Reflection, reflex, and folk intuitions

Bryce Huebner

Perhaps some day, we will find experimental data that deflates the philosophical presumption in favor of the ‘hard problem’ of consciousness. Unfortunately, that day is not today. Brian Talbot is right that data derived from commonsense psychology are unlikely to close the explanatory gap. More specifically, he is right that we must carefully investigate the source of commonsense judgments before employing them in the service of denying ‘‘the most central and manifest aspect of our mental lives’’ (Chalmers, 1995, p. 207). While no one should disagree with this claim, Talbot has used it to motivate an elegant case against the use of questionnaire-based studies that can only target—he claims—reflexive and associative processes that are likely to generate identical patterns of judgments regardless of whether qualia exist or not. In this brief commentary, I argue that Talbot’s claims about the processes responsible for producing commonsense judgments about mental states are unlikely to be correct. But, let me be clear: I do not think that this undercuts the force of Talbot’s broader conclusion; I agree that commonsense data are unlikely to suggest any plausible strategy for solving or dissolving the ‘hard problem of consciousness’.

Let me begin by briefly recounting what I see as the key details of Talbot’s argument. Talbot contends that System-1 processes are likely to play a determinative role in the production of commonsense judgments unless: (1) we have reason to mistrust our initial reactions, (2) we have the motivation to get the right answer, and (3) we are aware of a better decision strategy. In the case of mental state ascription, he argues that the relevant sort of System-1 process depend on associations between observable features of various entities and their mental states. In general, there are likely to be numerous associative links between non-qualitative states and the sorts of observable features that could be implemented in a simple robot. So, it comes as no surprise that people readily ascribe non-qualitative states to the simple robots described by Sytsma and Machery (2010). However, qualia—if they exist—probably have a more tenuous connection to the observable features of an entity; so, associations between qualitative states and observable features presumably derive from previous reflective thoughts about qualitative states, and from our reflexive simulations of the affectively valanced states of other people. Although a simple robot may posses the sorts of observable features that we have previously associated with seeing red or smelling a chemical, they are unlikely to possess the sorts of observable features that will trigger reflexive simulations of affectively valanced states. So, we should expect that commonsense psychology to resist ascriptions of affective states to simple robots.

1. Beyond System-1 and System-2 mechanisms

If Talbot is right that such associative System-1 mechanisms are ‘‘largely responsible for the folk judgments Sytsma and Machery (and by extension, other experimental philosophers in this area) study’’ (this issue), then he is also right that people will typically make judgments that are insensitive to the ontological status of qualia. This being the case, the data reported by Sytsma and Machery (2010) will offer no insight into the ‘hard problem’ of consciousness. While I concede that participants probably do begin from initial reactions that are driven by such associative mechanisms, I contend that questionnaire-based studies also leave substantial room for subsequent reflexive and reflective processing in the evaluation of these initial reactions. This being the case, I submit that while System-1 processes are sure to play some role in structuring the answers that are provided by participants in these studies, these initial reactions may not always be expressed as answers to the relevant experimental probes. I can't offer a fully elaborated justification for the claim that participants do revise their judgments on the fly in this paper.[1] Instead, I motivate the related claim that many participants are likely to mistrust their initial reaction, to carry out a targeted search of relevant possibilities, and then reflectively triangulate the output of these distinct processes in order to provide an answer to questions about the mental states and processes of non-human entities.

Talbot (this issue) contends that participants in experimental philosophy studies are in ‘‘a low stakes situation, a psychological experiment about a seemingly easy question’’. I agree that the stakes are low, but I strongly disagree with the claim that questions about the mental lives of non-human entities are easy to answer. Philosophers who have a well worked out and elaborated theory of the mind have clear intuitions; but asking whether a robot feels pain or smells bananas requires first getting clear (even if only implicitly) about the extent to which robots can have mental lives that are like our own in some critical respect (without this frame of reference, the question doesn't even make sense; cf., Dennett, 1997). However, it is not immediately obvious in which respect a mind must be like ours if it is to be in a particular mental state. Although ordinary people may initially react to a scenario by thinking that a simple robot probably doesn't have any mental states at all, they may quickly revise against this initial impression before offering their answer. Having listened to numerous participants who have preferred to carry out this process out loud, I assume that it proceeds roughly as follows: ‘‘I guess it seems like Jimmy wouldn't smell bananas. Oh, I don't know. Why would anyone ever build a robot that smells bananas! Whatever. That seems totally weird. I guess I'll go with a 2. . .no maybe a 3’’.

Where a person does not have a robust set of immediate expectations about another system's mental life, a new model of the situation has to be constructed. Since the scenarios that are used by experimental philosophers skirt the borders of fantasy and science fiction, it is reasonable to consider the kinds of systems that are employed in processing narrative representations of these kinds of worlds. The rich structure of a world that is presented in a science fiction film or a well crafted piece of fantasy literature can lead us to effortlessly interpret a wide range of non-human entities as having reasons, beliefs, intentions, goals, and feelings. By examining such worlds, we ‘learn’ that there are worlds where robots have mental lives that are quite similar to our own. However, even where people are unfamiliar with the structure of such science fiction worlds, they are likely to possess an array of cognitive capacities that allow them to construct a localized understanding of the world that has been presented in an experimental philosophy scenario.

There are numerous empirical questions about the precise nature of the mechanisms that allow us to evaluate counter-factual possibilities. However, it is clear that counterfactual scenarios often lead us to construct mental representations of what a world would be like in which the scenarios were true (cf., Fauconnier & Turner, 1998, 2003). Such representations begin from initial assumptions about the way that our world is, and then introduce modifications on the basis of ‘‘what the story tells us explicitly, what

[1] I develop this claim in Huebner (2011) in the context of moral psychology experiments; moreover, Bengson (in press) develops the conceptual point in a sympathetic criticism of the methodology of experimental philosophy.

we can directly deduce from specific conventions of the fictional genre, and, most importantly, how similar to the real world the fictional world is described as being'' (Skolnick & Bloom, 2006b, p. 77). By borrowing liberally from the structure of their everyday experience, and previously encountered and imagined worlds, people will reflexively construct a conceptual representation, and this representation provides a set of resources for evaluating the plausibility of an initial reactions to a claim about robotic mental states.2[2] In a slogan, questions about the mental life of a simple robot are likely to elicit the construction of 'blended mental spaces' that can be used to evaluate counterfactual possibilities that have not previously been considered (cf., Fauconnier & Turner, 2003).

Here, however, we come to the critical difference between this sort of theory and the theory that is advanced by Talbot. The blended mental spaces that are constructed in response to thought experimental prompts are simplified models, and as such they only provide a partial and incomplete representation of the described situation. These representations typically highlight a narrow range of salient features of the imagined situation, and they constrain the space of possibilities that will be considered in the evaluation of an experimental prompt (Fauconnier & Turner, 1998, 2003). This being the case, experimental prompts, like the philosophical thought experiments upon which they are based, are likely to succeed (where they do) by leading participants ''(either reflectively or unreflectively) to represent relevant non-thought experimental content in light of the thought experimental conclusion'' (Gendler, 2007, p. 69).

Returning to the 'smell' cases employed by Sytsma and Machery (2009), it seem that participants may have been lead to implicitly attend to features of the robot that are relevant to its purpose—after all, commonsense psychology often implicitly represents artifacts in terms of the purposes for which they are designed (cf., Bloom, 1996, 2000). Thus, a participant who initially judges that robots cannot smell anything might revise her judgment as she constructs a representation of a world in which she can consider the reasons why someone would build a robot that can detect strange chemicals like Isoamyl Acetate (which, unbeknownst to her, smells like bananas). It is not a stretch to think that military, industrial, or exploratory robots could be designed to do just this; so participants should be less apprehensive about claiming that a simple robot could smell Isoamyl Acetate. By contrast, there is no obvious reason why anyone would build a robot that was designed to smell bananas. While such a banana-sniffing robot may be conceivable, it is hard to imagine the justification for building one. So, this possibility suggests no obvious reason for revising the initial unwillingness to ascribe a capacity to smell to the robot. Thus if participants evaluate (perhaps subpersonally) different possibilities in light of the simplified mental models they construct, we should expect them to come to different conclusions about these two cases.

If this picture is approximately correct, then we should expect participants in such experiments to experience a conflict between their initial association-based reactions and the result of this process of mental space construction. But this yields a deep problem for the simple picture of cognitive processing that is advanced by Talbot. Building on a suggestion by Epley and Gilovich (2006), I contend that the difficulty of this type of question may lead participants

[2] Children as young as five-years-old have a sense of the possibilities that are allowed in different imaginary worlds that blends features of the actual world with shared and shareable features of previously imagined worlds (Skolnick & Bloom, 2006a,b). While there are numerous difficult empirical questions about the precise range of cases in which representations of merely possible worlds are likely to be constructed, I suggest that an understanding of mental spaces is likely to be critical for interpreting the results of the survey studies that have come to play a prominent role in experimental philosophy.

to carry out an effortful search of available possibilities, and this in turn may lead them to adjust their ‘gut response’ until they have arrived at a judgment that feels more plausible to them. Perhaps more importantly, if the difficulty of this question leads people to adjust their expressed response away from their initial reaction, then the conscious and deliberate nature of this adjustment will mean that telling participants not to adjust could even increase the extent of to which they adjust (cf., Epley & Gilovich, 2005). So, while participants who are presented with these kinds of experimental probe probably do yield a System-1 response, and the reflexive construction of a simplified mental models, these initial reactions are better understood as providing an initial assumption against which they can begin to reason about the proposed possibilities. While associative mechanisms play an integral role in structuring the range of possibilities that we consider, our capacity to construct and evaluate the structure of various fictional worlds also plays an important role in the way that we evaluate counterfactual possibilities.

References

Bengson, J. (in press). Experimental attacks on intuitions and answers. Philosophy and Phenomenological Research.

Bloom, P. (1996). Intention, history, and artifact concepts. Cognition, 60, 1–29.

Bloom, P. (2000). How do children learn the meaning of words. Cambridge, Mass: MIT Press.

Chalmers, D. (1995). Facing up to the problem of consciousness. Journal of Consciousness Studies, 2, 200–219.

Dennett, D. (1997). Kinds of minds: Towards an understanding of consciousness. New York: Basic Books.

Epley, N., & Gilovich, T. (2005). When effortful thinking influences judgmental anchoring: Differential effects of forewarning and incentives on self-generated and externally provided anchors. Journal of Behavioral Decision Making, 18, 199–212.

Epley, N., & Gilovich, T. (2006). The anchoring and adjustment heuristic: Why the adjustments are insufficient. Psychological Science, 17, 311–318.

Fauconnier, G., & Turner, M. (1998). Conceptual integration networks. Cognitive Science, 22(2), 133–187.

Fauconnier, G., & Turner, M. (2003). The way we think. New York: Basic Books.

Gendler, T. (2007). Philosophical thought experiments, intuitions, and cognitive equilibrium. Midwest Studies in Philosophy, 31(1), 68–89.

Huebner, B. (2011). Critiquing moral psychology from the inside. Philosophy of the social sciences, 41, 50–83.

Skolnick, D., & Bloom, P. (2006a). What does Batman think about SpongeBob? Children’s understanding of the fantasy/fantasy distinction. Cognition, 101(1), B9–B18.

Skolnick, D., & Bloom, P. (2006b). The intuitive cosmology of fictional worlds. In S. Nichols (Ed.), The architecture of the imagination: New essays on pretense, possibility, and fiction. Oxford: Oxford University Press.

Sytsma, J., & Machery, E. (2010). Two conceptions of subjective experience. Philosophical studies, 151(2), 299–327.

Talbot, B. (this issue). The irrelevance of folk intuitions to the ‘‘hard problem’’ of consciousness. Consciousness and Cognition. doi:10.1016/ j.concog.2010.12.005.