Reply to *Commentary on Buckwalter and Phelan*

Wesley Buckwalter (CUNY) and Mark Phelan (Lawrence)

We would like to thank Justin Sytsma for his insightful and comprehensive comments on our paper. They will surely help move the discussion about how people ordinarily sort and group mental states forward in a productive and empirically informed way. Nonetheless, we have a number of brief concerns about his objections to our work.

1. Positive vs. Negative Hypothesis Clarifications

In our paper we argued that the assumptions people make about the function of the S&M robot play an important role in their attributions of mental states to the robot. Sytsma's dialectical points about how this claim bears on the structure of S&M's paper are well taken. Perhaps we could have done a better job of clarifying and distinguishing our criticism of the evidence S&M use to support their *negative project* in their first study (that non-philosophers do not group mental states in the manner that philosophers do on the basis of phenomenal character) from the evidence used to support their *positive project* in subsequent studies (that the manner in which non-philosophers do sort mental states involves an appeal to a state's valence).

To clarify our approach, we begin our paper with a discussion of S&M's *positive project*. While we were able to detect differences in mental state attributions across our experiments, these differences were not correlated with people's judgments about valence. Sytsma seems perfectly willing to concede this point, and even provides his own compelling evidence in his response to our work that directly challenges the idea that people are sorting mental states based on valence. Later on, we will briefly discuss Sytsma's original data questioning the valence hypothesis. But for now, we will assume consensus on this point, and proceed under the assumption that the valence hypothesis is false.

However, this still leaves the further question about what to make of S&M's *negative hypothesis*. Does our evidence also challenge the broader claim that non-philosophers group mental states differently than philosophers? To review, the key piece of evidence to support S&M's claim comes from their first experiment, in which ordinary participants (but not philosophers) are willing to attribute seeing red, but not feeling pain, to a simple robot named Jimmy. S&M contend that these data demonstrate that the manifest dichotomy view is false, and that non-philosophers do not sort mental states as philosophers have assumed. On the other hand, we contend that, despite this evidence, the manifest dichotomy view may be true, and that the asymmetry in state attributions could be accounted for by the assumptions made on the part of non-philosophers about the *function* of the simple robot.

As we suggest, the difference between non-philosophers' and philosophers' responses could be due to the fact that non-philosophers (but not philosophers), lacking training in

how to assess thought experiments, are imputing likely functions to the robot, and thus attributing sensory states that seem necessary to achieve these functions. Specifically, while seeing colors or smelling a technical-sounding chemical seem as though they could play a role in helping a robot perform whatever functions a robot would be designed to perform, it is tough to imagine how feeling anger or pain could play any role in reasonable robot functions. So the former states are more likely to get attributed, and the latter states are not. And this fact may underlie the differences observed by S&M between philosophers and non-philosophers.

In his response, Sytsma points to an experimental confound in our studies, which he takes to bear on our criticism of S&M's negative project. He suggests that specifying a function will lead people to attribute greater complexity to the robot with respect to those capacities relevant to performing that function. And he suggests that these assumptions about complexity are really what guides mental state attribution. Since S&M suppose attributions of phenomenal states will increase with complexity, Sytsma claims that if our results are explained in virtue of this kind of function-specific complexity, this defeats any challenge to the negative view arising from our results. In the next section, we will review this challenge and give our response.

2. Function-specific Complexity and the Negative Project

In his comments, Sytsma rehearses our argument against the negative project, mentioned above: "It might be that people tend to assume different functions for the robot Jimmy across the probes we used, which could potentially explain the asymmetry we found…" (5). However, he thinks this is too quick. For one thing, as Sytsma points out, our studies, "do not establish that people assume different functions in different scenarios when none is provided" (5). For another:

> …perhaps more importantly, there is a confound in B&P's studies that casts doubt on the claim that the function assigned to Jimmy directly impacts mental state ascriptions. Specifically, function relates to complexity, and we expect complexity to impact mental state ascriptions. (5-6)

However, we think that this proposed confound is irrelevant to our argument against the negative project. And, though we concede that we have not decisively demonstrated that ordinary people *do* assume different functions across the different probes, the burden of proof seems to lie with S&M to show that this *is not* the case.

As we see it, there are two potential explanations of our results. On the one hand, it's possible that people attribute experiential states to the robot only when *and in virtue of* the fact that they think the states would play a certain functional role in whatever task they think the robot was designed to perform. On the other hand, it may be that people are making certain assumptions about the robot's design, such that, if they suppose that the robot is designed to perform a certain function, they will suppose that the robot was given sufficiently complex machinery to pull-off that function (i.e., they will assume

more function-specific complexity)—and in some cases this might involve the need to have certain mental states. But whichever of these explanations is correct, we suggest that S&M's negative hypothesis does not go through.

The worry we're raising with S&M's argument for the negative hypothesis is the following one. Their data reveal that ordinary people and philosophers exhibit different patterns of subjective state attribution when it comes to the robot in their studies. But that data supports the conclusion that philosophers and the folk have different conceptions of subjective experience only on the further assumptions that people are not willing to attribute subjective experiences to a simple robot and that people conceive of the S&M robot as a simple robot across all of S&M's studies. If these assumptions do not hold, then perhaps S&M's studies aren't getting at differences in how people think about mental states at all. Rather they may be due to differences in how people think about t*he robot*. That is, perhaps the robot isn't simple in the relevant ways for ordinary people, as it is for philosophers. So whether people are naïve teleofunctionalists and their assumptions about the function of the robot make a direct difference to their attributions, or whether their assumptions about the function of the robot make a difference to their attributions in an attenuated way, on the assumption that people are attributing functions to the robot while philosophers are not, S&M's argument for the negative hypothesis fails.

As noted above, we do not have direct evidence that ordinary people (but not trained philosophers) are making different assumptions about the robot's function across different probes. But we think the points mentioned above lend support to this conclusion. Given their lack of training in assessing thought-experiments, ordinary people are presumably more likely than philosophers to infer extra information beyond what is actually stated in the case of a thought-experiment-style vignette. And it is easy to imagine how sense experiences could be helpful in performing typical robot functions, whereas pain experiences would not. Thus, our proposal strikes us as a plausible rival explanation of S&M's results. Given their strong, revisionist aims—rejecting the manifest dichotomy view, a common-place of contemporary philosophy of mind by S&M's own lights—we think it's up to S&M to respond to this challenge.

In his comments, Sytsma attempts to offer some new evidence that avoids this criticism of the negative project. So in the next section, we'll assess this new evidence.

3. Sytsma's New Data for the Negative Project

In his response, Sytsma provides some new experimental evidence in favor of S&M's original negative project. Sytsma writes:

> If the asymmetry we found in the first study reported in Sytsma and Machery (2010)—that people are significantly more likely to say that Jimmy saw red than that Jimmy felt pain—simply reflects that we did not specify a function for the robot, then we would expect that asymmetry to

> disappear when Jimmy is described as having either the function of helping with household chores or lifting and moving heavy objects (9).

To test this prediction, Sytsma administered new vignettes in which the specified function of the simple robot was not obviously specific to one type of mental state being tested over the other. In this experiment participants were given new S&M robot cases where the robot's function was either (i) unspecified, (ii) specified as having a general function, or (iii) having a lifting function. Participants were then asked between-subjects "Did Jimmy see blue?" or "Did Jimmy feel pain?" And Sytsma found a significant difference between these two questions regardless of the function specified (means all around 3.5-4 of 7 for yes "feeling pain", and all around 5.5-6 of 7 for yes "seeing blue").

Sytsma takes this as evidence that obviates our rival explanation to S&M's original data, and thus as evidence in favor of the negative hypothesis that non-philosophers indeed do have a different concept of subject experience than philosophers. But do these data show this? It seems obvious (to us at least) that one clear explanation of these new results again relates to the functional specifications of the robot in question. All three functions in this new experiment have nothing to do with *feeling* (that is, it is tough to see how feeling pain could be involved in performing these functions). And thus, it is incredibly difficult to imagine how feeling pain would be functionally useful to the entity in all three cases. Conversely, it is very easy to imagine how *seeing* would be functionally useful in all three cases (for moving about, for succeeding in a general experiment where visual cues are essential, or for helping Grandma find her precious sapphires earrings). And of course, our hypothesis says that people will be much more likely to attribute sense experiences to entities in precisely the latter circumstances—just as Sytsma found.

So we think that these latest results don't do much to help S&M's negative hypothesis. They can also be straightforwardly explained by the function hypothesis.

4. Sytsma's Data Questioning the Positive, Valence Hypothesis

In our paper, we presented evidence that the valence associated with the smells used in S&M's original experiments was playing no role in people's state attribution judgments. More specifically, while we did detect differences in mental state attribution in cases very similar to the ones used originally by S&M, we were not able to detect a moderating relationship for this and valence ratings as measured by our general affect scale of positive and negative state preference judgments. (Indeed, we could not even detect simple correlations between highly significant differences in people's affect judgments of the states and state attributions to the entities, even though to do so would reflect a boring statistical fact given these differences).

So, we're very glad that Sytsma has volunteered evidence replicating the same finding on a larger scale—using 18 different smells of varying valences. We thank Sytsma for coming forward with these results, and agree that when joined with our prior findings, they present a compelling case that the valence hypothesis is false.

5. Conclusion

We think there are compelling reasons to reject S&M's positive view that non-philosophers group mental states by state valence. We also think that these results question whether the data in hand really do support the claim that non-philosophers group mental states differently than philosophers. And while more research will be essential to rule out our rival explanation, as it stands the dichotomy between mental states may be more manifest than Sytsma and Machery would have us believe