependent variable. The resulting model (Table 7b) was significant, R^2 = .28, $F(3,238)$ = 31, p < .001. Main effects again emerged for both luck and valence such that participants more strongly praised the Underground Railroad worker than they blamed the slave owner, d = 1.1, 95% CI: [0.79, 1.33]. Mirroring the patterning of essence-

Table 7
Linear models predicting essence-disclosure and responsibility ratings.

(a) Essence-disclosure

Predictor	*B*	*SE*	95% CI		*t*	*p*
Intercept	6.4	.14	6.2	6.7	46.2	< .001
Valence	1.6	.28	1.1	2.2	5.8	< .001
Luck	−0.68	.28	−1.2	−.13	−2.4	.016
Valence × Luck	1.7	.56	.59	2.8	3.0	.003

(b) Responsibility

Predictor	*B*	*SE*	95% CI		*t*	*p*
Intercept	5.5	.08	5.4	5.6	70.2	< .001
Valence	1.3	.16	1.0	1.6	8.5	< .001
Luck	−0.54	.16	−.84	−.23	−3.4	< .001
Valence × Luck	0.78	.31	.18	1.4	2.5	.012

disclosure ratings, the effect of luck × valence was also significant. When the agent was described as owning and mistreating slaves, participants who read that he had been adopted by Southerners gave significantly lower responsibility ratings than participants who were not provided with information about luck. Thus, emphasizing the role of luck in the agent's life also reduced the degree to which participants rated him as responsible for immoral deeds, $d = -0.61$, 95% CI: $[-0.97, -0.25]$, but had no meaningful effect on how praiseworthy he appeared for virtuous deeds.

4.3. Discussion

The substantial reduction in blame ratings in the Bad Luck condition indicates that participants in that condition were more likely to explain the Northern-born agent's immoral deeds in terms of his unlucky adoption by Southern plantation owners. However, because we see no reduction in praise ratings in the Good Luck condition, it seems participants did not similarly explain the Southern-born agent's virtuous deeds in terms of his fortunate adoption by Northern abolitionists. Thus, participants were more inclined towards a situation-focused explanation for an immoral action than for a virtuous action, even when the same explanation was made salient in the very same way.

The close symmetry of responses to the essence-disclosure and responsibility items is especially noteworthy as different participants responded to each measure. The hypothesis that beliefs about essence-disclosure mediate the effect of moral attitudes on explanatory preferences provides a natural explanation for this association: because we implicitly assume that the true self is good, immoral deeds appear less essence-disclosing than virtuous deeds. This suggests that the asymmetry in whether we perceive an agent's actions as essence-disclosing leads to an asymmetry in how we explain the agent's action: situation-focused explanations appear more appropriate when we explain why a person owned and mistreated slaves than they do when we explain why a person helped slaves escape. In turn, this explanatory asymmetry leads us to less strongly blame the slave-owning agent, but does not lead us to view the slave-helping agent as less praiseworthy because his actions do not challenge the presumption that his virtuous deeds expressed his true self.

This account is consistent with the mediation results from Studies 1–3 and provides further independent support for the hypothesis that beliefs about essence-disclosure mediate the effect of mismatching valences on explanatory preferences. However, while it seems very likely that working on the Underground Railroad is widely represented as morally better than owning human beings, it is plausible that it is also represented as having been rarer. Thus, participants are likely to know that in the antebellum South, for the purpose of categorization, owning slaves was less informative than working on the Underground Railroad. It is possible that this difference in the diagnostic value of the actions might interact with the presence or absence of luck. An ideal concluding study, therefore, would show that beliefs about essence-disclosure mediate the effect of moral attitudes on explanatory preferences in a context where the target actions are widely known to be equally diagnostic. Few actions split participants into two large and easily identifiable groups who attach equal-and-opposite moral valences to actions known to be equally common; however, recent research indicates that there are at least two: voting Democratic and voting Republican. Thus,

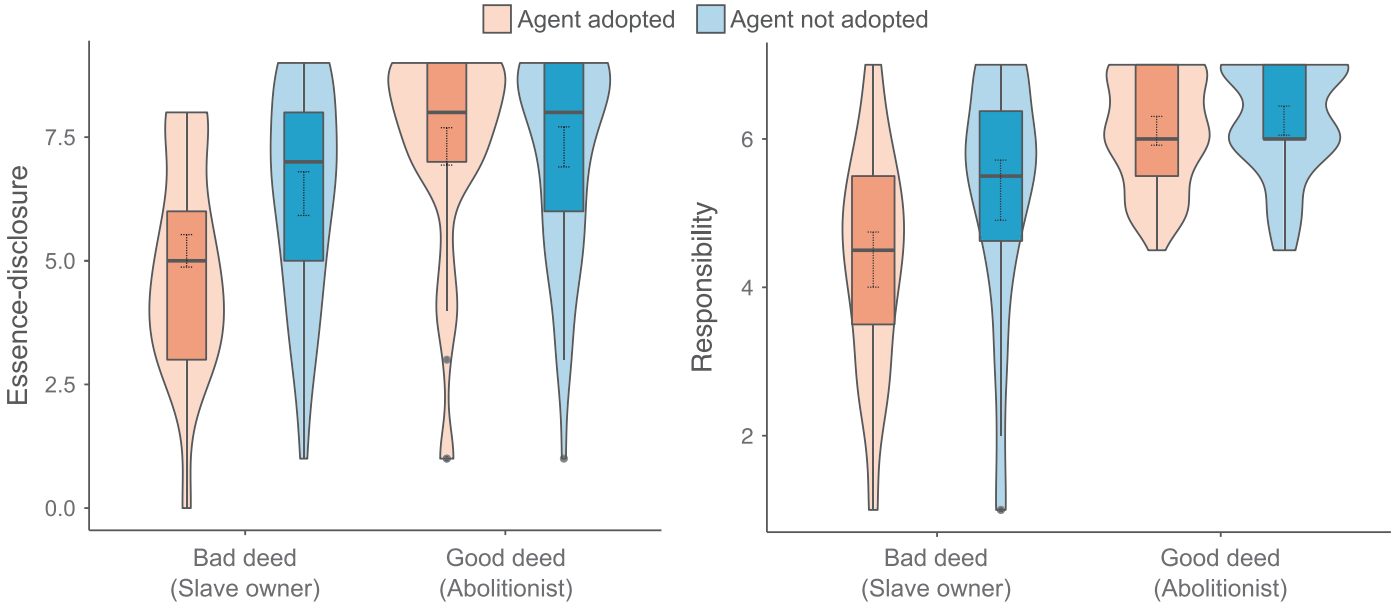

Fig. 5. The effect of luck and valence on ratings of essence-disclosure and responsibility (Study 4). Note: different participants rated the essence-disclosure and responsibility items. Dashed vertical-lines within boxes display 95% CIs around the means.

this paper's final study explores how partisans explain the political behavior of rivals as compared to fellow partisans.

5. Study 5: Partisan attribution

In discussing his conversion to Islam, Kareem Abdul-Jabbar commented that for people born into their religion, "it is mostly a matter of legacy and convenience." For converts, however, "it is a matter of fierce conviction and defiance … because we need a powerful reason to abandon the traditions of our families." This is understandable since religious disagreements within pious families are often costly (Boyatzis, Dollahite, & Marks, 2006); a fact which helps to explain why they are relatively rare (Kelley & Graaf, 1997). Because partisan conflict within families is also costly and rare, parallel considerations apply to political identities (Crawford & Pilanski, 2014; Jennings, Stoker, & Bowers, 2009; Kohn, Slomczynski, & Schoenbach, 1986). For example, we learn much more about a Liberal agent's values if we learn she was raised by Conservatives rather than by fellow Liberals.

The significant diagnostic value of a family-dissonant political identity suggests that people *should* prefer to explain such identities in agent-focused terms. However, against a backdrop of intense affective polarization (Clifford, 2017; Huddy, Mason, & Aaroe, 2015; Iyengar, Sood, & Lelkes, 2012; Iyengar & Westwood, 2015), the mismatch hypothesis predicts that moral attitudes will play an important role in modelling how people explain the actions of allies and rivals. For example, agent-focused attributions should seem *less* appropriate when explaining why someone defected for the rival party, despite the high diagnostic value of that action.

Researchers studying the moral-psychological bases of political ideology have reliably found that Conservatives tend to value ingroup loyalty and respect for authority more than Liberals (e.g., Feldman, 2003; Graham, Nosek, & Haidt, 2012; Gunther & Kuan, 2007; Haidt & Graham, 2007; Kohn, 1989; Schwartz, 2006; Schwartz, Caprara, & Vecchione, 2010). This suggests that Conservative participants may perceive a more serious moral violation when an agent is disloyal to his family. If so, the degree to which this highly diagnostic action is attributed to the agent himself may additionally depend on participants' own political identities. Study 5 therefore aimed to explore three questions: How do partisans' moral attitudes influence the explanations they provide for out-group political identities, and how strong is this effect relative to the effect of a highly diagnostic action such as converting to a family-dissonant ideology? Are Conservatives more reluctant than Liberals to accept agent-focused explanations for family-dissonant identities? Are these effects mediated by the degree to which an agent's political identity appears to express her true self?

5.1. Methods

5.1.1. Participants

Participants were 670 North Americans ranging in age from 18 to 74 years (M = 37; 59% female).

5.1.2. Procedure

The experiment employed a 2 (family ideology: liberal vs. conservative) × 2 (agent ideology) between-subjects design. Participants were asked to either imagine an agent who grew up in a conservative family or to imagine one who grew up in a liberal family. The agent was also described as either liberal or conservative, generating four experimental conditions.

In the *agent-family consonant* conditions, when agent and family shared a common political identity, participants read:

> Imagine a person, Sam, who grew up in a politically conservative [politically liberal] family. Like his parents, Sam often voted for Conservative [Liberal] candidates.

In the *agent-family dissonant* conditions, when agent and family were ideologically opposed, participants read:

> Imagine a person, Sam, who grew up in a politically conservative [politically liberal] family. However, unlike his parents, Sam often voted for Liberal [Conservative] candidates.

In all conditions, participants rated the reverse-coded statement "Voting liberal [conservative] did not reflect Sam's true self—the person he truly is deep down" on a scale ranging from 1 ('Strongly disagree') to 7 ('Strongly agree').

Because explanations typically seem more or less appropriate for different explananda, participants rated different attributions in the family-dissonant and family-consonant conditions. In the family-consonant conditions, participants rated a paradigm situation-focused attribution: "Sam voted conservative [liberal] because that's how he was raised." In the family-dissonant conditions, participants rated a paradigm agent-focused explanation: "Sam voted conservative [liberal] because of his most cherished values." The essence-disclosure and attribution items were counterbalanced and followed by an attention check.

On the next page (where the vignette was hidden), participants reported the ideologies of the agent and his family. Participants identified as Liberal, Conservative, or 'other'. Selecting 'other' prompted participants to enter their identification in their own words. Finally, participants provided basic demographic information.

5.2. Results

Five hundred and eighty participants passed the attention checks; 295 identified as Liberal and 193 as Conservative. Because our primary interest is in partisans, 'others' were not included here.[1] Participants who failed to correctly identify the political identity of the agent or his family were excluded.

5.2.1. Essence-disclosure

Dummy variables for agent, family, and participant identities were coded '1' for Conservative and '0' for Liberal. A dummy variable, *Consonance*, was coded '1' iff the agent and family supported the same ideology. To explore the effects of consonance and perceived valence side by side, essence-disclosure ratings were regressed against participant identity, agent identity, and agent-family consonance in a fully crossed model. The resulting model (Table 8) was significant, R^2 = .085, MSE = 2.2, $F(7,479)$ = 6.3, $p < .001$. As expected, relative to when the agent conformed with his parents' favored ideology, his political identity was seen as more essence-disclosing when he defected to a family-dissonant ideology (0.64 points). Critically, however, sharing an ideology with the participant also had a large, positive effect (0.92 points), indicating that perceived valence influenced essence-disclosure ratings across conditions.

An unexpected main effect of participant identity emerged such that Conservatives gave slightly lower essence-disclosure ratings.

5.2.2. Attribution

Moderation analysis was used to explore the participant-agent interaction. Participant identity was set as the focal predictor with agent identity and agent-family consonance as moderators. The resulting model was significant, R^2 = .10, MSE = 1.9, $F(7,479)$ = 7.4, $p < .001$. As Table 9a shows, sharing the agent's ideology led participants to emphasize the relevance of his values in the dissonant cases (+1.3 points) and to downplay the relevance of his upbringing in the consonant cases (−1.2 points) (Fig. 6). Table 9b describes in more detail how participants' ratings were conditioned by agent and family

[1] When data from Independents was analyzed separately, no significant effects emerged on either measure.

Table 8
Linear model predicting essence-disclosure ratings from participant ideology, agent ideology, and agent-family consonance (Study 5). Note: Reference categories are Liberal and dissonant.

Predictor	*B*	*SE*	*95% CIs*		*t*	*p*
Intercept	5.1	.075	5.0	5.3	74	< .001
Participant	−0.35	.14	−.62	−.085	−2.6	.010
Agent	0.04	.14	−.23	.31	0.31	> .5
Consonance	−0.64	.14	−.91	−.37	−4.7	< .001
Participant × Agent	0.92	.27	.38	1.5	3.4	< .001
Participant × Consonance	0.00	.27	−.54	.54	−.025	> .5
Agent × Consonance	0.18	.27	−.36	.72	0.64	> .5
Participant × Agent × Conso.	−0.82	.55	−1.9	.26	−1.5	.138

identity. When the agent defected to Liberalism, Conservative participants were highly reluctant to attribute his political identity to his values (−1.2 points). However, when the agent defected to Conservatism, both Liberal and Conservative participants rated his values as equally important to explaining his politics (0.021 points). Participants' responses were more evenly biased when the agent conformed with his family's favored ideology. Relative to Liberals, Conservatives emphasized the agent's upbringing when explaining Liberal conformism (0.54 points) and downplayed the agent's upbringing when explaining Conservative conformism (−0.65 points).

5.3. Conditional process analysis

The analyses above suggest that participant identity may moderate the effect of agent-family consonance *by* moderating the degree to which the agent's political identity appears to express his true self (Fig. 7). Conditional process analysis (Hayes, 2013) was used to explore this hypothesis. Participant identity was set as the focal predictor, with agent identity and agent-family consonance as moderators and essence-disclosure ratings as mediator.

The analysis indicated that the conditional effect of agent-family consonance was mediated by essence-disclosure ratings, $B = 0.12$, $SE = .11$, 95% CI (bootstrapped): [−.051, .43]. In a subsequent exploratory analysis, the size of this effect more than doubled when outliers were excluded, $B = 0.31$, $SE = .15$, 95% CI (bootstrapped): [.085, .69]. (Tukey's convention of $\pm 1.5 \times$ IQR classified 13 data points as outliers.) When the analysis focused on agent-family dissonant cases, the difference between the conditional indirect effects was large, indicating that the effect of valence on how participants explained a family-dissonant identity was substantially mediated by the degree to which the agent's identity appeared to reflect his true self, $B = 0.64$, $SE = .21$, 95% CI (bootstrapped): [.30, 1.1].

Table 9
(a) Effect of participant-agent interaction on ratings of each attribution (Study 5). Note: the attribution rated was agent-focused (values) in the agent-family consonant cases and situation-focused (upbringing) in the agent-family dissonant cases.

Agent-family relationship	*B*	*SE*	*95% CIs*		*t*	*p*
Dissonant	1.3	.37	.55	2.0	3.5	< .001
Consonant	−1.2	.36	−1.9	−.48	−3.3	.001

(b) Conditional effect of participant-identity.

Agent	Family	*B*	*SE*	*95% CIs*		*t*	*p*
Liberal	Conservative	−1.2	.27	−1.8	−.74	−4.7	< .001
Conservative	Liberal	0.02	.26	−.48	.53	.083	> .5
Liberal	Liberal	0.54	.26	.03	1.1	2.1	.037
Conservative	Conservative	−0.65	.25	−1.2	−.15	−2.6	.011

5.4. Discussion

This study found that partisans' explanatory preferences do not consistently reflect the insight that a family-dissonant identity provides more information about a person's values than the 'legacy and convenience' of a family-consonant identity. Rather, our explanatory preferences reflect the diagnostic value of defection only when we evaluate agents who defect *to* our own political ideology. If an agent instead defects *from* our ideology, the effect of his defiance is totally swamped by the effect of our disapprobation. In practice, this means that we end up seeing his values as not more relevant to explaining his defection, but less so. Indeed, our reactions are so biased that when an in-group member abandons our ideology, his values appear even less relevant to explaining his action than legacy and convenience appear to explaining out-group political conformity.

While Conservatives were particularly inclined to discount the importance of an agent's values when he rejected his parents' ideology, Liberals were reluctant to acknowledge the explanatory relevance of a Liberal agent's upbringing to his family-consonant identity. If political partisans represent both in-group and out-group members as essentially virtuous, we may be able to understand these differences as a reflection of the different values that Liberals and Conservatives tend to hold (e.g., Feldman, 2003; Graham et al., 2012; Gunther & Kuan, 2007; Haidt & Graham, 2007; Kohn, 1989; Schwartz, 2006; Schwartz et al., 2010). From the perspective of predominantly Conservative values, a family-dissonant Liberal identity is doubly bad: first, because it is Liberal, and second, because the agent defied authority and was disrespectful to his parents. But from the perspective of Liberal values, a family-dissonant Liberal identity is doubly good: first, because it is Liberal, and second, because the agent defied authority and was disrespectful to his parents. Hence, conditional on the mismatch hypothesis, values strongly associated with partisan identities appear to be consistent with the pattern of attributions that we see in the data.

These results suggest one way that belief in the virtuous true self may contribute to affective polarization. The more convinced we are of the virtuousness of an out-group member's true self, the more difficult it will be to recognize that her values are at work when she makes choices that are, from our perspective, morally wrong. Because we doubt that she is moved to action by her values, it may seem she is acting 'in bad faith' when she is really pursuing her deepest commitments. Indeed, belief in the good true self may help us to represent our rivals' projects as wrong not only from our own perspective, but also from theirs. If the way our minds represent agents biases us towards explaining moral dissent in non-agential terms, then perhaps this may lead coercion to seem easier to justify than it should. (This paper concludes with a brief discussion of other reasons to regret the virtuous true self bias.)

6. General discussion

Collectively, the studies presented here point to a large and robust effect of our moral attitudes and true-self beliefs on how we explain valenced behavior. This effect appears to arise from a general tendency to represent agents as possessing true selves that are, to a surprising degree, aligned with our own values (Newman et al., 2013; Strohminger et al., 2017). Because of this bias, actions of which we approve tend to seem more expressive of our own and others' true selves, and this causes us to prefer more agent-focused attributions.

6.1. Responsibility

6.1.1. Luck

Philosophers have long noticed that reflecting on the role that luck plays in human life tends to diminish our sense that we are morally responsible agents (e.g., Levy, 2011; Nagel, 1979; Williams, 1981). For example, but for the terrible misfortune of being born in Germany after the First World War, a boy who would in fact go on to become a German

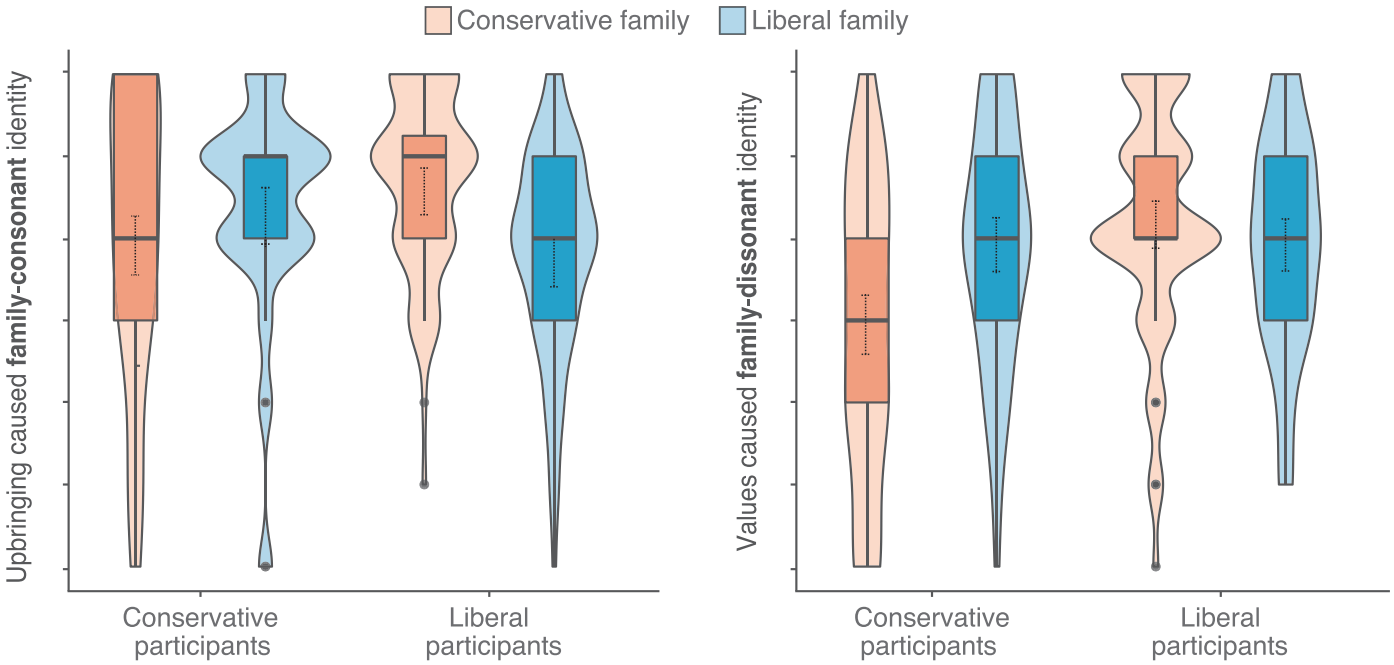

Fig. 6. Ratings of two attributions: the agent supported a family-consonant candidate "because that's how he was raised" (left), and the agent supported a family-dissonant candidate "because of his most cherished values" (right).

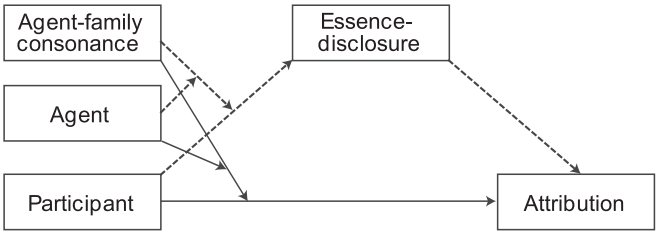

Fig. 7. Conceptual diagram for mediated conditional moderation analysis (Study 5). Lines represent causal pathways and line-arrow intersections represent moderation of one effect by another. Dashed lines show the indirect conditional effect of perceived valence on attribution.

soldier might instead have gone on to lead a morally innocuous life (perhaps in Australia). Reflecting on this fact and its terrifying dual might shake our sense that soldiers of the past were fully responsible.

Theorists have been tempted to explain the apparent blame-mitigating power of luck by positing 'control principles' which hold that agents are only considered responsible to the degree that their actions seem to be under their control (e.g., Nagel, 1979). Applied to the vignettes explored above, these principles suggest that we blame the adopted Northern-born agent less because we recognize that his actions are, to some significant degree, caused by the extreme misfortune of his being adopted by plantation-owning Southerners—something that was never under his control. However, the present results suggest that this plausible-sounding analysis is in fact incorrect. It predicts that learning about the parallel role that luck played in the Southern-born agent's coming to help people will reduce our sense that he is praiseworthy for his virtuous deeds, but this prediction was not supported by the data.

Theorists have often attempted to analyze the concept of luck in 'modal' terms—i.e., in terms of how events and actions causally covary within contextually specified sets of initial conditions, frequently modeled as possible worlds (e.g., Coffman, 2007; Levy, 2011; Pritchard, 2006). These analyses are all based on the intuitive idea that an event or action is lucky to the degree that it could easily have failed to occur. For example, on these views, winning the New Jersey State Lottery counts as enormously lucky because of how easily the winner *could* have ended up like any of an enormous number of losers, even holding fixed all (or most) of the conditions that in fact led to a windfall. In popular philosophical terminology: an event is influenced by luck to the degree that it does not occur in sufficiently many nearby possible worlds in which the relevant initial conditions obtain. Theorists disagree about how to elaborate this intuition into a satisfactory account of moral luck. However, reminiscent of classical accounts of the person/situation distinction, many philosophers have attempted to understand luck primarily in terms of the nature of the causal relations that condition outcomes on contingent features of the environment. The present studies indicate that ordinary people rely on a fundamentally different concept of luck, at least when they attempt to understand and explain morally valenced actions. According to this folk concept, the moral valences of an agent's true self and action are relevant to whether she was the beneficiary (or victim) of luck.

6.1.2. Determinism

Reflecting on the (putatively) deterministic nature of our universe affects people's responsibility judgments quite differently depending on whether the target action is represented as virtuous or immoral (Nelkin, 2011; Pizarro et al., 2003; Wolf, 1980). In particular, determinism appears to undermine blame far more readily than praise. (In this context, 'determinism' is the claim that the laws of nature and the state of the universe in the distant past jointly entail a unique future.)

Some theorists hypothesize that this determinism-valence asymmetry arises because people tacitly shift between two distinct conceptions of responsibility (e.g., Watson, 1996). On this view, when we judge that a determined agent is responsible for her virtuous deeds, we make a judgment about responsibility in the sense of 'attributability'—i.e., that the action is essence-disclosing. However, when we judge that an agent is less blameworthy because we believe her immoral deeds to be causally determined, we make a judgment about responsibility in the sense of 'accountability'—i.e., that the agent is less deserving of contempt or punishment on account of her action. This explanation is designed to preserve the idea that from within each perspective there is no determinism-valence asymmetry: causally determined agents can be responsible in the sense of *attributability* for both virtuous and immoral deeds; but in the sense of *accountability* they can be responsible for neither. The findings reported here tell against this ambiguity-based explanation of the asymmetry. People's judgments

about accountability–"How negatively or positively does [the agent] deserve to be treated?"—display the same asymmetry as their judgments about essence-disclosure (Fig. 5). As noted above, this is powerful evidence against ambiguity-based explanations, as different participants responded to the two outcome measures.

Another common explanation for the determinism asymmetry begins with the idea that responsible agents must possess the ability to 'do the right thing for the right reasons' (e.g., Nelkin, 2005; Wolf, 1980). On this account, when we assess whether an agent is responsible for ϕ-ing, we first assess whether there was most reason for the agent to ϕ. If we judge that there was *not*, we will only hold the agent responsible to the degree that we believe he *could* have done other than ϕ. (Since we do not have most reason to perform immoral deeds, if the agent was able at the time of action to do the morally right thing, then he must have been able to do something other than ϕ.) On the other hand, if we believe that the agent was ϕ-ing for the very reason in virtue of which it is right to ϕ, then we will *not* be interested in whether he could instead have done some other, immoral deed. After all, a cool-headed mother who can restrain herself while her children are helplessly trapped in a house fire does *not* seem, intuitively, more praiseworthy than one whose love renders her unable to resist a perilous rescue attempt—even if the cool-headed mother effortlessly wills herself to behave like the loving mother (Wolf, 1980).

Recent experimental research may appear to support this explanation of the determinism-valence asymmetry (e.g., Pizarro et al., 2003). However, because the asymmetry also appears to arise when we consider cases, like those in Study 4, where blameworthy agents are (presumably) able to do otherwise, the present studies suggest that this explanation is incorrect (also see Newman et al., 2014). Remarkably, the same asymmetry seen in our judgments about causally determined agents appears to arise even when there is no suggestion that the agents were overcome by irresistible emotions, that they are the denizens of deterministic universes, or that they lacked the ability to do the right thing for the right reasons. Hence, it seems implausible that the asymmetry arises because the folk theory of responsibility only requires that agents can do otherwise when they behave immorally.

By contrast, the determinism asymmetry makes sense if our explanatory preferences behave as the mismatch hypothesis describes. For, when we perceive a mismatch in the moral valences of the action and the agent's true self, we will tend to represent the action as concealing the true self, if there are suitable situation-focused explanations available. The mismatch hypothesis therefore suggests that the role of causal determinism is to make available a powerful, situation-focused explanation for any action at all—*the laws of nature and the state of the universe in the distant past made me do it.* If true, this helps to explain why we sometimes judge that an agent who behaves badly by ϕ-ing in a deterministic universe is less blameworthy than an otherwise-similar agent who ϕs in an indeterministic universe (Feltz & Millan, 2015; Nichols & Knobe, 2007; but cf., Murray & Nahmias, 2014).

6.2. The ego-syntonic/ego-dystonic distinction

In their important commentary on the actor-observer literature, Sabini et al. (2001) argued that the intuitive person/situation distinction is between *ego-syntonic* and *ego-dystonic* actions, i.e., between actions agents endorse reflectively and those they do not. The classic example is an 'unwilling addict' who may desire a substance while also hoping that this desire fails to lead her to use the substance (Frankfurt, 1971). Her occurrent desires seem not to reflect what she would want were she cool, calm, and collected. To the degree that this is so, her drug use is ego-dystonic. This understanding of the distinction parallels popular compatibilist theorizing and generates many of the same verdicts as the mismatch hypothesis. However, the ego-syntonic/ego-dystonic dichotomy cannot capture the full range of intuitions revealed by the present studies. For example, while most people in Studies 1–2 agreed that the agent's sexual attraction to men was not something he endorsed reflectively (i.e., that it was ego-dystonic), people with different attitudes towards homosexuality differed greatly in how they explained the agent's homosexual behavior. If we analyze the distinction in terms of the dichotomy between ego-syntonic and ego-dystonic actions, this variance must remain unexplained.

Sabini et al.'s insight is to understand situation-focused explanations in terms of causes that are "external not to the person, but to the person's self." However, to capture intuitions about cases like the ones studied here, the self must be understood as the true self, and the concept of the true self is not exhausted by what a person reflectively affirms—or indeed, by any other naturalistic, non-evaluative feature of his psychology. Rather, when explaining morally valenced actions, we will discount an agent's own affirmations when we disapprove of his reflectively endorsed attitudes. For example, although Mark's erotic feelings appear to be ego-dystonic (unlike his religiously motivated beliefs),[2] this fact does not prevent people with positive moral attitudes towards homosexuality from seeing his same-sex encounters as essence-disclosing. The folk concept of the true self allows that we can be mistaken in what we reflectively endorse: our true selves are something we must discover (Bench, Schlegel, Davis, & Vess, 2015). Thus, analyzing the person/situation distinction in terms of reflective endorsement appears to treat as constitutive what is really heuristic.

6.3. Strategic benefits of belief in the virtuous true self

Why do our minds represent agents as divided into true and superficial selves, and why do we tend to assume the true self is aligned with our own values, even when the agent belongs to a stigmatized outgroup? One approach to these questions (e.g., De Freitas et al., 2017; Newman et al., 2014; Strohminger & Nichols, 2014) begins with the idea that the true-self concept reflects a more general tendency of human minds to represent the surface features of animals (among other categories) as caused by hidden, underlying essences (Gelman, 2003). This hypothesis—*psychological essentialism*—is plausibly relevant to explaining why people represent the self as divided, but it seems less satisfying as an explanation for why members of stigmatized groups are represented as 'deep down' normatively aligned with the self. However, this surprising finding has emerged with participants from both independent and interdependent cultures (De Freitas et al., 2018, and Study 2 of the present work), suggesting that belief in the virtuous true self may be a species-typical feature of human psychology. Thus, it is worth considering (however speculatively) whether it may have helped to solve difficult sociobiological problems that our evolutionary ancestors would have faced repeatedly (Williams, 2008).

Consider someone who believes that gay sex is a grotesque moral wrong, for example, a high-prejudice participant from Study 1. This participant will surely have a powerful moral reaction to gay sex, yet this need not color her representation of the agent's true self (cf., De Freitas & Cikara, 2018). Indeed, even when participants were told that the agent *regularly* experiences erotic attraction to a *variety* of other men (in Study 2), the most highly prejudiced participants continued to explain the agent's same-sex encounter in non-agential terms (a bad upbringing, traumatic sexual experiences, stress-induced weakness, and so on). This illustrates one of the strange consequences of representing the self as divided: at least according to folk psychology, it seems you don't have to be good to be good deep down.

Belief in the virtuous true self may therefore have been adaptive because it allowed people's responses to norm violators to come apart from their responses to norm violations. This might have been useful for many reasons. To appreciate one possible strategic benefit, consider A

[2] Indeed, prior to 1987, Mark might have received treatment (probably, aversion therapy) for 'ego-dystonic homosexuality'—a diagnosis that had replaced 'sexual orientation disturbance' in the American Psychiatric Association's diagnostic manual, DSM-III.

and B, two members of a tight-knit, traditionally structured tribe. Because they live cooperatively in the same village, if A believes B shares his values and interests, and B believes the same of A, they will both be substantially correct—a fact which may lead to significant fitness benefits for each (Berkes, Colding, & Folke, 2000; Johnson & Earle, 2000). Now consider two members of two distinct tribes. They do not cooperate in a more than incidental way, and even this is limited to a small region where their respective territories overlap. Consequently, their values and interests are often in conflict (Diamond, 2013). The effect of the virtuous true self bias on these two agents will be to make each insensitive to the interests of the other, reducing the risk that either will incur harms or sacrifice benefits for someone in whom he has no fitness stake.

Shameless nepotism may seem to provide a more elegant solution. Why not make the moral value of an action depend, in part, on the identity of the actor? Perhaps humans use both strategies to focus their moral concern on members of their own ethnic, racial or language groups (Bernhard, Fischbacher, & Fehr, 2006). However, indiscriminately representing agents as, deep down, normatively aligned with the self may have better reconciled two opposing demands biology places on moral norms (McCullough, Kurzban, & Tabak, 2013). First, moral norms need to win the alliance of most members of a cooperating group, else morality will not serve one of its primary functions—effectively coordinating the group's behavior (DeScioli & Kurzban, 2009). Thus, effective moral norms must have at least the pretense of impartiality. This pressure helps to explain why the moral value of an action is often taken to depend on features of the action, rather than the identity of the actor. However, moral agents must also be somewhat parochial, lest morality lead them to emit benefits or absorb costs for people unlikely to return the favor. Belief in the virtuous true self might have been adaptive because it helped our ancestors to reconcile these seemingly inconsistent demands. More generally, defaulting to the belief that others are normatively aligned with the self may have helped to mitigate some of the biological costs induced by our sensitivity to moral norms. For example, belief in the virtuous true self may have helped vengeful agents to mend beneficial relationships after retaliating against severe norm violators (McCullough et al., 2013).

Speculations like these are famously difficult to test, yet they suggest clearly enough that there may have been surprising benefits to indiscriminately representing other agents as normatively aligned with the self. Thus, psychological essentialism may explain why people develop the true-self concept, while strategic sociobiological benefits may explain why the true self tends to be represented as virtuous.

6.4. Why it may be better not to believe people are virtuous

The preceding discussion and this paper's first two studies may also illuminate how representing the true self as virtuous can lead to painful inner conflicts and may slow the rate at which social norms change. As recently as the 1960s, law and social pressure led many gay men to undergo 'reparative' medical treatments—commonly, electric shock-based aversion therapy (Haldeman, 1991). Meanwhile, childhood abuse, parental neglect, mental illness, and demonic possession of the sort described by the Bible were all alleged to explain homosexuality.

The mismatch hypothesis suggests that this association may be explained, in part, by cognitive processes that preexist political ideology. It may also help to explain why, even today, the belief that sexual orientation is innate predicts support for gay rights (Jayaratne et al., 2006; Wood & Bartkowski, 2004). When people are asked a technical question, such as whether a human trait is innate, they will often respond by consulting their intuitions about a seemingly related question (Cullen, 2010), such as whether that trait expresses the agent's true self. This suggests that belief in the virtuous true self may incline people to reject genetic attributions for negative traits.

In the United States, attitudes towards homosexuality began to improve markedly in the 1970s, a period during which genetic explanations of homosexuality became more widely accepted (Hicks & Lee, 2006; Sherkat, Powell-Williams, Maddox, & De Vries, 2011). Perhaps people came to represent sexual orientation as 'innate' and this facilitated a broad change in their attitudes towards homosexuality (Sheldon, Pfeffer, Jayaratne, Feldbaum, & Petty, 2007; Wood & Bartkowski, 2004;); however, the mismatch hypothesis suggests that the causation may have traveled in the reverse direction. For example, cohort effects may have improved attitudes towards homosexuality (Andersen & Fetner, 2008; Treas, 2002) and this may have caused same-sex attraction to appear more essence-disclosing ('innate'). Thus, a tendency to represent the true self as virtuous may also help to explain the continued attraction of psychotherapies aimed at 'reorienting' non-heterosexuals (Dean Byrd, Nicolosi, & Potts, 2008; Haldeman, 2002).

The diversity of human values ensures that people who live in large, high-density societies will regularly interact with non-normative agents (Esmer & Pettersson, 2007; Haidt & Graham, 2007; Norris & Inglehart, 2011). Because we represent the true self as virtuous, we tend to view such agents as, in some deeper sense, normatively aligned with ourselves. But there is little reason to believe that this involves representing them as ethically competent or ourselves as having corresponding *pro tanto* reasons to respect their stated preferences. To the contrary, commitment to the hidden virtuousness of non-normative agents seems to involve representing these agents as the unwitting victims of external forces that mask their underlying ethical capacities.

7. Conclusion

A long tradition in psychology holds that our explanatory preferences are primarily driven by causal-statistical ('covariation') information (e.g., Kelley, 1967, 1973, 1987). In the present studies, however, the explanatory preferences of both North American and Indian participants were surprisingly unrelated to such information. By contrast, the mismatch hypothesis robustly predicted explanatory preferences across actions as varied as having consensual gay sex, aborting a pregnancy, converting to Islam, owning slaves in the antebellum South, and identifying as Conservative or Liberal today. Thus, the results reported here are plausibly general and should emerge whenever people explain valenced actions.

Theorists of responsibility often appear to assume that intuitive judgments about essence-disclosure primarily reflect facts about how actions are related to agents' mental states (e.g., Frankfurt, 1969; Smith, 2005; Sripada, 2016; Watson, 1996). However, the studies reported here support a strikingly different account of the processes underlying these intuitions. Researchers pursuing true self theories of responsibility should address the extent to which such processes provide appropriate inputs to normative theorizing.

Supplementary materials

Supporting data

Supporting data for this work are archived on the Open Science Framework: http://osf.io/mk8ft.

Argument visualization

An interactive map of this paper's argument is available online: http://wdce.simoncullen.org.

Acknowledgments

Thanks to Joshua Knobe, Gideon Rosen, Sarah-Jane Leslie, and Shamik Dasgupta for extensive feedback on this work, and to three anonymous reviewers for *Cognition* who provided many helpful comments. This publication benefited from a grant from the John Templeton Foundation; however, the opinions expressed in this

publication are those of the author and do not necessarily reflect the views of the John Templeton Foundation.

Appendix A. Vignettes

A.1. Study 1

Mark was raised in a large family. They went to church [temple] every week and were dedicated to charity. For most of Mark's childhood the family was very happy. But his parents began to fight, and eventually they divorced. Mark's family members grew uninterested in religion. This especially hurt him, since he had always felt a deep spiritual calling.

Soon after the divorce, Mark's mother found a new husband and started a new family. Mark's mother began to ignore him, and when Mark turned 16, his mother thought he should leave home. Out of home, Mark started hanging out with a new group of kids, some of whom were in same-sex relationships.

Mark believed that homosexuality is morally wrong, and he encouraged his new friends to resist their attractions to people of the same sex. However, Mark also realized that he himself was attracted to other men. He openly acknowledged this to his friends and discussed it as part of his own personal struggle. Mark believed that resisting his attraction to other men was his duty to God, and he vowed to live a morally decent life by remaining celibate.

However, Mark sometimes failed to live up to his values. For example, one day, after a bad fight with his father, Mark went to see his friend Bill. They talked for hours over a bottle of wine. That night, Mark hit on Bill and they ended up having sex.

A.2. Study 2

A.2.1. Baseline

As in Study 1.

A.2.2. Situation

As in Study 1, with the following appended:

Most people do not find Bill attractive, and in the past Mark himself has rarely felt sexually attracted to Bill. In fact, Mark has rarely experienced attraction to men other than Bill.

A.2.3. Person

As in Study 1, with the following appended:

Most people do not find Bill attractive, but in the past Mark has often felt sexually attracted to Bill. In fact, Mark has often experienced attraction to men other than Bill, too.

A.2.4. Substitutions for Indian participants

'Mark' → 'Aarav'.
'Bill' → 'Arjun'.
'duty to God' → 'duty'

A.3. Study 3a

Kate is a senior at college. Just like many of her friends, after graduation Kate plans to spend a year working for a charity organization before pursuing her dream of going to medical school.

Kate has recently discovered that she is pregnant. And what's worse, her boyfriend broke up with her just a week earlier. When she tells her parents, their reaction is clear: although they will support her no matter what she decides, they both think that she should abort the pregnancy. Kate's friends also agree that she should get an abortion.

Kate's local health clinic offers the procedure. After thinking it over, Kate decides to have an abortion.

A.4. Study 3b

Jane was raised in a large family. They went to church every week and were dedicated to charity. For most of Jane's childhood the family was very happy. But her parents began to fight, and eventually they got divorced. Jane's family members became uninterested in religion. This especially hurt her, since she had always felt a deep spiritual calling.

Soon, Jane's mother found a new husband and started a new family. Her mother began to ignore Jane, and when Jane turned 16, her mother thought Jane should leave home. Out of home, Jane needed money. She was good looking, so she started working as a model. But Jane's modeling agent pressured her to lose weight, which disgusted her and she soon quit modeling.

Jane started hanging out with a new group of kids, some of whom were Muslim. Jane was intrigued by their religion and decided to learn more about Islam. She discovered that many of these kids were also involved in charity through their mosque. She valued what she saw as their moral uprightness. And she perceived a kind of moral clarity in Islamic texts which she found reassuring. She came to believe that Islam is the one true path to God.

Eventually, Jane decided that she would convert to Islam.

References

Andersen, R., & Fetner, T. (2008). Cohort differences in tolerance of homosexuality: Attitudinal change in Canada and the United states, 1981-2000. *Public Opinion Quarterly, 72*(2), 311–330.

Asthana, S., & Oostvogels, R. (2001). The social construction of male 'homosexuality' in India: Implications for HIV transmission and prevention. *Social Science & Medicine, 52*(5), 707–721.

Batson, C. D., Darley, J. M., & Coke, J. S. (1978). Altruism and human kindness: Internal and external determinants of helping behavior. *Perspectives in interactional psychology* (pp. 111–140). Springer.

Bench, S. W., Schlegel, R. J., Davis, W. E., & Vess, M. (2015). Thinking about change in the self and others: The role of self-discovery metaphors and the true self. *Social Cognition, 33*(3), 169–185.

Berkes, F., Colding, J., & Folke, C. (2000). Rediscovery of traditional ecological knowledge as adaptive management. *Ecological Applications, 10*(5), 1251–1262.

Bernhard, H., Fischbacher, U., & Fehr, E. (2006). Parochial altruism in humans. *Nature, 442*(7105), 912.

Boyatzis, C. J., Dollahite, D. C., & Marks, L. D. (2006). The family as a context for religious and spiritual development in children and youth. *The Handbook of Spiritual Development in Childhood and Adolescence,* 297–309.

Clifford, S. (2017). Individual differences in group loyalty predict partisan strength. *Political Behavior, 39*(3), 531–552.

Coffman, E. J. (2007). Thinking about luck. *Synthese, 158*(3), 385–398.

Crawford, J. T., & Pilanski, J. M. (2014). Political intolerance, right and left. *Political Psychology, 35*(6), 841–851.

Cullen, S. (2010). Survey-driven romanticism. *Review of Philosophy and Psychology, 1*(2), 275–296.

Darley, J. M., & Batson, C. D. (1973). From Jerusalem to Jericho: A study of situational and dispositional variables in helping behavior. *Journal of Personality and Social Psychology, 27*, 100–108.

Dean Byrd, A., Nicolosi, J., & Potts, R. W. (2008). Clients' perceptions of how reorientation therapy and self-help can promote changes in sexual orientation. *Psychological Reports, 102*(1), 3–28.

De Freitas, J., & Cikara, M. (2018). Deep down my enemy is good: Thinking about the true self reduces intergroup bias. *Journal of Experimental Social Psychology, 74*, 307–316.

De Freitas, J., Cikara, M., Grossmann, I., & Schlegel, R. (2017). Origins of the belief in good true selves. *Trends in Cognitive Sciences, 21*(9), 634–636.

De Freitas, J., Sarkissian, H., Newman, G. E., Grossmann, I., De Brigard, F., Luco, A., & Knobe, J. (2018). Consistent belief in a good true self in misanthropes and three interdependent cultures. *Cognitive science, 42*, 134–160.

De Freitas, J., Tobia, K. P., Newman, G. E., & Knobe, J. (2016). Normative judgments and individual essence. *Cognitive Science*.

DeScioli, P., & Kurzban, R. (2009). Mysteries of morality. *Cognition, 112*(2), 281–299.

Diamond, J. (2013). *The world until yesterday: What can we learn from traditional societies?* Penguin.

Ellemers, N., Pagliaro, S., Barreto, M., & Leach, C. W. (2008). Is it better to be moral than smart? The effects of morality and competence norms on the decision to work at group status improvement. *Journal of Personality and Social Psychology, 95*, 1397–1410.

Epley, N., & Dunning, D. (2000). Feeling holier than thou: Are self-serving assessments produced by errors in self- or social prediction? *Journal of Personality and Social Psychology, 79*, 861–875.

Esmer, Y. R., & Pettersson, T. (2007). *Measuring and mapping cultures: 25 years of comparative value surveys, Vol. 104*. Brill.
Feldman, S. (2003). Values, ideology, and the structure of political attitudes. *Oxford Handbook of Political Psychology*.
Feltz, A., & Millan, M. (2015). An error theory for compatibilist intuitions. *Philosophical Psychology, 28*(4), 529–555.
Frankfurt, H. G. (1969). Alternate possibilities and moral responsibility. *The Journal of Philosophy, 66*, 829–839.
Frankfurt, H. G. (1971). Freedom of the will and the concept of a person. *The Journal of Philosophy, 68*(1), 5–21.
De Freitas, J., Cikara, M., Grossmann, I., & Schlegel, R. (2018). Moral goodness is the essence of personal identity (in press).
Gelman, S. A. (2003). *The essential child: Origins of essentialism in everyday thought.* New York: Oxford University Press.
Gelman, S. A., Heyman, G. D., & Legare, C. H. (2007). Developmental changes in the coherence of essentialist beliefs about psychological characteristics. *Child Development, 78*(3), 757–77 4.
Graham, J., Nosek, B. A., & Haidt, J. (2012). The moral stereotypes of liberals and conservatives: Exaggeration of differences across the political spectrum. *PLOS ONE, 7*(12), 1–13.
Gunther, R., & Kuan, H.-c. (2007). Value cleavages and partisan conflict. *Electoral Intermediation, Values, and Political Support in Old and New Democracies: Europe, East Asia, and the Americas in Comparative Perspective,* 255–320.
Haider-Markel, D. P., & Joslyn, M. R. (2008). Beliefs about the origins of homosexuality and support for gay rights: An empirical test of attribution theory. *Public Opinion Quarterly, 72*, 291–310.
Haidt, J., & Graham, J. (2007). When morality opposes justice: Conservatives have moral intuitions that liberals may not recognize. *Social Justice Research, 20*(1), 98–116.
Haldeman, D. C. (1991). Sexual orientation conversion therapy for gay men and lesbians: A scientific examination. *Homosexuality: Research Implications for Public Policy, 149*, 160.
Haldeman, D. C. (2002). Gay rights, patient rights: The implications of sexual orientation conversion therapy. *Professional Psychology: Research and Practice, 33*(3), 260.
Haslam, N., Bastian, B., & Bissett, M. (2004). Essentialist beliefs about personality and their implications. *Personality and Social Psychology Bulletin, 30*(12), 1661–1673.
Hayes, A. F. (2013). *Introduction to mediation, moderation, and conditional process analysis: A regression-based approach.* Guilford Press.
Heider, F. (1983). *The psychology of interpersonal relations.* Hillsdale, NJ: Lawrence Erlbaum.
Henrich, J., Heine, S., & Norenzayan, A. (2010). The weirdest people in the world? *The Behavioral and Brain Sciences, 33*(2-3), 61–83.
Herek, G. M. (1998). The attitudes toward lesbians and gay men (ATLG) scale. In G. F. Sanders, C. M. Davis, W. L. Yarber, R. Bauserman, G. Schreer, & S. L. Davis (Eds.). *Handbook of sexuality-related measures* (pp. 392–394). SAGE Publications.
Hewstone, M., & Jaspars, J. (1987). Covariation and causal attribution: A logical model of the intuitive analysis of variance. *Journal of Personality and Social Psychology, 53*(4), 663–672.
Hewstone, M., & Ward, C. (1985). Ethnocentrism and causal attribution in southeast asia. *Journal of Personality and Social Psychology, 48*, 614–623.
Hicks, G. R., & Lee, T.-T. (2006). Public attitudes toward gays and lesbians: Trends and predictors. *Journal of Homosexuality, 51*(2), 57–77.
Huddy, L., Mason, L., & Aaroe, L. (2015). Expressive partisanship: Campaign involvement, political emotion, and partisan identity. *American Political Science Review, 109*(1), 1–17.
Inbar, Y., Pizarro, D. A., & Bloom, P. (2009). Conservatives are more easily disgusted than liberals. *Cognition and Emotion, 23*(4), 714–725.
Inbar, Y., Pizarro, D. A., Knobe, J., & Bloom, P. (2009). Disgust sensitivity predicts intuitive disapproval of gays. *Emotion, 9*(3), 435–439.
Iyengar, S., Sood, G., & Lelkes, Y. (2012). Affect, not ideology: Social identity perspective on polarization. *Public Opinion Quarterly, 76*(3), 405–431.
Iyengar, S., & Westwood, S. J. (2015). Fear and loathing across party lines: New evidence on group polarization. *American Journal of Political Science, 59*(3), 690–707.
Jayaratne, T. E., Ybarra, O., Sheldon, J. P., Brown, T. N., Feldbaum, M., Pfeffer, C. A., & Pfeffer, C. A. (2006). White Americans' genetic lay theories of race differences and sexual orientation: Their relationship with prejudice toward Blacks, and gay men and lesbians. *Group Processes & Intergroup Relations, 9*(1), 77–94.
Jennings, M. K., Stoker, L., & Bowers, J. (2009). Politics across generations: Family transmission reexamined. *The Journal of Politics, 71*(3), 782–799.
Johnson, A. W., & Earle, T. K. (2000). *The evolution of human societies: From foraging group to agrarian state.* Stanford University Press.
Jones, E. E., & Nisbett, R. E. (1972). The actor and the observer: Divergent perceptions of the causes of behavior. In *Attribution: Perceiving the causes of behavior* (pp. 79–94).
Jones, E. E., & Davis, K. E. (1965). From acts to dispositions: The attribution process in person perception. In L. Berkowitz (Vol. Ed.), *Advances in experimental social psychology: Vol. 2*, (pp. 219–266). New York: Academic Press.
Kelley, H. (1967). Attribution theory in social psychology. In *Nebraska symposium on motivation* (Vol. 15, pp. 192–238).
Kelley, H. (1973). The processes of causal attribution. *American Psychologist, 28*(2), 107–128.
Kelley, H. (1987). Causal schemata and the attribution process. In E. E. Jones, D. E. Kanouse, H. Kelley, R. E. Nisbett, S. Valins, & B. Weiner (Eds.). *Attribution: Perceiving the causes of behavior* (pp. 151–174). Hillsdale, NJ: Lawrence Erlbaum.
Kelley, J., & Graaf, N. D. D. (1997). National context, parental socialization, and religious belief: Results from 15 nations. *American Sociological Review, 62*(4), 639–659.
Kim, J., Christy, A., Hicks, J., & Schlegel, R. (2017). Trust thyself: True-self-as-guide lay theories enhance decision satisfaction (in preparation).
Klein, N., & Epley, N. (2016). Maybe holier, but definitely less evil, than you: Bounded self-righteousness in social judgment. *Journal of Personality and Social Psychology, 110*(5), 660.
Kohn, M. L. (1989). *Class and conformity: A study in values.* University of Chicago Press.
Kohn, M. L., Slomczynski, K. M., & Schoenbach, C. (1986). Social stratification and the transmission of values in the family: A cross-national assessment. In *Sociological forum* (Vol. 1, pp. 73–102).
Kunda, Z., & Nisbett, R. E. (1986). The psychometrics of everyday life. *Cognitive Psychology, 18*(2), 195–224.
Leach, C. W., Ellemers, N., & Barreto, M. (2007). Group virtue: The importance of morality (vs. competence and sociability) in the positive evaluation of in-groups. *Journal of Personality and Social Psychology, 93*, 234–249.
Levine, R. A., & Campbell, D. T. (1972). *Ethnocentrism: Theories of conflict, ethnic attitudes, and group behavior.* New York: Wiley.
Levy, N. (2011). *Hard luck: How luck undermines free will and moral responsibility.* Oxford University Press.
Lewis, D. (1986). Postscript c to'causation': Insensitive causation. *Philosophical papers: Vol. 2*, (pp. 184–188). Oxford: Oxford University Press.
Lewis, G. B. (2009). Does believing homosexuality is innate increase support for gay rights? *Policy Studies Journal, 37*, 669–693.
Litman, L., Robinson, J., & Abberbock, T. (2017). Turkprime.com: A versatile crowdsourcing data acquisition platform for the behavioral sciences. *Behavior Research Methods, 49*(2), 433–442.
Malle, B. F. (2006). The actor-observer asymmetry in attribution: A (surprising) meta-analysis. *Psychological Bulletin, 132*(6), 895–919.
Malle, B. F. (2011). Time to give up the dogmas of attribution. An alternative theory of behavior explanation. *Advances in experimental social psychology: Vol. 44*, (pp. 297–352). Burlington: Academic Press.
Malle, B. F., Knobe, J., O'Laughlin, M. J., Pearce, G. E., & Nelson, S. E. (2000). Conceptual structure and social functions of behavior explanations: Beyond person-situation attributions. *Journal of Personality and Social Psychology, 79*(3), 309.
McCullough, M. E., Kurzban, R., & Tabak, B. A. (2013). Putting revenge and forgiveness in an evolutionary context. *Behavioral and Brain Sciences, 36*(1), 41–58.
Murray, D., & Nahmias, E. (2014). Explaining away incompatibilist intuitions. *Philosophy and Phenomenological Research, 88*(2), 434–467.
Nagel, T. (1979). Moral luck. *Mortal questions*. New York: Cambridge University Press.
Nelkin, D. K. (2005). Freedom, responsibility and the challenge of situationism. *Midwest Studies in Philosophy,* 181–206.
Nelkin, D. K. (2011). *Making sense of freedom and responsibility.* Oxford: Oxford University Press.
Newman, G. E., Bloom, P., & Knobe, J. (2013). Value judgments and the true self. *Personality and Social Psychology Bulletin*.
Newman, G. E., De Freitas, J., & Knobe, J. (2014). Beliefs about the true self explain asymmetries based on moral judgment. *Cognitive Science, 0*, 1–30.
Nichols, S., & Knobe, J. (2007). Moral responsibility and determinism: The cognitive science of folk intuitions. *Nous, 41*(4), 663–685.
Norris, P., & Inglehart, R. (2011). *Sacred and secular: Religion and politics worldwide.* Cambridge University Press.
Patel, V. V., Mayer, K. H., & Makadon, H. J. (2012). Men who have sex with men in India: A diverse population in need of medical attention. *The Indian Journal of Medical Research, 136*(4), 563.
Pettigrew, T. F. (1979). The ultimate attribution error: Extending allport's cognitive analysis of prejudice. *Personality and Social Psychology Bulletin, 5*(4), 461–476.
Phillips, J., Misenheimer, L., & Knobe, J. (2011). The ordinary concept of happiness (and others like it). *Emotion Review, 3*(3), 320–322.
Pizarro, D., Uhlmann, E., & Salovey, P. (2003). Asymmetry in judgments of moral blame and praise: The role of perceived metadesires. *Psychological Science, 14*(3), 267–272.
Prinz, J., & Nichols, S. (2016). Diachronic identity and the moral self. *The routledge handbook of philosophy of the social mind*. Routledge.
Pritchard, D. (2006). Moral and epistemic luck. *Metaphilosophy, 37*(1), 1–25.
Ross, L. (1977). The intuitive psychologist and his shortcomings: Distortions in the attribution process. *Advances in Experimental Social Psychology, 10*, 173–220.
Sabini, J., Siepmann, M., & Stein, J. (2001). The really fundamental attribution error in social psychological research. *Psychological Inquiry, 12*, 1–15.
Schlegel, R. J., Hicks, J. A., Arndt, J., & King, L. A. (2009). Thine own self: True self-concept accessibility and meaning in life. *Journal of Personality and Social Psychology, 96*(2), 473.
Schlegel, R. J., Hicks, J. A., King, L. A., & Arndt, J. (2011). Feeling like you know who you are: Perceived true self-knowledge and meaning in life. *Personality and Social Psychology Bulletin, 37*(6), 745–756.
Schwartz, S. H. (2006). Les valeurs de base de la personne: thorie, mesures et applications. *Revue Francaise de sociologie, 47*(4), 929–968.
Schwartz, S. H., Caprara, G. V., & Vecchione, M. (2010). Basic personal values, core political values, and voting: A longitudinal analysis. *Political Psychology, 31*(3), 421–452.
Sheldon, J. P., Pfeffer, C. A., Jayaratne, T. E., Feldbaum, M., & Petty, E. M. (2007). Beliefs about the etiology of homosexuality and about the ramifications of discovering its possible genetic origin. *Journal of Homosexuality, 52*(3-4), 111–150.
Sherkat, D. E., Powell-Williams, M., Maddox, G., & De Vries, K. M. (2011). Religion, politics, and support for same-sex marriage in the United States, 1988-2008. *Social Science Research, 40*(1), 167–180.
Smith, A. M. (2005). Responsibility for attitudes: Activity and passivity in mental life. *Ethics, 115*, 236–271.
Smith, G., & Cooperman, A. (2015). *America's changing religious landscape.* Washington, DC: Pew Research Center.

Sripada, C. (2016). Self-expression: A deep self theory of moral responsibility. *Philosophical Studies, 173*(5), 1203–1232.
Strohminger, N., Knobe, J., & Newman, G. (2017). The true self: A psychological concept distinct from the self. *Perspectives on Psychological Science*.
Strohminger, N., & Nichols, S. (2014). The essential moral self. *Cognition, 131*, 159–171.
Strohminger, N., & Nichols, S. (2015). Neurodegeneration and identity. *Psychological Science, 26*(9), 1469–1479.
Suhay, E., & Jayaratne, T. E. (2012). Does biology justify ideology? The politics of genetic attribution. *Public Opinion Quarterly, 77*, 497–521.
Tahmindjis, P. (2014). *Sexuality and human rights: A global overview*. Routledge.
Taylor, D. M., & Jaggi, V. (1974). Ethnocentrism and causal attribution in a South Indian context. *Journal of Cross-Cultural Psychology, 5*, 162–171.
Treas, J. (2002). How cohorts, education, and ideology shaped a new sexual revolution on American attitudes toward nonmarital sex, 1972-1998. *Sociological Perspectives, 45*(3), 267–283.
Watson, G. (1996). Two faces of responsibility. *Philosophical Topics, 24*(2), 227–248.
Williams, B. (1981). *Moral luck*. Cambridge: Cambridge University Press.
Williams, G. C. (2008). *Adaptation and natural selection: A critique of some current evolutionary thought*. Princeton University Press.
Wolf, S. (1980). Asymmetrical freedom. *Journal of Philosophy, 77*(3), 151–166.
Wood, P. B., & Bartkowski, J. P. (2004). Attribution style and public policy attitudes toward gay rights. *Social Science Quarterly, 85*(1), 58–74.
Woodward, J. (2006). Sensitive and insensitive causation. *The Philosophical Review*.