1

Huck vs. Jojo

Moral Ignorance and the (A)symmetry of Praise and Blame

David Faraci, University of North Carolina
David Shoemaker, Tulane University

The more I studied about this the more my conscience went to grinding me, and the more wicked and low-down and ornery I got to feeling. And at last, when it hit me all of a sudden that here was the plain hand of Providence slapping me in the face and letting me know my wickedness was being watched all the time from up there in heaven, whilst I was stealing a poor old woman's nigger that hadn't ever done me no harm, and now was showing me there's One that's always on the lookout, and ain't a-going to allow no such miserable doings to go only just so fur and no further, I most dropped in my tracks I was so scared. Well, I tried the best I could to kinder soften it up somehow for myself by saying I was brung up wicked, and so I warn't so much to blame; but something inside of me kept saying, "There was the Sunday-school, you could a gone to it; and if you'd a done it they'd a learnt you there that people that acts as I'd been acting about that nigger goes to everlasting fire." (31.19) It made me shiver. And I about made up my mind to pray, and see if I couldn't try to quit being the kind of a boy I was and be better. So I kneeled down. But the words wouldn't come. Why wouldn't they? It warn't no use to try and hide it from Him. Nor from ME, neither. I knowed very well why they wouldn't come. It was because my heart warn't right; it was because I warn't square; it was because I was playing double. I was letting ON to give up sin, but away inside of me I was holding on to the biggest one of all. I was trying to make my mouth SAY I would do the right thing and the clean thing, and go and write to that nigger's owner and tell

> where he was; but deep down in me I knowed it was a lie, and He knowed it. You can't pray a lie—I found that out. (31.20) [After writing the note,] I felt good and all washed clean of sin for the first time I had ever felt so in my life, and I knowed I could pray now. But I didn't do it straight off, but laid the paper down and set there thinking—thinking how good it was all this happened so, and how near I come to being lost and going to hell. And went on thinking. And got to thinking over our trip down the river; and I see Jim before me all the time: in the day and in the night-time, sometimes moonlight, sometimes storms, and we a-floating along, talking and singing and laughing. But somehow I couldn't seem to strike no places to harden me against him, but only the other kind. I'd see him standing my watch on top of his'n, 'stead of calling me, so I could go on sleeping; and see him how glad he was when I come back out of the fog; and when I come to him again in the swamp, up there where the feud was; and such-like times; and would always call me honey, and pet me and do everything he could think of for me, and how good he always was; and at last I struck the time I saved him by telling the men we had small-pox aboard, and he was so grateful, and said I was the best friend old Jim ever had in the world, and the ONLY one he's got now; and then I happened to look around and see that paper. (31.23) It was a close place. I took it up, and held it in my hand. I was a-trembling, because I'd got to decide, forever, betwixt two things, and I knowed it. I studied a minute, sort of holding my breath, and then says to myself: "All right, then, I'll GO to hell"—and tore it up. (31.24, 31.25)

In these famous scenes from *The Adventures of Huckleberry Finn*, Huck is trying desperately to decide what to do with Jim, the slave he has been traveling with down the river, the man who is, to Huck's mind, someone else's property. When given the opportunity to return Jim to his "rightful owner," Huck ultimately decides to go against his upbringing, his conscience, and societal norms in not turning Jim in. Consequently, by his own lights he is going to hell for doing what he sincerely believes is the wrong thing. By our lights, though, he is clearly doing the *right* thing, despite his morally deprived upbringing. Is he thus praiseworthy?

Compare the following case:

JoJo is the favorite son of Jo the First, an evil and sadistic dictator of a small, undeveloped country. Because of his father's special feelings for the boy, JoJo is given a special education and is allowed to accompany his father and observe his daily routine. In light of this treatment, it is not surprising that little JoJo takes his father as a role model and develops values very much like Dad's. As an adult, he

does many of the same sorts of things his father did, including sending people to prison or to death or to torture chambers on the basis of whim. He is not *coerced* to do these things, he acts according to his own desires. Moreover, these are desires 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.[1]

In torturing a peasant on a whim, JoJo goes along with his upbringing and conscience, doing what he sincerely believes is the right thing but what we know is the wrong thing. Is he thus blameworthy?

Both Huck and JoJo were raised in morally blinkered environments, and they have both consequently come to accept deeply mistaken moral views: Huck thinks he ought to return Jim to his "owner"; and JoJo thinks it is morally permissible (or perhaps obligatory) to beat peasants when he feels like it.

There have been (at least) two assumptions that theorists have thought it natural to make in cases such as these. First, it seems plausible to some that moral deprivation of this sort is a responsibility-undermining factor, that those who were raised in such morally blinkered environments are off the hook when it comes to assignments of moral responsibility. So JoJo, we might think, is excused for beating the peasants given the extent of his childhood moral deprivation. Call this the Moral Deprivation–Excuse Thesis (the MDE Thesis).[2] Second, it is widely believed that praiseworthiness is the positive analogue of blameworthiness, and thus that factors affecting judgments of moral responsibility should have symmetrical effects on judgments of praise- and blameworthiness. Call this the Symmetry Thesis.[3] On a natural interpretation of this thesis, if poor moral upbringing mitigates assignments of JoJo's blameworthiness, it ought likewise to mitigate assignments of *praise*worthiness in cases of equivalent moral deprivation.

In this essay, we examine how certain sorts of moral knowledge deprivations in an agent's upbringing bear on people's actual assessments of that

[1] Susan Wolf, "Sanity and the Metaphysics of Responsibility," in Gary Watson, ed., *Free Will*, 2nd edn (Oxford: Oxford University Press, 2003), p. 379.

[2] This is the view taken by Wolf, as we will detail below.

[3] The Symmetry Thesis is widespread. It is entailed by what Doris and Knobe call the more general assumption of *invariance* in moral responsibility assessments. For discussion and citation, see John Doris and Joshua Knobe, "Strawsonian Variations: Folk Morality and the Search for a Unified Explanation," in John Doris, ed., *The Moral Psychology Handbook* (Oxford: Oxford University Press, 2010).

agent's responsibility. At first blush, the data we have collected appears to cast doubt on both the MDE and Symmetry theses. First, our experimental results suggest that people do not, in fact, view deprivation as wholly morally excusing (though they do not view it as irrelevant, either). Second, our data appear to suggest that people have commitments to distinctly different conditions for blame- and praiseworthy agency.[4] This seems to cast doubt on the Symmetry Thesis.

In both cases, however, we argue that the correct response to our data is not simply to reject the theses in question. First, our results suggest that the relationship between deprivation and moral assessment is more nuanced than straightforward acceptance *or* denial of the MDE Thesis can account for. In earlier work, we discovered that while people do not judge JoJo to be *wholly* blameworthy, they do not judge him to be blameless, either. And the data we present here seems to confirm this for other cases of moral deprivation. This suggests that while the MDE Thesis cannot be accepted as it stands, it has captured an important aspect of blameworthiness judgments. As for the Symmetry Thesis, we argue that it can be interpreted in a way that renders it consistent with our data.

In both cases—especially if we are to vindicate our interpretation of the Symmetry Thesis—it is an important question *why* people judge as they do. In our earlier work, we hypothesized that people judge JoJo as they do because of the *difficulty* JoJo would have in overcoming his upbringing and doing the right thing. After considering some alternatives, we argue that the very same hypothesis can help vindicate our interpretation of the Symmetry Thesis. This, we take it, both buttresses the plausibility of our claims about difficulty and represents an independent advance in our understanding of the roots of praise and blame.

Blinkered Badness

In the previous work mentioned, we explicitly explored folk intuitions on the JoJo case.[5] In her original presentation and discussion of the case,

[4] Knobe and Doris include this asymmetry in their list, based on early experimental results generated by one of the researchers, but those experiments remained unpublished pending further refinement. The present essay (and the experiment done in its service) is an attempt at said refinement.

[5] David Faraci and David Shoemaker, "Insanity, Deep Selves, and Moral Responsibility: The Case of JoJo," *Review of Philosophy and Psychology* 1 (2010): 319–32.

Susan Wolf offers a strong version of what we have labeled the MDE Thesis, claiming that people's pretheoretic intuitions are that JoJo, because of his terribly deprived upbringing, is not morally responsible at all for his actions, despite their flowing from his deep self, i.e. his "true" or "real" evaluating or authenticating character.[6] Her diagnosis for this alleged reaction is that JoJo's deep self is normatively insane, i.e. he lacks the ability to recognize goodness, badness, and the difference between them.[7] She then introduces a sanity—normative competence—condition into her theory of moral responsibility to account for such cases.

Insofar as Wolf is claiming a result on behalf of our pretheoretic intuitions, we decided to test her claim. Our findings were that, while people did judge JoJo to be less blameworthy than a control (JoJo's presumably sane but nasty father), they still found him to be *seriously* blameworthy, assigning him an average blameworthiness score of 5 (where 7 was completely blameworthy and 1 was not at all blameworthy). Insofar as blameworthiness rides piggyback on responsibility (at least prima facie), JoJo is definitely viewed (pretheoretically) as morally responsible to some degree, contrary to Wolf's strong version of the MDE Thesis. We took this at the very least to indicate that her adoption of a sanity condition was unmotivated.

But—and this is the interesting bit—JoJo *was* viewed as less blameworthy than he would have been without such a deprived background, so deprivation of this sort does seem to excuse *to some extent*. But what precisely was the relation of this deprivation to reduced blameworthiness? And could it be overcome?

Our initial diagnosis of JoJo's reduced blameworthiness derived from his moral ignorance, not his normative insanity. It wasn't that he lacked the capacity to recognize right and wrong; it was, rather, that he had as a child been deprived of exposure to relevant moral alternatives. We attempted to explore these issues via the presentation of a third scenario in which JoJo was eventually exposed to moral alternatives but rejected them in favor of continuing to adhere to Daddy's value system. Here,

[6] On pp. 379–80, Wolf says "In light of JoJo's heritage and upbringing . . . it is dubious at best that he should be regarded as responsible for what he does." Later on p. 380, she says, "Our judgment *that JoJo is not a responsible agent* is one that we can make only from the inside" (emphasis ours).

[7] Wolf, pp. 379–85.

surprisingly, there was no statistically significant difference between people's reactions to the first and second JoJos. This might suggest, then, that it was their unfortunate formative circumstances that (slightly) mitigated their blameworthiness, and not their mere moral ignorance. But we remained unconvinced, arguing instead that the JoJos displayed a more fundamental and insidious type of moral ignorance than is usually discussed—an ignorance that expressions of ill will are wrong, not an ignorance of what specific act-tokens count as expressions of ill will—and that this sort of ignorance may not be displaced by mere exposure to moral alternatives (which would do nothing to counteract the thought that it's generally morally permissible to express ill will).[8] What reduces the degree of their blameworthiness, we suggested, was not any sort of incapacity, though; rather, it was the *difficulty* of overcoming their childhood moral deprivations as adults, a difficulty that nevertheless assumes the possibility of success and so grants a basic normative capacity to them (a necessary condition to their being judged to be seriously blameworthy in the first place).

In our current round of experiments, we aimed to do two things. First, we wanted to see if we could duplicate the JoJo-type results in cases in which normative capacities were not at issue, and neither were any thoroughgoing childhood deprivations regarding the general wrongness of expressions of ill will. This would enable us to focus solely on the work being done by ignorance with respect to a specific sort of moral value. Second, we wanted to explore whether structurally identical positive cases would yield analogous results. We will explain this latter point in the next section of the essay, focusing here solely on the former.

New subjects were presented, at random, with one of the following two scenarios:[9]

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

[8] We put this in terms of "ill will" in the original paper, but we recognize this to be a quite ambiguous notion. (For discussion, see David Shoemaker's "Qualities of Will," *Social Philosophy and Policy* 30 (2013): 95–120) We could have put this in less ambiguous terms as follows without losing the point. In standard cases of moral ignorance, the agent knows what properties make an act right or wrong but doesn't know whether some particular act-type instantiates those properties; JoJo's ignorance, on the other hand, is about what properties make actions right or wrong in the first place, and so is much more profound.

[9] n(A) = 84; n(B) = 84.

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.

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.

For each scenario, subjects were asked to rate Tom's blameworthiness for spitting on the black man on a scale from 1 ("not all blameworthy") to 7 ("completely blameworthy").

Our prediction, in keeping with our results from the earlier paper, were that Tom_B's upbringing and resultant ignorance with respect to the morality of racism would mitigate attributions of moral blameworthiness. The mean response to Tom_A was 6.68, a very robust blameworthiness score. The mean response to Tom_B was significantly lower at 5.4.[10] We conclude from this, again in keeping with earlier results, that certain sorts of moral blinders—i.e. childhood deprivations of exposure to moral truth—reduce blameworthiness assessments somewhat, but nowhere near completely.[11] And, once again, the fact that subjects viewed Tom_B as seriously blameworthy reveals that they likely believed him not to lack the relevant normative capacities, i.e., sanity was not an issue. Finally, our

[10] Results were subjected to an independent samples t-test: $t(166) = 6.54$, $p < 0.001$ (two-tailed), *SD* (ignorant) 1.53, *SD* (non-ignorant) 0.92, *Cohen's d* = 1.01.

[11] It might be thought that the vignettes do not do enough to establish sufficiently robust moral ignorance in Tom_B's case, as anyone growing up in the modern era will surely have been exposed to alternatives through TV, radio, the Internet, or travel. Given that this might be an assumption of the subjects who read the vignettes, then it might have grounded their being pretty punitive. (Thanks to an anonymous reviewer for pressing this point.) This is a fair point, and it is worth controlling for. Nevertheless, we doubt it is doing much work for subjects, given that we stressed JoJo's isolation and lack of exposure to alternatives in the original study, and we got nearly identical response levels to our JoJo doppelganger in the latest one.

speculative explanation remains quite plausible. We propose that the reason people judge Tom_B to be seriously blameworthy, despite the blameworthiness-reducing fact of his deprived childhood, is that, while it is more *difficult* for him to identify and do the right thing because of that childhood, it is nevertheless not overly demanding to expect him to do so.

Some new features we included buttressed this hypothesis by diminishing the plausibility of alternative explanations. First, we had Tom in both scenarios move to another town at the age of 25 in order to avoid the kind of ongoing isolated "preciousness" of the original JoJo case, and also to establish that he is an adult (with a fully-grown brain) who is making genuine moral decisions. His moving to another town would presumably also have given him opportunities to be exposed to moral alternatives. The question, then, was how that exposure would interact with what he had been taught as a child in people's assessments of him. As it turns out, if he was taught the moral truth and rejected that as an adult, he was viewed as particularly blameworthy for spitting on the black man; he was viewed as less so if he had been taught the racist moral lie but simply didn't reject it.

Second, we deliberately left open why it is that Tom either rejects or embraces what he's been taught. The reason was to allow for the possibility of any of the wide variety of sources of evaluations that occur in everyday life, and so not to privilege certain sources over others. Sometimes people evaluate after having pored over the reasons on both sides and determining which ones weigh more. Other times people evaluate after having seen a movie on the subject and without further deliberation. Other times people simply follow their intuitive hunches. Presumably all of these are methods that will preserve responsibility in assessors' eyes.

Of course, the extent to which our proposal is plausible depends not only on the paucity of plausible alternatives, but on the inherent plausibility of the proposal itself. As has become clear to us since first introducing the proposal, more details about the nature of the difficulty in question are required before plausibility can be adequately assessed.[12] However, we set this matter to the side momentarily in order to introduce the remainder of our data, which are relevant to our explication of the nature of the difficulty in question.

[12] Thanks to two anonymous referees and to Shaun Nichols for pressing us to say more about the idea of difficulty in this context.

Thus far, we have been considering the effects of moral deprivation on assignments of blameworthiness. Especially given our interest in the Symmetry Thesis, the obvious question now is whether similar results obtain in cases of *praiseworthy* action. It is to that question that we now turn.

Blinkered Goodness

The Symmetry Thesis avers that blameworthiness and praiseworthiness are structurally analogous. If so, then it seems we should expect that childhood moral deprivations will reduce praiseworthiness in people's eyes just as they do blameworthiness, at least for actions within the zone germane to the relevant deprivations. The basic thought here is rather compelling. It would seem that degrees of blame- or praiseworthiness both ought to track degrees of childhood-based moral ignorance in the same way: the less you know, the less you're "on the hook" in *either* case. If the Symmetry Thesis were true, childhood moral deprivations presumably ought to reduce moral responsibility all the way round.

The data, however, appear to undermine the Symmetry Thesis, at least when understood in this way. Further subjects were randomly assigned to one of the following two cases:[13]

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

[13] n(C) = 85; n(D) = 83.

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* 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.[14]

There are several surprising conclusions one might draw here. First, Tom_C, who was raised with moral awareness but adopted racism as an adult, is only viewed as "somewhat praiseworthy" for going against those moral beliefs in doing the right thing. Tom_A—Tom_C's twin who adhered to his adult-formed moral beliefs in spitting on the black man—was viewed as nearly completely blameworthy for doing so. Perhaps, then, the resistance to seeing Tom_C as more praiseworthy is due to his still being a racist?

This can't explain all the data, however, because Tom_D is also a racist, and yet he is viewed as significantly more praiseworthy than Tom_C in going against his moral beliefs. It seems the only thing that could ground the difference in people's assessments here is the difference in upbringings. But then here is the second surprising conclusion. As discussed above, it seems plausible that Tom_D's moral ignorance would be generally mitigating: that if he didn't know that his racism was wrong, then he couldn't know that his going against it was right. But apparently knowledge of the rightness of one's actions isn't viewed as necessary for praiseworthiness; indeed (and this is the truly surprising point) *moral ignorance seems to be viewed as a virtue*. Not only is the MDE Thesis being denied here, it is being turned on its head.

And thus we come to the third surprising feature. While moral ignorance reduces blameworthiness, it seemingly *increases* praiseworthiness. Importantly, this way of putting it suggests that the Symmetry Thesis as understood thus far is false, or at least highly problematic.

Resilient Symmetry

Nevertheless, we would like to explore the possibility of a different interpretation of the Symmetry Thesis that could allow for it to be maintained

[14] As before, results were subjected to an independent samples t-test: $t(166) = 4.18$, $p < 0.001$ (two-tailed), *SD* (ignorant) 1.69, *SD* (non-ignorant) 1.76, *Cohen's d* = 0.65.

in light of our results. To this point, we have understood the thesis to imply that factors reducing blameworthiness also reduce praiseworthiness, that negative and positive cases are symmetrical with respect to whether the relevant praise- or blameworthiness is reduced or increased (in comparison to some paradigm control case). It looks as if the Tom cases undermine this symmetry: the ignorance reducing Tom_B's blameworthiness score actually *increases* Tom_D's praiseworthiness score. Understood in this way, the relevant comparison point is the baseline at which Tom is not at all praise- or blameworthy. The right question to ask is: Did the feature in question move the degree of _____-worthiness away from or towards that baseline? (See Figure 1.1.)

Of course, one could simply accept the asymmetry. One might, for instance, attempt to explain it by drawing an analogy to the famous "Knobe effect." This effect (named after Joshua Knobe) occurs when subjects are asked whether the side-effects of some agent's actions were intentional.[15] Subjects tend to say "yes" when the side-effects include something harmful, but tend to say "no" when the side-effects include something helpful. This appears to reveal an asymmetry in factual judgments of intentional action. Knobe's own interpretation is that subjects' judgments depend on their antecedent assessments of the normative status of the action's side-effects, so that "bad" effects render the action intentional in subjects' eyes, whereas "good" effects don't. Applied to our Toms, then, it may look like a kind of *reverse* Knobe effect is revealed: moral ignorance via childhood deprivation seems to reduce attributions of responsibility when the actions are bad, whereas it increases such attributions when the actions are good. This could mean that factual judgments of responsible action also are dependent on antecedent assessments of the normative status of the action.

Given wide acceptance of the Symmetry Thesis, however, it seems that such a move would be overly hasty if it is possible instead to interpret the thesis in a way that is consistent with our data.[16] Indeed, such an

[15] See Joshua Knobe, "Intentional Action and Side Effects in Ordinary Language," *Analysis* 63 (2003): 190–3; and Joshua Knobe, "The Concept of Intentional Action: A Case Study in the Uses of Folk Psychology," *Philosophical Studies* 130 (2006): 203–31.

[16] While widely presupposed, the Symmetry Thesis is not accepted universally. For leading examples of asymmetricians, see Susan Wolf, *Freedom Within Reason* (Oxford: Oxford University Press, 1990); and Dana Nelkin, *Making Sense of Freedom and Responsibility* (Oxford: Oxford University Press, 2011). We have also already mentioned Doris and

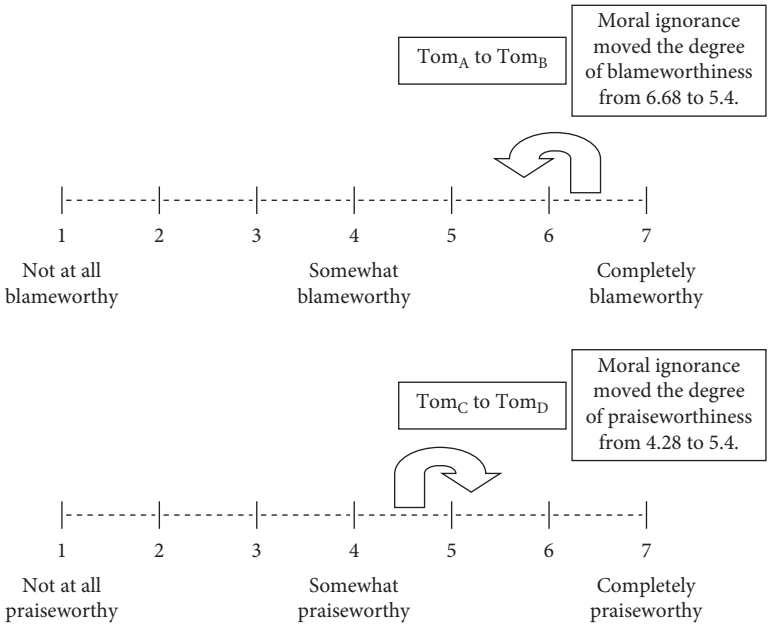

Figure 1.1 Moral ignorance, praise and blame.

interpretation exists: we propose, in contrast to the view represented by Figure 1.1, that negative and positive cases are symmetrically structured with respect to the direction in which various mitigating features shift the _____-worthiness judgments *in relation to their controls*. Understood in this way, the relevant comparison point would be the two endpoints of a continuous scale (see Figure 1.2).

So, with respect to the two endpoints ("completely blameworthy" and "completely praiseworthy"), the direction of movement from Tom$_A$ to Tom$_B$, and from Tom$_C$ to Tom$_D$, is symmetrical. What their sort of moral ignorance does, on this understanding of the relation, is move one more in the overall direction of complete praiseworthiness and away from complete blameworthiness.

If we understand the Symmetry Thesis in this way, then it might be preserved in light of our results. To do so, though, we must understand

Knobe's arguments against a version of the Symmetry Thesis. Importantly, though, all of these theorists take their burden of proof seriously, and so admit the need to provide arguments against symmetry in blame and praise.

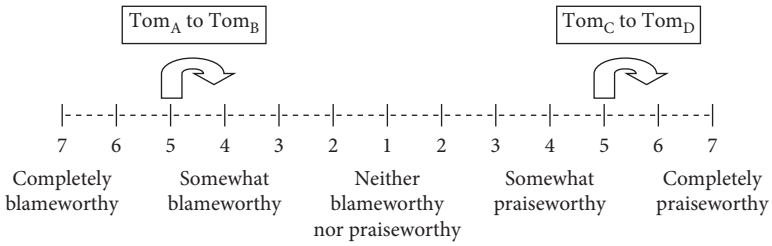

Figure 1.2 Symmetrical structure of negative and positive cases.

what the symmetry could consist in, i.e. what the explanation for the symmetrical movement would be. One possibility (continuing on the analogy with intentional action) stems from the work of Chandra Sripada, who has argued in favor of what he calls the "Deep Self Model" (DSM) of moral responsibility to provide a better explanation of the Knobe effect for action intentionality.[17] What Sripada argues is that intentionality attaches to agents' actions to the extent that their side-effects concord with the values and stable fundamental attitudes—the deep selves—that subjects attribute to them. If there is no concordance, goes the theory, subjects will be less likely to attribute intentionality to the agent's actions. On this interpretation, then, subjects view the harmful side-effect in the Knobe studies as in concordance with the agent's deep self and attribute intentionality to it thereby, but they view the helpful side-effect as in conflict with the agent's deep self and so tend not to attribute intentionality to it thereby. This is how the DSM maintains a unified symmetrical explanation of the Knobe effect. Normative judgments of goodness and badness don't do the relevant work here; the different responses instead flow from perceptions of different structural underpinnings.

We might, then, appeal to the DSM to explain our results regarding responsibility. In particular, it will be useful for our purposes to explore whether the DSM *alone* can explain the results while also preserving the Symmetry Thesis, just as it purports to do with respect to the Knobe effect. One immediate problem with this is figuring out just how the

[17] See Chandra Sekhar Sripada, "The Deep Self Model and Asymmetries in Folk Judgments About Intentional Action," *Philosophical Studies* (published online 2009), DOI: 10.1007/s11098-009-9423-5.

concordance condition should be specified with respect to responsibility. Here is one plausible possibility: Agents will be viewed as more or less responsible for some action or attitude A solely to the extent that A is viewed as concording more or less with their deep selves.[18] Of course, our scales were not put in terms of "responsibility"; rather, they were put in terms of praise- and blameworthiness. But perhaps we might articulate the DSM in terms of those labels as follows: agents will be viewed as more or less _____-worthy for A solely to the extent that A is viewed as concording more or less with their deep selves. Call this the DSM_{BP}. On this application of the DSM, one will be viewed as more blameworthy in correspondence with the perception that A more closely concords with one's deep self (and A is bad), whereas one will be viewed as more praiseworthy in correspondence with the perception that A more closely concords with one's deep self (and A is good). Can this explain the difference between our various Toms?

At issue is why moral ignorance via childhood deprivation moves one in a positive direction a full step up from control cases in the eyes of subjects. The DSM_{BP} would have us believe that this is because recognition of that sort of moral ignorance affects subjects' attribution of the relevant actions to the agents' deep selves. But this does not seem to fit the results of our study. According to the DSM_{BP}, given that Tom_B is less blameworthy than Tom_A, it must be that Tom_B's action concords less with his deep self—similarly for Tom_D vs. Tom_C. But neither view seems motivated, precisely because in each pairing *the Toms appear to have identical deep selves*. For example, both Tom_A and Tom_B fully embrace their racism as adults, and both act on it upon seeing the fallen black man. The only difference is in their upbringings. So why should we (or any subject) therefore think Tom_B's action is less reflective of his deep self than Tom_A's?

One response here would be to suggest that perhaps Tom_B lacks, or at least is deficient in having, a deep self, given the limited range of moral alternatives to which he was exposed in his upbringing. This might then explain why subjects judge him to be less blameworthy: his action is less concordant with his deep self than Tom_A's action insofar as Tom_B

[18] Indeed, this would be the view Susan Wolf labels the "Deep Self View" of responsibility that she draws from Harry Frankfurt, Charles Taylor, and Gary Watson. See Wolf, pp. 373–9.

doesn't have as robust a deep self as Tom_A. Unfortunately, making this move undermines the DSM_{BP} explanation of the *positive* cases, for if a deprived moral upbringing renders one's deep self less robust or more deficient, and so mitigates attributions of responsibility, then it ought to render Tom_D *less* praiseworthy than Tom_C in the eyes of subjects. But this is the opposite of our results.

Perhaps we can get a more plausible version of the DSM_{BP} if we start on the praiseworthy side of the map. How might we explain the differing assessments of Tom_C and Tom_D on the DSM_{BP}, given that both Toms were stipulated as identically embracing their racism and so would seem to have identical deep selves? Perhaps the thought is this: Tom_D's upbringing somehow made him less *committed* to the moral beliefs with which he was raised, such that when he goes against them, he is more likely to be expressing his actual deep self than is Tom_C. In other words, given their differences in upbringing, Tom_C's action is viewed as more "out of character," more anomalous, than Tom_D's. This thought could then be extended to the negative cases as follows: Tom_B is less blameworthy than Tom_A because his upbringing somehow renders his racist action more anomalous than Tom_A's. Tom_B isn't "really" committed to his racism, subjects might think, at least in the way that Tom_A is, so Tom_B is less blameworthy thereby.

This interpretation of the DSM_{BP} preserves the Symmetry Thesis, but how plausible is it? It is unclear to us why subjects would consider Tom_D's commitment to his moral beliefs to be any less serious than Tom_C's (or that Tom_B's is less serious than Tom_A's). Indeed, might not their kind of restricted ideological upbringing make them *more* committed to the belief system into which they had been indoctrinated? Might not this indoctrination create, if anything, a deep self *more* in line with its exclusively-taught, unquestioned principles than an upbringing without it? To the extent these core "moral" principles were instilled in a way that bypassed the ignorant Toms' rational, evaluative stance, they are likely to be resistant to such evaluations, much in the way religious belief with its source in childhood indoctrination is often difficult to expunge.

To be clear, we do not reject the *possibility* of reading a deep self view into the results here. It could well be that subjects really are viewing Tom_D's commitments as less attributable to his deep self than Tom_C's. Nevertheless, given the vignettes as stated, we have no evidence for this approach. Certainly, we welcome any future attempts to see whether

differential deep self commitments are doing any work here. Until then, though, we believe we are licensed in assuming either that the DSM_{BP} supports an asymmetrical approach to blame- and praiseworthiness or that its symmetrical approach is unmotivated and (at least initially) implausible. Focusing on attributions to the deep self *alone* does not look as if it will help us.

In our discussion of the MDE Thesis, we proposed that people's reactions to JoJo and Tom_B might stem from an appreciation of the *difficulty* those agents would have in doing the right thing. We believe that this idea can helpfully be extended to likewise explain our results regarding the Symmetry Thesis:

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.

On this hypothesis, as before, Tom_B is viewed as less blameworthy than Tom_A in light of how difficult it would be for him to go for a moral alternative not included in his morally blinkered upbringing. Our further suggestion is that Tom_D is viewed as more praiseworthy than Tom_C in light of how difficult it in fact *was* for him to go for a moral alternative not included in his morally blinkered upbringing. This way of viewing the matter easily explains the results while avoiding the dual implausibility of thinking that subjects think (a) there is a difference in the deep selves for either of the pairings, and (b) childhood indoctrination actually renders one's deep self more open or oriented toward moral alternatives than non-indoctrinated childhoods. If this is right, then the Symmetry Thesis may not after all be undermined by cases of moral ignorance in upbringing.

As noted earlier, however, we need to be clear about just what the Difficulty Hypothesis amounts to. In particular, we need to say something more about the nature of the difficulty in question. For one thing, one might worry that talk of difficulty just collapses into talk of *capacity*; perhaps, after all, Wolf was right to appeal to capacities. For another, it might be thought that we are suggesting that doing something inherently difficult (like, say, lifting something heavy) is, in itself, sufficient for praise. We agree that this would be implausible.

First, even if we grant that talk of difficulty is another way of talking about capacity, at the very least our results suggest that capacity talk must

be *scalar*. After all, both Tom_B and Tom_D are still viewed as _____-worthy for what they have done to a significant extent. This means that even though such _____-worthiness is thought to be affected by moral deprivation and ignorance, it is so only to some degree, so that if the Toms are viewed as being incapacitated in some respect it would only be an incapacity by degree. This does not sound, however, like traditional talk of capacity, which is typically taken as either obtaining or not; and it is not clear what the details of this take on capacity would consist in. At the very least, to reintroduce talk of capacity would require some difficult explicatory work about its re-envisioned nature that we are unable to assess sight unseen.

Regardless of whether our claims can somehow be adapted to a kind of capacity-talk, our interest is not in capacities themselves, but in the difficulty of *exercising* certain capacities, where we measure this against a baseline of what comparable agents might be expected to do. Though there is some historical precedence for including difficulty in exercising a capacity in the criteria for moral responsibility, the focus has typically been on *volitional* capacities.[19] It is difficult for the unwilling addict, for example, to resist taking the drug because his counter-desire, his craving for the drug, is so strong. To the extent we cut him some slack, then, we may do so because we think it was just too difficult for him to overcome that volitional obstacle, where the same would be expected of other similarly situated agents. (Of course, this way of putting it might raise capacity talk once more, for perhaps the unwilling addict really is incapacitated with respect to volitional obstacles *of that strength*, or perhaps he partially lacks a "meta-capacity" to exercise his volitional capacities.[20])

The difficulty we appeal to is analogous to that just mentioned, but differs in that it concerns *perceptual*, rather than volitional capacities.[21] Indeed, there is no reason to think that our Toms are struggling against their own desires and inclinations otherwise to do what's right. Rather, it is just harder

[19] See, e.g., a recent thread kicked off by Dana Nelkin on the agency and responsibility blog "Flickers of Freedom": http://agencyandresponsibility.typepad.com/flickers-of-freedom/2013/01/difficulty-and-degrees-of-blameworthiness-and-praiseworthiness.html.

[20] One might read Harry Frankfurt in the former way. See his "Freedom of the Will and the Concept of a Person," in Harry Frankfurt, *The Importance of What We Care About* (Cambridge: Cambridge University Press, 1988), pp. 11–25.

[21] By "perceptual" capacity, we mean whatever capacity allows apprehension of moral qualities in specific cases. We are not assuming that moral knowledge is fundamentally perceptual or empirical.

for our Toms to "see" what the right thing to do is, given their morally deprived upbringings. When they do, it is surprising, for we tend to think that most people from their background would not have seen the light.

As an analogy, suppose that I have been shown the famous image of the "duckrabbit" repeatedly since childhood (see Figure 1.3), and I have been taught over and over that what I am looking at is a duck, and only a duck. When, as an adult, I meet you, and you insist that the image can also be seen as a rabbit, it will be no surprise if, given my upbringing, I have a very hard time coming to see it as a rabbit. Certain features of the picture have been drilled into me as exclusively salient (e.g., the bill), so it is quite difficult for me to come to see other features (e.g., the little "rabbit mouth" indentation on the back of the duck's head) as salient in my assessment of the image. Suppose, however, that someone else present, who was raised just as I was to see the duck, *is* able to see the rabbit quite quickly once told about it. Especially if we have taken my inability to see the rabbit as representative of the baseline, we are likely to be impressed with this person. We will probably applaud his ability to see past what he is used to.

Of course, again, one might suggest that this is best understood in terms of capacities—perhaps he has a stronger "meta-capacity" to exercise his perceptual capacities than I do—but this is not necessary. Regardless of where we come down on capacity-talk, the point remains that this person's ability to see the rabbit is impressive, and speaks highly of him as a perceiver. Our suggestion is that, on analogy, we are more likely to praise Tom_D because we are impressed with his ability to "see" past what he is used to, *morally* speaking. This way of thinking about difficulty also responds to the worry that praiseworthiness might attach to any old difficulty. We are merely advancing the Difficulty Hypothesis to explain cases of moral ignorance given childhood deprivation, and

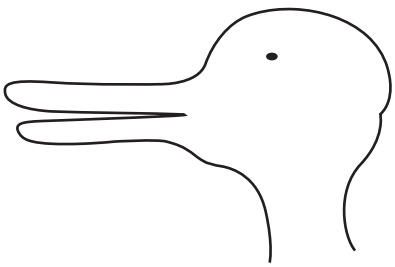

Figure 1.3 Duckrabbit.

what it appeals to is specifically perceptual difficulty in grasping moral reasons. We take no stand on whether it applies to volitional difficulty, or other sorts of difficulty for that matter.

Revisiting Huck

Our primary aim in this essay has been to bring some empirical results to bear on the MDE and Symmetry theses. While the MDE Thesis looks false in light of our results, there is a weaker, scalar, version of it that may be defensible. And while it may seem as if our results cause real worries for the Symmetry Thesis, it too may be defended. What inclines us toward the latter stance in both cases is the unified defense that may be provided them by the Difficulty Hypothesis.

We conclude where we began, with Huckleberry Finn. Huck has been raised to believe that stealing someone's property is immoral, and that slaves are people's property. As he grows he continues to mouth and even embrace the racist judgments instilled in him by his family and community since childhood. But when he is finally presented with an opportunity to turn Jim in, he balks, going against his moral beliefs and thus opting for hell instead. He is represented by our Tom_{D}, and people view him not only as praiseworthy, but as more praiseworthy than someone just like him but without the childhood moral deprivation.

There has been a lot of philosophical ink spilled (or at least a lot of words processed) on the Huck Finn case in recent years.[22] Our aim hasn't been to resolve the many issues raised in these discussions; rather, it has been the far more modest one of providing some much-needed empirical background to them. Far too often, discussion of the case proceeds by stating what our pretheoretic intuitions about Huck are. Ours is the only study we know of that attempts to determine just what those intuitions consist in. People do view him (or someone very much like him) as quite praiseworthy, although not completely so.

[22] For a tiny sampling, see Jonathan Bennett, "The Conscience of Huckleberry Finn," *Philosophy* 49 (1974): 123–34; Nomy Arpaly, *Unprincipled Virtue: An Inquiry Into Moral Agency* (Oxford: Oxford University Press, 2003); Nomy Arpaly and Timothy Schroeder, "Praise, Blame, and the Whole Self," *Philosophical Studies* 93 (1999): 161–88; Joel J. Kupperman, *Character* (Oxford: Oxford University Press, 1991); and Craig Taylor, "Moral Incapacity and Huckleberry Finn," *Ratio* 14 (2001): 56–67.

Further, we have attempted to show just how much of a role the moral deprivations of someone-like-Huck's upbringing contribute to people's assessments of his praiseworthiness. It indeed matters that he's been raised the way he has, moving him up the scale of praiseworthiness a full step over someone who does what he does without such a background. But one surprising feature of the study is that his increased praiseworthiness seems disanalogous to how people view our JoJo lookalike, who is less blameworthy—less responsible?—for his bad deeds than a morally undeprived doppelganger.

Indeed, here is the precise rub for determining whether the results of our study negatively impact or ultimately reinforce the Symmetry Thesis: childhood moral deprivations are often thought to mitigate responsibility itself. Being more or less blameworthy may simply be viewed as translating to being more or less responsible. If responsibility is indeed scalar in this fashion, then our results suggest that there *is* in fact an asymmetry between negative and positive cases: morally deprived upbringings are viewed as decreasing one's responsibility in the negative realm and increasing one's responsibility in the positive realm. This would truly be a surprising result.

Less surprising, but perhaps no less interesting, would be a different interpretation. Responsibility may not be scalar in the way just suggested; instead, perhaps it is either not scalar at all (one either is responsible for something or one isn't), or it is scalar, just not in line with the scalar nature of blame- and praiseworthiness (perhaps it is scalar in line with various capacities one may have or exercise by degrees, where this doesn't affect degrees of blame- or praiseworthiness). If this is the case, then the Symmetry Thesis might be reinforced. If we think of blame- and praiseworthiness on a single scale, then (where full-blown responsibility would attach to one's actions at any point on the scale, say) the moral deprivations of childhood could be viewed symmetrically in negative and positive cases as moving one away (by roughly the same amount) from the completely blameworthy endpoint.

While we haven't taken a definitive stand either way here, our previous discussion on the JoJo case does provide some explanatory ammunition if one adopts the second approach. What could well be moving the assessments of _____-worthiness up the scale away from complete blameworthiness is that people view childhood moral deprivations as making moral perception more difficult for their agents, so that adhering

to, or overcoming, the judgments ingrained by those deprivations will predictably yield assessments at a different degree than their nondeprived counterparts. Just as we cut JoJo some slack for doing what it would have been difficult for him not to see as wrong, we also admire Huck for having done what it was difficult for him to see as right. Or at least we admire Huck in a different way, or to a different degree, than we would those who more easily saw what he didn't because of their more privileged upbringing. As his internal monologue suggests, it's a serious and genuine struggle for him to reject "morality" in the way he does. That he did so and got it right (despite what he still thought) is worthy of real praise, apparently. It is a surprise.

There is, as always, much more work to be done on these issues. In particular, it would be valuable to know more about the reasons our positive Toms went against their moral beliefs. Again, we deliberately left this part open so as not to beg the question against any particular theory of relevant moral reasons, but in doing so we also left it open that the Toms might have changed their minds by accident, or under the influence of various nonrational forces. (We doubt this is how the cases were interpreted, but it's certainly possible.) It would also be interesting to know whether it's the mere isolation of their upbringing or the specific moral facts from which they have been deprived that is doing the work on people's different intuitions. To test this, one might have Tom_{C} and Tom_{D} raised in different degrees of isolation but now indoctrinated with the same set of moral facts, that people are to be treated equally, etc.; and then have them both act in accordance with their moral beliefs in helping the fallen man up. Whether or not there is a difference in people's moral intuitions, we would learn something interesting. In addition, the *degree* of praiseworthiness people would assign may be compared to the degrees assigned in our present cases, so that we may know more about how people assess right-doing in line with moral upbringing. Indeed, finding out how people assess these cases will likely tell us even more about how we ought to view the Symmetry Thesis, and so could ultimately generate real philosophical payoff.[23]

[23] We are very grateful to the folks at Yale's Experiment Month for carrying out our proposed study on these issues. We are also grateful to Don Callen for serving as advisor to the study, and to Tamler Sommers and Michael McKenna for discussion of some of the ideas herein.