Blame Mitigation: A Less Tidy Take and Its Philosophical Implications

Jennifer L. Daigle
Department of Philosophy, Yale University, New Haven, CT, USA

Joanna Demaree-Cotton*
Department of Philosophy, Yale University, New Haven, CT, USA
ORCiD: https://orcid.org/0000-0003-2428-7301

* Authors are listed alphabetically. The authors contributed equally to this paper.

Correspondence: jennifer.daigle@yale.edu; jodemaree@hotmail.com

This is an Accepted Manuscript of an article published by Taylor & Francis in Philosophical Psychology on 23rd November 2021, available online: https://www.tandfonline.com/doi/full/10.1080/09515089.2021.2000594

Blame Mitigation: A Less Tidy Take and Its Philosophical Implications

Why do we find agents less blameworthy when they face mitigating circumstances, and what does this show about philosophical theories of moral responsibility? We present novel evidence that the tendency to mitigate the blameworthiness of agents is driven both by the perception that they are less normatively competent—in particular, less able to know that what they are doing is wrong—and by the perception that their behavior is less attributable to their deep selves. Consequently, we argue that philosophers cannot rely on the case strategy to support the Normative Competence theory of moral responsibility over the Deep Self theory. However, we also outline ways in which further empirical and philosophical work would shift the debate, by showing that there is a significant departure between ordinary concepts and corresponding philosophical concepts, or by focusing on a different type of coherence with ordinary judgments.

Keywords: moral responsibility; blame; deep self; normative competence; moral philosophy; intuitions

1. Blame Mitigation and Theories of Moral Responsibility

Sometimes we deem others not very blameworthy, even when they cause harm. Consider a child who pulls her mother's hair; a grief-stricken widower who speaks hurtful words after losing his family in an accident; or a man with paranoid schizophrenia who violently lashes out while in the grip of a frightening delusion. Although each engages in harmful behavior, these agents seem less blameworthy than they would be were their circumstances different. Supposing that they *are* less blameworthy, what makes them so?

Different theories of moral responsibility provide different answers to this question, and in this paper we consider two such theories: Deep Self theory, on the one hand, and Normative Competence theory, on the other. In arguing for their view over Deep Self theory, Normative Competence theorists have sought to show that our tendency to judge agents less blameworthy in certain cases of mitigating circumstances is not only explained by considerations of diminished normative competence; it is not—nor could it be—explained by considerations pertaining to agents' deep selves. Because of this, many

philosophers have taken the "case strategy" to provide support for Normative Competence theory over Deep Self theory.

By contrast, recent work in experimental philosophy suggests that considerations related to an agent's deep self *are* involved in a variety of ordinary judgments, including judgments of mitigated blameworthiness (e.g. Sripada, 2010, 2012; Newman, De Freitas, & Knobe, 2015; Faraci & Shoemaker, 2019). Consequently, some have supported a dialectical shift favoring Deep Self theory over Normative Competence theory, taking *it* to exhibit greater coherence with everyday practices and judgments (e.g. Faraci & Shoemaker, 2010).

But, whereas both lines of argument are premised on the idea that judgments of mitigated blameworthiness are explained by *either* normative competence *or* the deep self, in this paper we argue for a less tidy take. Specifically, we provide evidence that judgments of mitigated blameworthiness are driven by *both* normative competence *and* the deep self. From this, we argue that neither Normative Competence nor Deep Self theorists can depend on the case strategy to support their view over the other.

In Section 2, we outline these theories of moral responsibility, explaining how the case strategy has been used by Normative Competence theorists to challenge Deep Self theory and support their normative competence requirement. In Sections 3 and 4, we present two studies showing that judgments of mitigated blameworthiness are driven by *both* perceptions of reduced normative competence *and* deep self considerations, while in Section 5 we provide evidence that only the former plays a distinctive role in people's attempts to justify their judgments. Finally, Section 6 discusses what this evidence means

for the debate between Normative Competence and Deep Self theorists, and how further empirical and philosophical work might shift the debate.

2. The Debate: Normative Competence versus Deep Self Theory

2.1. The Theories

Deep Self theory is based on the idea that there are certain elements within your psychology that are central to who you are as an agent. These elements comprise your "*deep self*," what might be described as your "*true character*," or "*who you really are*." These elements differ from superficial characteristics, including fleeting thoughts and physical attributes. Although Deep Self theorists disagree about which elements constitute the deep self (whether these are e.g. certain special desires, judgments, commitments, or cares), they pretty much all agree that

> Deep Self theory: an agent is morally responsible for ϕ-ing if, and only to the extent that, his ϕ-ing is *attributable* to his deep self.[1]

So, on this theory, the diminished responsibility of the child, grieving widower, or man with delusions might be explained by a disconnect between their behavior and who they are *deep down*. The widower's grief might overwhelm his deeply-held concern for others' feelings, so that his hurtful speech does not reflect a bad or selfish character; similarly, the man with delusions who behaves violently might not be a violent person at all. Finally, perhaps the child, owing to her psychological immaturity, lacks a deep self altogether, so that, no matter what she does, the right relation cannot obtain between her behavior and her deep self.

By contrast, Normative Competence theorists emphasize the relevance of a different kind of consideration, and thus offer a different kind of explanation for these agents' diminished responsibility. According to these theorists,

> Normative Competence theory: an agent is morally responsible for ϕ-ing if, and only to the extent that, she had the ability to do the right thing for the right reasons (henceforth, *the ability to act rightly*) at the time of her ϕ-ing.[2]

On this view, if you acted wrongly but could have acted rightly, you might be blameworthy, whereas you are not blameworthy if you could not have acted rightly.

Agents might lack the ability to act rightly for at least two kinds of reasons.[3] First, they might lack an *intellectual competence* that would enable them to grasp the difference between right and wrong and to know what ought to be done. For example, the child might be unable to understand that pulling her mother's hair will hurt her, and the deluded man might be unable to see that he really has no reason to lash out. Second, agents might lack a *volitional competence* that would enable them to act on such knowledge, including the ability to care about, and motivate themselves to do, what is right. So, the widower might grasp the wrongness of hurting others' feelings, though—being overcome by grief—he might be unable to motivate his behavior accordingly.

Thus, although both theories can deliver the same verdict of mitigated blameworthiness in these and other cases, their rationales diverge dramatically. For, whereas Deep Self theorists take moral responsibility to depend on deep self attributability, and so account for mitigated blameworthiness through diminished attributability of behavior to the deep self, Normative Competence theorists take moral responsibility to depend on normative competence, and so point to diminished normative competence in cases of mitigated blameworthiness.

Indeed, it is precisely because Deep Self and Normative Competence theory offer importantly different conditions for moral responsibility that they can also deliver importantly different verdicts in response to some of the same cases, satisfying the conditions of responsibility stipulated by one theory while failing those offered the other.

For example, an otherwise helpful and generous person might, against his better judgment and to his later regret, fail to extend a helping hand to a stranger he rightly saw to be in need. This person seemingly satisfies Normative Competence conditions—he is able, and normally willing, to know and to do what's right. Yet he plausibly does not satisfy Deep Self conditions, since this unhelpful behavior does not reflect an unhelpful or selfish deep self. Thus Normative Competence theory, but not Deep Self theory, might suggest that he is responsible and blameworthy for failing to help in this instance.[4]

On the other hand, agents might satisfy the conditions of responsibility stipulated by Deep Self theory while failing those offered by Normative Competence theory. For example, imagine that a voter deeply and wholeheartedly embraces the misogynistic values and sexist beliefs with which he grew up and which are similarly espoused by his community and local political leaders, where this particular voter is uneducated, a poor critical thinker, and economically marginalized. Now, during an election campaign, he is easily convinced by thinly-veiled smear campaigns and conspiracy theories directed against a female political candidate, and he widely shares cruel misinformation about her. Such an agent is seemingly morally responsible and blameworthy for doing so on the Deep Self theory, since he fully embraces the misogynistic values that these actions express; but, insofar as the deep entrenchment of his sexist worldview and his personal circumstances render him less able to critically evaluate the misinformation upon which the smear

campaign is based, he does not satisfy Normative Competence conditions for moral responsibility and blame.

So, although it is in principle open to a moral responsibility theorist to think that responsibility for some action requires *both* that one's action be attributable to one's deep self, *and* that, at the time of acting, one had the ability to act rightly, thus building a hybrid theory,[5] the Deep Self and the Normative Competence theories are best thought of as capturing independent conditions.

2.2. The Case Strategy

In arguing against Deep Self theory, Normative Competence theorists have primarily sought to target its claim that deep self attributability is itself sufficient for moral responsibility, and to support their normative competence requirement. In doing so, these theorists have relied upon what we will call the *case strategy*, which falls under a family of strategies premised on the idea that a theory's coherence with certain features of our actual moral responsibility practices and judgments counts in favor of that theory. The case strategy in particular focuses on the causal or explanatory basis of ordinary or pre-theoretic judgments of moral responsibility to agents in real or hypothetical cases. That is, that we make the judgments we do is taken to be explained by our sensitivity (whether unconscious or conscious) to some feature the case exemplifies, and a theory of moral responsibility is said to cohere with these responses just in case it implies that this feature is relevant to responsibility in a way that mirrors the sensitivity of our judgments. Assuming our judgments correctly capture something about the nature of moral responsibility, whether the theory in question coheres (or conflicts) with these judgments is then considered a defeasible epistemic reason for (or against) that theory. In this way, ordinary judgments in

response to cases are widely taken to be one important source of evidence for or against philosophical theories of responsibility.

In their own use of the case strategy, Normative Competence theorists have appealed to a wide range of cases—cases involving e.g. children (Nelkin, 2011, p. 8; cf. Nelkin, 2008, p. 498), childhood deprivation, compulsive disorders (Wolf, 1994, p. 81), mental illness, dementia, stress (Nelkin, 2013, pp. 12–13), and manipulation (Wolf, 1990, pp. 37–38)—in which the agents are judged to be less than fully responsible, arguing that this happens *because* we perceive these agents to exhibit diminished normative competence. Moreover, for a select subset of these cases, Normative Competence theorists have argued not only that our judgments are not, but that they *cannot be*, explained by deep self considerations, since in this subset it is explicitly stipulated that what the agents do is fully attributable to their deep self (e.g. Wolf, 1994, p. 37, pp. 80–81, pp. 85–87; Wolf, 2002, pp. 239–240; Wolf, 2003, pp. 382–383, p. 382 fn.7; Nelkin, 2015). In this way, the case strategy would seem to not only support Normative Competence theorists' rejection of the sufficiency of deep self attributability for moral responsibility, but their own normative competence requirement as well (Wolf, 1990, pp. 46, 68).

This strategy has been particularly strongly illustrated with cases involving agents with childhoods characterized by an impoverished moral education (henceforth bad upbringing or bad formative circumstances). Consider people whose parents consistently engaged in, and even encouraged, immoral behavior, and who had no alternative role models. When they go on to behave immorally, we tend to judge them less blameworthy compared to people who behave badly despite having had parents who taught them right from wrong. It seems highly sensible that an impoverished moral education of this kind

would inhibit the development of moral abilities, such as the ability to recognize that what one is doing is wrong; it is therefore extremely plausible that perceptions of diminished normative competence drive this tendency to mitigate blame.

On the other hand, it is much less clear why an impoverished moral education would lead to reduced deep self attributability. Why would the bad things someone does later in life be less attributable to her deep self if she had a bad upbringing? To the contrary, it might be thought that bad formative circumstances encourage bad behavior precisely because they lead agents to develop a bad deep self. For example, it might be thought that people whose childhood role models teach them racist values are exactly the sort of people most likely to endorse racism as adults. Consequently, if deep self considerations were driving our blameworthiness judgments, we would judge them more, not less, blameworthy for the racist things they do later in life.

Moreover, although bad formative circumstances need not always lead to the development of a bad deep self, cases in the literature often explicitly stipulate that the featured agents—due to their bad upbringing—have developed a bad deep self to which their subsequent immoral behavior is fully attributable (Wolf, 1980, pp. 155, 159–160; Wolf, 1990, pp. 37, 80–81; Wolf, 2012, pp. 333–334; Nelkin 2008; Nelkin 2015). It is argued that, therefore, any tendency to judge these agents less blameworthy cannot be explained by perceptions of reduced deep self attributability; rather, perceptions of reduced normative competence must be doing the work.

In this way, application of the case strategy to cases involving bad formative circumstances seems to provide powerful support both for the insufficiency of deep self

attributability for moral responsibility, and for the necessity of normative competence in particular.

Of course, Deep Self theorists are not without resources for response. One possibility would be to reject philosophers' preferred explanation of our judgments in cases involving these mitigating circumstances. Though even defenders of Deep Self theory have agreed that these judgments are explained by something *other* than deep self considerations (Watson, 1996, p. 240), it might be argued that this is a mistake, and that our judgments in these cases are explained by deep self considerations after all.

Current empirical work might be read as backing such a dialectical shift. As noted earlier, work in experimental philosophy suggests that intuitions about moral responsibility are often driven by intuitions about deep self attributability (e.g Sripada, 2010). And, importantly, recent work suggests that this may also be the case for judgments about diminished moral responsibility in cases of poor formative circumstances.

First, in a recent set of studies investigating judgments of mitigated blameworthiness in response to cases involving agents with bad formative circumstances, Faraci & Shoemaker (2010) found that even agents coming from bad formative circumstances were still judged to be blameworthy, and on this basis argue that diminished normative competence might not be at issue at all, and that whatever is driving our judgments might still be accommodated within Deep Self theory.[6] Moreover, in a still more recent set of studies, Faraci & Shoemaker (2019) found that the bad behavior of agents coming from bad formative circumstances *is* seen as less attributable to their deep selves, and that this partially explains judgments of mitigated blameworthiness.

Insofar as this recent work suggests that judgments of diminished blameworthiness in cases of poor formative circumstances are explained by perceptions of diminished deep self attributability, it might be thought that the case strategy favors Deep Self theory over Normative Competence theory after all. And indeed, this seems to be what Faraci and Shoemaker want to say. So, in discussing the results of their first study, Faraci and Shoemaker take this result to be explained by deep self considerations, and thus conclude that the addition of normative competence to a theory of moral responsibility would be "unmotivated" (2010, p. 326; 2014, p. 11) and that, "at least with respect to this particular objection, the [Deep Self theory] can survive Wolf's attack intact" (2010, p. 319). And, on the basis of their later results, they further conclude:

> One thing does seem clear from the data…judgments of blame- and praiseworthiness are intimately connected with judgments about the true self. [Deep Self] Theory remains (for now) on solid ground (Faraci & Shoemaker, 2019, p.16).

In fact, however, the empirical work conducted by Faraci and Shoemaker remains inconclusive as a test of whether Deep Self theory or Normative Competence theory better coheres with ordinary judgments of blameworthiness in cases of mitigating circumstances. First, existing studies have not directly measured the role of normative competence in these judgments. Consequently—though this empirical work challenges strong claims to the effect that deep self considerations are irrelevant to blameworthiness judgments—it remains possible that the influence of normative competence is stronger or more fundamental. Second, existing empirical work that tests the impact of deep self perceptions on blameworthiness judgments has yet to use cases that explicitly stipulate that the agent has developed a bad deep self, in the way that Normative Competence theorists have.[7] Thus, these findings do not affect the specific argument presented by Normative

Competence theorists that, once it is made clear that an agent's behavior is fully attributable to her deep self, normative competence considerations are required to explain attributions of mitigated blameworthiness.

3. Study 1: Blameworthiness and Bad Formative Circumstances

Our first study sought to address these empirical questions by extending Faraci and Shoemaker's research in two ways. First, we introduced measures of normative competence, allowing us to investigate the relationship between perceptions of deep self attributability and normative competence, as well as the relationship between these and judgments of blameworthiness. Second, following key cases in the philosophical literature, our study explicitly and unambiguously stipulates that the agent's behavior was fully attributable to his deep self.

A quick note to the reader about the format of our paper: consistent with reporting practices from experimental psychology, the "Results" subsections include the details of all of the results from our studies, including details of statistical analyses. Each of these subsections is followed by a "Discussion" subsection, in which the main findings as they pertain to the overarching philosophical questions of this paper are summarized. Finally, in the last section ("The Future of Deep Self and Normative Competence Theory"), these philosophical implications are developed and considered at length.

3.1. Method

3.1.1. Participants

All participants were recruited on Amazon Mechanical Turk and received compensation for participating. We set out to recruit 410 participants and ended up with a

sample of 413 due to random errors in the M-Turk software. The mean age of participants was 37.1 years, *SD* = 12.2 years (based on data from 408 participants; 5 participants declined to give their age). 207 (50.6%) of our participants identified as male, 198 (48.4%) as female, and 4 (0.01%) as another gender (4 participants declined to give their gender).

3.1.2. Materials and procedure

All participants read a version of a vignette adapted from the work of Faraci and Shoemaker about an agent named Tom. The vignette describes Tom's upbringing and then explains that later in life he commits a severely immoral act. Our version of the vignette explicitly emphasized that the immoral act is attributable to Tom's deep self.

Participants were randomly assigned to one of two upbringing conditions: good or bad. Those in the bad upbringing condition (*N* = 205) read a version of the vignette in which Tom suffers from a bad upbringing:

> Tom is a white male who was raised on an isolated island in the bayous of Louisiana. Growing up, Tom knew many bad people, people who taught him bad values. For example, he was taught to believe that all non-white people are inferior, unworthy of care, and that he should humiliate them when he gets the chance. At the age of 25, Tom moves to another town. Walking outside his home, he sees a black man who has tripped and fallen.
>
> Deep down, Tom truly is a racist. So, in keeping with who he truly is deep down inside, he spits on the black man as he passes by. The real Tom is simply a very bad person.

Participants in the good upbringing condition (*N* = 208) read a version that was identical in its characterization of the act and the agent's deep self but described a good upbringing:

> Tom is a white male who was raised in New Orleans. Growing up, Tom knew many good people, people who taught him right from wrong. For example, he was taught to believe that all people are equal, and that you should always care about others no matter their race. At the age of 25, Tom moves to another town. Walking outside his home, he sees a black man who has tripped and fallen. Tom spits on the man as he passes by.

> Deep down, Tom truly is a racist. So, in keeping with who he truly is deep down inside, he spits on the black man as he passes by. The real Tom is simply a very bad person.

After reading the vignette, all participants were asked to rate the degree to which they thought Tom was blameworthy for what he did on a seven-point Likert scale (1 = "not at all blameworthy", 4 = "somewhat blameworthy", 7 = "completely blameworthy").

On the next page, we asked participants to indicate their levels of agreement or disagreement with the following statements on 7-point Likert scales (1= "completely disagree", 4 = "neither agree nor disagree", 7 = "completely agree"). The statements were as follows, and were presented in a random order (the labels were not visible to participants):

Deep Self Attributability:	What Tom did expressed his *true self*—the person he really is deep inside.
Intellectual Competence:	Tom lacked the ability to recognize that what he was doing was wrong.
Volitional Competence:	Tom lacked the ability to motivate himself to do the right thing.

3.2. Results

Responses to the normative competence statements were reverse-coded, so that higher numbers indicate higher levels of agreement that the agent was normatively competent at the time of acting.

As residuals for our dependent variables were not normally distributed (Shapiro-Wilkes $p < .001$), non-parametric Mann-Whitney U tests were conducted to examine the effects of upbringing on judgments.

In line with philosophers' assessments of such cases, participants found the agent less blameworthy when he had a bad upbringing (see Table 1), $U = 15201.50$, $z = 5.99$, p

<.001, r = 0.30. Participants also, in line with the philosophical literature, judged the agent who had a bad upbringing to be less intellectually competent (U = 15749, z = 4.53, p <.001, r = 0.22). However, the agent was judged to be equally volitionally competent in both cases (U = 20372, z = 0.55, p = .585, r = 0.03.)

	Good Upbringing (N = 208)		Bad Upbringing (N = 205)	
	Mean	*SD*	Mean	*SD*
Blameworthiness	6.68	0.84	6.05	1.33
Deep Self	6.36	1.34	5.64	1.66
Intellectual Competence	5.13	2.33	4.19	2.23
Volitional Competence	3.90	2.58	3.72	2.19

Table 1. Means and standard deviations for blameworthiness and mediator ratings in Good and Bad Upbringing conditions respectively, Study 1.

Furthermore, and in contrast to claims made about such cases in the philosophical literature, whether the agent had a good or bad upbringing also affected the extent to which the agent's behavior was judged to express his deep self, U = 15043.50, z = 5.58, p <.001, r = 0.28. Specifically, when the agent had a bad upbringing, participants judged that his act was less attributable to his deep self (for all descriptive statistics, see Table 1).

So far, these results show that the moral quality of agents' upbringing can affect ascriptions of blameworthiness, as well as perceptions of their normative competence and the attributability of their later behavior to their deep self. What we wanted to know next was whether participants judged agents to be less blameworthy when they did *because* they perceived these agents as less normatively competent, or because they perceived their behavior as being less attributable to their deep self. To investigate this, we used a statistical method called *mediation analysis*, which tells whether the effect of the quality of agents'

upbringing on their blameworthiness is reduced, or even eliminated, when these other judgments are controlled for; if the effect of upbringing on blameworthiness is reduced when judgments of normative competence or deep self attributability are controlled for, this indicates that those judgments play a role in explaining the effect of upbringing on blameworthiness.

All mediation analyses reported in this paper were conducted using Process, Model 4, with 5000 bootstrapped samples (see Hayes, 2012; Preacher & Hayes, 2008), and 95% bias-corrected confidence intervals for indirect effects. Here, the model was for multiple mediation with Upbringing entered as the independent variable.

Results of the mediation analysis (see Figure 1 and Table 2) showed that the effect of the quality of upbringing on blameworthiness was partially mediated by intellectual competence judgments: the quality of upbringing significantly affected judgments of the agent's intellectual competence, which in turn significantly affected judgments of his blameworthiness. Again, this is in line with claims in the philosophical literature.

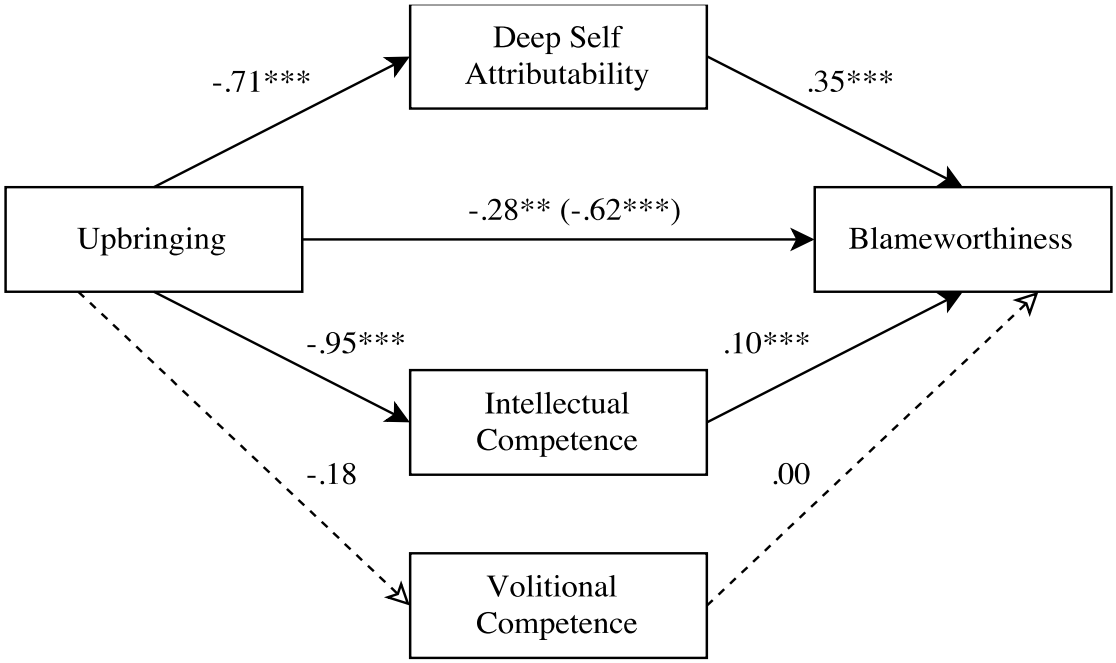

Figure 1. Mediation model showing that reduction in blameworthiness in cases of bad upbringing is partially mediated by both the effect of upbringing on deep self attributability and the effect of upbringing on intellectual competence. Unstandardized regression coefficients are shown for each path, with the total effect in parentheses. Significant predictive relationships are indicated by solid arrows, with *** indicating $p < .001$, ** indicating $p < .01$. Non-significant relationships indicated by dotted arrows.

But the analysis also showed that the effect of upbringing on blameworthiness was partially mediated by deep self judgments: quality of upbringing significantly affected deep self judgments, which in turn significantly affected blameworthiness judgments. Moreover, deep self judgments played a *greater* role in mediating the effect

of upbringing on blameworthiness than that played by intellectual competence judgments. This suggests that deep self judgments play an even bigger role in explaining the extent to which an agent with a bad upbringing is blamed less.

Mediator	Indirect Effect via Mediator (*b*)	Standard Error	95% Confidence Interval
Deep Self	-.25	.07	-.40, -.13
Intellectual Competence	-.09	.04	-.17, -.03
Deep Self—Intellectual Competence	-.16	.08	-.33, -.01

Table 2. Significant indirect effects of Upbringing condition on blameworthiness via the mediators, Study 1. A 95% confidence interval that does *not* overlap with zero indicates a statistically significant effect.

3.3. Discussion

Earlier, we suggested that philosophers were on strong ground in supposing that normative competence considerations, as opposed to deep self considerations, account for judgments of mitigated blameworthiness in cases involving bad formative circumstances.

Our first study challenges these assumptions. Not only did poor formative circumstances lead participants to judge that an agent was less blameworthy and less intellectually competent, it also led them to judge that the behavior was *less attributable to his deep self* compared to another agent who commits the same act after the benefit of a good upbringing. Moreover, statistical analysis indicated that these judgments were related: the evidence suggests that agents with a bad upbringing are judged less blameworthy in part *because* their behavior is seen as less attributable to their deep selves, and to a lesser extent because of their reduced intellectual competence.

It should be noted that participants in our study still had the intuition that the agent who suffered from poor formative circumstances was very blameworthy, albeit less blameworthy than an agent who committed the same action after a good upbringing. This

is consistent with both Deep Self and Normative Competence theories, since this pattern corresponded to participants' judgments that, despite the agent's upbringing, he still retained the ability to know that what he was doing was wrong, and his behavior was still somewhat attributable to his deep self.

This pattern of judgments is perhaps unsurprising when we note that the vignette concerned a severely immoral act (a spontaneous racist assault) and included only a limited description of the ways in which the agent's upbringing was morally impoverished. Thus, participants might have thought that anybody—regardless of upbringing—would know that the given behavior was wrong, or they might have assumed that the agent was exposed to at least some moral exemplars (either early on or in adulthood). Correspondingly, it seems that participants imagined that his morally impoverished upbringing reduced the attributability of his behavior to his deep self, but only to a certain extent: they may well have imagined that someone with a better deep self would have recognized the despicable values from his upbringing for what they are, and would not have endorsed and succumbed to them.

On the other hand, it remains unclear just *why* agents' immoral behavior would be seen as less attributable to their deep self when they had a bad upbringing, especially given the plausibility of the thought that upbringing shapes one's deep self (for discussion, see Faraci & Shoemaker, 2019). But, whatever the explanation for how information about formative circumstances features in reasoning about agents' deep selves, our study suggests that judgments of mitigated blameworthiness in cases of bad formative circumstances are indeed explained, in part, by intuitions about the attributability of behavior to the deep self.

Surprisingly, this was so even when the agent's behavior was explicitly described as being attributable to his deep self. Apparently, attempts to manipulate deep self judgments through such instructions are largely ineffective; although the reason for this is unknown, there seems to be resistance to taking agents' immoral behavior to be fully attributable to their deep self when they have had a bad upbringing.

Contrary to what defenders of Normative Competence and Deep Self theory alike have supposed, then, the upshot of this first study seems to be that application of the case strategy to cases involving bad formative circumstances favors neither Normative Competence nor Deep Self theory. Rather, both normative competence and deep self considerations together explain ordinary judgments in this context.

4. Study 2: Blameworthiness and Mental Illness

Although Study 1 suggests that blameworthiness judgments in cases of bad formative circumstances are driven by normative competence and deep self considerations, perhaps normative competence considerations *alone* account for such judgments in other contexts.

To explore this possibility, we turned to cases of mental illness. Mental illness is often defined partly in terms of impairments to normal psychological capacities,[8] and mitigated criminal responsibility due to mental illness is often explicitly framed in terms of normative competence in the law (a common legal insanity defense is defined in terms of not knowing what you are doing or that what you are doing is wrong).[9] So if normative competence is solely responsible for blame mitigation in certain cases, those involving mental illness are good candidates.

We decided to contrast two kinds of mental disorder: delusions and psychopathy. In moral philosophy and the legal system, suffering from delusions (e.g. due to a psychotic illness) is normally taken to significantly mitigate responsibility. On the other hand, psychopathy—a disorder characterized by a lack of empathy, callousness, and certain emotional deficits—is not considered to be legally mitigating, and it is highly controversial whether it undermines moral responsibility.[10] Moreover, whereas the relative reluctance to excuse individuals with psychopathy from blame might be widespread, it is not immediately clear whether the explanation for this reluctance is friendlier to Deep Self, or Normative Competence, theory. Is it because people with psychopathy are taken to be immoral, deep down, whereas those with delusions are not? Or is it because we tend to see delusions, but not psychopathy, as impairing people's ability to do the right thing for the right reasons?

4.1. Method

4.1.1. Participants

All participants were again recruited on Amazon Mechanical Turk and received compensation for participating. We set out to recruit 300 participants and ended up with a sample of 302 due to random errors in the M-Turk software. The mean age of participants was 34.5 years, *SD* = 10.2 years (based on data from 299 participants; two participants declined to give their age and one gave an obviously incorrect answer of "3" and was excluded). 161 (51.6%) of our participants identified as male, 138 (43.9%) as female, and 2 (0.6%) as another gender (1 participant declined to give their gender).

4.1.2. Materials and procedure

Participants were randomly assigned to one of three "Illness" conditions—No

Illness, Delusion, or Psychopathy—each with its own vignette. The vignettes began with a paragraph describing some psychological characteristics of the agents, including facts about their cognitive and emotional lives. Since we wanted to use these cases to examine what effect, if any, impairments to normative competence have on blameworthiness judgments, in both the Delusion and Psychopathy cases we emphasized the way that the disorder resulted in certain inabilities to grasp, or be motivated by, right reasons. By contrast, the third No Illness vignette described the agent's psychological features and habits in a way that involved no reference to mental disorder or to any kind of excuse. Instead, this vignette described the agent's investment in one of his hobbies: football.

In the second paragraph of each vignette, the agent commits a bad act (he kicks and spits on a man who has tripped and fallen), though his motivation varies: in the Delusion case, he is motivated by a paranoid delusion; in the Psychopathy case, by his abnormal callousness; and in the No Illness case, by a sports rivalry.

After reading the vignette, participants rated the agent's blameworthiness on a seven-point Likert scale (1 = "not at all blameworthy", 4 = "somewhat blameworthy", 7 = "completely blameworthy"). On the next page, using the same measures as in Study 1, they rated intellectual and volitional competence, as well as the attributability of the agent's behavior to his deep self.

4.2. Results

Again, responses to the normative competence statements were reverse-coded, so that higher numbers indicate higher levels of agreement that the agent was normatively competent.

As residuals for our dependent variables were not normally distributed, non-parametric Kruskal-Wallis tests were conducted to examine the effects of mental illness on judgments of blameworthiness, the deep self, and normative competence, with follow-up Mann-Whitney U tests for planned pairwise comparisons. (See Table 3 for a summary of the means and standard deviations for these judgments in the three conditions.)

Dependent Variable	Delusion (N =97)		No Illness (N = 104)		Psychopathy (N =101)	
	Mean	*SD*	Mean	*SD*	Mean	*SD*
Blameworthiness	4.33	(1.73)	6.75	(0.76)	6.06	(1.28)
Deep Self	3.60	(1.68)	5.87	(1.31)	5.61	(1.66)
Intellectual Competence	2.59	(1.58)	5.02	(2.31)	3.86	(2.26)
Volitional Competence	2.73	(1.43)	4.06	(2.34)	3.21	(1.82)

Table 3. Means and standard deviations for judgments of blameworthiness, deep self attributability, and normative competence for each Illness condition, Study 2.

Mental illness significantly affected judgments of blameworthiness, $\chi^2(2) = 121.39$, $p < .001$. In particular, the agent with delusions was judged to be significantly less blameworthy than the agent with psychopathy ($U = 2138$, $z = 7.06$, $p < .001$, $r = 0.50$) and the agent with no mental illness ($U = 1130$, $z = 10.32$, $p < .001$, $r = 0.73$). (See Figure 2.) Moreover, the Illness condition to which participants were assigned also affected judgments of the agent's intellectual competence ($\chi^2(2) = 53.86$, $p < .001$) and judgments of the attributability of the behavior to the agent's deep self ($\chi^2(2) = 89.46$, $p < .001$). In particular, mirroring the effect of these conditions on blameworthiness, the behavior of the agent with delusions was judged to be less attributable to his deep self compared to the agent with psychopathy ($U = 1890.50$, $z = 7.59$, $p < .001$, $r = 0.54$) and the agent with no mental illness ($U = 1526$, $z = 8.69$, $p < .001$, $r = 0.61$). Similarly, the agent with delusions was judged less intellectually competent—that is, less able to recognize that what he was

doing was wrong—compared to the agent with psychopathy (U = 3260.50, z = 4.14, p <.001, r = 0.29) and compared to the agent with no mental illness (U = 2214.50, z = 6.99, p <.001, r = 0.49).

Finally, the Illness conditions affected judgments of the agent's volitional competence (χ^2 (2) = 16.40, p <.001). The delusional agent was judged to be less volitionally competent than the agent with no mental illness (U = 3468.50, z = 3.88, p < .001, r = 0.27). However, there was no significant difference between the agent with delusions and the agent with psychopathy in this regard: both were regarded as equally unable to care about the harm they were causing (U = 4221, z = 1.71, p = .087, r = 0.12).

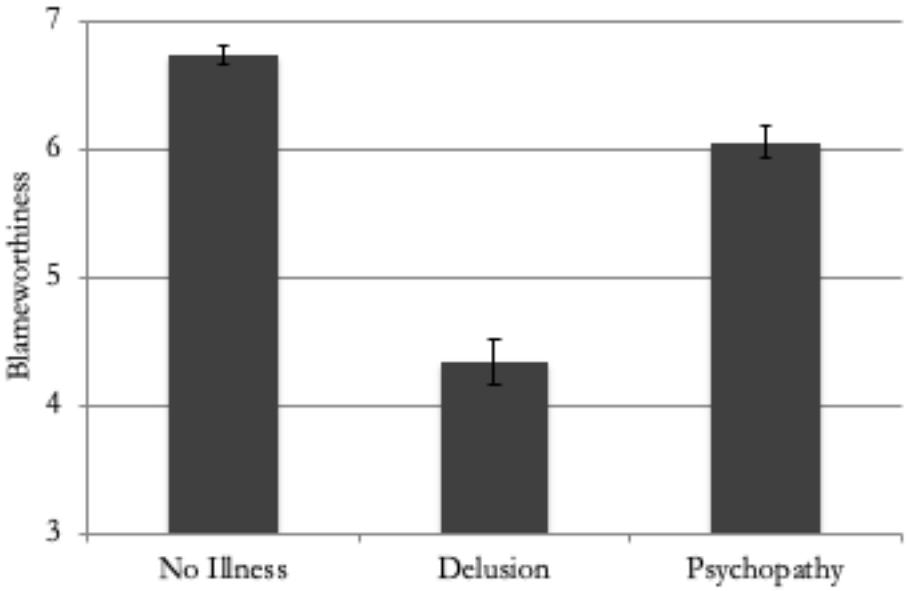

Figure 2. Mean ratings of blameworthiness as a function of Illness condition (No Illness vs. Delusion vs. Psychopathy). Error bars indicate standard errors of the mean.

As in Study 1, we conducted a mediation analysis (Figure 3) using Process, Model 4, with 5000 bootstrapped samples and 95% bias-corrected confidence intervals for indirect effects. The mediation analysis was conducted in line with guidelines for multicategorical independent variables from Hayes & Preacher (2014); indicator coding was used with the

Delusion condition as the reference group. The results uncovered a similar pattern to the mediation analysis of Study 1: deep self and intellectual competence judgments (but not volitional competence judgments) significantly predicted blameworthiness judgments (see Table 4 and Figure 3).

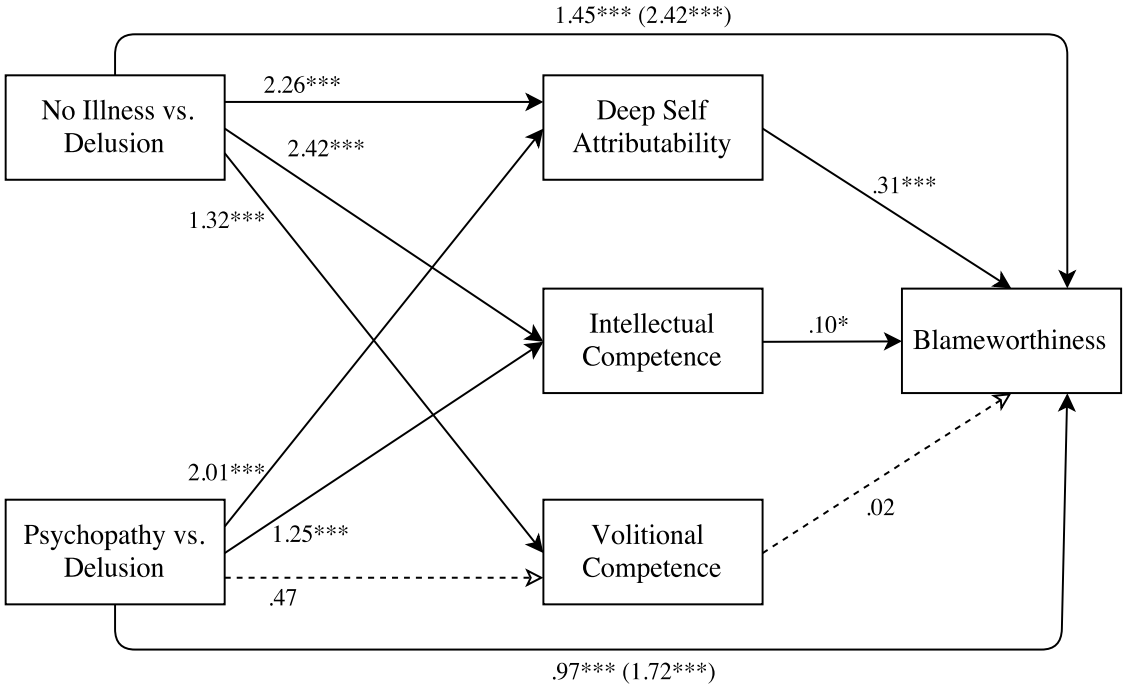

Figure 3. Mediation model showing that reduction in blameworthiness in Delusion relative to No Illness and Psychopathy is partially mediated by the effect of these cases on deep self attributability and on intellectual competence. Absolute values of unstandardized regression coefficients are shown for each path, with total effects in parentheses. Asterisks indicate statistically significant relationships, * $p < .05$, *** $p < .001$. Non-significant relationships indicated by dotted arrows.

Moreover, both sets of judgments partially mediated the effect of the Delusion case on blameworthiness; that is, the agent with delusions was judged to be less blameworthy than the others in part because his behavior was judged to be less expressive of his deep self, and in part because he was judged to be less intellectually competent.

	Mediator	Indirect Effect via Mediator (*b*)	Standard Error	95% Confidence Interval
No Illness vs. Delusion	Deep Self	.71	.15	.44, 1.02
	Intellectual Competence	.24	.08	.10, .41
	Volitional Competence	.02	.04	-.05, .11
Psychopathy vs. Delusion	Deep Self	.63	.14	.38, .93
	Intellectual Competence	.12	.05	.04, .24
	Volitional Competence	.01	.02	-.02, .06

Table 4. Indirect effects of Illness condition on blameworthiness via the mediators. A 95% confidence interval that does *not* overlap with zero indicates a statistically significant effect.

4.3. Discussion

An agent described as suffering from delusions differs in many respects compared to an agent with psychopathy or an agent with no illness. As such, this study alone tells us little about which aspects of delusional behavior cause participants to see the action as less attributable to the deep self and the agent as less intellectually competent.[11]

Nevertheless, this study provides further evidence that deep self and intellectual competence considerations are playing an important role in intuitive assignments of blame. In particular, we find those suffering from mental illness involving delusions to be less blameworthy than those with no mental illness or those with psychopathy in part be because we consider them to be less intellectually competent, *and* in part because we consider their behavior to be less attributable to their deep self.

Volitional competence, by contrast, plays little role in this pattern of blame. Even though the delusional agent was judged to be less able to motivate himself to do the right thing, statistical analysis suggested that this had no impact on blameworthiness judgments.

Thus, the results of our second study suggest a conclusion similar to those of our first. Whereas the first study found that application of the case strategy to cases involving

bad formative circumstances favors neither Normative Competence nor Deep Self theory, the second study extends this conclusion to cases involving mental illness. Just like blame of agents who had a bad upbringing, blame of those with a mental illness would seem to be mitigated due to perceptions of diminished deep self attributability and normative competence (specifically, intellectual competence).

5. Study 3: Explicit Justifications

Our first two studies raise serious problems for the claim that judgments of mitigated blameworthiness are explained only by normative competence, and not deep self, considerations. Our next study sought to investigate whether the folk appeal to normative competence or to deep self considerations when *justifying* their blameworthiness judgments.

5.1. Method

5.1.1. Participants

We recruited 91 participants on Amazon Mechanical Turk. They received compensation for participating. 89 participants provided demographic information. Participants had a mean age of 33.78 (*SD* = 10.13). 51 participants identified as male (57.3%) and 38 (42.7 %) as female.

5.1.2. Materials and procedure

Participants read a vignette in which three people (Kate, Jennifer, and Joanna) discuss the blameworthiness of someone who has, "for no good reason, spit on another man." Although initially they all agree that the agent, Tom, is "very blameworthy," this changes when they learn that he suffered from potentially mitigating circumstances. (Half

of our participants read that Tom had been experiencing hallucinations and delusions, and the other half read that he had a bad upbringing.) Whereas Kate says that these facts make Tom more blameworthy, and Jennifer explains that they make no difference at all, Joanna finds Tom less blameworthy because of them.

After reading the vignette, participants were asked whether they agreed more with Kate, Jennifer, or Joanna. Then they were instructed to reflect on why Tom's circumstances affect his blameworthiness in that way. For example, those who read that Tom had a bad upbringing, and who agreed that this made him less blameworthy, read the following instructions:

> Everyone agreed that Tom had a severely deprived childhood, and that this probably affected Tom's personality or the way he felt or was thinking at the time of the act. In your opinion, why does this make Tom less blameworthy for what he did? In your own words, please try to explain why the effects of Tom's childhood make a difference to how blameworthy he is. Remember, there is no right or wrong answer; just tell us what you think is relevant. You may take as much time as you need, but you must spend at least one minute reflecting on and answering this question.

The instructions were adapted accordingly for those who had answered that Tom's circumstances made no difference and for those who had answered that Tom's circumstances increased his blameworthiness. Participants were prevented from submitting their response until at least one minute had passed.

5.2. Results and Discussion

Based on discussion of the first 31 responses, we formulated a coding scheme for classifying the justifications cited by participants. As well as classifying responses under Normative Competence or Deep Self, our scheme included four "Other" categories (see Figure 4 below). (Each participant's response could involve multiple justifications and so could receive multiple code assignments.) The authors coded the remaining 60 participant

responses separately. These assignments exhibited high levels of interrater reliability (indicated by κ), agreeing 98% of the time on whether a participant cited a justification in the Deep Self category (κ = .90); 90% of the time on whether a participant cited a justification in the Normative Competence category (κ = .79); and 87% of the time on whether a participant cited a justification in one of the Other categories (κ = .72).

Any disagreements were then resolved. The responses of 12 participants were discarded for being incomplete or uninterpretable.

Any response suggesting that the agent's blameworthiness depends on his possessing relevant knowledge, understanding, or emotional/motivational capacities was classified as falling under Normative Competence. For example, one participant seeking to justify the agent's blameworthiness wrote:

> "I feel that even if someone has a bad childhood, they still have the ability to know whether something is wrong or right, regardless of the circumstances"

And another:

> "I don't think his illness affected his thoughts on right and wrong. He knows spitting is wrong but did it anyway."

Others cited normative competence considerations as grounds for excusing Tom. So one participant wrote:

> "Tom's mental illness might make him less capable of recognizing or distinguishing right from wrong. Especially because of his hallucinations, he might not have realized he was actually spitting on a person."

And another:

> "I think Tom is less blameworthy because i don't believe he knew what he was doing because of his severe mental illness"

By contrast, responses citing a dissociation between the behavior and the agent, or citing the agent's having positive moral or personal characteristics or intentions so as to justify mitigated blameworthiness (or, alternatively, the agent's having negative personal characteristics or intentions to justify judgments of blameworthiness) were classified in the Deep Self category. For example, the following participants referred to deep self considerations to mitigate the agent's blame:

> "I think that his childhood makes Tom less blameworthy because it is likely that he never had anyone to demonstrate proper behavior. Because of this, he… might not be acting entirely out of malice"
>
> "I don't think it necessarily makes Tom a bad person that he did this."
>
> "I believe he is less blameworthy because mental illness is something that causes individuals to act differently than they normally would."
>
> "He probably already feels guilty enough after realizing what he has done."

We attempted to apply the Deep Self category very generously and inclusively. For example, one participant seeking to justify blame wrote:

> "While Tom had a bad childhood, I don't think there is any reason why he can't choose to rise above it and become a better person."

We did not require that participants refer to the *relation* between Tom's self and the act, nor did we require language explicitly indicating a distinction between the "deep" and the "surface" self.

Nevertheless, results showed that a greater proportion of participants cited normative competence considerations in justifying blameworthiness ascriptions than any other category (42% of participants) (see Figure 4). By contrast, deep self justifications were strikingly rare (9% of participants). A McNemar test confirmed that participants were significantly more likely to cite a normative competence justification than a deep self

justification, rather than being equally likely to cite either one (χ^2 (1, N = 79) = 18.38, p < .001.)

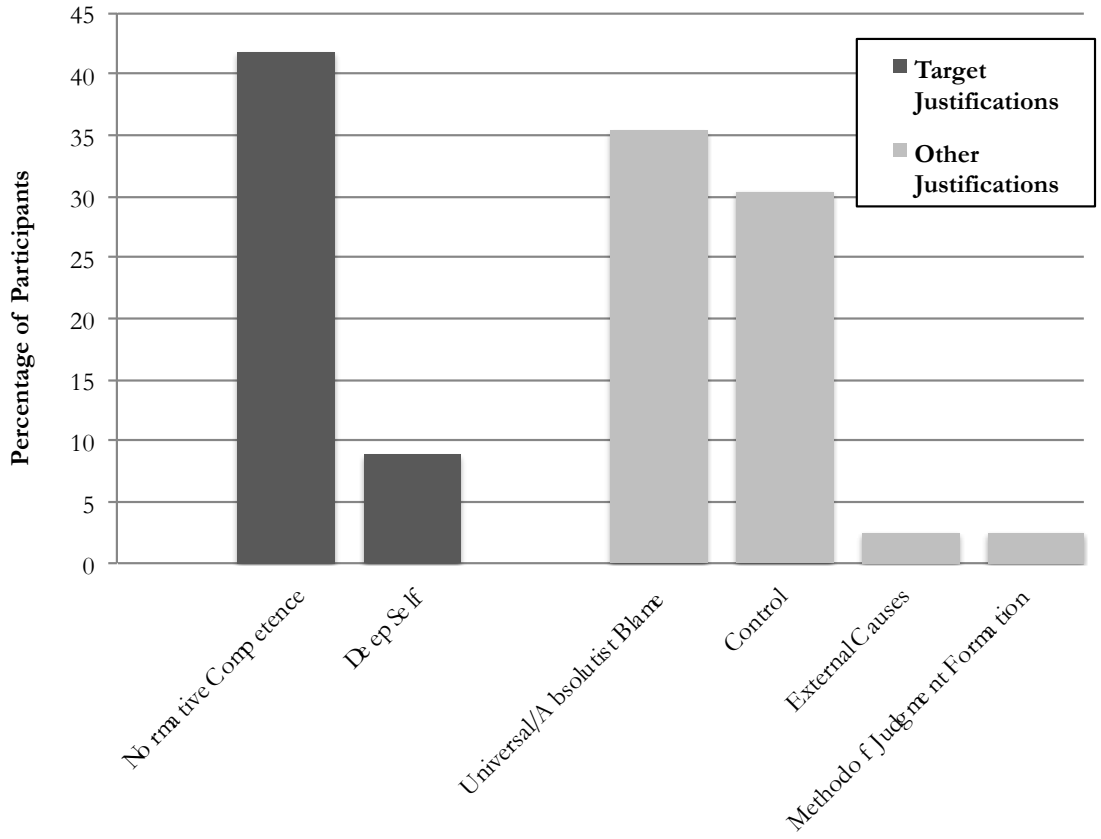

Figure 4. Results of Study 3. Bars indicate the percentage of participants who cited at least one justification in that category.

Notably, the only other types of justification that approached the frequency with which normative competence was cited were either irrelevant to the debate at hand, or conceptually close to normative competence. As shown in Figure 4, 35% of participants justified ascribing blameworthiness to the agent by reference to what we termed "Universal/Absolutist" standards of blame. This simply involved the *denial* of the very

possibility of mitigated blameworthiness (e.g., "A crime is a crime"; "There's no justification for spitting on someone, whether you're mentally ill or not"). Such claims are beside the point of a debate between competing accounts of moral responsibility, where both camps agree that moral responsibility can be undermined in certain circumstances. The second comparable category was Control (cited by 30% of participants), in which participants cited Tom's autonomy or control over his behavior as a requirement for blameworthiness. We did not include such responses within the Normative Competence category since Deep Self theory might also be able to account for control requirements (e.g. by suggesting that one's deep self is sometimes unable to control one's "surface" psychological states and behavior—as in the case of overwhelming desires). Nevertheless, there is an obvious sense in which an inability to control what one does when one acts badly *is* to be unable to do the right thing, and indeed Normative Competence theorists often refer to their responsibility requirements in terms of control (Nelkin, 2008; Nelkin, 2015; Wolf, 1994, pp. 44–45, 68). It is striking, therefore, that the second-highest category of justification that allows for variations in moral responsibility involves something very conceptually close to normative competence.

In summary, then, Study 3 suggests that the influence of deep self considerations does *not* extend to people's explicit, conscious reasoning about the basis of blameworthiness. Instead, people tend to appeal to normative competence considerations in justifying blameworthiness ascriptions.

6. The Future of Normative Competence and Deep Self Theory

In all, our studies provide consistent evidence that ascriptions of blameworthiness are sensitive to *both* normative competence *and* deep self considerations (Studies 1 and 2). When non-philosophers are challenged to explain *why* an agent is or is not blameworthy, considerations related to normative competence are commonly cited (Study 3), suggesting that normative competence presents as a relatively accessible ground of moral responsibility. Nevertheless, deep self considerations, too, have a role in actually driving ordinary blameworthiness assignments.

6.1. Assessing the Prospects of the Case Strategy

Thus, our first two studies suggest that, although Normative Competence theorists were right to think that normative competence considerations—specifically, those relating to "intellectual" competence—are relevant to ordinary judgments of blameworthiness in response to cases involving certain mitigating circumstances, they were wrong to think that deep self considerations aren't. Consequently, they were mistaken to think that the ability of Normative Competence theory to explain blameworthiness judgments in cases of mitigating circumstances—especially cases of poor upbringing—is a reason to reject Deep Self theory in favor of a theory that embraces normative competence as an independent requirement of moral responsibility.

On the other hand, our results also suggest that caution is warranted in drawing overly strong conclusions from recent experimental results in support of Deep Self theory over Normative Competence theory. While we did corroborate prior results from experimental philosophy showing that intuitions about the deep self play an explanatory role in a variety of ordinary judgments of blameworthiness and moral responsibility, we were the first to investigate whether intuitions in these cases are also explained by

considerations of normative competence. And we found evidence that they are: when agents are deemed less blameworthy due to mitigating circumstances, this is in part because they are judged less able to understand that what they are doing is wrong. Thus, although previous experimental work was interpreted as supporting the Deep Self theory, in fact it is not clear that ordinary judgments of cases provide Deep Self theory with a significant advantage over Normative Competence theory—at least, versions of Normative Competence theory according to which intellectual competence is required for an agent to be morally responsible.

For this reason, we suggest that neither Normative Competence nor Deep Self theorists can use the case strategy to support their theory over the other. This follows even if our deep self attributions *ought not* depend on information about facts concerning agents' formative circumstances or mental health (for instance, if we ought not think that poor formative circumstances obscure a truly good deep self, rather than simply altering the deep self). To see why, recall the way that philosophers previously argued that judgments about these cases cohere with Normative Competence theory. Their argument did not depend on whether an agent's level of normative competence *really* is affected by the quality of her upbringing, but on whether our blameworthiness ascriptions depend on what we take her level of normative competence to be. As such, their argument hinged on a claim, not about the actual cognitive and emotional impact of various kinds of childhood experiences, but about the explanatory basis of blameworthiness ascriptions. Our studies indicate that this explanatory claim is mistaken, and that ordinary blameworthiness ascriptions depend on the degree to which we *both* take the agent in question to be normatively competent *and* attribute her behavior to her deep self.[12]

Conflicting evidence could, of course, be uncovered by future work. One reason is that our studies only addressed very limited examples of immoral actions, and only addressed two types of mitigating context (bad upbringing and mental illness) and only particular manifestations of each (racist moral education and paranoid delusions/psychopathy, respectively). Future investigation could uncover that judgments in other mitigating circumstances are explained *just* by normative competence or *just* by the deep self, in which case application of the case strategy to cases in those circumstances would yield support for one theory over the other. But this possibility seems unlikely. Not only are formative circumstances and mental illness important mitigating contexts in their own right, there seems to be nothing that they have in common and that at the same time sets them apart from other types of mitigating circumstance. Thus, we might reasonably expect the role played by normative competence and the deep self in these cases to generalize to other mitigating circumstances.

Another way future research could challenge the conclusion that the case strategy offers comparable support to Normative Competence and Deep Self theories is through examining the folk concepts of normative competence and the deep self in more detail. For example, future research could uncover that the ordinary conception of the deep self is radically different from the way the deep self has been understood by philosophers. For instance, recent research has begun to explore the possibility that the deep self is conceived as an immutable "essence" at the core of the individual, which *gives rise to* (rather than being *constituted by*) his psychological characteristics (Christy, Schlegel, & Cimpian, 2016). If the deep self were conceived in this way, the ordinary deep self concept would represent a clear departure from the one endorsed by Deep Self theorists.

Supposing such a departure, it is less clear whether the case strategy would provide comparable support to Deep Self theory after all, which might then jeopardize its standing vis-à-vis Normative Competence theory. On the one hand, some theorists might see grounds for sufficient coherence, such that ordinary ascriptions would continue to offer comparable support to Deep Self theory, particular versions of which could then be seen as improving upon the folk concept. Conversely, others might argue that the ordinary concept depends on notions that are so implausible or irrelevant to questions of moral responsibility as to render the case strategy useless for Deep Self theorists, while leaving intact its support for Normative Competence theory.

Either way, as our studies say nothing conclusive concerning the existing degree of fit between the folk and philosophical concepts of the deep self, the current state of evidence indicates that the case strategy favors neither theory.[13]

Nevertheless, one of these theories might be favored on other grounds related to the empirical research presented here. Earlier, we said that the case strategy falls under a family of strategies premised on the idea that a theory's coherence with our moral responsibility practices and judgments might support that theory. But a theory might exhibit its fit with ordinary practices and judgments in other ways than by reflecting the causal basis of judgments about cases. To conclude, we consider two such possibilities, outlining the kind of empirical and philosophical work needed to shift the debate.

6.2. Capturing Ordinary Justifications of Moral Responsibility Attributions

Whereas the case strategy considers those features that causally contribute to our moral responsibility judgments, the activity of justifying these judgments might be considered a part of these practices, too. If so, then support for a theory of moral responsibility could also derive from accounting for agents' moral responsibility in a way that reflects ordinary justifications. The results of Study 3, indicating that normative competence considerations frequently feature prominently in such justifications, would thus uniquely support the Normative Competence theory.

Of course, whether this particular line of argument succeeds depends on a number of questions our studies leave unaddressed, not least of which is whether the activity of justification forms a part of ordinary responsibility practices, let alone a significant one. For example, our study does not tell us the extent to which people reflect on and discuss the grounds of particular moral responsibility ascriptions when not artificially prompted to do so; for all we know, these justificatory activities may not be part of "our practices" in everyday life. Of course, in the legal context, the standards for criminal responsibility *are* explicitly framed and discussed in terms of normative competence, but even here it remains to be seen whether practices of *criminal* responsibility form an important part of our practices of *moral* responsibility.[14]

Moreover, it is also unclear what it would mean for this line of argument were people to regularly rely on such normative competence justifications. For example, the mere frequency with which people engage in some behavior might give no indication of its significance; and, even assuming that it did, the fact that people seem to have just a partial understanding of what actually influences their judgments might undermine whatever significance this justificatory activity would otherwise possess.

So, although the results of our third study might bode favorably for Normative Competence, this particular line of argument depends on claims that require further defense.

6.3. Justifying Ordinary Moral Responsibility Attributions

The case strategy and the strategy just elaborated focus on different empirical claims: whereas the first relies on an empirical claim concerning the causal basis of ordinary attributions of moral responsibility, the second focuses on a claim concerning the considerations people cite in support of those attributions. Still a third strategy might depend on the simple fact that people make the moral responsibility attributions that they do—that people, for example, attribute less moral responsibility for bad actions to agents coming from bad formative circumstances or who suffer from delusions. This strategy might involve assessing the ability of different theories of moral responsibility to justify these attributions, regardless of their cause or the way that people ordinarily justify them.

Our results introduce different challenges for Normative Competence and Deep Self theories when it comes to the use of this strategy. Although both theories appear well-positioned to support the mitigated blameworthiness of agents with delusions, this does not seem to be the case with bad formative circumstances. For, on the one hand, it is clear how growing up under such circumstances would leave one with diminished normative competence, making Normative Competence theory well-equipped to justify the judgment that such agents are less blameworthy. By contrast, however, it is much less clear whether a plausible account of deep self attributability could do the same. As discussed in Section 2, there seems to be no good reason to think that deep self attributability is reduced under such circumstances, and good reason to anticipate the opposite. Thus, absent some

plausible explanation as to why such attributability would be diminished in the case of agents coming from bad formative circumstances, this seems to provide some support for Normative Competence theory over Deep Self theory.

Nevertheless, our studies also brought out one judgment that it might be *more* difficult for Normative Competence theory to justify: namely, the judgment that individuals with psychopathy are not excused in the way that agents suffering from other mental illnesses (e.g. delusions) are. Psychopathy might easily be understood in terms of an inability to truly grasp, or be motivated by, moral reasons, which is why a number of Normative Competence theorists have argued that those with psychopathy *lack* moral responsibility (Fine & Kennett, 2004; Nelkin, 2015). At the very least, then, it is difficult to see how this sort of theory could support the view that people with psychopathy *should* be held morally responsible. Deep Self theory, on the other hand, would seem to have little trouble justifying blaming agents with psychopathy: precisely because their behavior arguably results from a thoroughgoing lack of care for the wellbeing of others and for moral reasons quite generally, it seems plausible to attribute their behavior to a bad deep self (Talbert, 2008). Thus, even if Normative Competence theory were better positioned to support attributions of mitigated blameworthiness to agents coming from bad upbringings, whatever advantage it might thereby accrue over Deep Self theory would seem to balance out in light of its difficulty accommodating attributions of increased blameworthiness to individuals with psychopathy.

As with the case strategy, then, whether this line of argument ultimately favors one or the other of these theories cannot be resolved on the basis of our results alone, though future philosophical work could help make this determination. For instance, supporters of

either theory could try to show that—despite initial appearances—their theory *does* support the judgment in question. For Normative Competence theorists, this might involve rethinking the effects of psychopathy on normative competence, or showing that the competence at issue requires less than is often supposed. Deep Self theorists, on the other hand, might show that bad formative circumstances—like certain kinds of manipulation[15]—undermine the proper development of a deep self, so that the actions of agents with that background would lack full attributability.

Ultimately, our studies suggest that the answer to the question as to which of these theories—whether Normative Competence or Deep Self theory—better coheres with everyday practices and judgments of moral responsibility is a lot less clear than originally supposed. First, despite popular expectations coming out of the philosophical and empirical literature, our studies suggest that the actual explanation for blame mitigation is much less tidy, involving appeal to *both* normative competence *and* the deep self. Second, although other aspects of our results might be used in the development of at least two other lines of argument—one focused on how well these theories *capture* ordinary justifications of moral responsibility judgments and another on how well they can themselves *justify* such judgments—whether either favors Normative Competence or Deep Self theory is a question answerable only through more philosophical work.

Acknowledgements: We would especially like to thank Joshua Knobe for multiple discussions and comments on earlier drafts. In addition, we would like to thank Stephen Darwall, David Faraci, Eddy Nahmias, Kevin Tobia, and Gideon Yaffe for comments on drafts, and members of Joshua Knobe's Philosophical Psychology lab and audiences at the Buffalo Annual Experimental Philosophy Conference for comments. This project was funded by Yale University through its support of Joshua Knobe's Philosophical Psychology research lab.

Data Availability Statement: The data that support the findings of this study are available from the author J D-C upon reasonable request.

Declaration of Interest: None.

Notes

1. To be sure, the view we are identifying as "Deep Self theory" has come under various names, including the deep self view (Wolf, 1987; Sripada, 2016), the self-disclosure view (Benson, 1987; Watson, 1996), the real self view (Wolf, 1990), and attributability theory (Faraci & Shoemaker (2019). Its defenders have included, perhaps most prominently, Frankfurt (1971, 1975, 1987) and Watson (1987, 1996). More recent defenders include Scanlon (2008, pp. 180–181, 192–194); and Bratman (1996). Notably, though, even this account of the responsibility conditions fails to capture all those views that have been recognized as falling under Deep Self theory. This is because some instantiations of the view do not require that the actions of morally responsible agents *actually* reflect their deep self, only that these agents have the *ability* to make it so (e.g. Watson, 1987, 1996).
2. Normative Competence theorists include Wolf (1990) and Nelkin (2008, 2011, 2013). One thing separating Wolf and Nelkin from other theorists who have incorporated normative competence considerations into their views is their requirement that morally responsible agents have the ability to act rightly at the time, and in the circumstances, of their acting.
3. For discussion, see e.g. Wolf (1990, p. 88) and Nelkin (2011, p. 7).
4. Wolf does make claims to the effect that agents who satisfy conditions for moral responsibility on Normative Competence theory will satisfy conditions on Deep Self theory as well. For example, she claims that when an agent does the right thing for the right reasons, her action is also attributable to her deep self, writing, "If she has the right reasons, they are presumed to be among her values" (1990, p. 85). But Wolf does not explain why this should be presumed, and the above example provides reason to think that it should not (see also e.g. Crisp, 2015, who argues that agents might do the right thing for the right reasons without its being the case that those reasons align with their values, or are consistent with their character). Perhaps Wolf's thought is that normatively competent agents are ones who have the *ability* to govern their actions in accordance with their deep selves, even if they do not always exercise that ability (e.g. even if the professor who acted rightly in helping the student failed to act in accordance with his deep self, nevertheless he had the ability to do so). The trouble with this suggestion is that it conflicts with Wolf's claims to the effect that morally responsible agency does not require the ability to act badly or wrongly.
5. In her earlier article, Wolf (1987) presents a view that combines deep self and normative competence considerations. There, Wolf characterizes morally responsible agency, *not* in terms of possession of "the ability to do the right thing for the right reasons," but in terms of the possession of a "sane deep self"—one that is *both* "able to govern her (or his) actions by her desires and to govern her desires by her deep self," *and* able to "cognitively and normatively…recognize and appreciate the world for what it is" (p. 335) (cf. Wolf, 1990, p. 117, where she describes morally

responsible agents as ones who have "the ability to form, assess, and revise [their] values on the basis of a recognition and appreciation of… the True and the Good").

6. Faraci & Shoemaker (2010, 2014, 2019) allow that moral ignorance (e.g. knowledge that actions expressing ill will are wrong) might be partly responsible for judgments of mitigated blameworthiness in cases involving childhood deprivation, but they distinguish between such moral ignorance, on the one hand, and diminished normative competence, on the other. Normatively incompetent agents and morally ignorant agents both lack moral knowledge, but only normatively incompetent ones lack the ability to acquire such knowledge.

7. In Faraci and Shoemaker's cases, deep self-related information is limited (we are told only that the agent is a "proud racist" who "fully embraced what he'd been taught"), and the relation between these states and the deep self is left implicit. By contrast, in cases used by Normative Competence theorists, we are given more information about the agent's psychology. In the JoJo case, for instance, we learn not just that JoJo develops his father's values, but that these are values he "wholly *wants* to have. When he steps back and asks, 'Do I really want to be this sort of person?' his answer is resoundingly 'Yes', for this way of life expresses a crazy sort of power that forms part of his deepest ideal" (Wolf, 1987, p. 379).

8. See American Psychiatric Association (2013) for official diagnostic criteria, and Haslam and Giosan (2002) on the folk concept of mental illness.

9. Dubber, 2015, chapter 3. Legal criteria for *criminal* responsibility are regarded by some as part of our wider moral responsibility practices (though Shoemaker, 2013, disagrees), and that they are framed in terms of normative competence has been taken to support views of *moral* responsibility doing the same (Nelkin, 2008, p. 498; Nelkin, 2011, pp. 7–8; Wolf, 1987, pp. 55–56).

10. On the debate about the relationship between psychopathy and moral responsibility, see e.g. Watson (2011), Talbert (2008), and Nelkin (2015). For an overview of legal treatments of psychopathy, see Fine & Kennett (2004).

11. Indeed, it is possible that participants think of these differently motivated agents as committing importantly different action types.

12. If we were *never* justified in judging an agent's behavior as being less attributable to her deep self because of some fact about her circumstances, this might be more problematic for Deep Self theory insofar as the view would have little explanatory value.

13. Another possibility is that ordinary moral responsibility judgments and practices underwrite different conceptions of morally responsible agency, and thus different conceptions of blameworthiness, to which distinct considerations are relevant (e.g. Watson, 1996). If so, it might be that one or the other of these theories fares better in capturing the considerations relevant to one of these responsibility notions than does the other. Future research will have to address this possibility.

14. Shoemaker (2013) argues these practices are importantly different.

15. Although this development would represent a break with non-historical variants of Deep Self theory, it is not without some precedent (see Mele, 2013).

References

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders*, 5th Edition. Arlington, VA: American Psychiatric Publishing.

Benson, P. (1987). Freedom and value. *The Journal of Philosophy*, *84*, 465–486.

Bratman, M. (1996). Identification, decision, and treating as a reason. *Philosophical Topics, 24*,1–18.

Christy, A., Schlegel, R., & Cimpian, A. (2016). Why do people believe in true selves? The role of essentialist reasoning about personal identity and the self. *Journal of Personality and Social Psychology, 117,* 386–416.

Crisp, R. (2015). A third method of ethics?, *Philosophy and Phenomenological Research*, *90*, 257–273.

Dubber, M. (2015). *An introduction to the model penal code.* Oxford, UK: Oxford University Press.

Faraci, D., & Shoemaker, D. (2019). Good selves, true selves: Moral ignorance, responsibility, and the presumption of goodness. *Philosophy and Phenomenological Research, 98,* 606–622.

__. (2014). Huck vs. JoJo: Moral ignorance and the (a)symmetry thesis. In T. Lombrozo, J. Knobe, & S. Nichols (Eds.), *Oxford studies in experimental philosophy*, Vol. 1 (pp. 7–27). Oxford, UK: Oxford University Press.

__. (2010). Insanity, deep selves, and moral responsibility: The case of JoJo. *Review of Philosophy and Psychology, 1,* 319–332.

Fine, C., & Kennett, J. (2004). Mental impairment, moral understanding and criminal responsibility: Psychopathy and the purposes of punishment. *International Journal of Law and Psychiatry, 27,* 425–443.

Frankfurt, H. (1987). Identification and wholeheartedness. In F. Shoeman (Ed.), *Responsibility, character, and the emotions: New essays in moral psychology* (pp. 27–45). Cambridge, UK: Cambridge University Press.

__. (1975). Three concepts of action. *Proceedings of the Aristotelian Society, 49,* 95–125.

__. (1971). Freedom of the will and the concept of a person. *The Journal of Philosophy, 68,* 5–20.

Haslam, N., & Giosan, C. (2002). The lay concept of ‘mental disorder’ among American undergraduates. *Journal of Clinical Psychology, 58,* 479–485.

Hayes, A. (2012). PROCESS: A versatile computational tool for observed variable mediation, moderation, and conditional process modeling [White paper]. Retrieved from http://www.afhayes.com/public/process2012.pdf

Hayes, A., & Preacher, K. (2014). Statistical mediation analysis with a multicategorical independent variable. *British Journal of Mathematical and Statistical Psychology, 67,* 451–470.

Mele, A. (2013). Manipulation, moral responsibility, and bullet biting. *Journal of Ethics, 17,* 167–84.

Nelkin, D. (2015). Psychopaths, incorrigible racists, and the faces of responsibility. *Ethics, 125,* 357–390.

__. (2013). Desert, fairness, and resentment. *Philosophical Explorations, 16,* 117–132.

__. (2011). *Making sense of freedom and responsibility*. Oxford, UK: Oxford University Press.

__. (2008). Responsibility and rational abilities: Defending an asymmetrical view. *Pacific Philosophical Quarterly, 89,* 497–515.

Newman, G., De Freitas, J., & Knobe, J. (2015). Beliefs about the true self explain asymmetries based on moral judgment. *Cognitive Science, 39,* 96–125.

Preacher, K., & Hayes, A. (2008). Asymptotic and resampling strategies for assessing and comparing indirect effects in multiple mediator models. *Behavior Research Methods, 40,* 879–891.

Scanlon, T. M. (2008). *Moral dimensions*. Cambridge, MA: Harvard University Press.

Shoemaker, D. (2013). On criminal and moral responsibility. In M. Timmons (Ed.), *Oxford studies in normative ethics,* Vol. 3 (pp. 154–178). Oxford, UK: Oxford University Press.

Sripada, C. (2016). Self-expression: A deep self theory of moral responsibility. *Philosophical Studies, 175,* 1203–1232.

__. (2012). What makes a manipulated agent unfree? *Philosophy and Phenomenological Research, 85,* 563–593.

__. (2010). The deep self model and asymmetries in folk judgments about intentional action. *Philosophical Studies, 151,* 159–176.

Talbert, M. (2008). Blame and responsiveness to moral reasons: Are psychopaths blameworthy? *Pacific Philosophical Quarterly, 89,* 516–535.

Watson, G. (2011). The trouble with psychopaths. In R.J. Wallace, R. Kumar, & S. Freeman (Eds.), *Reasons and recognition: Essays on the philosophy of T. M. Scanlon* (pp. 307–331). Oxford, UK: Oxford University Press.

__. (1996). Two faces of responsibility. *Philosophical Topics, 24,* 227–248.

__. (1987). Free action and free will. *Mind, 96,* 145–172.

Wolf, S. (2012). Blame, Italian style. In R. J. Wallace, R. Kumar, & S. Freedom (Eds.), *Reasons and recognition: Essays on the Philosophy of T.M. Scanlon* (pp. 332–347). Oxford, UK: Oxford University Press.

__. (2002). The True, the good, and the loveable: Frankfurt's avoidance of objectivity. In S. Buss and L. Overton (Eds.), *Contours of agency: Essays on themes from Harry Frankfurt* (pp. 227–244). Cambridge, MA: MIT Press.

__. (1990). *Freedom within reason*. Oxford, UK: Oxford University Press.

__. (1987). Sanity and the metaphysics of responsibility. In F. Shoeman (Ed.), *Responsibility, character, and the emotions: New essays in moral psychology* (pp. 46–62). Cambridge, UK: Cambridge University Press.