Sentiment, Intention, and Disagreement: Replies to Blair & D'Arms[*]

Shaun Nichols

1. Introduction

I am most grateful to James Blair and Justin D'Arms for commenting on my work. I would be hard put to name two other moral psychologists whose reactions I'd be so keen to hear. There is a striking asymmetry in their commentaries. Blair prefers a minimalist story about moral judgment, maintaining that the appeal to rules is unnecessary. D'Arms, by contrast, maintains that the account I offer is overly simple and that children lack moral concepts despite their partial facility with moral language. It is tempting to treat my account as achieving the golden mean between Blair's austerity and D'Arms' extravagance. But it would be unfair to both. Blair is attracted to the sparse account for empirical reasons, and D'Arms is attracted to a richer account for philosophical reasons. Nonetheless, I still think that the account I offer is preferable to Blair's minimalism and to D'Arms neosentimentalism. Rather than give a point-by-point reply, which would likely be tedious, I'll try to say why I think that my account is still more plausible than the alternatives proffered by Blair and D'Arms.

2. The necessity of rules

In an earlier paper (Nichols 2002), I argued against Blair's account of moral judgment. Blair maintains that normal (i.e. nonpsychopathic) humans have a Violence Inhibition Mechanism (VIM) which is activated by cues of distress or suffering. To simplify his account somewhat, Blair suggested that moral judgment occurs when the VIM generates an emotional reaction. The gist of my objection was that Blair's account leads to the awkward conclusion that we make moral judgments when we witness accident victims. For in those cases, the VIM will generate emotional responses to the distress cues. As a result, I argued, Blair's VIM mechanism might lead to judgments that something is *bad,* but it doesn't get us all the way to judgments that something is *wrong.*

Blair is extremely gracious about my previous objection, and he now allows that VIM alone isn't sufficient for moral judgment. On his new proposal, VIM activation is still essential to moral judgment but so is the judgment that there was intent to cause harm. He writes, "actions are considered 'wrong' rather than merely 'bad' when there is intent to cause harm" (p. xx)

Blair recognizes two prima facie objections to his proposal. First, many clear cases of moral violations don't seem to count as moral violations on Blair's proposal. Sometimes an agent can lack the intent to cause harm but be just as much the target of moral condemnation. In his commentary, Blair discusses the case of a drunk driver causing fatalities, and Blair ultimately suggests that the drunk driver who kills receives our moral condemnation because he "intended to take an action that could be expected to harm the victims" (p. xx). That response works, but notice that Blair has now introduced a crucial change to his account. It isn't just "intent to cause harm"; moral judgment is

[*] Thanks to John Doris, Ron Mallon and Walter Sinnott-Armstrong for many helpful suggestions.

also generated when there is *intent to act in a way that can (or should) be expected to cause harm*.

Now that Blair has expanded the range in this way, the other prima facie problem becomes clear. There will be many cases of actions that are not judged morally wrong but in which the agent expects that the action will cause harm or suffering. Consider first a delightful example from Alan Leslie and Ron Mallon: cry babies. Two children, James and Tammy, are eating their lunch, and each has one cookie. But James wants to eat both cookies, and it's clear that he'll be very upset if he doesn't get his way. When Tammy eats her cookie, this makes James cry. Tammy might have known that James would be distressed, but her action isn't regarded as morally wrong (Leslie et al., forthcoming). For a second example, take boxing – a sport in which participants have the intent to hurt the opponent bad enough to render him unconscious. Most people, at least in many cultures, do not regard the boxer as committing a moral violation when he punches his opponent. Finally, consider punishment –we often engage in or endorse punishment of others, with the express intent to inflict suffering on the target.

Blair acknowledges the problem punishment might pose. And he tries to solve the problem by invoking competing factors in the judge. When we consider the propriety of an agent punishing a child, we "represent the punisher's internal state and represent two valenced goals: the aversive reinforcement of the child's distress as well as the appetitive reinforcement of the child's future well-being… According to the model, this judgment is determined according to whether or not the aversive reinforcement of the child's distress outweighs the appetitive reinforcement of the child's future well-being" (p. xx). So, when the appetitive reinforcement outweighs the aversive reinforcement, we judge that the action is permissible.

Blair's presentation of the appetitive/aversive model is very brief, but it seems to run into a serious problem. For sometimes appetitive reinforcement outweighs *moral judgment*. A child's desire for candy can be so great that the child decides to steal in the face of a countervailing moral judgment. In such a case, and many more like it, the individual will often say something like "I know I shouldn't have done that, but I just wanted candy so badly that I went ahead anyway." In short, the problem for Blair's proposal is that our appetites sometimes win out over our moral judgments. So it seems that Blair cannot account for the problematic cases of moral judgment simply by appealing to the victor of the competition between appetitive and aversive influences on our judgment.

The problem for Blair, then, is that often it is not a moral violation to intend to perform an action that is expected to cause harm. Cry babies, boxing, and punishment are the three examples that I suggested, though many more could be provided of course. It will be no trivial matter to fix the account in a way that matches up with our everyday practices of moral judgment. In contrast to Blair's proposed competition between aversion and appetite, I would again invoke moral rules to explain the complex pattern of moral judgment that we see surrounding permissible harmful actions. Harmful actions can be permissible because our rules are a complex lot, involving such considerations as fairness, consent, and retributive justice.

3. Concepts, content, and moral psychology: a prelude to disagreement

In the metaethical tradition that places emotions at the center of moral judgment, "neosentimentalism" is perhaps the major account on the contemporary scene. According to neosentimentalism, to judge that an action is morally wrong is to judge that it would be appropriate to feel guilt on doing the action.[1] In my chapter, I argue that neosentimentalism is threatened by the fact that we find a dissociation in children between the capacity for moral judgment and the capacity to judge the appropriateness of guilt. D'Arms offers a defense of neosentimentalism against this dissociation argument. One of the themes in D'Arms' response is that neosentimentalism is an account of the *content* of moral concepts. So I want to say something about the relationship between theories of content and neosentimentalism before proceeding to his main argument.

Issues about content are complex in ways that bear on the viability of neosentimentalism. Let's consider two extremely prominent approaches to content: informational accounts and functional role accounts. For present purposes we can be very crude about what these stories say. On prominent informational accounts, the content of a concept is characteristically determined by properties in the external world that cause the concept (Dretske 1981; Fodor 1990). So, familiarly, the concept WET refers to the property in the world that causes the tokening of that concept. If such an informational account is the globally correct theory of content, things don't look so good for neosentimentalism. For the typical external cause for the tokening of moral concepts is an *action*, not the appropriateness of an emotion. By contrast, if the informational account were globally true, that would weigh in favor of certain moral realist accounts, like Sturgeon (1988) and Brink (1989). On those views, the content of *wrong* is indeed the property in the world that causes the tokening of that concept.

Let's turn to the most important rival to informational accounts of content, functional role accounts. According to some prominent versions of the functional role approach, the content of a concept is given by the overall functional or causal role it plays in the psychology of the agent (e.g. Block 1986). Of course, when we look across individuals there will be differences in the functional roles of the concepts. So the theorist needs some account of how much (or what part) of the functional role has to be the same in order for two people each to have a token of a concept with the same content. If we think that even small differences in functional roles mean a difference in content, then it will turn out that people rarely use the same concepts, and we typically talk past each other. Such fine-grained functional role theories would certainly give us reason to say that a young child and I do not share a common concept of morality. But those theories would also lead us to say that it would be exceedingly rare for any two adults to have the same concept of morality. At the other end of the spectrum, one might take a very coarse-grained functional role approach, on which two people have the same concept so long as there is some significant overlap in functional role,. At least prima facie, that approach would favor the view that a young child and I have the same concept

[1] I will focus here, following my paper and D'Arms' reply, on guilt as the crucial emotion for neosentimentalism. But I should note that D'Arms' own view is more complex. He has argued against the view that an adequate neosentimentalist account of moral judgment can be provided by appealing to guilt alone (see D'Arms & Jacobson 1994). None of the discussion in D'Arms' commentary or in this reply, though, hangs on the additional complexities.

MORALLY WRONG, since is significant overlap in the functional roles of our concept tokens (as indicated, say, by performance on the moral/conventional task). Neither of these outcomes would be favorable to D'Arms' neosentimentalism. D'Arms wants to maintain that people typically don't talk past each other in moral conversation, but also that children do not have the same concepts as adults. Fine-grained and course-grained accounts wouldn't let him have it both ways.

Thus, if we approach the issue of the content of moral concepts by considering different theories of content, there is no reason to think that neosentimentalism gives us the right semantics for moral concepts. My own suspicion is that there is no single correct theory of content, even for a given class of concepts, like proper name concepts or natural kind concepts (see e.g., Machery et al. 2004; Mallon et al. in prep.). If that's right, then we need to be very cautious about claims like that of neosentimentalism. For neosentimentalists are promoting a univocal account of content for moral terms, and that might be a fundamentally misguided undertaking. It might turn out that *there is no* univocal account of content for moral terms.[2]

But what I suspect is really going on is that D'Arms thinks that by looking at various facts about moral judgment, we can glean constraints on the semantics that push in favor of neosentimentalism. So, while D'Arms' neosentimentalism would be doomed if it turned out one of the above theories is the universally correct theory of content, D'Arms will presumably say that these cannot be globally correct theories of content, because at least for evaluative concepts, the neosentimentalist account is right. The informational theories are incorrect as accounts of the content of moral concepts. For functional role accounts, when we individuate evaluative concepts, the grain is neither very fine nor very coarse. Rather, the grain must be neosentimentalist. Similarly, if D'Arms' argument succeeds, it will relieve my skepticism about the univocality of content, at least for moral terms. The arguments in favor of neosentimentalism will not only tell us something important about moral judgment, they will also inform the theory of content itself. That is, of course, a bold thesis, so let's look at the argument.

4. Disagreements

According to neosentimentalism, to judge that it is morally wrong to *A* is to judge that it is appropriate to feel guilty for doing the action. As D'Arms notes, this judgment of appropriateness needn't be occurrent whenever one makes a moral judgment. Rather, the key point is that people who make moral judgments have the *disposition* to judge that it's appropriate to feel guilty for doing the action. It is that disposition that makes a given judgment an instance of moral judgment.

Obviously there is a huge set of dispositions surrounding occurrent moral judgment. For instance, when people judge that it is wrong to *A*, they have dispositions like the following:

-the disposition to be attentive to such actions,

-the disposition to think that it is appropriate to be so attentive,

-the disposition to show high galvanic skin response,

[2] See the contributions by Gill and Loeb, this volume, for arguments that moral semantics will not be univocal.

-the disposition to respond more quickly in recognizing moral vocabulary in word/nonword tasks.

In addition, of course, when people judge that it is wrong to *A*, they have the disposition to make judgments about the appropriateness of guilt. Some of these dispositions are likely present in children, e.g. GSR response; and some are likely absent in children, e.g., dispositions to make judgments about the appropriateness of attentiveness or guilt. Why do neosentimentalists insist that the key disposition, the one that makes or breaks moral judgment, is the disposition to judge that guilt (or some other emotion) is appropriate? Apart from neosentimentalism, presumably the standard view is that judgments about the appropriateness of guilt are *consequences* of moral judgment. So why opt for the complex neosentimentalism view?

The answer is disagreement. According to D'Arms, we need a neosentimentalist account to accommodate moral disagreement. D'Arms is committed to the superiority of neosentimentalism here generally (see D'Arms 2005), but for present purposes, all D'Arms needs to argue is that neosentimentalism offers a better explanation of moral disagreement than the affect-backed normative theory account that I've proposed. So I'll restrict the discussion to these two accounts.

As I say in my chapter, on the affect-backed theory account, much of the moral disagreement that we see in everyday life can be explained by adverting to the content of a widely shared normative theory. For in most moral disagreements the disputants share certain basic moral principles that guide their judgments. Some philosophers suggest that once we clear away all the diversity in factual judgments, there will be no remaining cases of moral diversity (e.g. Boyd 1988). I'm unwilling to take *that* strong a line. It seems likely that there are some cases of "fundamental" moral disagreement that would remain even after all factual disagreements are resolved. For instance, I think it quite possible that Brandt was right that there was fundamental moral disagreement between his Hopi informants and suburbanites on whether it's okay to harm animals for sport (Brandt 1954). In addition, I think that Doris and Stich make a good case that there is fundamental moral disagreement in the US between southerners and northerners on the moral propriety of violent reprisals for insults (Doris & Stich 2005). And Doris & Plakias (this volume) make a nice case that there are fundamental disagreements between East Asians and Westerners on familiar cases like the magistrate and the mob. Nonetheless, I do think it's important to recognize that fundamental moral disagreement might be a comparatively rare form of moral diversity.

To see whether neosentimentalism provides a better account of fundamental moral disagreement than the account I've offered, I would like to begin by noting two central *apparent facts* about moral disagreement. The point here is not to insist that moral disagreement really has these characteristics, but merely that this is how, at first glance, moral disagreement seems.

i. Many people *seem* to treat some moral transgressions as objectively wrong. They say things like "Ethnic cleansing is wrong even if another culture thinks something else" and "Ethnic cleansing would be wrong even if we didn't have the emotions we do."

ii. Moral disagreements *seem* to be fundamentally about right and wrong, and not fundamentally about emotion.

On the neosentimentalist account of moral disagreement, things are not as they seem. According to neosentimentalism, moral disagreements are fundamentally about the appropriateness of certain emotions. D'Arms writes, "what makes moral disputes univocal… is that they concern the appropriateness of responses to which all parties to the dispute are susceptible" (p. xx). As far as I can tell, no one innocent of sentimentalist metaethics has ever thought that the content of their moral disagreements with others about, say, abortion, animal experimentation, circumcision, corporal discipline, or capital punishment, were, most fundamentally, about the appropriateness of feeling various emotions. Although moral disagreements don't seem fundamentally to be about emotions, neosentimentalists would maintain that this superficial feature of moral disagreement doesn't reflect the deep truth about moral disagreement. The deep truth is that, despite appearances to the contrary, moral disagreements are really about the appropriateness of feeling certain emotions. As a result, the other apparent fact about moral disagreement is also false. Although it seems like some people are objectivists, this appearance is misleading. People don't really think that moral claims are true in some objective way. Rather, moral claims are about which emotional responses are appropriate, given the emotional repertoires that we have.

Now, of course, even if it's true that we don't think of our moral claims as having a certain content, it's possible that they do have that content nonetheless. Many philosophers think that sometimes we don't know the contents of our concepts. For instance, if we focus on the *extensional content* of our thoughts, then many would say that important elements of the extensional content of WATER were missed by ancient thinkers. They didn't realize that water is a combination of hydrogen and oxygen. But it's crucial to note that the neosentimentalist is not merely saying that judgments of appropriateness of emotions capture the *extensions* of moral concepts. Rather, as D'Arms (2005) writes, "A shared sentiment supplies a shared element in the *intensions* of our evaluative thoughts" (D'Arms 2005, MS 17, emphasis added). Although it's a familiar claim that we often lack access to the *extensional content* of our concepts, it's quite controversial to maintain that we sometimes don't even know the *intensional* content of our everyday concepts.[3] Even a slave to Nisbett & Wilson (1977) like me thinks that we typically do have access to the intensional contents of our thoughts. As a result, I submit that the neosentimentalist needs a very good argument that the intensional content of people's moral concepts and moral disagreements are not accessible to them.

How does my theory fare with respect to the apparent facts about moral disagreement listed above? On the account I've proposed, moral judgment derives from an affect-backed normative theory. On that approach, we can easily allow that things are

[3] There are some approaches to intensional content on which we might lack such access. For instance, on a possible worlds semantics approach to intensions, an intension is just the extension of an expression across all possible worlds. But I should think that neosentimentalists would not want to tie their fate to a particular approach to intensions. And, in any case, it would seem ad hoc to cast about for an approach to intensions that will assist the neosentimentalist in avoiding this problem.

as they seem. Many people seem committed to moral objectivism. To accommodate this, we can simply allow that one feature of the normative theory (at least for many people) is that it carries the presupposition that moral claims like "it's wrong to force sex on a stranger" are true *simpliciter*. Given the background of assumed moral objectivity, it's easy to accommodate these as disagreements about right and wrong rather than about emotions. The problem of capturing moral disagreement only emerges when we deny that people are moral objectivists. Now, I also doubt, as do many neosentimentalists, that morality really is objective. But that does nothing to undermine the claim that lay moral disagreements are crucially underwritten by the presumption of objectivity.

So the affect-backed theory account can easily accommodate the two apparent facts about moral disagreement. What my theory *cannot* easily accommodate is fundamental ethical disagreement between individuals who fully reject moral objectivism about the target action. That is, if Mark thinks that it is wrong to *A* and Eric thinks that it's not wrong to *A*, and if both of them reject that idea that it is *objectively wrong (or not wrong)* to *A*, then my account has no obvious story to tell on which Mark and Eric are locked in fundamental ethical disagreement. But it's not clear to me that this is much of a cost. For it's not clear to me that in such cases there *is* fundamental ethical disagreement.

There's another obvious gloss available, and it's one that D'Arms acknowledges at the end of his response. I can maintain that those who reject moral objectivism treat moral claims as true only relative to some community, culture, or arbitrary feature of human constitution. In fact, this seems fairly plausible to me. Many people seem to treat some moral transgressions as wrong in a relativistic way. They say things like "Unprovoked hitting is wrong in our culture, but it's not wrong in other cultures." These cases are somewhat controversial, but it is worth noting that there are utterly clear cases of normative relativism if we look to non-moral transgressions. Consider the following scenario (from Nichols 2002): "Bill is sitting at a dinner party and he snorts loudly and then spits into his water before drinking it." The vast majority of subjects said that this was not okay and would not have been okay even if the host had said that such actions were permissible. However, Western undergraduates also say that the disgusting violations that are wrong for us need not be wrong *simpliciter* (Nichols 2004). In these cases, we find normative diversity without fundamental evaluative disagreement. Although I wouldn't count these judgments about disgusting transgressions as cases of *moral* diversity, the evaluative diversity surrounding disgusting violations can serve as canonical cases of relativism about norms. This gives us a natural place to fit our moral nonobjectivists – they treat moral transgressions the way most people (in our culture) treat disgusting violations.

In summary, the one thing that neosentimentalism can do that I can't is explain fundamental moral disagreement among nonobjectivists. But this does not seem a sufficient advantage to favor the neosentimentalist approach to moral disagreement. For (i) in order to explain the disagreement the neosentimentalist needs to claim that the appearances of moral disagreement are quite misleading – that people don't really understand what their moral disagreements are about. And (ii) it is independently dubious that moral nonobjectivists typically do engage in fundamental moral disagreement with each other. Neither of these reasons counts as a refutation of neosentimentalism. But they do, I think, suffice to show that the neosentimentalist lacks

any powerful advantage over my theory when it comes to the phenomena of moral diversity.

5. What is moral?

Both Blair and D'Arms have a tidy story that picks out the moral domain. For Blair, the moral domain is given by his VIM mechanism. For D'Arms, the moral domain is carved out by the appropriateness of guilt or anger. I don't have any comparable proposal, as D'Arms notes: "I am not sure what distinguishes moral prohibitions from other prohibitions on the Sentimental Rules account – which I take to be an important problem" (p. 8).

My failure to give an account of the moral domain wasn't just oversight on my part. One of the few things I recall from graduate school is that proposing definitions is not for the risk averse. Rather, definitions are for those willing to philosophize dangerously. Most proposed definitions in philosophy have been manifest failures. Indeed it's hard to come up with any examples of successful definitions, despite the considerable energy philosophers have exerted. When chemists meet with so little success, their jobs are in jeopardy.

As with other philosophically important concepts, I think it unlikely that MORAL will have a nicely delineated definition. Although it might seem tempting to claim that the moral/conventional distinction provides a crisp line for characterizing the moral domain, the temptation should be resisted. The different factors in the moral/conventional task (e.g. generalizability, authority contingence, seriousness), have been shown to come apart in various ways (e.g. Kelly et al. forthcoming).[4]

Even though I refuse to adopt an account of the moral domain, this does not present an insurmountable problem for investigating moral psychology. For we can recognize obvious cases that fall in the domain of moral prohibitions (e.g. rape, murder) and obvious cases that fall outside the domain of moral prohibition (e.g. table settings). By coming to understand the clear cases, we can at least get the beginnings of an account of the nature of moral judgment, even if that account will not deliver anything like a sharp delineation of the moral domain.

References:

Block, N. 1986. "Advertisement for a Semantics for Psychology," in P.A. French, T.E. Uehling and H.K. Wettstein, eds., *Midwest Studies in Philosophy, Vol. X*, Minneapolis: University of Minnesota Press: 615-678.

Boyd, R. 1988. "How to Be A Moral Realist." In G. Sayre-McCord (ed.), *Essays on Moral Realism*. Ithaca and London: Cornell University Press.

[4] These results pose a problem for trying to use the moral/conventional test to define the moral domain. But the results do not, in my view, undermine the interest of the basic finding that there are important differences in judgments about intuitively immoral actions and judgments about intuitively convention-defying actions.

Brandt, R. 1954. *Hopi Ethics: A Theoretical Analysis*. Chicago: The University of Chicago Press.

Brink, D. 1989. *Moral Realism and the Foundations of Ethics*. Cambridge: Cambridge University Press. Doris, J. & Stich, S. 200x. As a Matter of Fact: Empirical Perspectives on Ethics.

D'Arms, J. 2005. "Two Arguments for Sentimentalism", *Philosophical Issues*, 15, 1-21.

D'Arms, J. and Jacobson, D. 1994. "Expressivism, Morality, and the Emotions," *Ethics*, 104, 739-63.

Doris, J. and Stich, S. 2005. "As a Matter of Fact: Empirical Perspectives on Ethics." In F. Jackson and M. Smith (eds.), *The Oxford Handbook of Contemporary Philosophy*. Oxford: Oxford University Press.

Dretske, F. 1981. *Knowledge and the Flow of Information*, Cambridge, Mass.: The MIT Press.

Fodor, J. 1990. *A Theory of Content and Other Essays*, Cambridge, Mass.: The MIT Press.

Kelly, D., Stich, S., Haley, K., Eng, S. and Fessler, D. forthcoming. "Harm, affect and the moral / conventional distinction." *Mind and Language*.

Leslie, A., Mallon, R., & Dicorcia, J. forthcoming. "Transgressors, victims, and cry babies: Is basic moral judgment spared in autism?"

Machery, E., Mallon, R., Nichols, S., and Stich, S. 2004. "Semantics, Cross-Cultural Style." *Cognition*, 92, B1-B12.

Mallon, R., Machery, E., Nichols, S., and Stich, S. in prep. "Cross-Cultural Semantics and Arguments from Reference."

Nichols, S. 2002. "Norms with Feeling: Towards a Psychological Account of Moral Judgment." *Cognition*, 84, 221-236.

Nichols, S. 2004. "After Objectivity: An Empirical Study of Moral Judgment." *Philosophical Psychology*, 17, 5-28.

Nisbett, R. & Wilson, T. (1977): "Telling More than We Know: Verbal Reports on Mental Processes". *Psychological Review* 84: 231-59.

Sturgeon, N. 1988. "Moral Explanations." In G. Sayre-McCord (ed.), *Essays in Moral Realism*. Ithaca and London: Cornell University Press.