Blame for Hum(e)an beings
The role of character information in judgments of blame

*Forthcoming in *Social Psychological and Personality Science*

Samuel Murray[a,b], Kevin O'Neill[c,d,e], Jordan Bridges[f], Justin Sytsma[g], and Zachary C. Irving[h]

[a] Laboratorio de Emociones y Juicios Morales, Universidad de los Andes, Bogotá, Colombia
[b] Department of Philosophy, Providence College, Providence, RI, USA
[c] Department of Psychology and Neuroscience, Duke University, Durham, NC, USA
[d] Duke Institute for Brain Sciences, Duke University, Durham, NC, USA
[e] Center for Cognitive Neuroscience, Duke University, Durham, NC, USA
[f] Department of Philosophy, Rutgers University, New Brunswick, NJ, USA
[g] School of History, Philosophy, Political Science, and International Relations, Victoria University of Wellington, Wellington, NZ
[h] Corcoran Department of Philosophy, University of Virginia, Charlottesville, VA, USA

Author Note

Correspondence may be addressed to: Samuel Murray, 125 Siena Hall, 1 Cunningham Sq., Providence, RI 02918. Email: smurray7@providence.edu. All authors approved the final version of the manuscript. Materials, preregistrations, data, and analysis script can be found at the OSF repository for the project: https://osf.io/fuhte/.

Abstract

How does character information inform judgments of blame? Some argue that character information is indirectly relevant to blame because it enriches judgments about the mental states of a wrongdoer. Others argue that character information is directly relevant to blame, even when character traits are causally irrelevant to the wrongdoing. We propose an empirical synthesis of these views: a Two Channel Model of blame. The model predicts that character information directly affects blame when this information is relevant to the wrongdoing that elicits blame. Further, the effect of character information on blame depends on judgments about the true self that are independent of judgments of intentionality. Across three pre-registered studies (N = 662), we found support for all three predictions of the Two Channel Model. We propose that this reflects two distinct functions of blame: a social regulatory function that encourages norm compliance and a pedagogical function that encourages personal improvement.

Relevance

When we blame someone for wrongdoing, we care about their character. Why is this? Some argue that we use character information to get a better sense of what a wrongdoer intended. Others argue that blame fundamentally targets character. We propose that character information serves two different functions, thereby synthesizing these different viewpoints. Our studies show that character information directly affects blame, but only when the information is relevant to wrongdoing. The relationship between character information and blame depends on inferences about whether the wrongdoing reflects a person's true self. This clarifies an important role for character information in judgments of blame, namely that character information indicates whether wrongdoing manifests the kind of people we are deep down.

Keywords: blame, character, attribution, true self, intentionality

Introduction

In his *Treatise on human nature*, David Hume argued that character information is relevant to judgments of blame:

> Actions are by their very nature temporary and perishing; and where they proceed not from some cause in the characters and disposition of the person, who perform'd them, they infix not themselves upon him, and can neither redound to his honour, if good, nor infamy, if evil (1955 [1739], p. 107).

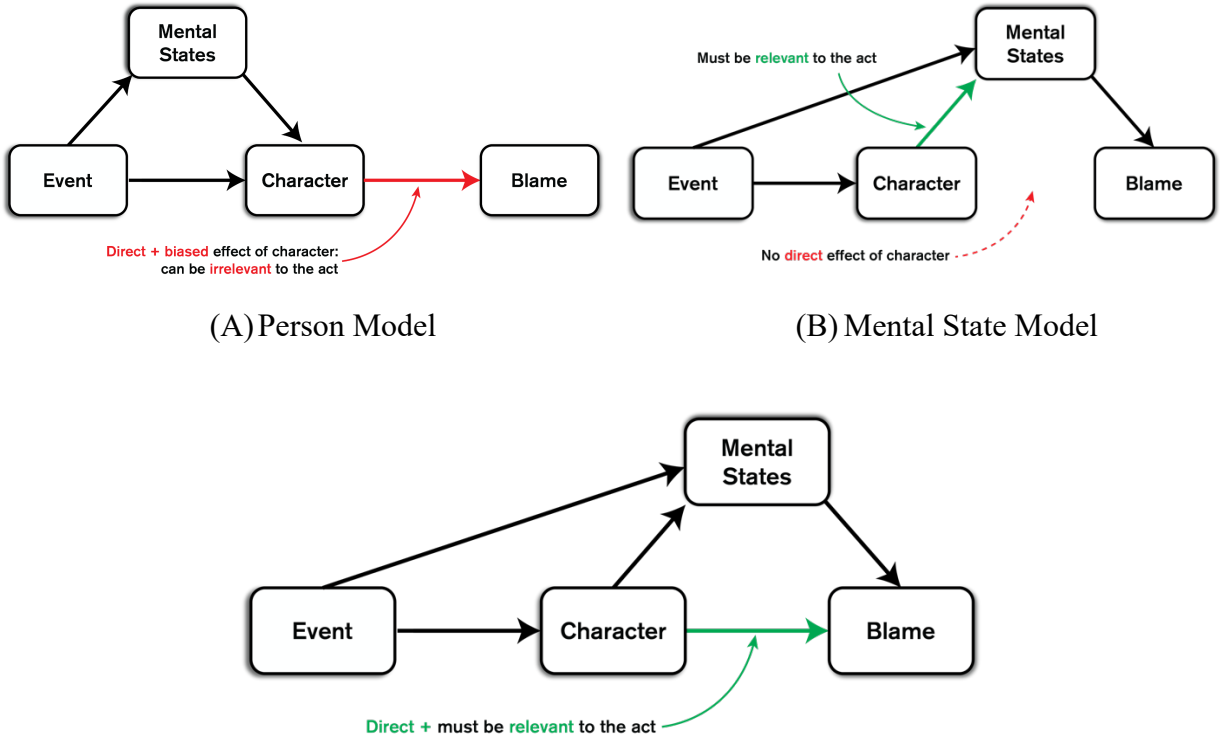
Contemporary models of blame in moral psychology follow Hume in acknowledging the effect of character information on blame judgments. However, these models ascribe fundamentally different roles to character. On *Person Models*, character is the focus of blame judgments. On *Mental State Models*, character merely provides *evidence* about blame-relevant mental states (e.g., intentions). We propose—and find experimental support for—an empirical synthesis on which *Person* and *Mental State Models* each identify *distinct channels* by which character information influences blame. This *Two Channel Model* better captures how hum(e)an beings assign blame.

Psychological models of blame attribution. On Person Models, blame judgments primarily evaluate *people's character* (Nadler, 2012; Rai, 2017; Uhlmann et al., 2015). Bad character elicits blame, whereas good character elicits praise. We use praise to decide which social partners to approach (virtuous persons) and blame to decide which to avoid (vicious persons) (Pizarro & Tannenbaum, 2012). Person Models therefore make two predictions about how character affects blame judgments (Figure 1a). First, character should have a *direct* effect on blame: vicious character makes people worse social partners, and thus more blameworthy, independent of anything else about the person (Alicke & Zell, 2009; Siegel et al., 2017). Second, character judgments should affect blame even when a perpetrator's vice (e.g., uncleanliness) is causally and explanatorily irrelevant to her harmful action (e.g., stealing). As Rai puts it, "The existence of negative character traits will make moral judgments more severe even when the traits had no effect on the outcomes that occurred" (2017, p. 193).

On Mental State Models, blame judgments primarily evaluate the moral quality of *actions*: this depends on (a) the action's harms and (b) the agent's mental states (Malle et al., 2014). This is why intentional wrongdoing is more blameworthy than unintentional wrongdoing (Cushman, 2008). Mental State Models make two predictions that conflict with Person Models (Figure 1b). First, character should have an *indirect* effect on blame, insofar as it provides evidence about the agent's mental states (Alicke, 1992; Cushman, 2015; Guglielmo & Malle, 2017; Koster-Hale et al., 2013). Suppose a frequently dishonest person leaves a store without paying for her ice cream. Her bad character provides evidence that she took the ice cream on purpose rather than by accident. Second, character judgments should affect blame only when the traits in question are *relevant* to

the harmful action (Hughes & Trafimow, 2012). Whereas dishonesty makes one likely to intentionally steal, for example, uncleanliness does not. In this way, the influence of character is sensitive to relevance: "People...integrate any and all information given to them, including clues about potentially relevant general dispositions, to interpret the causal-mental facts of a naturally ambiguous situation" (Malle et al., 2014 p. 17).

Mixed evidential support. Past evidence provides mixed support for either model. Some studies seem to support Person Models. Woolfolk et al. (2006) found that people assign blame for wrongdoing that aligns with someone's character, even when they were coerced (with a mind-control drug!), and thus lacked mental states like intention and volition. Similarly, Uhlmann et al. (2013) found that people tend to attribute blame to *morally correct actions* that provide evidence for an underlying vice.



(A) Person Model

(B) Mental State Model

(C) Two Channel Model

Figure 1: How does character inform blame judgments? Every model assumes that detecting norm-violating events initiates processes related to blame attribution. (A) Person Model: character informs blame directly and does not depend on whether the character trait is relevant to the wrongdoing. (B) Mental State Model: character informs blame by providing evidence about mental states such as intentions, which requires that the character trait is relevant to the wrongdoing. (C) Two Channel Model: character informs blame in two ways: one indirect and another direct. On the indirect channel, character provides evidence about mental states. On the direct channel, wrongdoing is more blameworthy when it manifests a relevant character flaw and less blameworthy when it manifests a relevant character strength.

Other evidence supports Mental State Models. Consistent with the above studies, Sytsma (2021) found that character predicts judgments of blame about an ambiguous norm-violating event. However, character no longer affects blame if we resolve ambiguity by clarifying the agent's motives and intentions. Character therefore seems to affect blame only indirectly when it provides evidence about ambiguous mental states. Furthermore, apparent evidence for Person Models is difficult to interpret because character elicits *many* inferences that typical experiments do not control for (Malle, 2021; Royzman & Hagan, 2017).

The two-channel model. We propose a new empirical synthesis: the Two Channel Model (Figure 1c). Character influences blame through two channels: one direct and another indirect. Character influences blame *indirectly* by providing evidence about mental states. Dishonestly makes you more likely to intentionally steal, for example. Character also has a *direct* effect on blame: people are more blameworthy when actions manifest an underlying vice than when those same actions are uncharacteristic. If a dishonest person steals an ice cream, for example, this manifests something about the kind of person she is. If an honest person steals ice cream, in contrast, I may excuse her misdeed as an uncharacteristic lapse.

The Two Channel Model makes distinct empirical predictions (Irving et al., 2023). Some of our predictions integrate those of past theories. Like the Person Model, we predict that character has a *direct* effect on blame. Yet like the Mental State Model, we predict that this direct effect is sensitive to whether the agent's character (e.g., dishonesty) is relevant to her wrongdoing (e.g., stealing). Other predictions are novel. Unlike Person or Mental State Models, we predict that the direct effect of character on blame depends on judgments about whether *an action manifests your true self*, rather than whether you have (a) some relevant or irrelevant vice or (b) bad intentions.

Experimental approach. Three experiments confirmed the Two Channel Model's predictions (Table 1). Experiment 1 (n=151) found that character information has direct effects on blame after fixing mental state information in an earlier stage of the experiment. Experiment 2 (n=160) found that this direct effect of character on blame requires that the character trait is relevant to wrongdoing. Experiment 3 (n=351) found that the direct effect of character on blame depends on whether the wrongdoing manifests a vice that is partly constitutive of someone's true self.

Table 1. Summary of predictions for different models of blame attribution and findings across three studies.

	Person	Mental State	Two Channel
Direct (*Study 1*)	+	X	+
Relevance (*Study 2*)	X	+	+
Direct: True Self (*Study 3*)	X	X	+

Note. Green plus (+) represents evidence that supports a prediction. X represents lack of evidence in support of a prediction. Whereas Studies 1 and 2 test predictions that integrate past theories, Study 3 tests a novel prediction.

Data availability. Materials, preregistrations, and data for all studies are available at the OSF repository for the project: https://osf.io/fuhte/. In the interest of transparency, each study was preregistered. All studies reported below were approved by [redacted for peer review] Institutional Review Board under [redacted for per review].

Transparency and openness. All experiments were preregistered to clearly establish sample size justification, design, and analysis plans. Data were analyzed using R 4.3.0 (R Core Team, 2023)

Study 1: Character has a direct effect on blame.

All participants provided electronic consent following the procedures approved by [omitted for review].

Participants. Preregistration site is https://aspredicted.org/Q1K_ST8. 152 participants were recruited on Academic Prolific. Sample size was determined using G*Power software (Faul et al., 2007). For an ANOVA test to be 95% powered to detect the smallest effect size of interest (f = 0.33) at standard error thresholds ($p < 0.05$), 122 participants were recommended. Effect sizes were estimated using lower-bound of 90% confidence intervals for effects measured during pilot testing (N = 159). We over-recruited by 25% to account for exclusions. 1 participant was excluded for failing an attention check (final N = 151, M_{age} = 34.8, SD = 12.2, 49% female).

Methods and procedure. Study 1 used a 2(vignette)x2(valence)x2(intentionality) between-subjects design. The study consisted of two stages to test whether character has a *direct* or *indirect* effect on blame. Stage one presented participants with a story where someone harms his roommate by taking their ice cream or spoiling their TV show. We resolved any ambiguity about the wrongdoer's mental states: the harm was either accidental or it was intentional and motivated. People typically attribute belief, desire, and foresight to intentional actions, which means that intentional behaviors encompass a range of mental states implicated in action (Kirfel & Lagnado, 2021; Quillien & German, 2021). Participants registered initial blame judgments using a 7-pt. Likert scale ("How much should Bob blame Jim for [condition-dependent action]?" 1 = Not at all, 4 = Somewhat, 7 = Very much, anchored at 1).

Stage two updated the stories with character information: the wrongdoer either had a relevant vice (e.g., dishonesty) or virtue (e.g., honesty). Participants were then allowed to update their blame judgments on a 7-point scale from -3 (less blameworthy) to 3 (more blameworthy) (the scale was anchored at 0 = the same amount), and to register judgements about the intentionality of the wrongdoing as a manipulation check. Participants then indicated the social desirability of 10 different traits presented at the end of the survey. Participants saw either the virtuous or vicious member of the pair and rated whether it was undesirable or desirable for someone to have the trait using a 7-pt. scale (1 = Very undesirable, 4 = Neither desirable nor undesirable, 7 = Very desirable, anchored at 4). Participants always saw one part of the selfless/selfish and honest/dishonest pairs.

Since we resolved any ambiguity about the agent's mental states in stage one, participants should *not* update their blame judgments if the effect of character on blame is only indirect. In contrast, participants *should* update their blame judgments if this effect is direct.

Results. There was no evidence for an effect of vignette on judgments of revised blame (p = .51). Because of this, there was no term for vignette in the models summarized below.

Three results strongly suggest that the effect of character on blame is direct. First, character information had a large effect on blame updating (Figure 2; $F(1, 145) = 161.8$, $p < .001$, $\eta^2_p = .53$, 90% *CI*[.44, .60]). Blame ratings *increased* after participants learned about a wrongdoer's vice (M = 1.76, 95% *CI*[1.48, 2.05]) but *decreased* after learning about a wrongdoer's virtue (M = -0.52, 95% *CI*[-0.81, -0.24]). Second, there was no evidence that the degree of updating was different across intentionality conditions for either the dishonest ($t(145) = 1.44$, p = .15, d = 0.37, 95% *CI*[-0.14, 0.88]) or honest conditions ($t(145) = -1.40$, p = .17, d = -0.35, 95% *CI*[-0.84, 0.15]). Third, judgements about the intentionality of the wrongdoing did not mediate the effect of character on blame updating (p = .25). Participants also rated the social desirability of different traits and rated dishonesty as highly socially undesirable (M = 1.28, 95% *CI*[1.09, 1.48]) and honesty as highly socially desirable (M = 6.60, 95% *CI*[6.40, 6.80]).

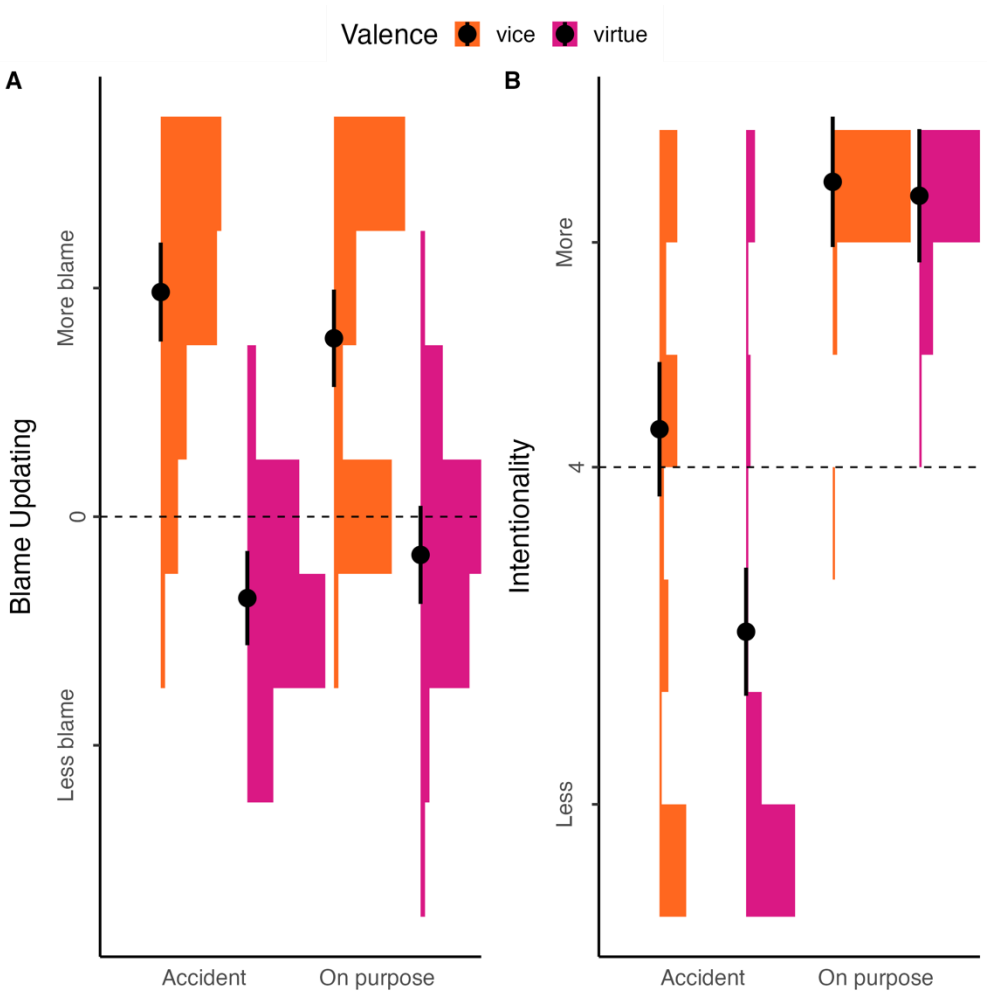


Figure 2. Judgments of revised blame (Panel A) and intention (Panel B) by valence and intentionality conditions. Error bars represent 95% confidence intervals on estimated marginal means.

Discussion. When mental state information is fixed, participants tend to update blame judgments after receiving character information. Vicious character information amplifies blame while virtuous character information mitigates blame. Neither intentionality information nor causal judgments seem to explain the effect of character information on blame updating. Thus, character information seems to have a direct effect on blame. To see how different kinds of character traits influence blame updating, we conducted another study.

Study 2: *Relevance* alters the effect of character on blame

Participants. Preregistration site is https://aspredicted.org/ZF5_XZ2. 160 participants were recruited on Amazon's Mechanical Turk. Sample size was determined using an *a priori* power analysis for a regression model with three predictors to detect a large effect size ($f = 0.3$) at 95% power with standard error thresholds ($p < .05$). 67 participants were recommended. Because we used different vignettes, we recruited full samples for each vignette. Effect sizes were estimated with the results of pilot studies. 1 participant was removed for not finishing the study ($N = 159$, $M_{age} = 30.7$, $SD_{age} = 11.2$, 39% female).

Methods and procedure. The Two Channel and Person Models both predict that there is a direct effect of character on blame, consistent with Study 1. However, the Two Channel Model further predicts that this direct effect is sensitive to whether the character trait is causally relevant to the wrongdoing. Whether one is blameworthy for theft, for example, should depend on whether one is honest or dishonest, because being dishonest might lead one to steal. But it should not depend on whether one is slobby or hygienic, because those are not causally relevant to stealing. In contrast, the Person Model predicts that character will bear on blame regardless of whether it is relevant (Rai, 2017).

To test these competing predictions, we adapted the two-stage paradigm from Study 1. We used a 2(vignette)x2(valence)x2(relevance) design. As in Study 1, after registering an initial judgment of blame for some wrongdoing, participants were told about the character of the wrongdoer. We manipulated both the valence of the character information and whether the traits were relevant (e.g., selfish/selfless) or irrelevant (e.g., unhygienic/clean) to the wrongdoing. Participants were then given the opportunity to update their judgment of blame and rated the extent to which the wrongdoing caused a bad consequence. Because all models predict that blame updates in opposite directions across vice and virtue conditions, we reverse-coded the virtue condition so that all blame judgments were on the same scale (-3 = inconsistent blame updating, 0 = no blame updating, 3 = consistent blame updating).

Results. We found that the effect of character on blame depends on whether character is relevant to the wrongdoing in question (Figure 3a; each model controlled for the effect of vignette). We found a moderate main effect of relevance on judgments of revised blame ($F(1, 154) = 14.27$, $p < .001$, $\eta^2_p = .08$, 90% *CI*[.03, .16]). People update blame judgments significantly after they receive character information that is relevant ($M = 0.76$, 95% *CI*[0.50, 1.03]) but not after receiving character information that is irrelevant ($M = 0.14$, 95% *CI*[-0.13, 0.40]). There was no evidence for an interaction between relevance and valence ($p = .49$). When testing revised blame scores against the indifference point ($M = 0$), there was no evidence that revised blame in the irrelevant condition was significantly different from 0 ($t(154) = 1.18$, $p = .24$, $d = 0.10$, 95% *CI*[-0.06, 0.25]).

Proponents of Person Models might object that our results are due to the social desirability of character traits, rather than their relevance to wrongdoing. Compared to unhygienic people, for example, dishonest people may have a more socially undesirable trait, be worse social partners, and thus receive more blame. To test this prediction, participants rated the social desirability of ten different traits, including those from our studies. There was no evidence of a difference in social desirability between relevant and irrelevant virtues ($t(313) = -2.14$, $p = .14$, $d = -0.35$, 95% *CI*[-0.68, -0.03]) or vices ($t(313) = -0.14$, $p = .99$, $d = -0.02$, 95% *CI*[-0.32, 0.28]) (see Figure 3b). Virtues were considered highly socially desirable, regardless of whether they were relevant ($M = 6.55$, 95% *CI*[6.39, 6.71]) or irrelevant ($M = 6.31$, 95% *CI*[6.14, 6.47]) and vices were considered highly socially undesirable, regardless of whether they were relevant ($M = 1.29$, 95% *CI*[1.14, 1.44]) or irrelevant ($M = 1.28$, 95% *CI*[1.13, 1.43]). Thus, our results are inconsistent with this alternative explanation.

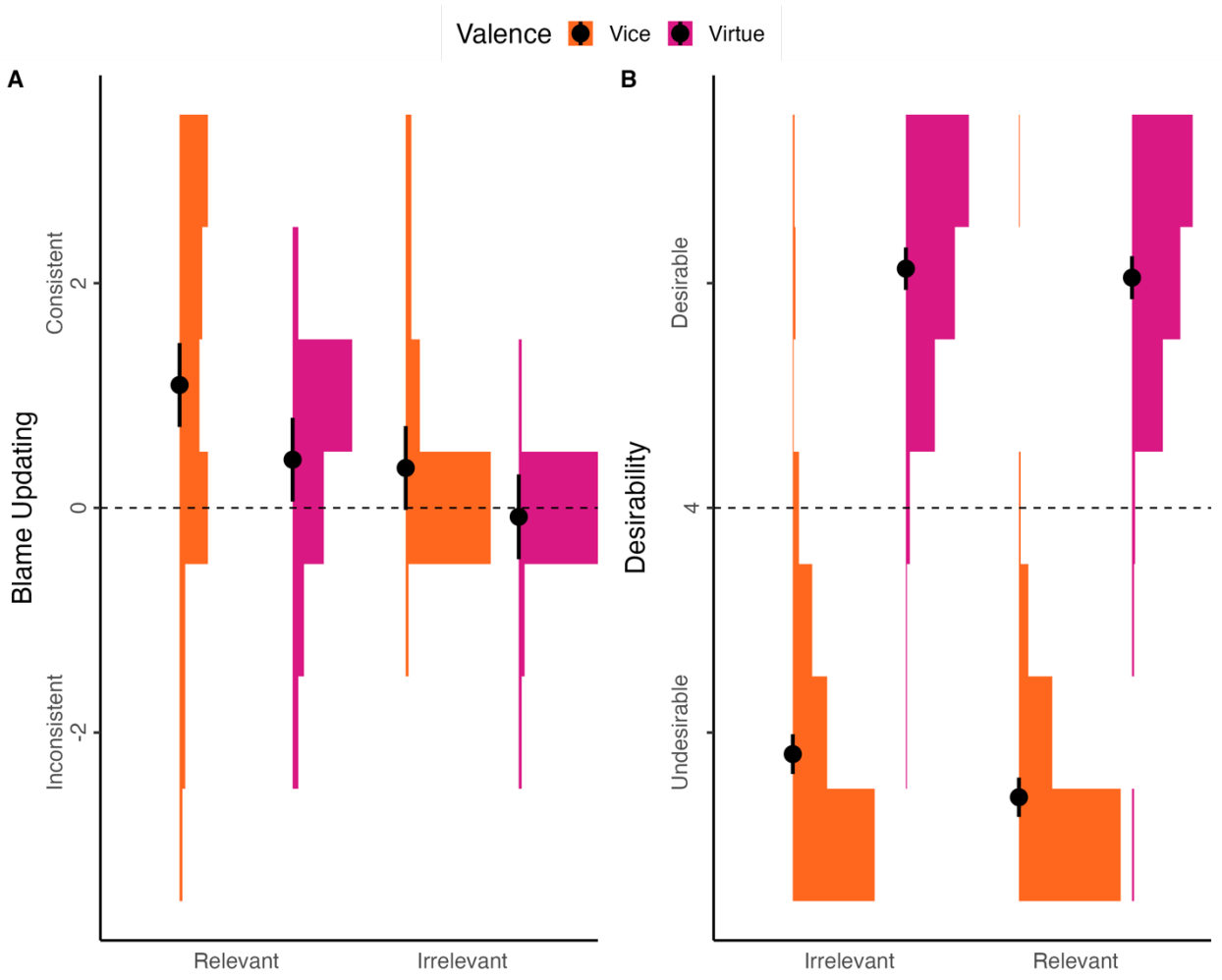


Figure 3. Judgments of revised blame (Panel A) and judgments of social desirability (Panel B) across relevance and valence conditions. Error bars represent 95% confidence intervals on estimated marginal means. Judgments of revised blame in the Virtue condition were reverse-coded. *Consistent* updating reflects revised judgments in the predicted direction, given only trait valence (i.e., consistent updating in vice conditions indicates stronger judgments of blame, while consistent updating for virtue conditions indicates diminished blame).

Discussion. The Two Channel model predicts that character information prompts blame updating when the underlying trait is relevant to the wrongdoing in quest. Consistent with this prediction, we found that people tend to update in predicted directions when they receive relevant character information but do *not* update when the receive irrelevant character information. This is not due to differences in the perceived undesirability of the traits in question, as relevant and irrelevant traits were rated as similarly desirable or undesirable.

Study 3: **The direct effect of character on blame depends on whether an action manifests one's true self.**

Participants. Preregistration site is https://aspredicted.org/9YD_CD4. 351 participants were recruited on Academic Prolific. Sample size was determined with a Monte Carlo simulation app

for mediation model power analyses (Schoemann et al., 2017). 1000 replications were performed, with 20,000 draws per replication. Standardized coefficients, mediator covariance, and standard deviations for the model were estimated from a pilot study (N = 100). For a multiple mediation model with two parallel mediators to achieve 95% power to detect differences in indirect effects, 330 participants were recommended. To account for exclusions, we over-recruited by 7% based on exclusion rates in previous studies. No participants were excluded according to our pre-registered exclusion criteria (described below). No data were analyzed prior to exclusions (final N = 351, M_{age} = 37.40, SD = 12.4, 50% female).

Methods and procedure. Hume predicts that actions are more blameworthy when they "proceed… from some cause in the characters and disposition of the person" (1955 [1739], p. 107). In contemporary parlance, people are more blameworthy for wrongdoing that reflects what kind of person they are deep down, compared to uncharacteristic lapses. The Two Channel Model predicts that the direct effect of character on blame depends on these Humean judgments about whether our actions reflect what kind of person we are. Mental state models, on the other hand, predict that relevant character information affects blame by supporting inferences about intentionality.

We again used a two-stage experiment to test this prediction. Participants first read a story where a character (Jim) intentionally takes his roommate's (Bob) ice cream. We made the action intentional to block indirect effects of character. Participants were then told that the roommate was either typically honest or dishonest. After this, they had the opportunity to update their judgment of blame. Participants responded to a novel item indexing judgments about the true self using a 7-pt. scale: "How much do Jim's actions reflect the kind of person he is deep down inside?" (1 = Not at all, 4 = Somewhat, 7 = Very much, anchored at 4). The Two Channel Model predicts that these "true self" judgments should mediate the effect character on revised blame. We also measured the intentionality of Jim's actions to control for possible indirect effects of character. Intentionality judgments should not mediate the effect of character on blame because intentionality does not vary across conditions.

Results. To test these predictions, we fitted a model to predict revised blame judgments by valence with two parallel mediators: intentionality and true self. As predicted, there was a significant indirect effect of condition on blame partially mediated by judgments of the true self (**β** = -0.22, SE = 0.05, p < .001) but not intentionality (**β** = -0.01, SE = 0.01, p = .28). There was also a remaining direct effect of condition on revised blame (**β** = 0.24, SE = 0.06, p < .001) (see **Figure 5**). Consistent with Study 1, our results indicate that character effects blame through a channel that bypasses the wrongdoer's mental states (intentionality). Consistent with the Humean perspective that *relevant* character information affects blame, this channel depends on whether the wrongful actions reflect one's true self.

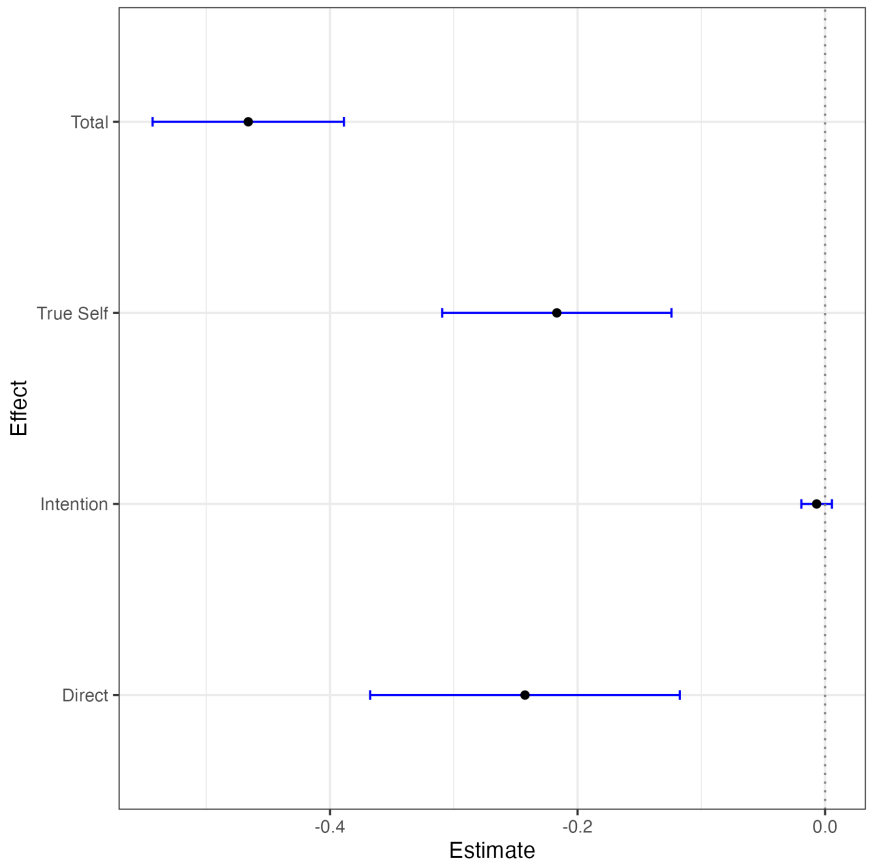


Figure 5. Standardized coefficients from mediation analysis in Study 3. Error bars represent 95% confidence intervals.

Discussion. The Two Channel model predicts that people are more blameworthy for wrongdoing that manifests the deep self. Insofar as character information sheds light on what one is like deep down inside, such information can prompt blame updating. The results of Study 3 confirm this prediction. Judgments about an individual's true self—who they are deep down—mediates the effect of character information on blame updating.

General Discussion

Summary of results. Hume argued that bad acts are blameworthy to the extent that they manifest objectionable character. Judgments of blame, then, are sensitive to character information. We find support for this prediction across three studies: participants consistently revise their initial blame judgments after learning about the wrongdoer's character.

Previously articulated models captured only *part* of how character informs blame. Study 1 found that character affected blame even after fixing the agent's intentions and motives. This is inconsistent with *Mental State Models*, which predict that character affects blame only by providing evidence about ambiguous mental states (intentions and motives). Study 2 found that character affected blame only when those traits are relevant to the wrongdoing in question. This is inconsistent with *Person Models*, which predict that irrelevant traits still provide evidence about who is a good partner, and thus should affect blame judgments. Study 3 found that character affects

blame independently of intentionality judgments (contra the Mental State Model) and instead depends on whether an action reflects the kind of person one is deep down inside (contra the Person Model). This is consistent with a Humean view, on which acts are more blameworthy when they reflect one's true self.

The *Two Channel Model* predicts all three findings. This model posits two channels through which character informs blame. One channel involves mental state inference: people use character information to resolve ambiguities about mental states. The second channel involves person perception: people integrate character information into their picture of the wrongdoer's true self.

In some ways, the Two Channel Model is therefore an empirical synthesis of Person and Mental State Models. Like Person Models, the Two Channel Model predicts that character information bears on blame independently of its role in mental state inference. This is why character has a direct effect on blame that bypasses intentions. Like Mental State Models, the Two Channel Model predicts that character information influences blame only when it is relevant to the wrongdoing. Selfishness makes one more blameworthy for selfish actions, for example, but bad hygiene does not.

In another way, the Two Channel Model makes novel predictions. Following Hume, the Two Channel Model also predicts that blame partly depends on whether an action reflects the kind of person one is deep down inside. Such deep-self judgments are not determined by judgments about whether someone is vicious or virtuous in general (unlike Person Models) or had mental states like intentions (as per Mental State Models).

The demands of morality. Our studies may appear limited, insofar as the stimuli focus on (a) minor moral transgressions that (b) occur within a close interpersonal relationship. Jim might spoil his roommate's favorite TV show, for example, rather than kill a stranger's cat. We restricted our stimuli in this way for methodological and theoretical reasons, which have crucial—but somewhat neglected—implications for moral judgment in everyday life.

The methodological point is that moral transgressions had to be somewhat minor for our vignettes to be believable. Given the factorial nature of our manipulations, specifically, it had to be believable that a *virtuous person* could intentionally commit the wrongdoing in question. This constrains the severity of wrongdoing. But this generates a theoretical constraint, since some behaviors are highly diagnostic of underlying character: doing them *even once* prompts attributions of the relevant trait (Chadwick et al., 2006). In particular, some acts are so heinous that virtuous people would *never do them*. Consider: participants would likely not believe that Jim is a genuinely kind person who simply happened—just this once!—to intentionally murder Bob's cat. In contrast, being virtuous is compatible with occasional relevant, but minor, transgressions. For example, even honest and selfless people occasionally act dishonestly and selfishly (Miller, 2013).

This points to an important difference between two domains of morality: one tolerates imperfection, whereas the other demands perfection. Much of what we do as agents admits of imperfection: careful people occasionally make mistakes, committed dieters sometimes backslide, and so on (Amaya, 2013; Sripada, 2018). The same is true of morality. Even virtuous people sometimes make (small) moral mistakes. Within limits, honest people can lie and caring people

can lash out. Such actions are still morally wrong; but it matters that they are *uncharacteristic wrongs*. In contrast, morality demands perfection with respect to norms prohibiting killing, child abuse, extreme violence, and so on. Murdering someone's pet, for example, is an area where morality demands perfection; taking ice cream or spoiling a movie is not. Within the imperfect domain, it matters whether the transgression is an uncharacteristic slip or part of a pattern. Within the perfect domain, the pattern is less important.

This illustrates an important point about moral psychology: there are major advantages to focusing on everyday cases of moral wrongdoing. Certain features of morality become evident when we focus on the kinds of moral situations people encounter in their daily lives, namely minor transgressions that occur within close relationships. We've already mentioned one such feature: everyday cases are more likely to involve situations where imperfection is tolerated, which would allow for certain patterns to emerge that might otherwise be obscured. Another feature relates to the function of blame judgments, which we discuss in detail below. We posit that blame can have a pedagogical function, for example, where we call upon (somewhat) vicious social partners to grow as people. After discovering that your dishonest roommate took your ice cream, you might stage an intervention and ask her to work on her disrespect for your property. The capacity to call for personal change is a crucial feature of everyday, interpersonal moral relationships. Yet it's unclear whether there's any equivalent for impersonal, serious moral harms.

There is another independent reason to focus on everyday moral transgressions: moral intuitions may be most reliable and stable in everyday contexts. Some have argued that serious, impersonal moral transgressions sometimes distort the logic of moral judgment by focusing on cases where people lack relevant experience (Kahane, 2015; Schein, 2020). People likely have highly refined intuitions about cheating detection in interpersonal relationships due to repeated exposure or testimony. People can draw on this experience to determine, say, how to treat a dishonest roommate. In contrast, it is unclear whether people have highly refined intuitions about how to treat anonymous murderers or cat killers. Most people have the good fortune to never encounter people so vicious, so may lack stable and reliable intuitions about such serious transgressions.

The functions of blame. *Why* are there two distinct channels from character to blame? Again, the Two Channel Model synthesizes insights from competing views. There is broad consensus that blame serves the function of *social regulation* (Malle et al., 2014; Vargas, 2013). Yet proponents of Mental State and Person Models focus on different kinds of social regulation.

Proponents of Mental State models argue that blame serves to enforce norms and facilitate norm internalization (Malle et al., 2014; Malle, 2020). Blame has a *protestive function*: we protest norm-violating acts so that wrongdoers and other community members recognize the importance of these norms and commit to complying with them in the future (Pereboom, 2021; McKenna, 2012). We blame people for stealing, for example, so that the community will respect personal property. Fulfilling the protestive function requires sensitivity to the wrongdoer's mental states. Intentional actions are controllable and guided by norms; accidents are neither controllable nor guided. So, protests of accidents are unlikely to change one's commitments or behavior going forward. Blame thus demands either an *apology* (which signals that the wrongdoer has internalized a normative protest) or excuse (which may signal that the harm was accidental) (Jefferson, 2019; Vargas,

2021). We agree that blame has this protestive function: this is what explains the indirect channel from character to blame.

Proponents of Person Models argue that blame is targeted at potential social partners rather than norm-violating actions (Uhlmann et al., 2015; Rai, 2017; Pizzarro & Tenenbaum, 2012). Blame, then, also has a *protective function*: blame identifies those to be avoided (Pizarro & Tannenbaum, 2012). For example, we blame the selfish to mark out undesirable social partners (Schwartz et al., 2022). This explains the direct channel from character to blame.[1]

A limitation of Person Models is that they assume that we have only two options with potential social partners: approach or avoid. This misses a third option. Wrong acts can be *teachable moments* used to illustrate more general character flaws. Blame has a *pedagogical function*, as well: blame can be a call for our social partners to grow as people. Blaming a roommate for taking your ice cream isn't just telling them to respect your stuff; it is also a call to be the kind of person that values respecting others and their things. This intervention is *much* more likely to be successful if you can point to (a) a particular action that (b) reflects an underlying character deficit. For this reason, the direct channel is sensitive to relevance. This, in turn, explains why the effect of character information is likely limited to the "imperfect" domain of morality. Serious moral transgressions strongly incline people toward the avoidance option. As severity increases, the preference for avoidance likely swamps any pedagogical considerations.

Character supports the protestive, protective, and pedagogical functions of blame. Understanding someone's character helps to interpret what they knew, wanted, or intended in acting badly. We need this mental state information to know whether—and how much—to protest their action. Understanding someone's character also lets us determine whether their wrongdoing reflects the kind of person they are deep down inside. When wrongdoing reflects a deep character deficit, blame presents us with a choice. We can protect ourselves and begin to wind down a relationship. Or we can take a pedagogical stance and call upon our social partner to improve. The fact that blame has these distinct functions explains why there are two channels from character to blame (Figure 6).

[1] People may *blame victims* who fail to protect themselves from vicious transgressors (thanks to an anonymous reviewer for proposing this hypothesis). People may blame Bob, for example, if Bob's dishonest roommate steals his ice cream (*you should have known better than to keep a dishonest roommate!*). Whether such victim blaming occurs––and how it interacts with the pedagogical function of blame—is an important question for future research. But such a possibility does not confound our results. If anything, victim blaming should *diminish* responsibility for vicious transgressors. But in every study, we find the opposite: people blame vicious transgressors more than virtuous ones. Victim blaming therefore (a) cannot explain our results and (b) if anything, may lead us to underestimate the direct effect of character on blame.

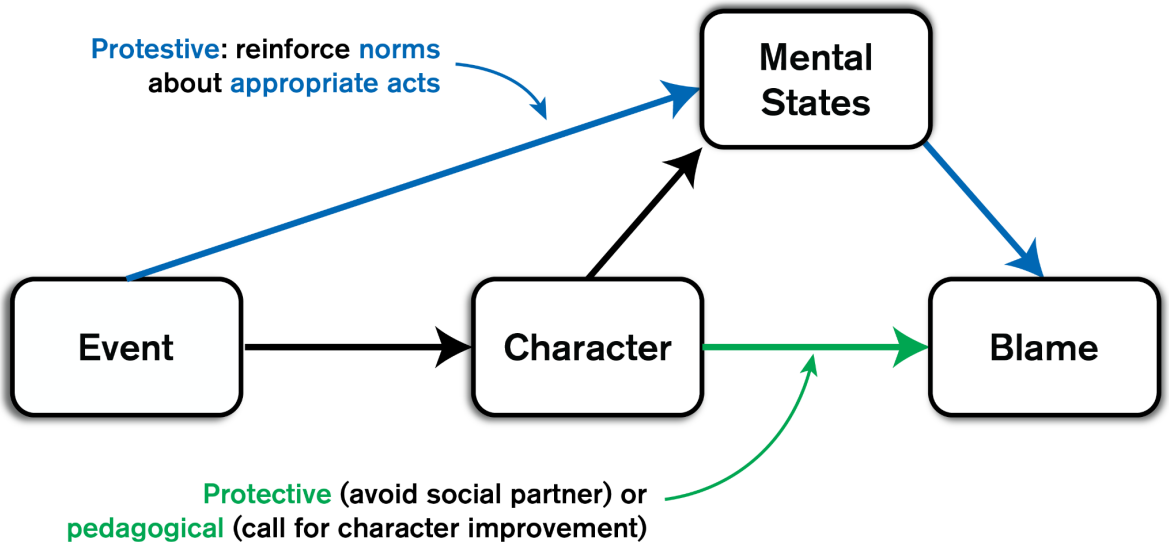


Figure 6. Summary of protestive, pedagogical, and protective functions of blame

References

Alicke, M. D. (1992). Culpable causation. *Journal of personality and social psychology*, *63*(3), 368.

Alicke, M. D., & Zell, E. (2009). Social attractiveness and blame. *Journal of Applied Social Psychology*, *39*(9), 2089-2105.

Amaya, S. (2013). Slips. *Noûs*, *47*(3), 559-576.

Chadwick, R. A., Bromgard, G., Bromgard, I., & Trafimow, D. (2006). An index of specific behaviors in the moral domain. *Behavior research methods*, *38*, 692-697.

Cushman, F. (2015). Deconstructing intent to reconstruct morality. *Current Opinion in Psychology*, *6*, 97-103.

Faul, F., Erdfelder, E., Lang, A. G., & Buchner, A. (2007). G* Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior research methods*, *39*(2), 175-191.

Guglielmo, S., & Malle, B. F. (2017). Information-acquisition processes in moral judgments of blame. *Personality and Social Psychology Bulletin*, *43*(7), 957-971.

Hughes, J. S., & Trafimow, D. (2012). Inferences about character and motive influence intentionality attributions about side effects. *British Journal of Social Psychology*, *51*(4), 661-673.

Hume, D. 1955 [1739]. *Treatise on human nature*, ed. L.A. Selby-Bigge (Oxford: Clarendon Press).

Irving, Z. C., Murray, S., Glasser, A., & Krasich, K. (2023). The catch-22 of forgetfulness: Responsibility for mental mistakes. *Australasian Journal of Philosophy*, 1-19.

Jefferson, A. (2019). Instrumentalism about moral responsibility revisited. *The Philosophical Quarterly*, *69*(276), 555-573.

Kirfel, L., & Lagnado, D. (2021). Causal judgments about atypical actions are influenced by agents' epistemic states. *Cognition*, *212*, 104721.

Koster-Hale, J., Saxe, R., Dungan, J., & Young, L. L. (2013). Decoding moral judgments from neural representations of intentions. *Proceedings of the National Academy of Sciences*, *110*(14), 5648-5653.

Malle, B. F. (2020). Graded representations of norm strength. In S. Denison, M. Mack, Y. Xu, & B. C. Armstrong (Eds.), Proceedings of the 42nd Annual Meeting of the Cognitive Science Society. (pp. 3342–3348). Cognitive Science Society.

Malle, B. F. (2021). Moral judgments. *Annual Review of Psychology*, *72*, 293-318.

Malle, B. F., Guglielmo, S., & Monroe, A. E. (2014). A theory of blame. *Psychological Inquiry*, *25*(2), 147-186.

McKenna, M. (2012). *Conversation & responsibility*. Oxford: Oxford University Press.

Miller, C. B. (2013). *Moral character: An empirical theory*. Oxford: Oxford University Press.

Nadler, J. (2012). Blaming as a social process: The influence of character and moral emotion on blame. *Law and contemporary problems*, *75*(2), 1-31.

Pereboom, D. (2021). Undivided Forward-Looking Moral Responsibility. *The Monist*, *104*(4), 484-497.

Pizarro, D. A., & Tannenbaum, D. (2012). Bringing character back: How the motivation to evaluate character influences judgments of moral blame. In M. Mikulincer & P. R. Shaver (Eds.), *The social psychology of morality: Exploring the causes of good and evil* (pp. 91–108). American Psychological Association. https://doi.org/10.1037/13091-005

Quillien, T., & German, T. C. (2021). A simple definition of 'intentionally'. *Cognition*, *214*, 104806.

R Core Team (2023). _R: A Language and Environment for Statistical Computing_. R Foundation for Statistical Computing,Vienna, Austria. <https://www.R-project.org/>.

Rai, T. S. (2017). Exile of the accidental witch. *Moral inferences*, 191.

Royzman, E., & Hagan, J. P. (2017). The shadow and the tree. In *Moral inferences* (pp. 56-74). London: Routledge.

Schoemann, A. M., Boulton, A. J., & Short, S. D. (2017). Determining power and sample size for simple and complex mediation models. *Social Psychological and Personality Science*, *8*(4), 379-386.

Schwartz, F., Djeriouat, H., & Trémolière, B. (2022). Agents' moral character shapes people's moral evaluations of accidental harm transgressions. *Journal of Experimental Social Psychology*, *102*, 104378.

Siegel, J. Z., Crockett, M. J., & Dolan, R. J. (2017). Inferences about moral character moderate the impact of consequences on blame and praise. *Cognition*, *167*, 201-211.

Sripada, C. (2018). Addiction and fallibility. *Journal of Philosophy* 115:11, 569 – 587.

Sytsma, J. (2021). Causation, responsibility, and typicality. *Review of Philosophy and Psychology*, *12*, 699-719.

Uhlmann, E. L., Zhu, L. L., & Tannenbaum, D. (2013). When it takes a bad person to do the right thing. *Cognition*, *126*(2), 326-334.

Uhlmann, E. L., Pizarro, D. A., & Diermeier, D. (2015). A person-centered approach to moral judgment. *Perspectives on Psychological Science*, *10*(1), 72-81.

Vargas, M. 2013. *Building better beings*. Oxford: Oxford University Press.

Vargas, M. (2021). Constitutive instrumentalism and the fragility of responsibility. *The Monist*, *104*(4), 427-442.

Woolfolk, R. L., Doris, J. M., & Darley, J. M. (2006). Identification, situational constraint, and social cognition: Studies in the attribution of moral responsibility. *Cognition*, *100*(2), 283-301.