GOOD SELVES, TRUE SELVES: MORAL IGNORANCE, RESPONSIBILITY, AND THE PRESUMPTION OF GOODNESS

David Faraci
Georgetown University

David Shoemaker
Tulane University

In a remarkable series of studies, George Newman, Joshua Knobe, and colleagues have been making a strong case that most of us believe in the better angels of our nature, that is, we believe that our fellows are, essentially, *good*. According to the *Good True Self* (GTS) theory, if an action is deemed good, its psychological source is typically viewed as more reflective of its agent's true self, of who the agent really is "deep down inside"; if the action is deemed bad, its psychological source is typically viewed as more *external* to its agent's true self.[1]

In previous work, we discovered a related asymmetry in judgments of blame- and praiseworthiness with respect to the mitigating effect of *moral ignorance via childhood deprivation*. Inspired by work motivating the GTS theory, we ran a new study to discover whether our asymmetry likewise reflected judgments about the true self. It did. However, it is unclear whether our results fit with the *good* part of the GTS theory: some of our and others' results suggest that, in certain contexts, *wrong* actions are taken to be more expressive of agents' true selves than right ones.

[1] Though our focus is moral responsibility, relevant effects have been found in several different arenas, from what psychic states count among one's values, to assessments of happy lives among self-described contented people, to determinations of weakness of will, to sexual orientation. For summary discussion of this literature and relevant references, see Newman, Bloom, and Knobe 2013; and Newman, De Freitas, and Knobe 2014.

In this paper, we propose that our and others' data can be explained by the hypothesis that we are inclined to judge as the GTS theory predicts when there is a *readily available* external explanation for an agent's action. In short, we give people *the benefit of the doubt*. There are a number of possible explanations for this tendency, possibly including that we are inclined to see others as good "deep down," as the GTS theorist holds (but that this is blocked when no external explanation is readily apparent). Thus, further thought and study are called for to determine whether our hypothesis is best viewed as providing a substantial amendment to the GTS theory or the seeds of a replacement theory.

We proceed as follows. First, we briefly introduce *Attributability Theory*—the view that responsibility is a function of the causal and concordance relations between an action and its agent's true self—as a philosophical framework for thinking about moral responsibility. Second, we discuss in detail some of the experimental results that lend support to the GTS theory, ending with our own studies and new results. Third, we discuss the implications of these new results for our previous work and for Attributability Theory. Fourth, we explain how our new results might challenge the GTS theory, especially in light of one plausible version of Attributability Theory. Fifth, we motivate our *benefit of the doubt* hypothesis and discuss its relationship with the GTS theory. We conclude with some worries about the methodology adopted thus far in the relevant literature.

Attributability Theory

The core thought behind Attributability Theory is that the objects of moral approval and disapproval are agents, enduring entities. Because of this, for an agent to be the proper target of praise or blame *for* the action of a particular moment, that action must be expressive *of* that agent, an agential fingerprint, as it were, on the window of the world (see Hume 1739/1969 and, for more contemporary framing Sher 2006: Ch. 2). But not just any action or attitude is expressive of the

agent in a way that aptly grounds blame or praise; rather, the action or attitude must have its source in some privileged, inalienable psychic feature of the agent, something that represents who the agent really is deep down inside. This privileged psychic domain, whatever it consists in, is sometimes known as the "deep self," other times as the "true self." We will use both terms interchangeably.[2]

The nature of the deep self is a predictable source of dispute amongst deep self theorists. Several theories have cropped up over the years, but they basically fall into two camps: non-cognitive and cognitive. The former point exclusively to features like desires or emotional dispositions (cares) as the ultimate location of the deep self (Frankfurt 1988 and 1999; Shoemaker 2003; Sripada 2010), whereas the latter point exclusively to evaluative judgments, or a general evaluative stance, as its home (Scanlon 1998; Watson 2004; Smith 2005).

Both camps face difficulties, which we will not rehearse here. In light of relevant problems with both kinds of exclusivist accounts of the deep self, one of us has recently developed a pluralistic account, maintaining that as long as actions or attitudes flow from either one's emotional dispositions (cares) *or* one's evaluative stance (commitments), they express one's deep self (OMITTED; also see Sripada 2015). This is an attempt to capture all of those things that *matter* to us under the rubric of the deep self,[3] and mattering, it may be thought, is best captured by both cognitive and non-cognitive elements.[4]

What all these theories have in common is that they attempt to provide a way to distinguish the "psychic junk"—random thoughts, images, impulses, and compulsive urges—from the

---

2 In the psychological literature, "true self" is the term deployed most often, whereas in the philosophical literature, it is "deep self."

3 See Wolf (1990: 31) for mention of 'mattering' in this context.

4 What happens in cases of conflict? In such cases, *we* are in conflict, ambivalent between the warring parts of our deep selves. For discussion of *attributable ambivalence*, see (OMITTED).

psychological elements that warrant attribution of an action or attitude to an agent for purposes of holding the agent responsible. An action or attitude is attributable to an agent in virtue of its expressing something truly *of the agent,* whether that deep self consists in second-order desires, evaluative judgments, cares, or general evaluative commitments.

This expression relation itself isn't always explicitly spelled out, but one necessary condition is invariably thought to be causal, i.e., an action or attitude expresses the deep self only if it causally depends on the deep self. There may be reasons to doubt this causal requirement, but we will set those aside here (see OMITTED). What Chandra Sripada adds to the mix is a *concordance* requirement, namely, that any attributable action or attitude must also be in harmony with the values in which the deep self consists (Sripada 2010). This is to block attributability of actions or attitudes disharmonious with the deep self that are nevertheless produced via some malfunctioning causal mechanism. Nearly all attributability theorists would take these two necessary conditions to be jointly sufficient for expression.

We thus arrive at the generally agreed-upon schematic view that the deep self is specified in terms of some privileged subset of psychic elements (e.g., cares and/or commitments), and an action or attitude is attributable to an agent in virtue of its expressing (being both caused by and in concordance with) that deep self. This renders determinations of attributability a simple matter of tracing the relevant action or attitude to those particular causal and concordant sources: either the relation obtains or it doesn't.[5]

With Attributability Theory in hand, we now turn to consider some old and new experimental results lending support to the GTS theory. These results represent something of a

---

[5] Though the *extent* to which the action expresses the agent's deep self vs. other sources may determine in more complicated fashion the *level* of responsibility.

double-edged sword for Attributability Theory. On the one edge, they seem to confirm that judgments about the true self do mediate judgments of blame- and praiseworthiness. On the other, they threaten to undermine a basic assumption of the theory, the assumption that *all* that matters is whether the expression relation obtains between certain psychological features of the agent and the actions or attitudes in question. In other words, the normative status of those actions or attitudes is typically assumed to be irrelevant to whether that expression—and thus responsibility—obtains.

Previous Data

Three sets of moral responsibility results to date lend support to the GTS theory. First, Pizarro et al. (2003) showed that people view emotional swamping as mitigating responsibility primarily where the emotionally influenced action is viewed as bad (e.g., an enraged person smashes the window of a car he perceives as having been parked too close to him vs. someone who smashes the car window calmly and deliberately). When it is viewed as good, subjects assign agents just as much praiseworthiness as they do to agents who did what they did in a sober and deliberate fashion (e.g., giving a homeless man one's coat calmly vs. doing so overwhelmed by sympathy). What Newman, De Freitas, and Knobe found in addition is that

> beliefs about the true self explain this effect. In the case where the agent's emotions draw him to do something morally bad, these emotions are seen as lying outside his true self and, in turn, he is given less blame. However, in the case where the agent's emotions draw him to do something morally good, the emotions are seen as part of his true self and so he is given as much praise as if there were no conflict (Newman, De Freitas, and Knobe 2014).

Notice the sequence: (1) the assessment of an action as good/bad looks to be the source of (2) the belief that the emotions causing it lie inside/outside the true self, which then explains (3) the

assignment of unmitigated praise/mitigated blame. So with respect to the connection between (2) and (3), the results indicate that the *reason* people show this pattern of praise/blame judgments is just that there is a similar pattern to their true self judgments.[6]

The second asymmetry was not presented as a moral responsibility asymmetry, but it is explicitly about attributability, which, as we saw earlier, purportedly grounds responsibility judgments. Here the relevant asymmetry goes to whether attitudes about homosexuality are attributed to the agent's true self: they tend to be when the attitude in question is deemed good; they tend not to be when the attitude in question is deemed bad. The basic idea was to present self-declared liberals and conservatives with one of two scenarios in which someone experiences a tension between his feelings and his beliefs about same sex relationships. The first scenario went as follows:

> Mark is an evangelical Christian. He believes that homosexuality is morally wrong. In fact, Mark now leads a seminar in which he coaches homosexuals about techniques they can use to resist their attraction to people of the same-sex [sic.]. However, Mark himself is attracted to other men. He openly acknowledges this to other people and discusses it as part of his own personal struggle.

Here was the second scenario:

> Mark is a secular humanist. He believes that homosexuality is perfectly acceptable. In fact, Mark leads a seminar in which he coaches people about techniques they can use to resist their negative feelings about people who are attracted to the same sex. However, Mark himself has a negative feeling about [the] thought of same-sex

---

6 Thanks to OMITTED for discussion.

> couples. He openly acknowledges this to other people and discusses it as part of his own personal struggle (Newman, Bloom, and Knobe 2013: 7).

Subjects from different political backgrounds were inclined to attribute Mark's feelings to his true self, despite their being in tension with his evaluative beliefs. (This in itself is rather remarkable, as the currently dominant version of Attributability Theory views the true self as the domain of evaluative judgment (e.g., Watson 2004; Scanlon 2008; Smith 2005 and 2012).) The relevant asymmetry here arises from the comparison between the response that the feelings were more representative of Mark's true self than his beliefs and the response that *both* were representative of his true self. In responding to the first scenario, 57% of liberals thought Mark's feelings expressed his true self while his beliefs were "peripheral," whereas only around 30% thought that both his feelings *and* his beliefs expressed his true self. When it came to the second scenario, however, the trend reversed: more of the liberals (43%) thought that both attitudes expressed his true self than that only his feelings did (38%) (Newman, Bloom, and Knobe 2013: 8).

For the conservatives, while a general asymmetry was also in place, it was starkly reversed. In responding to the first scenario, only about 26% of conservatives thought Mark's feelings expressed his true self, whereas around 42% thought that both his feelings and his beliefs expressed his true self. In the second scenario, however, there was a huge disparity: 68% of the conservatives thought that only Mark's feelings really expressed his true self, whereas only around 20% thought both attitudes did so. The results here strongly suggest that what subjects judge to be attributable to true selves—and so, presumably, what attitudes are attributable to agents for purposes of moral responsibility—is a function of those subjects' antecedent normative stances toward the attitudes in question.

The third moral responsibility asymmetry is displayed in our own studies. In previous collaborations, we explored the effect of moral ignorance stemming from "morally blinkered"

formative circumstances on attributability judgments, as illustrated by two classic cases in the literature. The first is Susan Wolf's fictional case of JoJo, the son of a brutal dictator, who grows up to be just like his dad and fully embraces and endorses his dad's values. When he tortures a peasant on a whim, Wolf claims, our "pretheoretical intuitions" (Wolf 1987: 56) are that JoJo is not responsible, as it "is unclear whether anyone with a childhood such as his could have developed into anything but the twisted and perverse sort of person that he has become" (Wolf 1987: 54). We found to the contrary that subjects actually assign significant blameworthiness to JoJo for what he does, although he is viewed as somewhat less blameworthy than a control without his morally deprived background (OMITTED).[7]

The asymmetry was revealed when we looked at positive actions performed despite moral ignorance via childhood deprivation, modeled on the much-discussed case of Huck Finn (see, e.g., Arpaly 2003). In testing what people thought of someone like this, we once again started with a randomly assigned pair of negative cases:

> A. Tom is a white male who was raised in New Orleans. Growing up, he was taught to respect all people equally. Nevertheless, as an adult, he decided to become a proud racist, someone who believes that all non-white people are inferior and that he has a moral obligation to humiliate them when he gets a 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. In keeping with his moral beliefs, Tom spits on the man as he passes by.

---

[7] In OMITTED, subjects assigned JoJo a mean of around 5 out of 7 on a blameworthiness scale (where 7 was "completely blameworthy" and 1 was "not at all blameworthy"), whereas the non-deprived control who tortured the peasant was assigned a mean of around 6 out of 7.

B. Tom is a white male who was raised on an isolated island in the bayous of Louisiana. Growing up, he was taught to believe that all non-white people are inferior and that he has a moral obligation to humiliate them when he gets a chance. As an adult, he fully embraced what he'd been taught, becoming a proud racist. At the age of 25, Tom moves to another town. Walking outside his home, he sees a black man who has tripped and fallen. In keeping with his moral beliefs, Tom spits on the man as he passes by.

Subjects were each given one of these two scenarios and asked to rate Tom's level of blameworthiness for spitting on the man, on a scale from 1 ("not at all blameworthy") to 7 ("completely blameworthy"). The mean assignment of Tom_A's blameworthiness was 6.68. The mean assignment to Tom_B was 5.4. This mirrored our earlier results about JoJo.

But we also simultaneously surveyed additional subjects with one of the following two randomly assigned positive versions of Tom:

C. Tom is a white male who was raised in New Orleans. Growing up, he was taught to respect all people equally. Nevertheless, as an adult, he decided to become a proud racist, someone who believes that all non-white people are inferior and that he has a moral obligation to humiliate them when he gets a chance. At the age of 25, Tom moves to another town. Walking outside his home, he sees a black man trip and fall. Usually, Tom would spit on the man. But this time, Tom goes against his current moral beliefs, and helps the man up instead.

D. Tom is a white male who was raised on an isolated island in the bayous of Louisiana. Growing up, he was taught to believe that all non-white people are inferior and that he has a moral obligation to humiliate them when he gets a chance. As an adult, he decided to become a proud racist, embracing what he was taught. At

the age of 25, Tom moves to another town. Walking outside his home, he sees a black man trip and fall. Usually, Tom would spit on the man. But this time, Tom goes against his current moral beliefs, and helps the man up instead.

This time, subjects were asked to rate Tom's level of *praiseworthiness* for helping the man up, on a scale from 1 ("not at all praiseworthy") to 7 ("completely praiseworthy"). The mean response to Tom_C was 4.28. The mean response to Tom_D was 5.40. (See Figure 1) People not only thought that moral ignorance via childhood deprivation didn't reduce praiseworthiness relative to a deliberate counterpart; they also thought that such ignorance made one *more* praiseworthy than the deliberate counterpart.

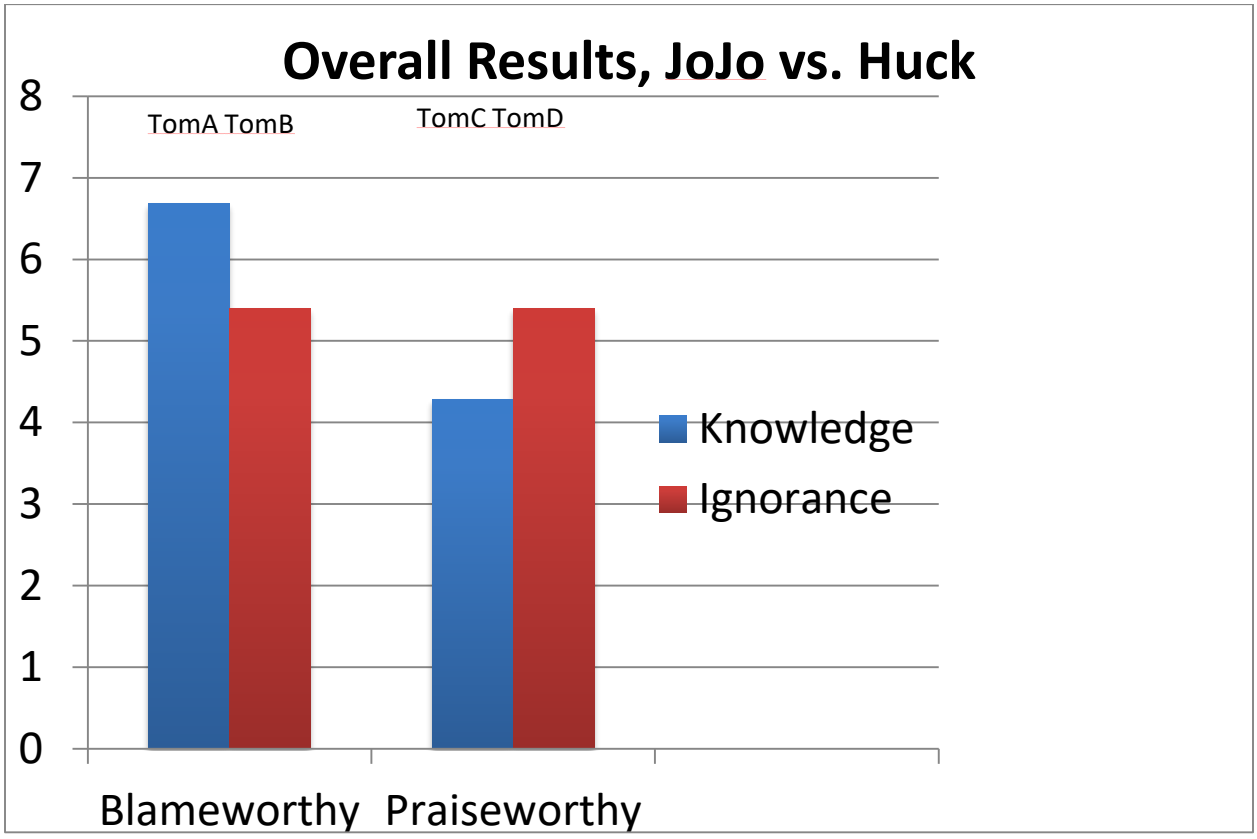

Figure 1

On its face, the results do seem to display an asymmetry: moral ignorance of this sort, according to subjects, decreases blameworthiness but not praiseworthiness.

New Results

Inspired by the work mentioned earlier, in a brand new study we set out to discover whether true self judgments mediated moral responsibility judgments in our Tom cases. We did so by running, with both design and financial assistance from [OMITTED], the exact same four scenarios as above, adding the following question to each[8]:

> On the scale below, please circle the number that best represents the extent to which you agree that what Tom did expressed his *true self* – the person he really is deep down inside.
>
> |---------|---------|---------|---------|---------|---------|
> 1 2 3 4 5 6 7
> Disagree Completely — Neither Agree Nor Disagree — Agree Completely

As it turns out, the responses lend credence to *both* a normative asymmetry *and* a true self explanation. Arguably, this is a GTS explanation at work.

Our data suggest three important results. First, our original effect was replicated yet again (see Figure 2).[9] People assign less blameworthiness for a bad action when the agent is morally ignorant because of childhood deprivation, but they don't assign less praiseworthiness for a good action to the same sort of agent (indeed, they tend to assign slightly more).[10, 11]

---

8 Participants were recruited using Amazon's MTurk, N = 307, mean age 28.9 years.

9 At least with respect to the negative cases. More on the positive cases further on.

10 The data were analyzed using a 2 (moral valence: good vs. bad) x 2 (moral knowledge: ignorant vs. knowledgeable) ANOVA. There was a main effect of moral valence, $F(1, 303) = 80.6$, $p < .001$, and a main effect of

moral knowledge, $F(1, 303) = 4.1$, $p < .05$. Most importantly, there is a significant interaction effect, $F(1, 303) = 25.5$, $p < .001$.

[11] An anonymous referee pointed out an elision in our earlier studies: Only Tom_A makes a clear-eyed decision in opposition to his upbringing, coming to racism on his own and then acting on it. It would thus have been useful to consider clear-eyed, upbringing-opposed *right*-doing, such as a Tom who was raised racist, but comes to believe in equality for all, and as a result helps up the black man on the street. The question, then, is whether this additional information about Tom_A is a better explanation for why people reacted so strongly to him, generating the greatest assignments of attributability (much more than Tom_C). We are hesitant to think there's much explanatory work being done by this disanalogy, however, and suspect that more information of that sort wouldn't have made much difference on the assignment of attributability. We deliberately left open how the positive agents come by their decisions to help the man up, so it may be, for instance, that some subjects already filled in such information. Nevertheless, we take the point, and any future studies will attempt to correct for this oversight (or at least add the relevant cell for survey purposes).

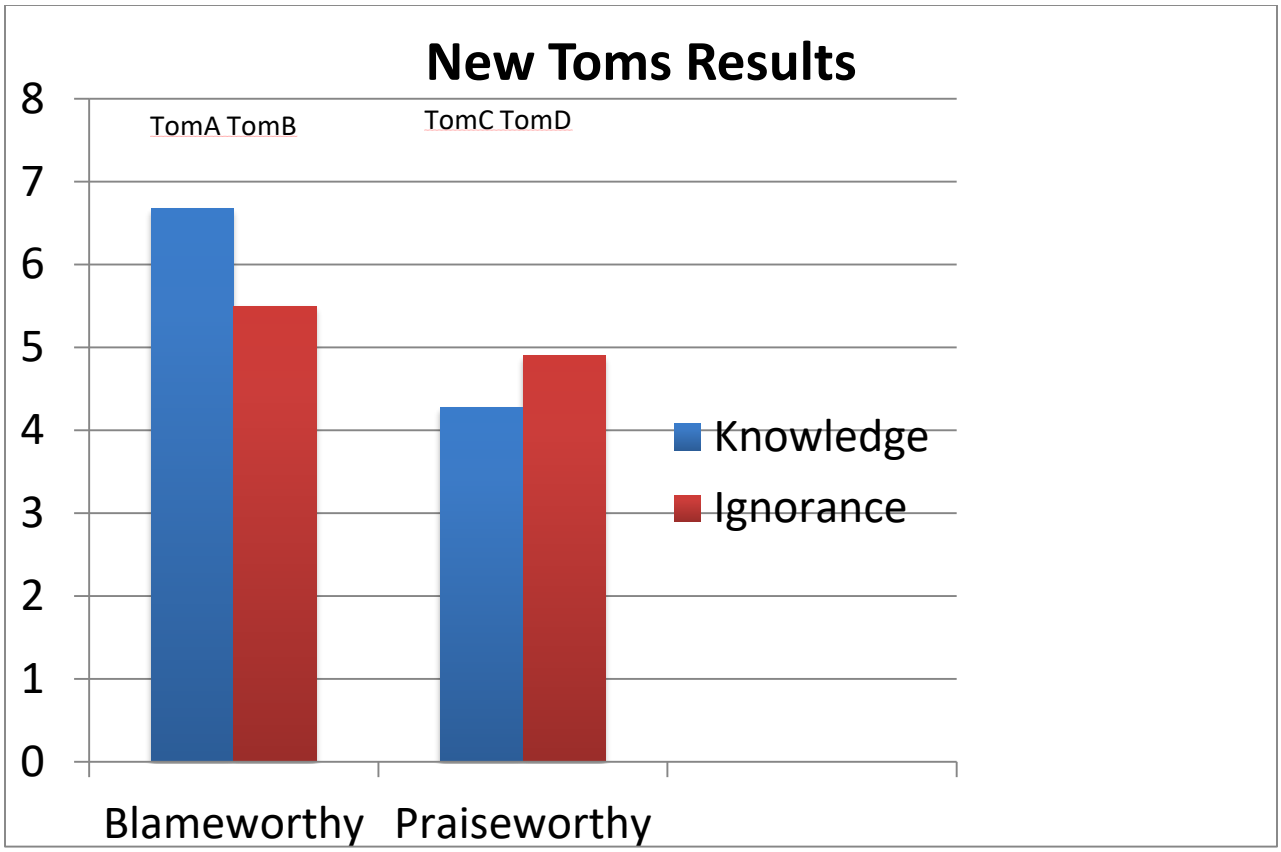

Figure 2

Second, the new true self results display just the same pattern (see Figure 3). People say that the morally bad action less expresses the agent's true self when he is morally ignorant because of childhood deprivation, but they do not think that the morally good action less expresses his true self when he is ignorant in this respect.[12]

[12] The data were analyzed using a 2 (moral valence: good vs. bad) x 2 (moral knowledge: ignorant vs. knowledgeable) ANOVA. There was a main effect of moral valence, $F(1, 303) = 26.1$, $p < .001$ but no main effect of moral knowledge, $F(1, 303) = 1.9$, $p = .17$. There is a significant interaction effect, $F(1, 303) = 4.4$, $p < .05$.

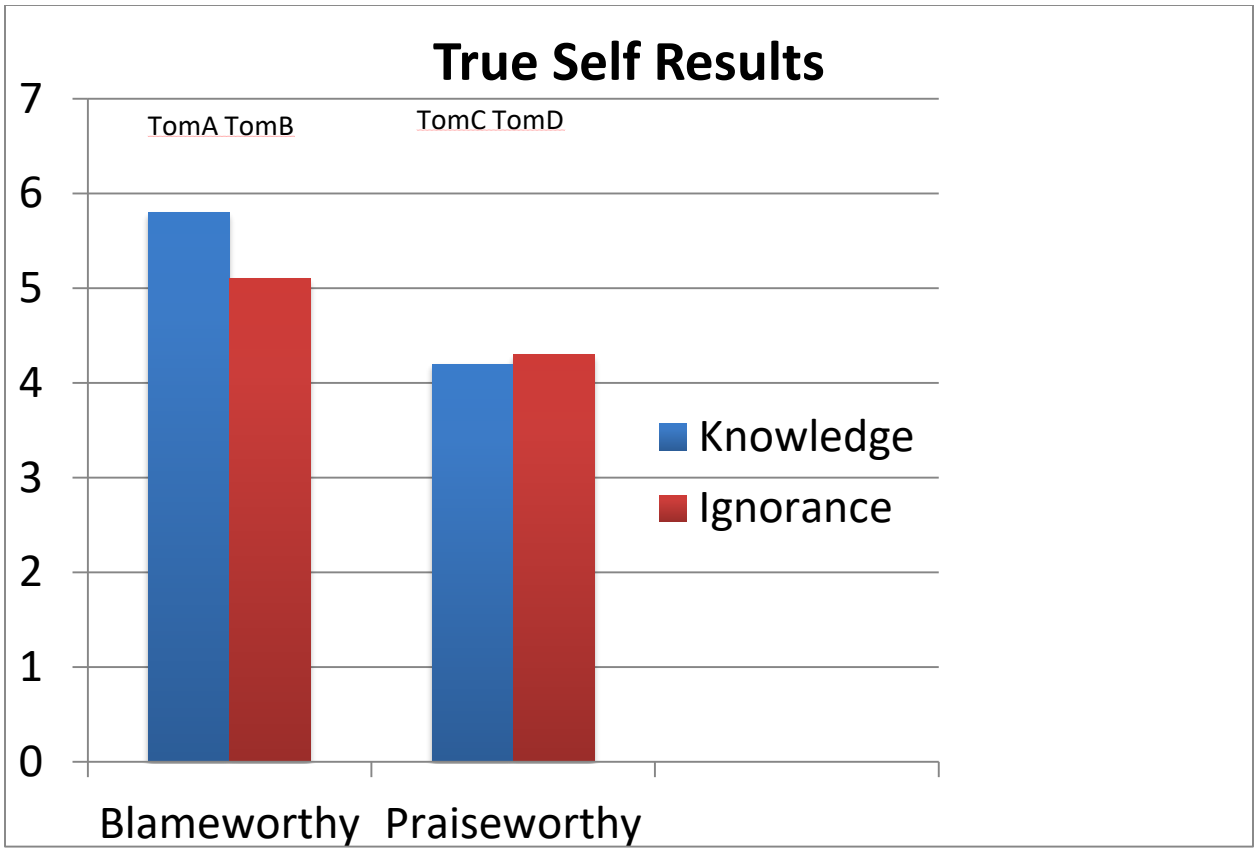

Figure 3

Third, and most importantly, the first result is *mediated* by the second (see Figure 4). In other words, the results indicate that the reason why people show such a pattern on the blameworthy/praiseworthy judgments is precisely that they show the pattern they do on the true self judgments.[13]

[13] The data were analyzed using bootstrap mediation (cf., Preacher and Hayes 2008) with the interaction term (moral valence x moral knowledge) as the independent variable, appraisal (praiseworthy or blameworthy) as the dependent variable, and true self judgments as the mediator. The analysis showed significant mediation of the interaction by true self judgments (95% CI = .04 to .21).

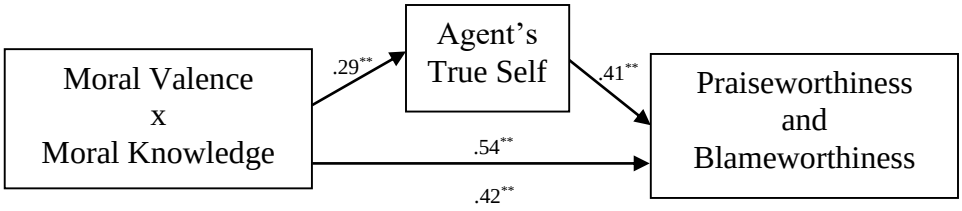

Figure 4

Implications for Our Previous Philosophical Work

In our earlier work, we discussed possible explanations for the apparent praise/blame asymmetry discovered via our Tom cases. We maintained, though, that one might preserve a kind of symmetrical understanding of the results by pointing to the difference between negative and positive assessments, not with respect to each other, but with respect to their *controls* on a complete assessment scale. In other words, think of all the assessments taking place on the following sort of scale, from completely blameworthy to completely praiseworthy (see Figure 5):

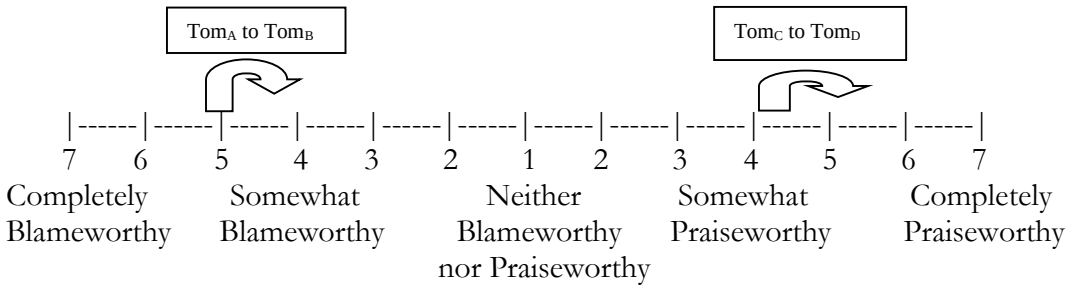

Figure 5[14]

On this scale, the direction of subjects' assessments of Tom_A and Tom_B in the first study (from 6.68 to 5.4) is precisely *symmetrical* to the direction of their assessments of Tom_C and Tom_D (from 4.28 to

14 Taken from [OMITTED].

5.4). That is, moral ignorance via childhood moral deprivation—of both negative and positive sorts—tends to get us to view the actions caused in the absence of the relevant moral knowledge as moving uniformly *away* from the "completely blameworthy" end of the scale and their controls.

But what explains such a directional movement in subjects' assessments? We offered the following *Difficulty Hypothesis*:

> Moral ignorance resulting from childhood deprivation functions symmetrically in both negative and positive cases (moving assessments up the single scale of blameworthiness to praiseworthiness in relation to the control) in virtue of the *difficulty* agents are viewed as having in overcoming their morally deprived upbringing to grasp the relevant moral reasons [OMITTED].

This model takes seriously the moral deprivations in childhood, as they are precisely what would be thought to make it more difficult for the various Toms to do the right thing. Consequently, Tom_B is viewed as less blameworthy than Tom_A in virtue of its being thought more difficult for him to recognize that he should not spit on the black man *given his upbringing*. And Tom_D is viewed as more praiseworthy than Tom_C, goes the explanation, in virtue of its being thought more difficult for him to recognize that he should do what he actually *did*, namely, help the black man up, given his upbringing.

We advanced this explanation while explicitly welcoming future attempts to undermine our hypothesis or buttress the alternatives. It now might seem we ourselves have done just that. We took ourselves to have strongly implied in our prompts that the deprived Toms had, at the time of action, the *same deep selves* as their controls. But given the True Self results, our subjects obviously did not interpret things this way. It might thus look like the correct interpretation of our results is now more in line with the GTS theory. However, as we argue below, that theory may be in tension with other elements of our data, as well as data from other studies. As we explain, resolving this tension may likewise resolve these worries about the Difficulty Hypothesis.

Potential Problems for the Good True Self Theory

Our results clearly support one aspect of the GTS theory: When presented with an agent with a morally deprived upbringing, subjects were indeed more likely to interpret the agent's action as expressive of his true self when that action was good. So our results do strongly support the *true self* part of the GTS theory. This is not an insignificant result.

But it is only one of our results. Another is the response set to the control cases (Toms $_A$ and $_C$), those in which the Toms were not morally deprived as children. There is, in these cases alone, an asymmetry between blame and praise: the mean of Tom$_A$'s blameworthiness was 6.6 (out of 7), whereas Tom$_C$'s praiseworthiness was 4.3 (out of 7). Why is deliberate and knowledgeable badness viewed as more blameworthy than deliberate and knowledgeable goodness is viewed as praiseworthy?

Attributability Theory tells us that the difference should be in virtue of (a) the degree to which the action performed is viewed as good or bad (this goes only to the extent to which the action is viewed as something for which the agent is potentially praise- or blameworthy), and (b) the degree to which the action performed is viewed as expressive of the agent's true self. In our study, Tom$_A$'s bad action produced very significant blameworthiness scores, whereas Tom$_C$'s good action produced middling praiseworthiness scores. Prima facie, this is the *opposite* of what we would expect the GTS theory to predict.

The GTS theorist might attempt to accommodate this result, however, by holding that our subjects judge Tom$_A$'s bad action (spitting on the man) to be so much more bad than Tom$_C$'s action (helping the man up) was good that this swamped the GTS effect. In other words, GTS theorists

could hold that were Tom_C to have done something that is as good as spitting on someone is bad, he would have been praised more than Tom_A was blamed.[15]

We cannot deny this possibility from the armchair. Nor, in fact, is it obvious that we can retest to rule it out, for it is unclear how we could gauge the relative goodness vs. badness judgments of our participants.[16] Nevertheless, other experimental results undercut the plausibility of this general response. Indeed, the most famous example of normative judgments' impacting agential appraisals exhibits this same pattern. In the original study supporting "the Knobe effect," when a CEO doesn't care how his decision for his company's policy will, as a side effect, impact the environment, people tend to think of that side effect as intentional when it hurts the environment, but unintentional when it helps the environment (see, e.g., Knobe 2006).

Generally, when what one does is truly accidental—nonculpably unintentional—one is excused from responsibility. In the CEO case, though, whether his decision is seen as intentional actually depends on whether subjects view the side effect (helping/hurting the environment) as good or bad: if bad, the action is viewed as intentional; if good, it's not. But this is the very *opposite* of what the GTS theory would predict, that the CEO ought to be off the hook (or less on the hook) for

---

15 Our thanks to an anonymous reviewer for highlighting the need to address this possibility. Notice that similar things could be said about the morally deprived Tom_B and Tom_D considered relative to one another. In the most recent results, Tom_B's average blameworthiness score was 5.5. Tom_D's average praiseworthiness score was 4.9. Again, were the GTS theory applicable here, we would expect (at least) a reverse relation: insofar as people's true self is thought to be default good, bad actions would be thought less a part of that self than good actions, and so bad actions would be deemed less blameworthy than good actions are praiseworthy.

16 We could, of course, simply ask them. But even as professional ethicists, it is not clear to us that we would trust our own judgments about such matters, let alone those of non-philosophers'.

blame given that the side effect is viewed as *bad*. Indeed, in our previous discussions, we referred to the purported asymmetry in the Toms cases as a "*reverse* Knobe effect."[17]

In attempting to accommodate as above, the GTS theorist would have to maintain that all subjects take hurting the environment to an unspecified degree to be more bad than helping the environment to an unspecified degree is good.[18] But it is unclear what grounds there would be for such a claim.[19]

---

[17] It is important to note that the GTS theory doesn't mandate that participants always see the true self as good. Sometimes they see the true self as bad, and when they do, they in fact attribute certain bad actions and attitudes to it (see Newman, De Freitas, and Knobe 2014; thanks to an anonymous referee for reminding us of this). But these are in fact cases in which subjects were explicitly told that, deep down, the character in the scenario was "fundamentally evil" (Newman, De Freitas, and Knobe 2014: 22). When not told this fact explicitly, however, subjects in the cases under discussion reduced attributability. So given that there was no such explicit wording about fundamental evil in the bad Chairman prompt, we've not been given any reason to believe that subjects in that case were thinking of him as having a bad true self.

[18] Note that this is different from the somewhat similar, and perhaps more plausible, claim that people tend to blame more for things that are bad to some degree than they praise for things that are good to the same degree. If true, this might recommend amending the GTS with a claim about the asymmetry of praise and blame. It is unclear how this would interact with other aspects of the theory. For instance, if we tend to blame people *more* for doing bad things than we praise them for doing good things, it is hard to see how this could be reconciled with the claims that we praise or blame to the extent that we judge agents' deep selves to concord with their actions, and generally take their deep selves to be good.

[19] An anonymous referee recommends a different response. The GTS theorist could appeal to the *spectral* nature of the deep self. Perhaps Tom_A's racism, which stems from childhood indoctrination, is seen as part of his "shallower deep self," while Tom_B's racism, which he has come to on his own, is seen as part of his "deeper deep self." We are inclined away from this suggestion, given that it is not clear how it would fit the Knobe effect cases. But even if correct,

We have offered evidence that people do not always, or even typically, judge good actions to be more attributable to an agent's deep self. This undercuts the GTS theory as stated. But it does not follow from this that the core thought behind the GTS theory—that people are prone to think that others are good "deep down"—is false. Rather, what follows is that we view good actions as representative of an agent's true self *only in certain contexts*. The obvious question going forward is the nature of those contexts and why the effect appears in some of them but not others. In the next section, we offer a proposal for marking out these contexts. As we'll see, this proposal can be used either to amend the GTS theory or to replace it with an alternative version of Attributability Theory.

The Benefit of the Doubt

We know that people view Tom_B's bad actions as less attributable to his true self than Tom_A's bad actions to his, but we still lack a full explanation for the difference. The GTS theory tells us that the explanation is simply "because Tom_B's action is a bad action but people view Tom_B as good deep down inside." But this explanation obviously doesn't help differentiate between Tom_A and Tom_B, as *both* are doing bad things. The GTS theory also predicts the wrong results in other cases (e.g., Knobe's Chairman case). So it seems that something specific to childhood moral deprivation likely explains the difference (which is what we had been insisting on in our earlier advocacy of the Difficulty Hypothesis). We suggest that it may well be one or another species of the more general stance people often take when evaluating others, namely, they give others *the benefit of the doubt*, assuming, when given the right opportunity, that others' bad actions concord with their deep selves less than their evil counterparts' actions concord with their own. More precisely, our hypothesis is

---

there remains work for our hypothesis in the next section, namely, explaining why upbringing impacts attribution of an action to an agent's deeper or shallower self.

that subjects tend to give agents the benefit of the doubt when there is any readily available (partial) *external* explanation for their actions, an explanation that involves something outside of their agency. Let us elaborate.

Many of us can understand the experience of being influenced, sometimes quite heavily, by emotions or other psychic forces that feel like agency-derailing invasions, or at least enormous impediments to doing what we want or ought. We tend to think that these are cases in which we should get off the hook in certain respects for what we do or feel. For example, in the grips of mild depression or exhaustion or stress, we cite these factors as having gotten the better of us, as having prevented us from meeting the demands and expectations of others. Sometimes others excuse us thereby, and when this occurs, we feel vindicated: "Yes," we think, "they got it."

When assessing the conduct of others, then, we may be alive to these kinds of excusing conditions in a way that tends to have us looking for them on others' behalf, with respect to conduct that would otherwise ground indignation and blame. This is particularly the case, we stress, when evaluating the conduct of others *where we ourselves are not involved.* Cases where we ourselves are wronged are cases in which our anger may tend to hold sway regardless of the excusing conditions present. But as it turns out, *all* of the discussed experimental conditions are cases in which the subject isn't involved, where the subject is a mere observer, and so they likely tend to generate fairly bloodless third-personal judgments about the degree to which someone is worthy of blame from the comfort of subjects' computers or classrooms.[20] But bloodless judgments of blameworthiness are quite different from engaged blame itself, and so may tend to produce very different assignments.

---

[20] To be clear, our claim here is not that all third-personal judgments are bloodless, for there are clearly third-personal judgments where we are involved, or at least have a stake, such as third-personal judgments about our friends and loved ones. Our thanks to an anonymous reviewer for encouraging us to be clearer about this point.

One can imagine, for instance, that the differences in bloodless assignments of Tom_A's and Tom_B's degree of blameworthiness might be erased were we to have asked, "Were you in the black man's position, how angry would you be at Tom for spitting on you (assuming you knew the facts of Tom's upbringing)?"

Of course, first-personal responses may be distorted or disproportionate; people tend to be quite retributive. But the point is simply that putting these prompts in the terms that we and others have been putting them—requesting third-personal assignments of blameworthiness (or how much blame the target is worthy of)—opens the door to the types of more lenient assignments we think may explain the results here. These are cases, after all, in which it is easy—costless, really—to extend the benefit of the doubt, and so to shift at least some responsibility to a cause readily construed as external to the agent.[21]

---

[21] One might think that this talk of engaged and disengaged blaming responses could actually buttress the GTS theory. For when we *do* directly blame someone (second-personally), surely we are assuming that the blamed agent does have a good true self, for otherwise it would make no sense to demand via our blame that he join us in condemning his action and to make things right. If he weren't good, what grounds could he have to do so? (Thanks to an anonymous referee for raising this concern.) This is an interesting consideration, but there are several reasons to resist it. For one, someone's ability to judge as to the worth of certain sorts of reasons (e.g., the worth of condemnation or making things right) doesn't necessarily make that person good; it may just make that person able to read the writing on the wall, as it were. Presumably bad people can judge some reasons of this sort worth acting on, if only to get along with others (for further nefarious purposes). Second, there are plenty of cases in which we blame others *proleptically*, that is, with an eye toward getting them to eventually have access to the reasons they didn't have access to when engaged in wrongdoing (for the term, see Williams 1995: 41–44). This will hold paradigmatically with respect to bad people (we want them to become good). Finally, my blame of you may be a demand for empathic acknowledgment, a demand that you come to recognize and acknowledge what you did to me from my perspective [OMITTED]. But bad people can make a sudden

In light of this, let us reconsider the original emotional swamping case. In one scenario, an enraged agent smashes someone else's window (not the subject's). His emotions "got the better of him," we tend to say. We ourselves have been in similar circumstances, enraged in a way that felt surprising, alien, and as a result we might have done something regrettable. So too, we tend to think, it might be with this agent. But in the scenario in which the window-smashing agent is calm and reasonable, no such alien force is readily available to explain what he's doing. (And again, note just how differently subjects might react if told it was *their* car window that was smashed.)

Similarly, in the studies concerning homosexuality, it is not hard to imagine uninvolved observers' giving Mark the benefit of the doubt in *either* direction. For conservatives, the issue would be akin to emotional swamping: Mark has been taken over by alien and unruly emotions, as sometimes happens to us all. For liberals, the issue is similar to that raised by our own studies: both Mark and the deprived Toms have trouble escaping what they have been raised to believe, as, again, sometimes happens to us all. We may think we've overcome some bug of our upbringing when, in a heated moment, it takes over yet again.

By contrast, there is no readily available external explanation for the CEO's ignoring environmental harms. Rather, the obvious (internal) explanation is his selfishness, which is far from mitigating (at least in the absence of some story about, say, his upbringing). Indeed, the fact that we view selfishness as *bad* may well explain why the CEO case seems to exhibit a sort of *Bad* True Self effect. At any rate, the point is that there is no external factor ready to hand that might (largely) explain what the CEO is doing. And so what he's doing is naturally taken to be attributable to *him*.

---

turnaround when presented with a dramatically new perspective on what they've done. See, e.g., Biblical Saul's sudden conversion to Paul.

For the Toms, then, the idea is that subjects naturally give Tom_B the benefit of the doubt, viewing his wrongdoing as partially explained by his morally-deprived upbringing, something external to his current agential features. By contrast, no such easy mitigating element would lead one to give Tom_A the benefit of the doubt relative to Tom_C. He just broke bad.

So, what of the GTS theorist's idea that people tend to judge that others are good "deep down?" We allow that that idea is one possible explanation for our hypothesis. Perhaps subjects are inclined to believe the best of others, and that's why they give them the benefit of the doubt. It's just that in many cases the lack of an obvious external explanation makes that possibility seem less likely, in which case subjects fail to judge as the GTS theory predicts. Now we the authors are dubious about this explanation, but we take no official stance against it here. Our position is that an inclination to give people the benefit of the doubt is a plausible explanation of our and others' data, and that further thought and study are needed to uncover the deeper explanation (if any) for this tendency.

Before moving on, however, we wish to note an alternative explanation that has not yet been explored in the literature. On this alternative, there may be a good true self present in these interactions, but it would actually be in the heart of the *beholder*. When witnessing norm violations where only others are affected or wronged, good people do sometimes look first for an excuse for the wrongdoer, a way of explaining the behavior via appeal to a feature external to the wrongdoer's agency. Sometimes the external source is emotional swamping. Sometimes the external source is a compulsion or disorder. And sometimes, as could be true in our studies, the external source is childhood moral deprivation. But the beholder's goodness extends only so far: If there is no such readily available external explanation for bad behavior, then the badness is taken to belong to the

agent. Again, we are not claiming that this *is* the explanation for our and others' results, only that it is a possible explanation deserving investigation.[22]

If our hypothesis about the tendency to give others the benefit of the doubt is correct, it opens the way for a return to our Difficulty Hypothesis. It is too strong, of course, to say that subjects think of Tom_B's actions as being wholly due to his upbringing. If it were merely that, they would presumably take him to not be responsible *at all*—if, for instance, they saw him as a sort of brainwashed automaton. Instead, as we suggested, it may be that people judge that for the morally ignorant, it is more difficult to overcome the deprivations of their upbringing and see the reasons in favor of the non-bad option than it is for those who do not have such an upbringing. They have to work harder to machete their way through the jungle of deprivation and, in this case, they couldn't cut it. They are thus given the benefit of the doubt in a mitigating way: The action performed is thought to be *less* attributable to them. This would explain why such ignorance is mitigating, but not wholly so.[23]

So what about *good* actions and difficulty? What does Tom_D's alleged difficulty in recognizing moral reasons have to do with the roughly equivalent patterns of attributability and praiseworthiness

---

[22] Another possibility is that this tendency, too, is context-dependent. Perhaps we are only inclined to give the benefit of the doubt to certain people (e.g., along in-group/out-group lines). Again, this possibility sows the seeds for further study. Our thanks to an anonymous reviewer for raising this point.

[23] Above, we suggested that the benefit of the doubt is given to agents for whom there is some readily available external explanation for their actions. The suggestion here implicitly relies on attributability's being a matter of degree: an action is seen as more or less attributable to an agent's deep self (and therefore, the agent is seen as more or less praise- or blameworthy) dependent on the *extent* to which that action can be explained by some readily available external source.

assignments given to both Tom_C and Tom_D? If people attribute both Toms' helping actions equally to their true selves, then it looks as if difficulty makes no difference at all.[24]

If the praiseworthiness scores for Tom_C and Tom_D are indeed roughly equivalent, then we would suggest that while difficulty *mitigates* attributability, it does not *augment* it. The obvious reason for this has to do with the motivation for giving the benefit of the doubt in negative cases: The violation of moral norms often renders harsh treatment (e.g., stinging words, sanctions, punishment) appropriate, and so people think that such treatment cranks up justificatory standards for the fairness of doing so in a way that is unnecessary for positive cases, where the rewards of praise are much less significant for the target's well-being, and where it seems less "unfair" to praise someone who doesn't deserve it. In other words, there may be a number of features that matter for mitigation in negative cases that do not matter (or matter far less) for positive cases, given how much negative responses can *hurt*.[25] We would thus have much stronger reason to extend the benefit of the doubt in negative cases than in positive cases.

---

[24] It's actually not so clear to us that there is equal attribution in the positive cases. Our most recent experimental results are different from previous ones. In our earlier study, we did get a significant difference between praiseworthiness in the two cases, with Tom_C assigned 4.3 and Tom_D assigned 5.4. This result suggested the Difficulty Hypothesis might be relevant in the positive cases too, as subjects might be assigning significantly higher praiseworthiness to Tom_D in light of how difficult it was for him to have been moved by the moral reasons he ostensibly acted on. It's not clear which set of studies yields the most accurate results, but we will proceed in the text on the assumption that our most recent results are the ones that hold, as they are the only ones that pose possible trouble for us.

[25] We should also acknowledge a third possibility, that as, per the discussion above, it could be that because what the good Toms did is viewed as less good than what the bad Toms did was bad, the effect of the relevant normative judgments on attribution was stronger in the latter case than the former.

Conclusion: A Slight Scold Regarding the Ongoing Ambiguity of "Blameworthiness" in Experimental Work

Needless to say, much more work needs to be done here, including a direct exploration of the role thoughts about difficulty might be playing in assignments of attributability. But we want to close by exposing a crucial ambiguity that runs through the work done in this area thus far (including our own), an ambiguity that needs to be recognized and eliminated in future work. Indeed, this ambiguity might also serve to provide some explanation for the results we have seen thus far. The ambiguity is in the terms "blame" and "blameworthy" (and "praise" and "praiseworthy" as well, though we will just focus on the negative). The problem is that "blame" cuts across multiple types of moral responsibility, demarcated in terms of distinct agential capacities.

Suppose I have seen you over and over again being amused by injustice, and so I have contempt or disdain for you. This seems a type of blame, regardless of whether I express it to you or not. Now suppose a chess coach sees her otherwise excellent pupil make a foolish move in a competition, so she shakes her head in disapproval and criticizes him vociferously afterwards for having done it. This too seems a type of blame. Finally, suppose I am a department chair who, at a meeting, ignores the voices of all the female members of the faculty, and so they become resentful of me and the male members grow indignant with me. All are blaming me. Now in each of these three cases, the different attitudinal responses of the blamers pick out very different agential qualities. In the first case, my contempt or disdain for you is *aretaic*, and so goes to your poor quality of *character*, as expressed by your pattern of amusement at injustice. In the second case, the coach's disapproval and criticism goes to her pupil's poor quality of *judgment* or *decision-making* in that particular instance of acting. In the third case, everyone's angry responses go to my (the department chair's) poor quality of *regard* for the women in the department, my failing to take their voices seriously (see [OMITTED]). Asking subjects to assign a degree of "blame" or "blameworthiness,"

then, could yield triply ambiguous results, as subjects might be assigning it along any (or all) of these three dimensions.

We ultimately have no idea which sense of the terms subjects have had in mind. And that's a serious problem in this literature. What we suggest, then, going forward, is that prompts be designed to cut through this ambiguity. One promising start would be to see if results about "blameworthiness" thus far generated might somehow be subject to common translation. Take all the prompts previously presented, in other words, and present in each case new prompts using the specified terminology above, as in: "To what extent does this behavior reflect poorly on the agent's *character*?" and "To what extent does this behavior reflect poorly on the agent's *judgment*?" and "To what extend does this behavior reflect poorly on the agent's *regard for others*?" One pattern of response may better correspond than the others to the patterns of response we already have with respect to the "blameworthiness" prompts. Alternatively, we might ask explicitly about the emotional responses that subjects think would be most appropriate in each case (e.g., disdain, disapproval, or anger). But at any rate, we urge that theorists take seriously this ambiguity in designing future studies in this arena.

We have offered one possible explanation for why the GTS effect appears in some contexts but not others, having to do with subjects' tendency to give the benefit of the doubt in bloodless third-personal appraisals of the scenarios. But far more work needs to be done before we can confidently say we understand the nature of these judgments, including whether they represent a general tendency to see others as good. One thing does seem clear from the data, however: Whatever the fuller explanation for what's going on, judgments of blame- and praiseworthiness are intimately connected with judgments about the true self. Attributability Theory remains (for now) on solid ground.

REFERENCES

Arpaly, Nomy. 2003. *Unprincipled Virtue*. Oxford: Oxford University Press.

Frankfurt, Harry. 1988. *The Importance of What We Care About*. Cambridge: Cambridge University Press.

-----. 1999. *Necessity, Volition, and Love*. Cambridge: Cambridge University Press.

Hume, David. 1739/1969. *A Treatise of Human Nature*. Harmondsworth, Middelesex, England: Penguin Books.

Knobe, Joshua. 2006. "The Concept of Intentional Action: A Case Study in the Uses of Folk Psychology." *Philosophical Studies* 130: 203–231.

Lippert-Rasmussen, Kasper. 2003. "Identification and Responsibility." *Ethical Theory and Moral Practice* 6: 349–376.

Nelkin, Dana. 2011. *Making Sense of Freedom & Responsibility*. Oxford: Oxford University Press.

Newman, George E., Bloom, Paul, and Knobe, Joshua. 2013 (online). "Value Judgments and the True Self." *Personality and Social Psychology Bulletin*. DOI: 10.1177/0146167213508791.

Newman, George E., De Freitas, Julian, and Knobe, Joshua. 2014 (online). "Beliefs About the True Self Explain Asymmetries Based on Moral Judgment." *Cognitive Science*. DOI: 10.1111/cogs.12134.

Phillips, J., Misenheimer, L., Knobe, J. 2011. "The Ordinary Concept of Happiness (and Others Like It." *Emotion Review* 3: 1–3.

Phillips, J., Nyholm, S. & Liao, S. 2014. "The Good in Happiness." *Oxford Studies in Experimental Philosophy* 1: 253–293.

Pizarro, D.A., Uhlmann, E., & Salovey, P. 2003. "Asymmetry in Judgments of Moral Blame and Praise: The Role of Perceived Metadesires." *Psychological Science* 14: 267–272.

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

Scanlon, T.M. 1998. *What We Owe to Each Other*. Cambridge, MA: The Belknap Press of Harvard University Press.

Sher, George. 2006. *In Praise of Blame*. Oxford: Oxford University Press.

Smith, Angela M. 2005. "Responsibility for Attitudes: Activity and Passivity in Mental Life." *Ethics* 115: 236–271.

-----. 2012. "Attributability, Answerability, and Accountability: In Defense of a Unified Account." *Ethics* 122: 575–589.

Sripada, Chandra Sekhar. 2010. "The Deep Self Model and Asymmetries in Folk Judgments about Intentional Action." *Philosophical Studies* 151: 159–176.

-----. 2015. "Moral Responsibility, Reasons, and the Self." In *Oxford Studies in Agency and Responsibility* 3: 242–264.

Watson, Gary. 2004. *Agency and Answerability*. Oxford: Oxford University Press.

Williams, Bernard. 1995. *Making Sense of Humanity*. Cambridge: Cambridge University Press.

Wolf, Susan. 1987. "Sanity and the Metaphysics of Moral Responsibility." In Ferdinand Schoeman, ed., *Responsibility, Character, and the Emotions* (Cambridge: Cambridge University Press, 1987), pp. 46–62.

-----. 1990. *Freedom Within Reason*. Oxford: Oxford University Press.

OMITTEDs