






The Catch-22 of Forgetfulness: Responsibility for Mental Mistakes

Zachary C. Irving ^a, Samuel Murray ^b, Aaron Glasser^c and
Kristina Krasich ^d

^aUniversity of Virginia; ^bProvidence College; ^cUniversity of Michigan Ann Arbor; ^dDuke University

ABSTRACT

Attribution theorists assume that character information informs judgments of blame. But there is disagreement over why. One camp holds that character information is a *fundamental* determinant of blame. Another camp holds that character information merely provides *evidence* about the mental states and processes that determine responsibility. We argue for a *two-channel view*, where character simultaneously has fundamental *and* evidential effects on blame. In two large factorial studies ($n = 495$), participants rate whether someone is blameworthy when he makes a mistake (burns a cake or misses a bus stop). Although mental state inferences predict blame judgments, character information does not. Using mediation analyses, we find that character information influences responsibility via two channels (Studies 3–4; $n = 447$), which are sensitive to different kinds of information (Study 5; $n = 149$). On the one hand, forgetfulness increases judgments of responsibility, because mental lapses manifest an objectionable character flaw. On the other hand, forgetfulness decreases judgments of state control, which in turn decreases responsibility judgments. These two channels cancel out, which is why we find no aggregate effect of forgetfulness on responsibility. Our results challenge several fundamental assumptions about the role of character information in moral judgment, including that good character typically mitigates blame.

ARTICLE HISTORY Received 2 July 2021; Revised 30 May 2022

KEYWORDS responsibility; character; mental action; forgetfulness; blame

1 Introduction

I, your lead author, am deeply forgetful. I leave fridges open, miss appointments, and lose all manner of possessions. In contrast, my conscientious wife rarely has such mental lapses. Yet when mistakes happen, people blame us in perplexing ways. In one way, people blame me more. My mental lapses are part of a general pattern, the manifestation of a character flaw. In another way, people blame my wife more. People hold her to a higher standard: they assume that she is in control of her thoughts and actions, and thus is responsible for their consequences. It's a catch-22 [Heller 1961]: the forgetful are blamed because they *lack* control and the conscientious are blamed because they *have* control. You're blamed no matter what.

This catch-22 has implications for debates about how people use character information in making judgements of blame.¹ In this debate, there are (at least) two families of competing philosophical views [Pizarro *et al.* 2012]. On act-based views, people make blame judgements by considering the proximal causes of wrongful action [Cushman 2015; Lagnado and Gerstenberg 2015]. That is, when we blame someone, we fundamentally aim to make inferences about her mental states (for example, intentions) and causal contributions to a bad outcome [Young and Tsoi 2013; Malle *et al.* 2014]. Thus, character information merely provides evidence about the wrongdoer's mental states. It is those mental states that are directly relevant to blame [Malle *et al.* 2014: 17]. On person-based views of blame, people justify blame judgements by considering the character of the wrongdoer. Ultimately, when we blame someone, we aim to make inferences about the wrongdoer's character from their actions [Uhlmann *et al.* 2015]. Thus, character information is directly relevant to blame: when deciding whether to blame someone, the fundamental question is whether her actions manifest deficient or bad character [Brandt 1958; Nozick 1981; Woolfolk *et al.* 2006; Pizarro *et al.* 2012; Sripada 2016].

This disagreement raises an empirical question: How does character information inform ordinary judgements of responsibility? Empirical research on this question delivers mixed results. Some find that character information directly affects responsibility attributions. For instance, one study compared individuals who played either fairly or unfairly during an economic cooperation game (which presumably reflected on their character). People blamed the unfair participants more for an unrelated harm, which suggests that character has a direct effect on responsibility judgements [Kliemann *et al.* 2008], as predicted by the person-based view. However, recent studies found that effects of character on judgements of blame disappear when suitable mental state information is provided [Royzman and Hagan 2017]. This suggests that character assessments merely function to provide additional evidence about an agent's mental states. Further support for the act-based picture comes from findings that character information anchors spontaneous inferences about an individual's motives and desires, which then guide attributions of responsibility [Koster-Hale *et al.* 2013; Cushman 2015].

The catch-22 of forgetfulness suggests that person- and act-based views each provide only part of the story. In one way, people take character to be fundamental: they blame forgetful people for mental lapses because those mistakes manifest a character flaw. In another way, people treat character as evidence: they excuse forgetful people because they assume that our negligence does not reflect malicious intent. Character may therefore simultaneously influence responsibility through two channels—one fundamental and one evidential. These channels can push in opposite directions, which generates the catch-22 of forgetfulness. If character affects responsibility through two channels with counteracting effects, this may explain why past empirical research on character has found seemingly inconsistent results.

In the course of our investigation, we empirically tested person-based, act-based, and two-channel views in the context of control. There are three reasons for this

¹ We focus on responsibility as *accountability*, as do our experiments. Some individual is responsible in the accountability sense when their conduct makes them an appropriate target of negatively-valenced reactive attitudes, such as resentment [Shoemaker 2015]. Thus, accountability involves some culpability-imputing judgement that is non-trivially connected to moral emotions. Moreover, while some forms of responsibility mark out deserved credit or praise for right action, we limit our discussion to blame and the reactive attitudes associated with blameworthiness.

approach. First, philosophers have argued on theoretical [Fischer and Ravizza 1998; Vargas 2013; Murray and Vargas 2020] and empirical [Malle *et al.* 2014; Murray *et al.* 2019; Murray *et al.* forthcoming] grounds that one's responsibility for an outcome depends on whether the outcome is under one's control—you wouldn't blame someone if she bumped you because of an uncontrollable twitch. Second, evidential views claim that character information is used to make inferences about factors that constitute control. Third, traits that are related to control (such as forgetfulness or conscientiousness) can be manipulated in experimental contexts without inviting further inferences about the moral character of the individual. Traits related to self-governance are not necessarily moral [Kupperman 1995: 7], so these manipulations do not introduce further confounds that might make results difficult to interpret.

We also believe that previous research on the relationship between control and responsibility has been narrowly focused on what we call 'state control'. State control refers to whether one can bring about or prevent a particular outcome *here and now*. But responsibility may also depend on 'trait control'—that is, whether someone tends to be in control of her thoughts and actions—which manifests in character traits like forgetfulness and conscientiousness. By discussing trait control, we hope to make progress on three crucial topics in moral psychology: responsibility, control, and character.²

We use a vignette-based method to test whether and how information about trait control influences blame. In two large factorial studies ($n = 495$), participants read stories where someone makes a mistake (burns a cake or misses a bus stop) because their mind is elsewhere. We systematically manipulated the character's state control and trait control (character) and asked whether they were responsible for burning the cake. In both studies, we found that while state control significantly predicted blame, trait control had no effect. Thus, character information had no overall effect on responsibility judgements.

One possible explanation for this null effect is that trait control influences responsibility only when it provides evidence for state control. If so, character may become insignificant when participants are explicitly told about state control [Sytsma, *in preparation*]. We tested this prediction in three other studies ($n = 596$), where we manipulated only trait control and examined whether its effect is mediated by assumptions about state control. In these studies, mediation analysis suggested that character control influenced responsibility via two channels: (i) Forgetfulness decreases judgements of state control, which in turn decreased responsibility judgements; and (ii) forgetfulness also increased judgements of responsibility because mental lapses manifest an objectionable character flaw. These two channels cancelled out, which is why we found no aggregate effect of forgetfulness on responsibility. This is the catch-22 of forgetfulness.

2 Study 1: Character and Trait Control

2.1 Methods

The materials, data, and analyses for all studies reported here are available on the OSF page for the project: <https://osf.io/eqb2f/>. All participants provided electronic consent

² Previous research in the moral psychology of character has largely ignored trait control. This is part of a general trend. Past research has neglected character traits that concern one's capacity for *self-governance* (e.g., forgetfulness), and instead focused on *other-directed* traits (e.g., fairness and kindness). We argue that both kinds of character traits affect responsibility, but in different ways (see discussion).

following procedures approved by the University of Virginia's Institutional Review Board for the Social and Behavioral Sciences.

2.1.1 Participants

256 participants were recruited through Academic Prolific. We determined sample size with *a priori* power calculations using G*Power [Faul *et al.* 2007]. For a 2×2 ANOVA to have 95% power to detect the predicted effect sizes ($f = 0.27$) at standard error thresholds ($p < .05$), 236 participants were recommended. We over-recruited by 10% to account for exclusions. 16 participants failed an attention check ($N = 240$, $M_{\text{age}} = 31.6$ years; $SD_{\text{age}} = 11.0$, 35% female). We used two qualification conditions to restrict participation: participants needed to be fluent in English, and be based in the United States.

2.1.2 Materials and Procedures

The materials included written vignettes in which a character (Randy) burns his friend's birthday cake because his mind is elsewhere. In a 2×2 design, the vignettes varied Randy's state (high vs. low) and trait (forgetful vs. conscientious) control.

Box 1: Sample vignette from Study 1. The state control manipulations included high (SMALL CAPS) and low control (bold). The trait control manipulation included forgetful (italics) vs. conscientious (underline).

Randy typically has [*little*] control over his thoughts: he is a [conscientious / *forgetful*] person who [*rarely* / *frequently*] gets distracted, [*even*] when he is doing something important. Today, he puts a cake in the oven, which he promised to bake for a close friend's birthday party, when he finds himself [LEISURELY THINKING / **repeatedly worrying**] about various things: starting a new job, buying a car, going on a date tomorrow ... Randy has [A LOT OF / **little**] control over his thoughts today: Randy's mind is [WANDERING AND HE COULD / **racing, but he could not**] easily pull himself back to what he's doing (baking the cake). Because Randy's mind is [WANDERING / **racing**], he forgets to take the cake out of the oven when it's ready. The cake is burned and it's too late for Randy to buy another one from the store. Randy's friend will be sad because now she won't have any dessert on her birthday.

Participants read one vignette and then answered the following question: 'How much should Randy's friend blame him for burning the cake?' (1 = not at all, 7 = very much; midpoint not labelled).

Some previous studies manipulated character indirectly, by altering the subject's current mental states (specifically her motives; [Alicke 1992]). For example, Alicke [1992] alters a driver's character by manipulating whether he is speeding in order to hide cocaine or an anniversary gift. We instead manipulated character directly, by specifying the agent's *typical* behaviour and thoughts (rather than her *current* state of mind), to disentangle the effects of state control and character, which would be confounded if we manipulated character by altering the subject's mental states.³

Participants also answered two questions to assure that our manipulations altered perceptions of state control ('How much control does Randy have over his thoughts?') and trait control ('What type of person is Randy typically?'). Participants

³ Participants had to infer trait information based on short text descriptions. Previous research has shown that people reliably infer dispositions from small amounts of information [Willis and Todorov 2006], including short narrative descriptions or single terms [Peabody and Goldberg 1989]. Moreover, these inferences are relatively stable within social groups [Stolier *et al.* 2020].

chose between two options: ‘A little’ and ‘A lot’ for state control and between ‘Conscientious’ and ‘Forgetful’ for trait control. Participants then assessed the causal relevance of Randy’s thinking (‘How much do Randy’s thoughts lead him to burn the cake?’) using a 7-point Likert scale (1 = not at all, 7 = very much; midpoint not labelled).

2.2 Results

Results are summarized in Figure 1a and Table 1. A 2 (state control: high vs low) x 2 (trait control: conscientious vs forgetful) between-subjects ANOVA showed a main effect of state control on blame ($F(1, 236) = 36.64, p < 0.001, \eta^2 = .13$), but no effect of trait control on blame ($F(1, 236) = 1.8, p = .179, \eta^2 = .01$) and no interaction between state and trait control ($F(1, 236) = .00, p = .953, \eta^2 = .00$). Participants attributed more blame to Randy in the high state control condition relative to the low state control condition. However, participants attributed statistically equivalent degrees of blame to Randy in both the high (conscientious) and low (forgetful) trait conditions. This null effect was not due to participants misunderstanding our trait control manipulation, as most participants in both conditions correctly answered comprehension questions about Randy’s trait (90% correct) and state control (78.7% correct).

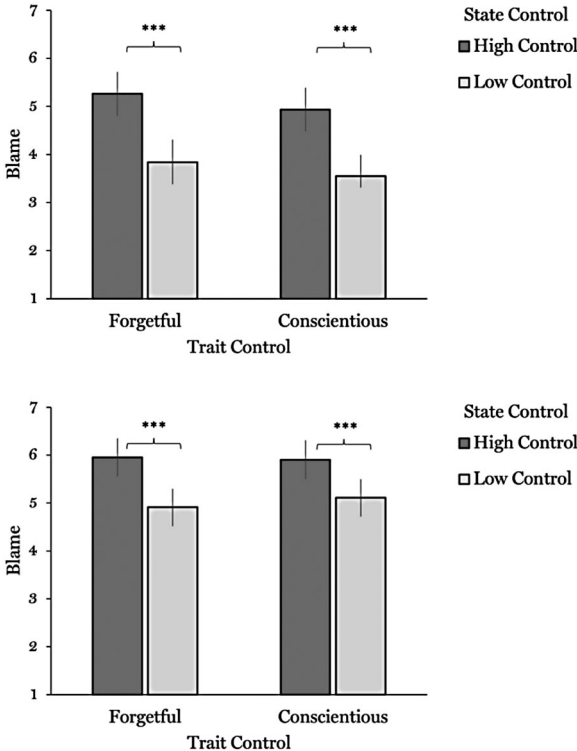


Figure 1: Mean Blame Ratings by Trait Control and State Control for Studies 1 and 2. Error Bars Represent 95% Confidence Intervals. *** $p < .001$.

Table 1: Mean Blame Ratings for Studies 1 and 2 by Trait Control (Conscientious vs Forgetful) and State Control (High vs Low).

A: Study 1					
State Control	Trait Control	Mean	SE	95% Confidence Interval	
				Lower	Upper
High Control	Forgetful	5.26	0.235	4.80	5.72
	Conscientious	4.93	0.231	4.48	5.39
Low Control	Forgetful	3.84	0.235	3.38	4.31
	Conscientious	3.55	0.224	3.31	3.99
B: Study 2					
High Control	Forgetful	5.95	0.204	5.55	6.35
	Conscientious	5.90	0.204	5.50	6.31
Low Control	Forgetful	4.91	0.201	4.51	5.30
	Conscientious	5.11	0.199	4.71	5.50

One might worry that blame ratings are biased by *causal* judgements. To control for this confound, we tested whether causal judgements mediate the effect of state control on responsibility. The mediation analysis first calculates an overall effect of state control on responsibility (this should be significant, given that our ANOVA found a significant effect of state control). It then decomposes that overall effect into indirect and direct effects. Indirect effects of state control on responsibility are a function of the degree to which causal judgements alter responsibility judgements. Having isolated the indirect effect, we are then able to assess direct effects of state control on responsibility.

If judgements of blame reflect judgements that Randy is merely *causally* responsible for burning the cake, then we should find a large (and significant) *indirect* effect and a small or non-significant direct effect. If judgements of responsibility are distinct from causal judgements, then we should find a large (and significant) *direct* effect with a small (or insignificant) indirect effect. Mediation analyses found a significant overall effect of state control on blame ($p < .001$) and a significant direct effect of state control on blame ($p < .001$), but no significant indirect effect of state control on blame, as mediated by causal judgements ($p = 0.195$).⁴ This suggests that the effect of state control on blame was not driven by causal judgements.

3 Study 2: Replication

Study 1 showed no effect of trait control (forgetfulness vs. conscientiousness) on judgements of blame. This null effect runs counter to previous research, which suggests that character information informs blame [Pizarro *et al.* 2012; Malle *et al.* 2014; Sripada 2016]. We therefore aimed to replicate the effects from Study 1 with different materials.

3.1 Methods and Results

256 participants were recruited through Amazon's Mechanical Turk. We used three qualifications to restrict participation: (a) participants needed to be located in the

⁴ To carry out our mediation analysis, we used a bootstrapping procedure with 5000 samples to compute bias-corrected confidence intervals. Unsurprisingly, the overall effect of state control on blame was significant ($p < .001$, $b = 1.46$, 95% CI [1.02, 1.90]). Importantly, the indirect effect, mediated by causal judgements, was not significant ($p = .065$) whereas the direct effect was significant ($p < .001$, $b = 1.41$, 95% CI [0.97, 1.85]). This is the opposite of what one would predict if causal judgements drove the effect of state control on responsibility.

United States, (b) have earned the ‘Masters’ label⁵ and (c) have an approval rate above 90%. Sample size was based on the same power analysis as Study 1. One extra participant was collected because of simultaneous study enrolment. Two participants failed an attention check ($N = 255$, $M_{\text{age}} = 39.5$ years; $SD_{\text{age}} = 11.3$, 43.3% female). Participants read the following vignette:

... Today, [Ray] is riding the bus to meet a friend for lunch, when he finds himself repeatedly worrying about various things: buying a new cellphone, booking a trip, joining a gym ... Because Ray’s mind is racing, he misses his stop. By the time he notices, he no longer has time to backtrack and meet his friend. Ray’s friend is sad because she spent her lunch break waiting for Ray.

All other elements of our design were identical to Study 1.

The findings are illustrated in Figure 1b. A 2×2 between-subjects ANOVA (state control X trait control) showed a significant and medium-sized effect of state control ($F(1, 249) = 20.71$, $p < .001$, $\eta^2 = .08$), with greater judgements of blame for high state as opposed to low state control. Just like in Study 1, judgements of blame did not vary across high and low trait control, ($F(1, 249) = .14$, $p = .705$, $\eta^2 = .00$) and there was no significant state by trait interaction ($F(1, 249) = .38$, $p = .537$, $\eta^2 = .00$). As with Study 1, most participants correctly answered questions about trait control (96.5% correct) and state control (85% correct).

We conducted mediation analysis to test whether the effect of state control on blame is confounded by causal judgements. As in Study 1, we found a significant direct effect of state control on responsibility ($p < .001$) but no significant indirect effect of state control on blame, mediated by causal judgements ($p = .267$).⁶ Our results suggest that judgements of responsibility are not biased by judgements of causal relevance.

4 Study 3

Studies 1 and 2 investigated the factors that inform judgments of blame for mental lapses. Blame varied significantly as a function of whether someone has control over their thoughts. However, contrary to the predictions of the person-based view of blame, blame did not vary significantly as a function of character information. This null effect runs contrary to the received wisdom in moral psychology, which says that character information informs responsibility judgments.

The act-based model offers one potential explanation of this null effect. For the act-based theorist, trait control should affect responsibility judgments only when no explicit information about state control is available. When faced with this informational deficit, people use character information to make inferences about the agent’s state control, which they subsequently use to attribute responsibility. When you learn that Randy is habitually forgetful, for example, you may infer that he had little control over his thoughts when he burned the cake today. You may then judge that his mistake did not stem from malicious intent, thereby mitigating blame [Alicke

⁵ The Masters label is a performance-based distinction given to Mechanical Turk workers who demonstrate exemplary performance. Masters workers must maintain a high level of performance to retain the label.

⁶ We used a bootstrapping procedure with 5000 samples to compute bias-corrected 95% confidence intervals. We again found significant overall ($p < .001$, $b = .92$, 95% CI [.50, 1.29]) and direct effects ($p < .001$, $b = .96$, 95% CI [.56, 1.35]) of state control on blame. The indirect effect of state control on blame, mediated by causal judgements, was not significant ($p = .267$).

2000; Young *et al.* 2011]. In contrast, the act-based model predicts that character information is no longer useful when people are *explicitly* provided with information about state control [Sytsma *In preparation*]. Consider participants in our previous studies, who are explicitly told whether Randy was in control over his thoughts today. Our participants needn't use Randy's character to *infer* whether he was in control today; they already know! Under these circumstances, the act-based model suggests that character information should not predict blame. But character should become predictive when information about state control is absent. To test this, we ran an additional study where we removed any explicit information about state control while manipulating trait control.

4.1 Methods

4.1.1 Participants

234 participants were recruited through Amazon's Mechanical Turk. The same qualifications as Study 2 were used to pre-screen participants. Sample size was determined with *a priori* power calculations using G*Power. For a linear regression to have 95% power to detect predicted effect sizes ($f^2 = .06$) at standard error thresholds ($p < .05$), 219 participants were recommended. We recruited an additional 15 participants to account for exclusions. 7 participants did not complete the study and 2 failed attention checks ($N = 225$, $M_{\text{age}} = 39.6$ years; $SD_{\text{age}} = 10.9$ years, 39% female).

4.1.2 Materials and Procedures

In a between-subjects design, each subject read about Randy burning his friend's birthday cake and were asked, 'How much should Randy's friend blame him for burning the cake?' Participants indicated their response with a 7-pt. Likert scale (1 = not at all, 7 = very much; midpoint not labelled). This time, there were only two conditions, which varied in terms of trait control (forgetful vs conscientious) (**Box 2**). State control was not explicitly manipulated or specified in the vignettes. Instead, participants were asked the following about state control: 'Could Randy have easily stopped his mind from wandering (and remembered the cake)?' (1 = not at all, 7 = very much; midpoint not labelled). This was done to test whether assumptions about state control mediate the effect of forgetfulness on blame. As in Studies 1 and 2, we included a manipulation check to ensure that our conditions altered beliefs about trait control.

Box 2: Manipulations in Study 3. Trait control manipulation: forgetful (italics) vs. conscientious (underline). No information about state control was provided.

Randy typically has [*little*] control over his thoughts: he is a [*forgetful/conscientious*] person who [*frequently/rarely*] gets distracted, [*even*] when he is doing something important. Today, he puts a cake in the oven, which he promised to bake for a close friend's birthday party, when he finds himself thinking about various things: starting a new job, buying a car, going on a date tomorrow ... Because Randy's mind is wandering, he forgets to take the cake out of the oven when it's ready. The cake is burned and it's too late for Randy to buy another one from the store. Randy's friend will be sad because now she won't have any dessert on her birthday.

4.2 Results

The act-based theory makes two predictions. First, assessments of trait control should significantly predict blame when we do not account for state control (the person-based

model makes the same prediction). Second, the act-based model predicts that this effect should disappear when we control for state control [Livengood, Sytsma, and Rose 2017]. Our findings revealed a pattern of results inconsistent with both predictions.

To test the first prediction, we followed the typical approach in the literature [Kliemann *et al.* 2008; Siegel *et al.* 2017; Lynch *et al.* 2019]. We measured the effect of character assessment on blame by examining overall blame as a function of condition. An independent samples t-test found no significant overall effect of character on blame ($p = 0.449$). This null effect cannot be attributed to participants misunderstanding our trait control manipulation, since a chi-square test of independence confirmed that this manipulation altered beliefs about trait control ($p < .001$; $\chi^2 = 148$). This disconfirms the act-based model's first prediction (which is also a core prediction of the person-based model) and contrasts with past empirical studies that have found an overall effect of character on blame ratings [Young and Tsoi 2013]. As with Studies 1 and 2, most participants correctly answered a manipulation check (91.1% correct).

Act-based theorists have rightly critiqued this traditional analytic approach [Livengood, *et al.* 2017; Sytsma, *in preparation*]. Researchers have often studied the effect of character on blame *without* controlling for beliefs about the agent's mental states (as we did in our t-test). But if people use character information to make inferences about mental states, the traditional analytic approach may distort the effect of character on blame.

To correct this distortion, we analysed whether state control mediates the effect of character on blame.⁷ Similar to the mediation analyses used in Studies 1 and 2, this analysis decomposes the overall effect of character assessments on responsibility judgments into two channels: one indirect and the other direct. The indirect channel measures the extent to which character influences blame ratings by altering judgments about state control. After we remove this indirect channel, we are left with the direct effect of character on judgments of responsibility.

If the act-based model is correct, then we should observe a large (and significant) indirect effect of character on blame mediated by state control, as this would indicate that people are using information about trait control to infer information about mental states. If the person-based model is correct, then we should observe a large (and significant) direct effect of character on blame, as this would indicate that people are making judgments of blame in accordance with perceptions of underlying character traits. However, if the two-channel model is correct, we should observe both significant (and roughly equivalent) direct and indirect effects of character on blame, because the two-channel model posits that character information is used to inform judgments of blame both indirectly—to infer mental states—and directly.

To carry out our mediation analysis, we used a bootstrapping procedure with 5000 samples to compute bias-corrected 95% confidence intervals (CI). We found a significant indirect effect of character on blame, mediated by state control ($b = -0.45$, $p < .001$, 95% CI $[-0.23, -0.71]$). The direct effect of character on blame was also significant, but in the opposite direction ($b = 0.63$, $p = 0.006$, 95% CI $[-0.18, 1.08]$). Because these effects were in opposite directions and of similar size, the total effect of forgetfulness on blame was insignificant ($p = .451$) (Figures 2a and 3a).

⁷ We modelled state control as a mediator, rather than a covariate, because we hypothesized that state control judgments should depend on our character manipulation.

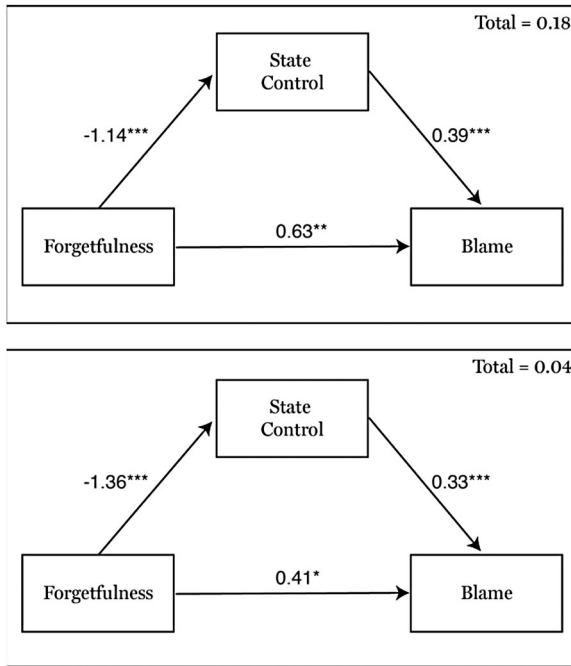


Figure 2: Mediation analysis from Studies 3 and 4 with state control mediating the effect of forgetfulness on blame. * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$

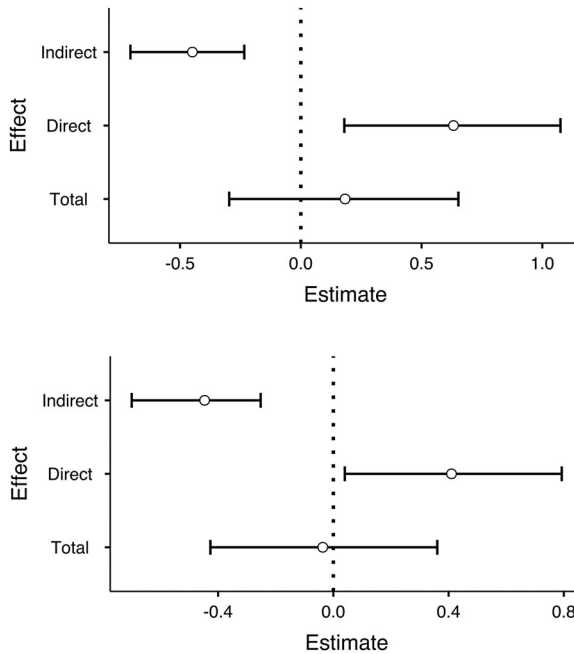


Figure 3: Results from mediation analysis, with bias-corrected 95% confidence intervals (Studies 3 and 4). The indirect effect of forgetfulness on blame, mediated by state control, is significantly negative. The direct effect is significantly positive. Because these effects are of similar in size and in opposite directions, the aggregate effect of forgetfulness on blame is insignificant.

Results from our mediation analysis are not consistent with the act-based model. Act-based theorists are technically correct that varying character (in our case, trait control) without controlling for mental states (in our case, state control) substantially distorts the effect of character on blame. But we found that this distortion goes in the *opposite direction* from what is predicted by the act-based theory. If character assessments affect blame only because it provides *evidence* about state control, character should have no direct effect when we include state control as a mediator. We found the opposite: the effects of character become evident only when we *include* state control as a mediator. The act-based theory therefore cannot explain our results. Yet results from our mediation analysis are also not consistent with the person-based model, which predicts that there should be no significant indirect effect.

In contrast, our results support the two-channel model, which predicts that character information both (i) directly increases blame and (ii) decreases assessments of state control, which indirectly decreases blame. Because the direct and indirect effects of forgetfulness on blame are predicted to be in opposite directions, the overall effect of character on blame is insignificant. This is the catch-22 of forgetfulness.⁸

5 Study 4: Replication

The results of Study 3 revealed that character information had significant indirect and direct effects on blame. However, as these effects are in opposite directions, the overall effect of character on blame is insignificant. Since these results run contrary to standard models of responsibility attribution, we aimed to replicate them in a different sample with a new vignette.

5.1 Methods and Results

We used the sample-size calculations from Study 3 and recruited a new sample of 234 participants on Amazon's Mechanical Turk. One participant failed attention checks, for a total of 233 participants (35% female; $M_{\text{age}} = 36.9$ years; $SD_{\text{age}} = 11.2$ years). Participants were presented with the vignette that was used in Study 2. Materials and procedures are otherwise the same as those used in Study 3.

An independent samples t-test found no significant overall effect of character on blame ($p = 0.858$). Furthermore, a chi-square test of independence confirmed that this manipulation altered beliefs about trait control ($\chi^2(1, n = 235) = 171, p < .001, 92.7\%$ correct responses). We then performed a mediation analysis to determine whether the two-channel theory can explain this null effect. We used a bootstrapping procedure with 5000 samples to compute bias-corrected confidence intervals. We found significant direct ($b = 0.41; p = 0.029; 95\% CI [0.03, 0.78]$) and indirect ($b = -0b = -0.45; p < .001; 95\% CI [-0.70, -0.24]$) effects of forgetfulness on blame. Again, these effects cancelled out so the total effect of forgetfulness on blame was insignificant ($p = .86$) (Figures 2b and 3b).

⁸ Using a multiple mediation analysis, we confirmed that neither the direct nor indirect effects of character on blame were fully mediated by judgments of causal responsibility. We present these additional analyses in the OSF page for the project: <https://osf.io/eqb2f/>.

6 Study 5

Studies 3 and 4 found that character information influences blame through two channels, one indirect and the other direct. Character indirectly influences blame by providing evidence about state control: forgetful people tend to have less control over their thoughts, and thus are less blameworthy for mental lapses. Character also exerts a *direct* influence on blame, which runs in the opposite direction. The two-channel view predicts that this direct effect represents a fundamental effect of character on blame. That is, we blame people more for actions that reflect the kind of person they are than we do for actions that are outside of their usual character. Study 5 tests this explanation of the direct channel.

6.1 Methods

6.1.1 Participants

149 participants were recruited through Academic Prolific. Sample size was determined with a Monte Carlo simulation app for mediation model power analyses [Schoemann *et al.* 2017]. 2000 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, 145 participants were recommended. We recruited 149 participants to account for exclusions using the same attention check as previous studies. However, no participants were excluded ($N = 149$, $M_{\text{age}} = 38.01$; $SD_{\text{age}} = 12.7$; 49.7% female).

6.1.2 Materials and procedure

We used the vignette from Study 1, although we used a two-stage updating paradigm to isolate how character informs responsibility judgments [Monroe and Malle 2019]. In stage 1, all participants initially saw the same vignette where Randy burned a cake and had minimal state control: ‘While Randy was baking the cake today, he had little control over his wandering mind: he was lost in thought and could not easily pull himself back to what he was doing (baking the cake)’. At this stage, participants were not provided with information about Randy’s character. Participants’ initial judgments about blame and state control were recorded, using questions adapted from Study 3. Participants then answered a new question: ‘Do Randy’s actions reflect the kind of person he is deep down inside?’ This question was designed to test whether Randy’s actions reflect his deep self. All answers were provided on a 7-point sliding scale (1 = Not at all, 4 = Somewhat, 7 = Very much) anchored at the midpoint.

Participants were then instructed, on a separate screen, that they would be given more information about the situation. The new information specified whether Randy was forgetful or conscientious, and was presented underneath the previously viewed scenario. In the forgetful condition, participants were told that: ‘Randy typically has little control over his thoughts: he is a forgetful person who frequently gets distracted, even when he is doing something important’. In the conscientious condition, participants were told that: ‘Randy typically has control over his thoughts: he is a conscientious person who rarely gets distracted when he is doing something important’. After seeing the new information, participants were asked whether, given what they

know now, they would change any of their previous judgments, via the following three questions:

1. *Revised blame*: Do you think Randy's friend should blame him less, more, or the same amount?
2. *Revised control*: Do you think it would be less easy, more easy, or just as easy for Randy to stop his mind from wandering?
3. *Revised deep character*: Do you think Randy forgetting the cake reflects less, more, or the same amount about the kind of person he is deep down inside?

To prevent ceiling effects, participants were provided with three 7-point scales ranging from -3 to 3 (-3 = A lot less; 0 = The same amount; 3 = A lot more) anchored at the midpoint. As in Studies 1–4, participants answered manipulation checks to ensure that our conditions manipulated beliefs about Randy's character.

6.2 Results

Independent samples *t*-tests found no evidence for significant differences in judgments of initial blame, control, or deep character across conditions (all $p > 0.84$).

To examine revised judgments, we fitted a multiple mediation model with two parallel mediators using the *lavaan* package in R [Rossee 2012]. Revised control and revised deep character judgments were coded as mediators, with Condition as a predictor of Revised blame. We found significant indirect effects of Condition through both revised control ($b = -0.23$, $p = .036$, 95% CI [0.01, 0.44]) and revised deep character ($b = 0.30$, $p = .008$, 95% CI [-0.53, -0.08]) (Figure 4). There is no remaining direct effect of Condition on Revised blame ($b = 0.01$, $p = .967$, 95% CI [-0.49, 0.47]), suggesting that these indirect effects capture the entire impact of C on blame. Because the indirect effects were of similar magnitude and in opposite directions, the overall effect was non-significant ($b = -0.07$, $p = .772$, 95% CI [-0.53, 0.39]) (Figure 5).

To assess whether these indirect effects are independent, we fitted a separate model that assumed the indirect effects are equivalent. We then compared model fit statistics

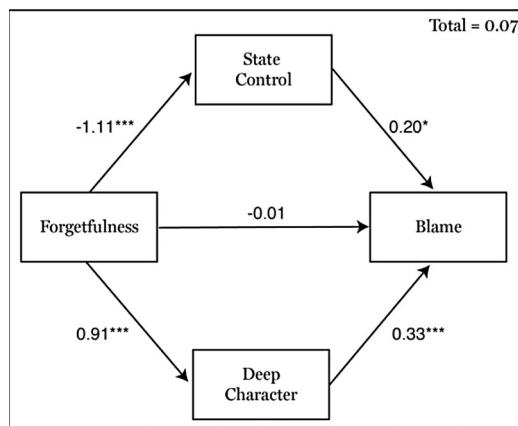


Figure 4: Mediation analysis from Study 5 with state control and deep character mediating the effect of forgetfulness on blame. * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$

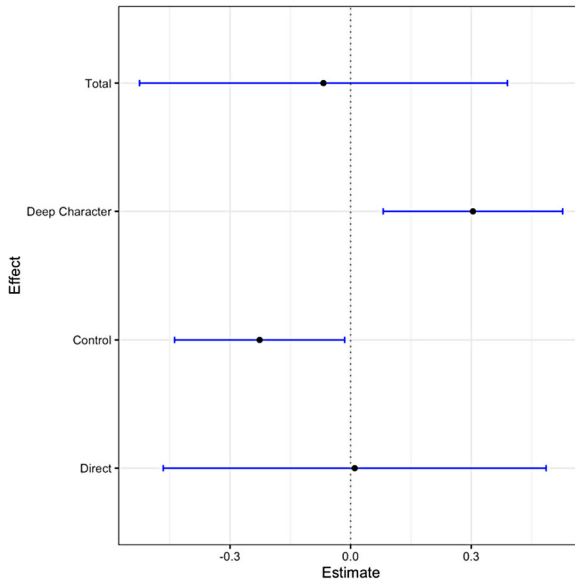


Figure 5: Results from Study 5's mediation analysis, with bias-corrected 95% confidence intervals

for the constrained and two-channel models. A chi-squared difference test indicated that the two-channel model displayed better fit ($\chi^2(1) = 15.37, p < .001$). This suggests that the two indirect effects independently mediate revised judgments of blame.

7 General Discussion

Attribution theorists widely assume that people rely on character assessments to assign blame. But there is substantial disagreement over why. Act-based views hold that character has a *fundamental* effect on blame: we blame people for actions that manifest bad character and excuse people for uncharacteristic lapses in judgment. You might excuse an honest friend's lie as a one-off, for example, while full-throatedly blaming a duplicitous friend who tells the same lie. Person-based views hold that character merely provides *evidence* about the mental states and processes that determine responsibility. Suppose that you agree to evenly split the dinner bill with a friend, but he gives you only 30%. You might rely on character assessments for evidence about the mental states that led to under-paying. You might assume that an honest friend made a harmless mistake in mental math, for example, whereas a duplicitous friend tried to cheat you. Both views, as noted in the Introduction, have intuitive pull and enjoy modest (though conflicting) empirical support.

We aimed to unify these frameworks by providing empirical evidence for a *two-channel view*, where character has both fundamental and evidential effects on blame. We investigated the effects of trait and state assessments on judgments of blame, in situations where a person makes a mistake because his mind is elsewhere. Studies 1 and 2 found that blame increased with *state control*: the ease with which someone can bring his mind back on task. In contrast, blame was unaffected by *trait control* (character): whether a person is conscientious or forgetful. This null effect is surprising, given the widespread assumption that character assessments affect responsibility

judgments [Uhlmann *et al.* 2015]. Studies 3 through 5 showed that the two-channel view can explain this null effect. Mediation analyses revealed two channels by which character modulated blame: forgetfulness directly *increased* blame, but also decreased state control, which indirectly *decreases* blame. These channels cancel each other out, which leads to an overall null effect of character on blame judgments.

Our results have four important implications for how character information informs judgments of responsibility. First, our results suggest that neither act-based nor person-based accounts fully capture the role of character information in judgments of blame. Each theory identifies only one channel through which character assessments affect responsibility judgments.

Second, our results may explain why moral psychology has yielded apparently inconsistent results about character and responsibility, where some findings support a person-based view [Kliemann *et al.* 2008] and others support an act-based view [Cushman 2015]. We find that character can have fundamental *and* evidential effects on the same responsibility judgments, which can lead to puzzles like the catch-22 of forgetfulness. Similarly, seemingly inconsistent results across studies may speak to different channels by which character affects responsibility: one channel is fundamental and the other evidential.

Third, our study points to the importance of choices about research design and statistical analysis. Studies that manipulate character without measuring mental state information may appear to support the fundamental model. Studies that measure only the indirect effect of character on responsibility may appear to support the evidential model. Our study suggests that the true view (the two-channel model) may become evident only when using appropriate research designs (that manipulate character and measure mental state information) and statistical models (mediation analyses that disentangle the direct and indirect effects of character).

Fourth, our results put pressure on the widespread assumption that good character always alleviates blame, whereas bad character always increases blame. This assumption has prior empirical support (see [Young and Tsoi 2013] for review). Furthermore, it is intuitive: shouldn't we cut good people some slack? We find this is not always the case. In one way, we hold conscientious people to a higher standard than forgetful people, because we assume they have more control over their thoughts. As such, conscientiousness can indirectly *magnify* judgments of blame. With power comes responsibility.

Our results also bear on a broader debate within moral psychology about the role of control in responsibility judgments. Some theorists assume that control is the *fundamental* determinant of responsibility. If you have the capacity to control your actions, you are responsible for their consequences [Shaver 1985; Schlenker *et al.* 1994; Fischer and Ravizza 1998]. Previous empirical research also suggests that judgments of blame for mental lapses are driven solely by perceived control [Murray *et al.* 2019; Murray *et al.* forthcoming]. Others assume that control is a distinctive kind of expression of the deep self, a manifestation of an integrated network of moral motivations and concerns in one's actions [Doris 2015; Sripada 2016]. Some empirical research indicates that judgments of responsibility are more sensitive to the structure of an agent's *cares* rather than her *intentions* [Woolfolk, Doris, and Darley 2006].

The results reported here are difficult to square with either of these frameworks. Let's start with control-based frameworks. Recall that we examined two different character traits that are intimately connected to control: conscientiousness and forgetfulness. We defined conscientiousness as a tendency to 'have *control* over [one's] thoughts

and [to] rarely get distracted when doing something important', and forgetfulness as the opposite. Within a control-based framework, the degree of responsibility for some outcome scales with one's control over that outcome. We confirmed this assumption for state control. But trait control had the opposite effect: conscientious people, who tend to control their minds, are directly blamed *less* for their mental lapses than forgetful people. Our results therefore suggest that a greater capacity for mental self-control can mitigate, rather than amplify, blame. Control-based frameworks are therefore hard to square with direct effect of trait control on judgments of responsibility.

Can deep self-theorists do better? Not as currently constructed. Deep self-theorists typically claim that the key ingredient in responsibility judgments is whether an outcome manifests one's true moral motivations (for example, one's concerns, projects, values, and so on). Our results reveal another crucial ingredient in judgments of responsibility: whether one is the kind of person who typically controls one's thoughts and actions. Trait control however, is likely not fixed by one's moral motivations. Genuine moral concern is no bulwark against forgetfulness.

Our results therefore point to an ingredient of responsibility judgments that has been overlooked by control-based *and* deep-self frameworks. Whether you possess control here and now is relevant, as per control-based theories. Your moral motivations are also relevant, as per deep-self theories. But these theories miss another relevant feature: whether you are the *kind of person* who tends to be capable of effective self-governance, of controlling your mind. This trait control is like state control in that it concerns one's overall capacity for self-governance. But it also concerns a deep, persistent characteristic of yourself as a person. Participants in Study 5, for example, judged that when a forgetful person has a mental lapse, this 'reflects something about the kind of person he is deep down inside'. Hence, trait control is somewhat of a hybrid between the factors that drive responsibility judgments, according to standard philosophical theories.

Unlike previous literature, we found both evidential and fundamental effects of character on responsibility. One might ask why this dual effect appears nowhere else in moral psychology. We consider two potential explanations of this difference, both of which raise important questions for future research. First, we focused on moral evaluations of *negligence*, which contrast in systematic ways with other moral judgments. Moral psychologists typically study intentional actions that include straightforward descriptions of an individual's mental states such as desires and beliefs and how these produce the individual's decision to act. Whether an agent is responsible then depends on how people interpret these action-producing states. For example, consider someone who puts poison in their friend's coffee instead of sugar. Whether we blame (and how much we blame) the individual depends on whether we think they *believe* they were adding sugar and whether they *wanted* or *intended* to harm their friend [Young *et al.* 2007].

During negligent action (or omission), in contrast, the mind's characteristic furniture of occurrent beliefs, desires, and intentions is often absent. When Randy burns his friend's birthday cake, for example, his mistake doesn't result from a malicious intention, a false belief, or an immoral desire. Randy may genuinely *intend* and *want* to bake a good cake and *believe* that the cake will burn after 30 minutes in the oven. Randy's negligence results from his failure to *activate* those mental states at the appropriate time because his mind is elsewhere. Negligence results from a failure of *control*, rather than vicious beliefs, desires, and intentions. Judgments about responsibility for negligence may therefore elicit different attributional processes than those used to evaluate non-negligent moral failures.

Second, we study a different kind of character trait than past research. Past studies focused on character traits such as kindness [Gill and Andreychik 2014], fairness [Kliemann *et al.* 2008; Siegel, Crockett, and Dolan, 2017], and honesty [Sandry *et al.* 2011], which reflect one's policies about *what one owes to others*. These traits contrast with other traits that reflect one's capacity for *effective self-governance*. The former kind of trait is outward facing, while the latter is inward facing and reflects the ability to align one's actions and commitments over time. We studied conscientiousness, the self-governmental capacity to implement and follow through on one's plans and commitments. Similar traits may include vigilance, courage, self-control, and punctuality.

This distinction has roots in positive psychology, which distinguishes three categories of virtue: intellectual, moral, and self-regulatory (Wright [1907: 156] distinguishes between virtues of the intellect, the will, and the affections). While the precise boundary between these domains remains a matter of dispute, many agree that virtues cluster into three groups that track these distinctions [Worthington and Hampson 2011; McGrath 2020]. Traits like honesty are moral virtues, whereas traits like conscientiousness are self-regulatory virtues. Our work might provide an empirical way to distinguish these kinds of virtues from one another. Intuitively, different traits may bear different relationships to blame. Moral virtues may always mitigate blame since people always get credit for having a policy of treating others well. In contrast, self-regulatory virtues such as conscientiousness may sometimes *increase* blame since they reflect an increased level of control over one's thoughts and action. Future research might systematically investigate the differential impact of different kinds of traits on blame.

Future research can extend our contributions in (at least) two ways. First, we asked how self-regulatory character traits effect *blame* for mental lapses but did not examine *praise* for conscientious acts. For example, does a conscientious person deserve extra praise, when she follows every step in a complex recipe? Second, future research can investigate whether blame judgments depend on whether someone has *control over their character* and/or *identifies with their character*. For example, if a forgetful person cannot change and laments their forgetfulness, does this mitigate blame?

Perhaps our most intriguing finding is the catch-22 of forgetfulness. Forgetful people like me, your lead author, are in one way held accountable for our wandering minds. When we leave fridges open or burners on, it's not just a one-off; it's a character flaw. When conscientious people (occasionally!) make the same mistakes, we give them leeway: such mistakes are uncharacteristic lapses, not manifestations of their personalities. But in another way, forgetful people like me get off the hook. People assume we have little control over our thoughts, that we couldn't help ourselves. We expect more of conscientious people: we assume they are in control, and thus responsible for the consequences of their thoughts and actions. No one gets around it: you are either blamed for bad character or high expectations. There is only one catch and that is catch-22.

Acknowledgements

We presented an earlier version of this paper at the Australasian Experimental Philosophy Group and the University of Michigan. We are grateful to everyone who participated in these discussions, especially Justin Sytsma, Jonathan Weinberg, Chandra Sripada, Peter Railton, and Laura Soter. For comments on our draft, we thank Jordan Bridges, Elise Dykhuis, Santiago Amaya, Rebecca Stangl, Nate Adams, and anonymous reviewers at this journal. Most importantly, we thank Sheisha Kulkarni, without whom this paper would have failed at multiple stages. She gave us key empirical advice

(visualize your descriptive statistics!!!), philosophical feedback, and inspired the conscientious character in our Catch-22 of Forgetfulness.

Disclosure Statement

No potential conflict of interest was reported by the author(s).

Funding

This project was supported by University of Virginia's 3 Cavaliers program, grant #164 ('Harnessing the Wandering Mind'), the John Templeton Foundation, grant #60845 ('Getting better at simple things'), and Duke University's Summer Seminars in Neuroscience and Philosophy.

ORCID

Zachary C. Irving  <http://orcid.org/0000-0002-6621-5202>

Samuel Murray  <http://orcid.org/0000-0002-4959-3252>

Kristina Krasich  <http://orcid.org/0000-0001-6350-1855>

References

- Alicke, M. D. 1992. Culpable Causation, *Journal of Personality and Social Psychology* 63/3: 368–78.
- Alicke, M. D. 2000. Culpable Control and the Psychology of Blame, *Psychological Bulletin* 126/4: 556–74.
- Brandt, R. 1958. *Ethical Theory*, Englewood Cliffs, NJ: Prentice Hall.
- Cushman, F. 2015. Deconstructing Intent to Reconstruct Morality, *Current Opinion in Psychology* 6: 97–103.
- Doris, J. 2015. *Talking to our selves: Reflection, Ignorance, and Agency*, Oxford: Oxford University Press.
- Faul, F., E. Erdfelder, A. G. Lang, and A. Buchner 2007. G*Power 3: A Flexible Statistical Power Analysis Program for the Social, Behavioral, and Biomedical Sciences, *Behavioral Research Methods* 39/2: 175–91.
- Fischer, J. M., and M. Ravizza 1998. *Responsibility and Control*, Cambridge: Cambridge University Press.
- Gill, M. J., and M. R. Andreychik 2014. The Social Explanatory Styles Questionnaire: Assessing Moderators of Basic Social-Cognitive Phenomena Including Spontaneous Trait Inference, the Fundamental Attribution Error, and Moral Blame, *PLoS One* 9/7. doi: [10.1371/journal.pone.0100886](https://doi.org/10.1371/journal.pone.0100886).
- Heller, Joseph 1961. *Catch-22*, New York: Simon and Schuster.
- Kliemann, D., L. Young, J. Scholz, and R. Saxe 2008. The Influence of Prior Record on Moral Judgment, *Neuropsychologia* 46/12: 2949–57.
- Koster-Hale, J., R. Saxe, J. Dungan, and L. Young 2013. Decoding Moral Judgments from Neural Representations of Intentions, *Proceedings of the National Academy of Sciences of the United States of America*. doi:[10.1073/pnas.1207992110](https://doi.org/10.1073/pnas.1207992110).
- Kupperman, J. 1995. *Character*, Oxford: Oxford University Press.
- Lagnado, D. and T. Gerstenberg 2015. A Difference-Making Framework for Intuitive Judgments of Responsibility, in *Oxford Studies in Agency and Responsibility: Volume 3*, ed. David Shoemaker, Oxford: Oxford University Press: 213–41.
- Livengood, J. M., J. Sytma, and D. Rose 2017. Following the FAD: Folk Attributions and Theories of Actual Causation, *Review of Philosophy and Psychology* 8/2: 273–94.
- Lynch, J. M., J. D Lane, C. M. Berryessa, and J. Rottman 2019. How Information about Perpetrators' Nature and Nurture Influences Assessments of their Character, Mental States, and Deserved Punishment, *PLoS One* 14:10. Doi: [10.1371/journal.pone.0224093](https://doi.org/10.1371/journal.pone.0224093).
- Malle, B. F., S. Guglielmo, and A. E. Monroe 2014. A Theory of Blame, *Psychological Inquiry* 25/2: 147–86.

- McGrath, R. E. 2020. Darwin Meets Aristotle: Evolutionary Evidence for Three Fundamental Virtues, *The Journal of Positive Psychology* 16/4: 431–45. doi: [10.1080/17439760.2020.1752781](https://doi.org/10.1080/17439760.2020.1752781).
- Monroe, A. E. and B. F. Malle 2019. People Systematically Update Moral Judgments of Blame, *Journal of Personality and Social Psychology* 116/2: 215.
- Murray, S., E. D. Murray, G. Stewart, W. Sinnott-Armstrong, and F. De Brigard 2019. Responsibility for Forgetting, *Philosophical Studies* 176/5: 1177–201.
- Murray, S., K. Krasich, Z. C. Irving, T. Nadelhoffer, and F. De Brigard **Forthcoming**. Mental Control and Attributions of Responsibility for Negligent Wrongdoing, *Journal of Experimental Psychology: General*. doi: [10.1037/xge0001262](https://doi.org/10.1037/xge0001262).
- Murray, S. and M. Vargas 2020. Vigilance and Control, *Philosophical Studies* 177/3, 825–43.
- Nozick, R. 1981. *Philosophical Explanations*, Cambridge: Belknap Press.
- Peabody, D. and L. R. Goldberg 1989. Some Determinants of Factor Structures from Personality-trait Descriptors, *Journal of Personality and Social Psychology* 57/3: 552–67.
- Pizarro, D. A., D. Tannenbaum, and E. Uhlmann 2012. Mindless, Harmless, and Blameworthy, *Psychological Inquiry* 23/2: 185–88.
- Rossee, Y. 2012. Lavaan: An R Package for Structural Equation Modeling, *Journal of Statistical Software* 48/2:1–36.
- Royzman, E., and J. P. Hagan 2017. The Shadow and the Tree: Inference and Transformation of Cognitive Content in Psychology of Moral Judgment, in *Moral Inferences*, ed. J. F. Bonnefon and B. Trémolière, Routledge: 56–74.
- Sandry, J., G. Hunt, S. Rice, D. Trafimow, and K. Geels 2011. Can Priming Yourself Lead to Punishing Others?, *Journal of Social Psychology* 151/5: 531–34.
- Schlenker, B. R., T. W. Britt, J. Pennington, R. Murphy, and K. Doherty 1994. The Triangle Model of Responsibility, *Psychological Review* 101/4: 632–52.
- Schoemann, A. M., A. J. Boulton, and S. D. Short 2017. Determining Power and Sample Size for Simple and Complex Mediation Models. *Social Psychological and Personality Science* 8/4: 379–86.
- Siegel, J. Z., M. J. Crockett, and R. J. Dolan 2017. Inferences About Moral Character Moderate the Impact of Consequences on Blame and Praise, *Cognition* 167: 201–11.
- Shaver, K. 1985. *The Attribution of Blame*, Dordrecht: Springer.
- Shoemaker, David 2015. *Responsibility at the Margins*, Oxford: Oxford University Press.
- Sripada, C. S. 2016. Self-expression: A Deep Self Theory of Moral Responsibility, *Philosophical Studies* 173/5: 1203–32.
- Stolier, R. M., E. Hehman, and J. B. Freeman 2020. Trait Knowledge Forms a Common Structure Across Social Cognition, *Nature Human Behaviour* 4/4: 361–71.
- Sytsma, J. **In Preparation**. The Character of Causation: Investigating the Impact of Character, Knowledge, and Desire on Causal Attributions. Available at: philsci-archive.pitt.edu/16739.
- Uhlmann, E. L., D. A. Pizarro, and D. Diermeier 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.
- Willis, J. and A. Todorov 2006. First impressions: Making Up Your Mind After a 100-ms Exposure to a Face, *Psychological Science* 17/7: 592–98.
- Woolfolk, R. L., J. M. Doris, and J. M. Darley 2006. Identification, Situational Constraint, and Social Cognition: Studies in the Attribution of Moral Responsibility, *Cognition* 100/2: 283–301.
- Worthington, E. L., and P. J. Hampson 2011. Forgiveness, Reconciliation, and the Hard March to Peace, in *Explaining Evil: Vol. 3 Approaches, Responses, Solutions*, ed. J. Harold Ellens, Santa Barbara, California: ABC-CLIO: 109–124.
- Wright, H. W. 1907. The Classification of the Virtues, *Journal of Philosophy, Psychology, and Scientific Methods* 4/6: 155–60.
- Young, L., and L. Tsoi 2013. When Mental States Matter, When They Don't, and What That Means for Morality, Social and Personality, *Psychology Compass* 7/8: 585–604.
- Young, L., F. Cushman, M. Hauser, and R. Saxe 2007. The Neural Basis of the Interaction Between Theory of Mind and Moral Judgment, *Proceedings of the National Academy of Sciences of the United States of America* 104/20: 8235–40.
- Young, L., J. Scholz, and R. Saxe 2011. Neural Evidence for 'Intuitive Prosecution': The Use of Mental State Information for Negative Moral Verdicts, *Social Neuroscience* 6/3: 302–15.