Phenom Cogn Sci (2010) 9:133–155
DOI 10.1007/s11097-009-9126-6

Commonsense concepts of phenomenal consciousness: Does anyone *care* about functional zombies?

Bryce Huebner

Published online: 14 April 2009

Abstract It would be a mistake to deny commonsense intuitions a role in developing a theory of consciousness. However, philosophers have traditionally failed to probe commonsense in a way that allows these commonsense intuitions to make a robust contribution to a theory of consciousness. In this paper, I report the results of two experiments on purportedly phenomenal states and I argue that many disputes over the philosophical notion of 'phenomenal consciousness' are misguided—they fail to capture the interesting connection between commonsense ascriptions of pain and emotion. With this data in hand, I argue that our capacity to distinguish between 'mere things' and 'subjects of moral concern' rests, to a significant extent, on the sorts of mental states that we take a system to have.

Keywords Folk-psychology · Consciousness · Cyborgs · Phenomenal consciousness

Philosophical debates about the nature of consciousness have traditionally centered on competing *a priori* intuitions about which properties are necessary for a system to count as a locus of consciousness. But, who would have thought otherwise? If armchair reflection is going to be useful anywhere, it had better be useful in helping us to understand *what it is like to sit in an armchair*! Recently, however, some philosophers and cognitive scientists have attempted to shift these debates away from individual reflection, focusing instead on commonsense intuitions about consciousness.[1] In this paper, I offer a further contribution to this alternative approach by examining the ascription of mental states to humans, cyborgs, and

[1]The experimental philosophy of consciousness is currently blossoming (for exampels, see Arico 2007; Arico et al. submitted; Gray et al. 2007; Haslam et al. 2008; Huebner 2009; Knobe and Prinz 2008; Robbins and Jack 2006; Sytsma and Machery, submitted, 2009)

B. Huebner (✉)
Center For Cognitive Studies, Tufts University, 111 Miner Hall, Medford, MA 02155, USA
e-mail: huebner@wjh.harvard.edu

B. Huebner
Cognitive Evolution Laboratory, Harvard University, 33 Kirkland St, Cambridge, MA 02138, USA

robots. I begin with a brief overview of the recent literature in the experimental philosophy of consciousness. I then report the results of two experiments designed to examine the commonsense understanding of subjective experience. I close by arguing that disputes over the philosophical notion of 'phenomenal consciousness' are misguided and that they fail to capture the important role of moral consideration in determining whether an entity is a locus of subjective experience.

Philosophical and commonsense theories of the mind

The emerging literature in experimental philosophy, beginning with Knobe and Prinz (2008), has tended to target the similarities and differences between philosophical and commonsense understandings of various mental states. From this perspective, there are three philosophical views that might be thought to garner the allegiance of commonsense:

1. A *functionalist* view of the mind, according to which mental states can be realized in any sufficiently complex information-bearing medium (cf., (Chalmers 1996; Dennett 1978a; Lewis 1980).
2. A *neuronal* view of the mind according to which "intentionally characterized capacities are realizable only as neurophysiological capacities" (Cummins 1983).[2]
3. A hybrid view according to which phenomenal consciousness must be implemented in neural wetware, but non-phenomenal states can be functionally realized.

At various points in the philosophical literature on the metaphysics of mind, each of these views has derived intuitive force from thought experiments intended to show that two entities that are functionally equivalent can differ in their capacity for subjective experience.[3]

Most of the work in experimental philosophy has targeted the third theory. Since this theory requires a distinction between the *psychological states* that explain behavior (e.g., beliefs, desires, and memories) and the *phenomenal states* that it is like something to experience (Chalmers 1996), clear experimental hypotheses seem to follow—and they seem to require nothing more than a brief questionnaire and a few null-hypothesis significance tests to show that commonsense psychology

[2] Searle (1992) argues that this view is closest to commonsense

[3] In presenting a version of this paper at the Society for Philosophy and Psychology, it was quickly made clear that many philosophers of mind object to the claim that their views rely on mere intuition. I found this surprising, and a quick look through the most important papers in the philosophy of mind reveals a pervasive appeal to intuition. Lewis (1980) motivates his analytic functionalism with the claim that he has a deep intuition that both mad-pain and Martian-pain are possible; similarly, in developing his homuncular functionalism, Lycan (1987) claims that the ascription of mental states to a tinfoil man controlled by MIT scientists is "absurd, because 'he' is a mere mock-up, largely empty inside". Searle's (1980) neuronal view is motivated by the claim that the Chinese-room could not understand Chinese. Analogously, Block (1978) claims that there is a *prima facie* doubt that there is anything that it is like to be the nation of China, so a functionalist theory of the mind that allows for such a system must be mistaken. Chalmers (1996) argues that a philosophical theory that takes consciousness seriously must accommodate the conceivability of zombies; and, Frank Jackson (1982) claims that it is "inescapable" that the color-blind Mary had incomplete knowledge before she left her room. Such appeals pervade the philosophy of mind. But this should be enough to convince skeptical readers that *many* philosophers of mind rely on intuition to motivate their positions.

includes a concept of phenomenal consciousness. But, unfortunately, some philosophical blockades stand in the way of an easy deployment of this methodology. To begin with, it has proven difficult to offer a clear account of what distinguishes phenomenal from non-phenomenal states. Ned Block (1995), one of the chief defenders of this theory, has gone so far as to suggest that no non-circular definition of phenomenal consciousness is likely to be on the horizon. Thus, in philosophy, it has become common practice to adopt an ostensive strategy in identifying phenomenal states (like obscenity, you are supposed to know phenomenal consciousness when you see it). A person is phenomenally conscious when she has a visual, auditory, or tactile experience, when she feels a pain, or when she experiences an emotion.[4] Of course, merely appealing to a priori intuitions about which states should belong to the same class has never been a good strategy (see the familiar worries about classification offered by Goodman 1955, 1972). However, recent work in experimental philosophy and cognitive science has begun to suggest that there are distinctions, perhaps implicit in the structure of commonsense judgments, that seem to map something like a distinction between phenomenal and non-phenomenal mental states.

The first systematic examination of commonsense ascriptions of various mental states to various sorts of entities was carried out by Heather Gray et al. (2007), using a web-based survey to examine a series of contrastive judgments comparing the extent to which various entities (including humans, nonhuman animals, God, and a sociable robot) were judged to have various cognitive capacities.[5] These data suggest that judgments about mental states tend to cluster into two distinct categories, which Gray and her colleagues label 'agency' (which included the capacity for self-control, morality, memory, emotion recognition, planning, communication, and thinking) and 'experience' (which included experiencing hunger, fear, pain, pleasure, rage, pride embarrassment, joy, and desire; as well as being conscious and having a personality). More interestingly, in examining the judgments about particular entities in relation to these clusters of mental states, it seems that people tend to see God as a locus of agency with little experience; a dog as a locus of experience with very little agency; ordinary human beings as having both a lot of experience and a lot of agency; and robots as having some agency but little experience.

These data seem to sit well with a view that claims that phenomenal states must be realized neurally while non-phenomenal states need not be so understood. After all, God doesn't have a body (I assume) and an ordinary robot has the wrong sort of body (or so say the philosophers who defend this view). However, dogs have biological bodies like humans even though they are not likely to be functionally complex enough to warrant the ascription of many non-phenomenal states. So, on the basis of these data, it may seem that commonsense must distinguish between experiential and agential states. However, on closer examination, these data cannot license this conclusion.

First, although these data demonstrate that commonsense psychology allows for comparative judgments about the *relative* plausibility of ascriptions of mental states,

[4] Chalmers (1996) claims that developing an adequate account of phenomenal consciousness is the most pressing problem in the philosophy of mind. However, he offers little more than a list of paradigmatic experiences that includes sensory experience, pain, emotion, and mental imagery.

[5] http://www.wjh.harvard.edu/~wegner/pdfs/Gray_et_al._(2007)_supportive_online_material.pdf

such contrastive judgments cannot discriminate between 1) a commonsense understanding of the mind that allows for robotic experience that is impoverished relative to human experience, and 2) a commonsense theory that only allows for experience in biological entities. Second, although the clustering of mental states makes for interesting psychological data, it also introduces a number of confounding variables that make it difficult to see how *this distinction* between 'agency' and 'experience' is related to the philosophical distinction between phenomenal and non-phenomenal states. Gray and her colleagues do not examine the sensory capacities that have been central to debates over the nature of phenomenal consciousness (Haslam et al. 2008; Sytsma and Machery, 2009). Moreover, the classification scheme reported by Gray et al. (2007) treats 'desire' as an experiential state while many of the philosophers who are party to these disputes often treat 'desire' as a non-phenomenal state (though there is room for dispute here). Thus, although these data suggest important differences in the strategies that people use in ascribing mental states to various entities, they do not precisely target the presence, or lack thereof, of a commonsense concept of phenomenal consciousness.

In the first experimental philosophy paper directly targeting the question of whether there is a commonsense concept of phenomenal consciousness, Joshua Knobe and Jesse Prinz (2008) hypothesized that ascriptions of phenomenal states would be sensitive to facts about the physical make up of a system in a way that ascriptions of non-phenomenal states would not. To examine this hypothesis, they asked participants to rate the acceptability of ascriptions of familiar phenomenal states (experiencing a sudden urge, experiencing great joy, vividly imagining, getting depressed, and feeling excruciating pain) and non-phenomenal states (deciding, wanting, intending, believing, and knowing) to ACME Corporation on a 7-point scale (1-sounds weird; 7-sounds natural). Knobe and Prinz found that participants judged ascriptions of non-phenomenal states to ACME to sound natural and ascriptions of phenomenal states to ACME to sound weird. Knobe and Prinz (2008) argue that because a group is just as capable of having a state with the functional role of depression, for example, as it is of having a state with the functional role of intention, these data are most plausibly explained by positing two processes of mental state ascription: one which checks functional roles, and one which is driven by a concept of phenomenal consciousness.

Unfortunately, these experiments also fail to license the strong conclusion that commonsense psychology relies on a concept of phenomenal consciousness. As Justin Sytsma and Edouard Machery (2009) convincingly argue, Knobe and Prinz failed to control for the obvious differences between individuals and corporations in functional architecture and expected behavior. To establish the claim that commonsense psychology includes a concept of phenomenal consciousness that lies beyond the scope of functional organization, it must be shown that two entities can be functionally equivalent while varying in their capacity for being in phenomenal states. Once again, while the experiment carried out by Knobe and Prinz (2008) does demonstrate that there are differences in the sorts of mental states that people are willing to ascribe to different entities, it does not demonstrate the existence of a commonsense concept of phenomenal consciousness that is not beholden to facts about the entities functional organization.

Following Sytsma and Machery (2009), it seems clear that the only way to demonstrate that there is a commonsense concept of phenomenal consciousness is to

show that ordinary people ascribe different sorts of mental states to entities that are equivalent in their behavioral and functional capacities. The key question that must be addressed in the attempt to establish the existence of a commonsense concept of consciousness is, "Do people rely on the structural properties of an entity in making judgments about the acceptability of a mental state ascription, or are they more concerned with that entity's functional organization?" In hopes of providing a clearer answer to this question, I developed a simple experiment that would test the extent to which commonsense psychology relied on structural cues in evaluating the ascription of mental states to entities that were functionally equivalent.

Robot beliefs and cyborg pains: Experiment 1

In examining the extent to which commonsense psychology relies on something like the philosophical notion of phenomenal consciousness, it is important to examine ascriptions that pick out *qualitative states* (e.g., the color of an afterimage, the smell of brewing coffee, the pitch of a sound, and the taste of a raw onion). However, experimentally targeting judgments about such *qualia* presents a difficulty. If participants are asked to make judgments about whether something *looks red* to various entities, such responses may be confounded by the fact that 'looks' is polysemous. On one interpretation of 'looks', seeing red requires only that a system has detected red things and reported having done so. However, on a more 'phenomenal' interpretation of 'looks' the entity would also have to be the subject of an *immediate experience* of red. But there seems to be no obvious way to guarantee that all participants in a psychological experiment will adopt the intended reading! Similar considerations appear to hold for questions about other sensory modalities as well. Thus, an experiment designed to examine judgments about sensation seems ill advised.[6] However, one qualitative state that seems immune to this worry is the feeling of pain—a clearly phenomenal state that allows for the addition of the 'feels' locution to disambiguate the intended ascription.

In Experiment 1, I thus targeted the acceptability of ascriptions of belief and pain to humans, robots and cyborgs. Ninety-five (95) participants were recruited from introductory philosophy classes at The University of North Carolina, Chapel Hill and were randomly assigned to one of four conditions (a human with a human brain; a Human with a CPU; a robot with a human brain; and a robot with a CPU). Each condition included a questionnaire with 1) a picture, 2) a brief description of a target entity, and 3) two sentences that ascribed mental states to the targeted entity. In the Human-Human brain condition, for example, participants saw a picture of a male human being and read the following description (for the pictures and the complete

[6] Sytsma and Machery (unpublished data) found that non-philosophers tend to say that a simple robot can 'see red' but philosophers find such ascriptions odd. However, the responses of non-philosophers are not "bimodal as would be expected if they found 'seeing red' to be ambiguous between distinctive behavioral and phenomenal readings" (Sytsma, personal communication). So, my worry may be misplaced. Whether there are important differences between ascriptions of sensory states and ascriptions of pain and happiness remains a question for further empirical investigation.

text of all scenarios, see the online supplementary materials at http://www.wjh.harvard.edu/~huebner/robosupp.pdf:

> This is a picture of David. David looks like a human. However, he has taken part in an experiment over the past year in which his brain has been replaced, neuron for neuron, with microchips that behave exactly like neurons. He now has a CPU instead of a brain. He has continued to behave in every respect like a person on all psychological tests throughout this change.

Each participant then rated her or his agreement with two statements about David (He believes that 2+2=4; He feels pain if he is injured or damaged in some way) on a 5-point scale (1-strongly disagree; 5-strongly agree). All data for this and the subsequent experiment were collected in accordance with the policies of the Institutional Review Board of UNC, Chapel Hill.

The mean responses to these two sentences are reported in the preceding table (standard deviations included in parentheses). These data yield the following significant effects.[7] First, participants' judgments regarding the acceptability of the mental state ascriptions depended, to a significant extent, on the *type of entity* to which the mental states were ascribed. Overall, participants tended to find the ascription of mental states more acceptable for humans than for a cyborg or a robot. Second, the acceptability of mental state ascriptions depended, to a significant extent, on the *type of mental state* that was being ascribed. That is, participants tended to find the ascription of belief more acceptable than the ascription of emotion. Of greater theoretical interest is the fact that these data failed to reveal any significant differences in the acceptability of belief ascriptions for the four entities (although participants did judge ascriptions of belief to be *slightly* less acceptable for robots as compared to humans). However, there were statistically significant differences between the four entities insofar as the ascription of pain was concerned. Specifically, participants tended to judge that pain ascriptions were more acceptable for a normal human than for any other entity.

The fact that participants tended to judge it acceptable to ascribe beliefs across the board, while ascriptions of pain were attenuated for any entity that was not an ordinary human, seems to offer a further confirmation of the view that is advanced by Knobe and Prinz (2008), against the criticism offered by Sytsma and Machery (2009). That is, these data seem to confirm that commonsense psychology does draw a distinction between phenomenal and non-phenomenal states—and this distinction seems to be dependent on the structural properties of an entity in a way that ascriptions

[7] A 4 (entity) x2 (state) Mixed Model ANOVA revealed that participants' judgments differed significantly as a result of the type of entity being considered, F (3, 91)=7.24, p<.001, η_p^2=.193. There was also a significant difference between the acceptability of ascribing beliefs and pains, F (1, 91)=11.03, p=.001, η_p^2=.108; but no significant interaction between entity and mental state, F (3, 91)=1.93, p=.130, η_p^2=.060. Planed comparisons using Univariate ANOVAs failed to reveal any significant difference between the ascriptions of beliefs to the various entities, F (3, 94)=2.30, p=.083; in fact, Bonferroni corrected post-hoc tests revealed only one marginally significant difference (Human-Brain vs Robot-CPU, p=.086), with all remaining comparisons failing to reach significance, p>.40. However, the analogous ANOVA revealed a significant difference between the various entities insofar as pain was concerned, F (3, 94)=8.755, p<.001; and, Bonferroni corrected post-hoc tests revealed significant differences between the Human with a Human Brain and every other entity (Human-CPU, p=.014; Robot-Brain, p=.027; Robot-CPU, p=.000), with all remaining comparisons failing to reach significance, p>.15.

of non-phenomenal states are not. However, as with the data reported by Knobe and Prinz (2008), it is unclear that these data can really lend credence to this hypothesis. The thing that it is important to note at this point is that there is an incredibly high degree of variation across participants (demonstrated by the wide standard deviations for every ascription and every entity). This, I suggest leaves open an alternative interpretations of these initial results.

To further examine the extent to these data can be seen as lending credence to the claim that commonsense psychology incorporates a concept of phenomenal consciousness, I carried out planned comparisons of the relative frequency with which participants 'agreed', 'disagreed', of 'neither agreed nor disagreed' with each of the mental state ascriptions for each of the entities.[8] In the case of the ascription of belief, these data (represented graphically in Fig. 1) reveal that the majority of participants found belief ascriptions acceptable regardless of the entity to which the belief was being ascribed: Robot-CPU=64%; Robot-Brain=65%; Human-CPU=64%; Human-Brain=73%.[9] More interestingly, these data also show 1) that no participant *disagreed* with the ascription of a belief to an ordinary human, and 2) that no participant was ambivalent in their judgments about the capacity of a robot to have beliefs.

The pattern of responses for the ascriptions of pain (represented graphically in Fig. 2) reveals a very different pattern of judgments. Here, a majority of participants agreed that the ordinary human felt pain (82%); a majority of participants judged that the ordinary robot did not feel pain (56%); but judgments about the pain ascriptions for each of the cyborg entities were essentially at chance.[10] These results sit far less comfortably with the claim that the commonsense understanding of the mind includes a concept of *phenomenal consciousness*.

Previous data have suggested that pain ascriptions were sensitive to the biological features of an entity (e.g., the presence of soft biological tissue that includes pain receptors).[11] However, appeal to such a theory fails to explain why participants' judgments concerning the ascription of pain to both of the cyborg entities (i.e., the

[8] Despite long-standing worries about null hypothesis significance testing (Cohen 1994; Dixon 2003; Gigerenzer 2002, 2004), psychologists and experimental philosophers tend to focus almost exclusively on whether their p-value reaches the critical value of $p=.05$. Moreover, it is typically just assumed that Likert scales should be treated as interval data and analyzed using the measurements of an ANOVA. But, such statistical methodologies often mask important differences in the data—suggesting to readers and experimenters alike that the lack of a significant difference implies the lack of an important difference. Although a mean response of 3 on a 5-point scale could be the result of every participant being ambivalent about their response, it could also be the result of some participants offering 5 s and others offering 1 s. However, since my concern in this paper concerns the strategies we adopt in ascribing mental states, this data calls for a further analysis that is both more intuitive and that makes the structure of the data more transparent. I, thus, examined the relative frequency of affirmative responses (either 'agree' or 'strongly agree'), as compared to negative ('disagree' or 'strongly disagree') and ambivalent responses ('neither disagree nor agree') to the questions that I presented. This collapses the data from a 5-point scale to a 3-point scale for ease of presentation. However, it preserves the relevant distinctions for answering questions such as "How frequently did participants agree with the sentence 'David believes that 2+2=4?".

[9] Robot-CPU, $\chi^2(2, N=25)=15.44$, $p<.001$; Robot-Brain, $\chi^2(2, N=23)=10.78$, $p<.01$; Human-CPU, $\chi^2(2, N=25)=10.64$, $p<.01$; Human-Brain, $\chi^2(2, N=22)=17.82$, $p<.001$.

[10] Robot-CPU, $\chi^2(2, N=25)=5.84$, $p=.059$; Robot-Brain, $\chi^2(2, N=23)=.35$, $p=.84$; Human-CPU, $\chi^2(2, N=25)=1.52$, $p=.47$; Human-Brain, $\chi^2(2, N=22)=24.36$, $p<.001$.

[11] Sytsma (personal communication) asked participants to justify their claim that a simple robot could not experience pain. His participants tended to appeal to the physical features of the entity: being "metal, mechanical, non-biological, and so on".

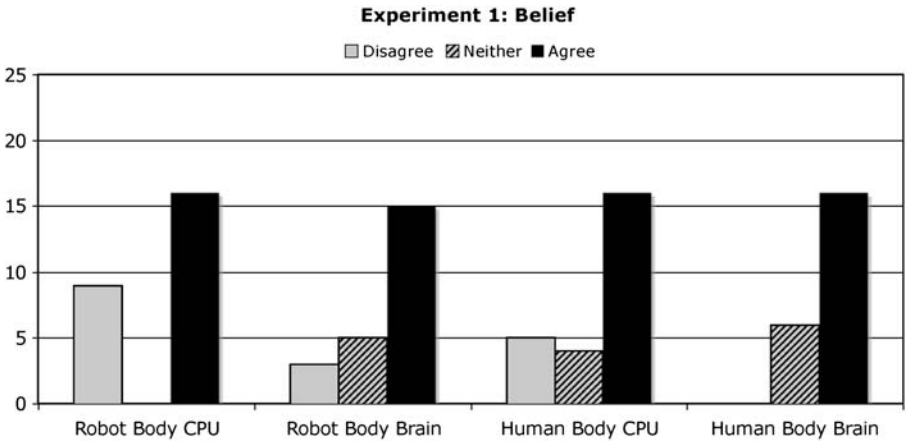

Fig. 1 Experiment 1: Belief

human with a CPU and the robot with a human brain) were nearly identical.[12] Even if people do distinguish humans from robots by appealing to the presence of biological structures, such considerations alone seem to be insufficient to yield a definitive answer concerning the acceptability of pain ascriptions to cyborgs.

What, then, could drive the difference in willingness to ascribe pain to these entities? One possibility is that while people have fairly clear intuitions about the capacity of ordinary humans and ordinary robots to feel pain, they seem unsure about what to say about entities that share some of their structural features with humans and other structural features with robots. To put this hypothesis in its strongest form, the commonsense understanding of the mind is completely non-committal for the most philosophically interesting entities that have been thought to yield important intuitions about the nature of consciousness—a possibility to which I return below.

Another possibility, suggested by Sytsma and Machery (submitted) is that affective valence plays an integral role in the folk understanding of subjective experience. They contend that there is not commonsense concept of phenomenal consciousness; instead judgments about the presence of a particular mental state in a system are likely to track the extent to which an entity is a locus of affect. Surprisingly, Murat Aydede (2001) has offered conceptual arguments for the claim that the commonsense understanding of pain typically relies on the unpleasant affective sensations typically associated with real or perceived physical damage. On this view, although having a hard metallic body would typically count against the capacity to experience pain, having a human brain may be, at least for some people, sufficient to justify the ascription of affective sensations. Perhaps the robot with a human brain is understood as a complex brain-in-a-vat, experiencing pain with no tissue to be damaged. Parallel considerations apply to the human with a CPU: while a human body is typically sufficient to warrant the ascription of pain, it may be unclear whether a CPU can generate the relevant affective response to bodily damage.

[12] Note the frequency of responses for these two cases: Human-CPU (Agree = 11; Neither agree nor disagree = 6; Disagree = 8); Robot-Brain (Agree = 9; Neither agree nor disagree = 7; Disagree = 7).

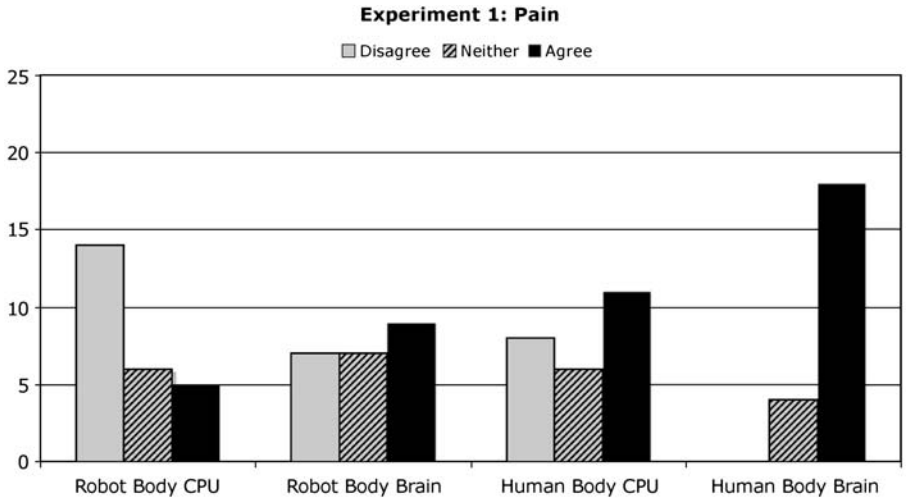

Fig. 2 Experiment 1: Pain

The emotional life of robots and cyborgs: Experiment 2

With these empirical and conceptual considerations in mind, I hypothesize that the ambivalence about cyborg pain may be driven by a corresponding ambivalence about the possibility of affective sensations in cyborgs. This raises a second empirical question, "Are people willing to ascribe affective states to cyborgs?" To examine this question, I designed a second experiment targeting the generality of belief ascriptions and the commonsense understanding of an emotional state. In this second experiment, I turned to the ascription of the experience of 'feeling happy'.

There are a number of reasons for focusing on happiness. First, targeting an emotion allowed me to further examine the commonsense understanding of affective states. Second, to develop a broader understanding of the ascription of affective states, I chose a state with positive valence as opposed to negative valence that is assumed to be important for the experience of pain. Third, Georges Rey (1980) has argued that unless emotional ascriptions are constrained by their neurophysiological realizers, we risk advocating an implausibly liberal theory of the mind (Block 1978). Fourth, Knobe and Prinz (2008) advance their claim that commonsense psychology includes a concept of phenomenal consciousness on the basis of judgments about emotional states. Finally, just as the experience of pain can be determinately picked out as a qualitative state by adding the 'feels' locution, similar effects can be achieved by adding a 'feels' locution before emotional terms.[13]

[13] I assumed that 'feels' would disambiguate phenomenal and non-phenomenal states because Knobe and Prinz (2008) found that people were willing to ascribe emotions-*sans*-'feels' to a corporation, but unwilling to ascribe the same states when the 'feels' locution was included. Sytsma and Machery (submitted), however, found no significant difference in judgments about the acceptability of the sentences 'Microsoft is feeling depressed' and 'Microsoft is depressed'. Sytsma and Machery (unpublished data) report a similar finding for 'anger'. However, as the inclusion of 'feels' was merely a safeguard, this in no way impugns my results.

One-hundred-and-nine (109) participants were recruited from introductory philosophy classes at UNC, Chapel Hill. The methodology of this experiment paralleled the methodology of Experiment 1. Participants were randomly assigned to the same four conditions, and were provided with questionnaires that included the same pictures and the same brief scenarios that were used in Experiment 1. In all four of the conditions, volunteers were asked to rate their agreement with two claims about David on a 5-point scale ranging from '1-strongly disagree' to '5-strongly agree': "He believes that triangles have three sides", and "He feels happy when he gets what he wants".

Mean responses are reported in the preceding table. Mirroring the analyses carried out in Experiment 1, participant responses to these two sentences yielded the following significant effects.[14] As in Experiment 1, participants' judgments depended to a significant degree on *the sort of entity* to which the mental states were ascribed. However, although participants found ascriptions of mental states to humans *slightly* more acceptable than the ascription of mental states to cyborgs and robots, the size of this effect was incredibly small. Second, participants' judgments were sensitive to the *sort of mental state* that was being ascribed, and again participants found belief ascriptions more acceptable than emotion ascriptions. Paralleling the key results from Experiment 1, these data failed to reveal any significant differences in the acceptability of belief ascriptions to the four entities, and participants tended to judge emotion ascriptions to be significantly more acceptable for a human with a human brain than for any other entity.

To further examine these results, I once again carried out planned comparisons to examine the relative frequency with which participants 'agreed', 'disagreed', or 'neither agreed nor disagreed' with each of these mental state ascriptions. For the ascription of belief, these data (represented graphically in Fig. 3) once again revealed that the majority of participants found belief ascriptions acceptable regardless of the entity to which the belief was being ascribed (Robot-CPU=75%; Robot-Brain=73%; Human-CPU=71%; Human-Brain=82%).[15] In line with the results obtained in Experiment 1: ordinary humans, cyborgs, and ordinary robots were all seen as plausible believers. However, things once again look very different for the ascription of happiness (represented graphically in Fig. 4). A majority of participants agreed with the ascription of emotion to an ordinary human being (76%), and no participant disagreed with the claim that an ordinary human would feel happy when he got what

[14] A 4 (entity) x 2 (state) Mixed Model ANOVA revealed that participants' judgments differed significantly as a result of the type of entity being considered, F (3, 105)=3.05, p=.032, η_p^2=.080; though the size of this effect was incredibly small. This analysis also revealed a significant difference between the acceptability of ascribing beliefs and emotions to the various entities, F (1, 105)=19.82, p<.001, η_p^2=.159; however, there was no significant interaction between entity and mental state, F (3, 105)=1.76, p=.159, η_p^2=.048. Planed comparisons using Univariate ANOVAs failed to reveal any significant difference between the ascriptions of beliefs to the various entities, F (3, 105)=.423, p=.737; in fact, Bonferroni post-hoc tests revealed no significant difference for any comparison, all p=1.00. However, the analogous ANOVA revealed a statistically significant difference between the ascription of pain to the various entities, F (3, 105)=4.51, p=.005; Bonferroni post-hoc tests revealed significant differences between the Human with a Human Brain and both the Human with a CPU (p=.041) and the Robot with a CPU (p=.004), but all other comparisons failed to reach significance p>.299.

[15] Robot-CPU, χ^2(2, N=28)=22.36, p<.0001; Robot-Brain, χ^2(2, N=26)=19, p<.0001; Human-CPU, χ^2(2, N=28)=18.29, p<.0001; Human-Brain, χ^2(2, N=27)=28.22, p<.0001.

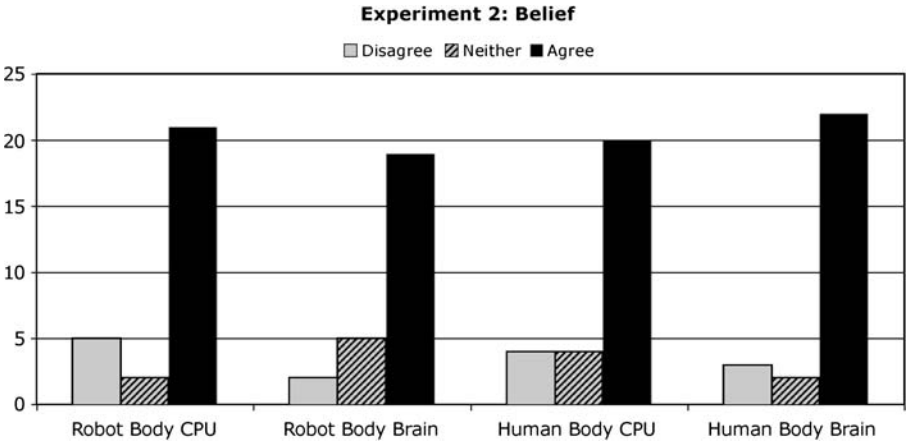

Fig. 3 Experiment 2: Belief

he wanted. However, a small, but significant number of participants found the ascription of happiness to a robot with a human brain to be acceptable (41%); and participant responses for the ordinary robot and the Human with a CPU were essentially at chance.[16]

These data further suggest that the commonsense understanding of the mind is non-committal for many of the most philosophically interesting cases of mental state ascription. People seem to have fairly clear intuitions about the capacity of ordinary humans to be in emotional states; however, their intuitions are much fuzzier when we move away from this paradigmatic case and begin to address entities that differ, in their structural features, from ordinary humans.

Three failed views?

Returning to the question of how the commonsense understanding of mental states is related to the philosophical views that have dominated recent work in the philosophy of mind, I suggest that my data are not straightforwardly predicted by any of the dominant views in the philosophy of mind. First, participants tended to treat the ascriptions of belief as plausible regardless of the sort of entity to which it was being ascribed. This suggests that people who have not been trained in philosophy or cognitive science are not likely to have *neuronal* intuitions. Of course, some participants in each of my experiments did find ascriptions of beliefs to robot unacceptable, and other participants found ascription of beliefs to cyborgs unacceptable. But, by and large, these data suggest that commonsense includes a broadly functionalist understanding of beliefs, treating them as largely insensitive to

[16] Robot-CPU, $\chi^2(2, N=28)=1.79$, $p=.409$; Robot-Brain, $\chi^2(2, N=26)=7.46$, $p=.024$; Human-CPU, $\chi^2(2, N=28)=1.36$, $p=.507$; Human-Brain, $\chi^2(2, N=29)=26.14$, $p<.001$.

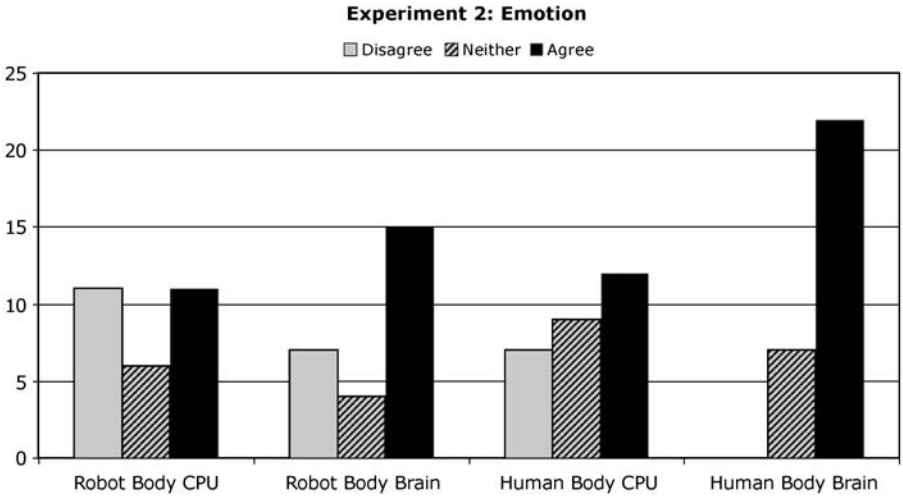

Fig. 4 Experiment 1: Emotion

differences in implementation.[17] Thus, any philosophical defense of a neuronal theory of the mind must be treated as a revisionary project grounded in empirical evidence about the realization of mental states and not taken to be a commonsense theory. *Pace* Searle (1980), appealing to the implausibility of ascribing belief to an aggregation of beer cans cannot double as an argument for the claim that there are biological constraints on cognition.[18]

Unfortunately, this is cold comfort for the functionalist. Commonsense ascriptions of happiness and pain introduce complications for anyone who defends the intuitive plausibility of functionalism. While behavioral and functional considerations may be enough to warrant the ascription of belief, they do not exhaust the perceived mental life of every sort of entity. That is, some entities also tend to be seen as possessing capacities for subjective experience that are not captured by the behavioral and functional properties of an entity. But, this should not be too surprising. Following Aydede (2001), the commonsense understanding of pain seems to rely both on considerations of damage to bodily tissue and on the unpleasant affective sensation

[17] An anonymous referee has argued that this claim outstrips the data reported in this paper. Demonstrating that commonsense psychology is functionalist regarding belief would require that ascriptions of belief were acceptable wherever an entity was computationally organized in the right way. Fortunately, a number of recent studies (Arico 2007; Arico et al. submitted; Gray et al., 2007; Haslam et al., 2008; Huebner et al. 2009; Knobe and Prinz 2008; Sytsma and Machery, submitted) demonstrate that ascriptions of belief are ascribed to a wide variety of entities including humans, non-human animals, robots, supernatural entities, and groups. On the basis of this data, I feel warranted in the stronger claim—though future data could show that there are contexts in which people do place biological constraints on cognition.

[18] This is not to claim, in an ontological tone of voice, that there are no biological constraints on cognition; these data cannot warrant such claims. I merely wish to note that reliance on thought experiments to establish this conclusion may fail to make parallels in functional organization sufficiently clear. This is clearly the case with Searle's aggregations of beer cans, network of windpipes, and Chinese rooms. My suggestion is merely that the intuitive pull of such thought experiments suggests little more than a failure of imagination (Dennett 1988), a point to which I return below.

associated with real or perceived physical damage. Aches are typically treated as being in the muscles (and analogously, hunger is in the stomach, and the shooting pain of a migraine as directly behind the eyes); however, pain can also emerge even when the relevant physical structures are lacking if the unpleasant sensation is being experienced. Analogously, commonsense psychology seems to hold that where happiness is concerned physiology is important, but the production of an affective sensation might, in some case, be enough to justify the ascription of emotion. Yet again, this provides only cold comfort for the view according to which commonsense psychology distinguishes phenomenal and non-phenomenal states on the basis of physiological features (Knobe and Prinz 2008).

Pace Knobe and Prinz, the fact that people rely more heavily on physiological cues in ascribing affective states does not show that commonsense psychology relies on a concept of phenomenal consciousness. A view according to which the physiological features of an entity are necessary for ascribing affective states fails to explain why there is little difference between ascriptions of pain to the two different types of cyborgs. Moreover, although a statistically significant proportion of participants judge a robot with a human brain to be likely to experience happiness, judgments about happiness in non-human systems tended to range widely with some participants agreeing and some disagreeing with each of the ascriptions to non-human entities. Thus, rather than supporting the existence of a commonsense concept of phenomenal consciousness that is not beholden to facts about functional organization, the results that I have presented seem to be more consistent with the emerging consensus that existing philosophical accounts of 'phenomenal consciousness' fail to do justice to important features of the commonsense understanding of minds.

Sytsma and Machery (submitted) present data suggesting that although philosophers do not typically ascribe perceptual states (like seeing red) to simple robots, ordinary people who have not been trained in philosophy or cognitive science do. They also found that participants were divided in their willingness to ascribe the capacity to smell a banana to a robot (approximately half of their participants agreed, approximately half disagreed); and, most participants were willing to ascribe the capacity to smell isoamyl acetate to a simple robot. Returning to the results reported by Knobe and Prinz (2008), participants judged that it was acceptable to say that ACME Corporation could regret its recent decision, but not that ACME could feel upset; however, Arico (2007) found that the difference is attenuated by the inclusion of contextual information (e.g., Microsoft feels sad when it looses customers).[19] Moreover, although Huebner et al. (2009), have shown that the difference reported by Knobe and Prinz can be recovered even where the contextual information is included, they have also shown that the effect here is

[19] Adam Arico and his colleagues (unpublished data) asked participants to categorize clearly figurative sentences (e.g., "Einstein was an egghead") and clearly literal sentences (e.g., "Carpenters build houses"), as well as sentences attributing different mental states to individuals (e.g., "Some millionaires want tax cuts") and groups (e.g., "Some corporations want tax cuts") These sentences were rated on a 7-point scale (1 = 'Figuratively True' 7 = 'Literally True'), and judgments regarding different types of mental states were compared. Arico et. al found that participants tended to treat the ascription of non-phenomenal mental states to collectivities as 'literally true'.

culturally variable. Huebner and his colleagues presented participants at a large American university with the following two sentences:

"Pepsi is feeling upset about the recent decline in profits."
"Apple is happy about the increase in software sales."

Replicating the results from Knobe and Prinz (2008), participants tended to judge that the Pepsi sentence (is feeling upset) sounded less natural than the Apple sentence (is happy). However, when these sentences were presented to people in Shanghai, using stand ard back-translation techniques to present these sentences in Mandarin Chinese, participants tended to judge that the Apple sentence sounded far less acceptable than the Pepsi sentence. This suggests that even the considerations that underwrite the ascription of affective states may be sensitive to cultural pressures. Huebner and his colleagues suggest that the ascriptions of subjective experience are not driven by an implicit theory of what minds are, but are instead likely to vary as a result of perceived *entitativity*.

Building on the numerous complications in the existing data, the data that I have reported in this paper, license a further methodological conclusion, and raise a theoretical question that calls for further analysis. First, the methodological conclusion (primarily for philosophers—experimental philosophers included): In interpreting disagreements about subjective experience, experimental philosophers and cognitive scientists must be careful not to make the simple mistake of deriving illegitimate ontological conclusions from epistemological premises. Ontological conclusions about the plausibility of materialism have often been derived from the purported possibility of functional zombies lacking in experience despite being functionally equivalent to an ordinary person. True to form, Dennett (1988) has countered by arguing that such claims show nothing more than a failure to imagine a system that is functionally equivalent to an ordinary person. The key point to notice here is that even if it seems that functional zombies are possible, this does not—by itself—license ontological claims of any sort. To show that something follows from this purportedly widely shared intuition, philosophical arguments would have to be marshaled to show that functional zombies actually are possible (Dennett 1988). My claim is that a parallel consideration is important in interpreting the results of studies like the one that I have reported: even if a psychological experiment demonstrated conclusively that functional zombies were inconceivable, nothing of ontological import would follow.

While examinations of ordinary people's judgments about thought experiments is not without precedent, this methodology still raises eyebrows in many philosophical circles—and it does so for at least two good reasons. First, metaphysical questions about consciousness will be answered, if they are answered at all, with a philosophically sophisticated account of the dynamic neural processes that implement our mental states. There is little reason to suppose that insights into these processes are likely to be provided as responses to thought experiments. So, commonsense psychology has little to offer in support of this project! Second, there is no reason to suppose that people who have not studied philosophy will have privileged access to the meaning of our concepts (Kauppinen 2007). So, even if there were *a priori* insights to be gained by reflecting on consciousness, there is no reason to suppose that responses to surveys would be revealing in this matter either. To put

the point briefly, the commonsense intuitions that are being catalogued by experimental philosophers and cognitive scientists interesting in commonsense ascriptions of mental states are unlikely to provide insight into highly technical questions about the nature of the mind.

Why do intuitions diverge?

What, then, are we to make of the commonsense ambivalence about ascriptions of pain, happiness, and sensations. People who have not been trained in academic philosophy or cognitive science do not seem to have clear commitments about the features that justify ascriptions of pain or happiness to entities that are neither paradigmatically human nor paradigmatically machines. One way to read this data is by noting that 'philosophy is hard', and claiming that only those who have thought long and hard about phenomenal consciousness will have clear intuitions about what it is. But, this response is troubling. While 'intuition pumps' can lead us to better ways of theorizing, they are equally likely to solidify philosophical prejudice, reaffirming the theoretical commitments that were foisted upon us in graduate school, and further entrenching the dominant trends in our discipline.

To take one key example from contemporary philosophy of mind, consider the purported implausibility of thinking that the system constituted by John Searle and his Chinese room is incapable of understanding Chinese. Searle won many converts to the neuronal view of subjective experience with this thought experiment. Moreover, this thought experiment spawned the homunculi-head objections to functionalism typified by Block's Nation of China thought experiment. However, when some philosophers and cognitive scientists were presented with this thought experiment (e.g., Dan Dennett, Marvin Minsky, Seymour Papert, and Herbert Simon), they tended to think that it was perfectly plausible to treat a system like this—on the assumption that such a system could be constructed—as a clear locus of understanding. The fact that these people were deep in the midst of research projects in artificial intelligence and computational modeling no doubt played an important role in the production of their judgments! However, it is not obvious whether the intuitions that result from taking artificial intelligence seriously ought to be treated as features or bugs in the operation of the theories advanced by such philosophers and theoretically minded cognitive scientists.

I suggest that to appreciate the value of the experimental data on commonsense ascriptions of mental states, we need a shift in philosophical perspectives. These experiments cannot tell us much about the unobservable properties that confer mentality on an entity. However, they can help us to grapple with the strategies that are typically adopted in *treating an entity as a locus of mentality.* But this is just to recognize that the experimental philosophy of consciousness is nothing more nor less than a branch of heterophenomenology (Dennett 1978a, 1987, 1992), the interpretive method of cataloguing overt speech acts, systematizing them as far as possible, and generating an account of how things hang together from the commonsense perspective or perspectives. I suggest that we should withhold ontological commitment about the states posited in these experiments, while taking the responses of participants to be sincere expressions of their beliefs about how the

world seems. However, doing careful heterophenomenology also means that we must be "prepared to discover unresolved and undismissible disagreements" (Dennett 1978a) 182).[20] The key claim that I develop below is that the differences between participants offer an insight into the interpretative strategies that are being used by participants in these experiments to make judgments about the capacities of various entities to be in different mental states.

Recall that it takes very little to be treated as a locus of belief. As philosophers have often noted, ascriptions of belief cast a broad net, allowing us to make a variety of behaviors comprehensible by assuming that a social target is a rational agent. Numerous psychological studies have shown that the appearance of goal-directed behavior (Heider and Simmel 1944), contingent interaction (Johnson et al. 1998; Shimizu and Johnson 2004), or the apparent adoption of efficient means to achieving a goal (Gergely and Csibra 2003) are all sufficient to yield the judgment that something is a rational agent.[21] Of course, not every agent will always succeeds in adopting efficient means to its goals, and we do not fault entities for occasional lapses of irrationality. However, this heuristic strategy offers a useful strategy for navigating most of our social environments. As this is a point that is often discussed, I will not dwell on it here. Instead, I will focus on the fact that ascriptions of subjective experience seem to require more.

Sytsma and Machery (Sytsma and Machery submitted) have argued that commonsense psychology relies on affective valence—rather than a concept of phenomenal consciousness—to determining whether a mental state should be ascribed to a robotic entity. On their view, robots don't have the capacity for affective experience, so they don't dislike bad smells, feel pain, or fear the future. But, even if this is the case, we should still wonder why we adopt the standards that we do for ascribing affective states to various entities. Thinking through this issue begins to push us toward at an important fact about the operation of commonsense psychology.

In ascribing the capacity for pain or happiness to an entity, we thereby recognize that it matters *to this entity* how things turn out for her.[22] As Philip Robbins and Anthony Jack (2006) put this point, we adopt a 'phenomenal stance' in ascribing

[20] I leave the defense of experimental philosophy as heterophenomenology for another paper.

[21] However, as Griffin and Baron-Cohen (2002) note: "While the vast majority of six-year-olds cannot take the intentional stance on the Republican party or the Roman Catholic church, they do pretty well with people, so we can expect the core mechanisms to be in place, with success on these other entities dependant largely on experience." There is much to say about the development of this capacity for ascribing beliefs; however, in this paper I can only note that the mature understanding of beliefs is highly promiscuous.

[22] In collaboration with Jesse Prinz (unpublished data), I examined the relationship between judgments about subjective experience and moral concern. Participants read a brief scenario where a scientist turned-off an android and erased his program without asking permission and were asked about the moral status of the action (completely permissible—1; morally forbidden—7). In one condition, the android was described as having the capacity to feel pain; in the other he was described as having the capacity for emotional experiences of various sorts. In both conditions, the android was described as having the appearances and behaviors of an ordinary human. In the pain condition, the android was described as feeling pain in response to various sorts of physical damage; in the emotion condition, the android was described as having hopes, fears, and other emotional feelings. Surprisingly, there was no statistically significant difference in participant's judgments about the permissibility of shutting down either android (Emotion M=3.21, SD=1.96; Pain: M=3.10, SD=1.72), t (87)=.30, p=.7654, d=.06, r^2=.001—just one-tenth of the between groups variance can be explained by the difference between pain and emotion. Perhaps this offers further confirmation of the claim that the affective component of pain and emotion are important to determining whether an entity should be treated as a subject of moral concern.

experiential states that allows us to see an entity *as a subject of moral concern*. By adopting the phenomenal stance, we both see an entity as having concerns and adopt a sympathetic appreciation of what it is like to feel what that system is feeling (Robbins and Jack 2006). But when we feel sympathetic concern for an entity, we are perceiving the entity as an entity that could care how things go for it, and this yields an inclination to treat that entity as a subject of moral concern. This much may seem unsurprising. After all, you would seem to show no moral failing if you dismembered the computer with whom you were playing chess or deactivated the robot with whom you had been having intriguing philosophical conversations (Dennett 1978b). However, you would clearly show a moral failing if you were to disassemble your housemate or permanently deactivate your neighbor. At least in part, the reason why these latter actions seem to be obviously wrong is that neighbors and housemates are concerned about how things go for them; robots, it may seem are not so concerned.

While you might disagree with my particular way of framing the cases, my guess is that you will think that there are cases where it is obviously morally objectionable to intentionally frustrate an entity's concerns without its permission. However, I would also guess that there are other cases where it seems that reasonable people could disagree about the moral status of an entity. Perhaps, then when it seems to a person that it is impermissible to shut down a robot, this is because the robot is seen as having concern for how things go *for her*. In this case, strategies of dehumanization and anthropomorphizing strategies are likely to play an integral role not only in our judgments about whether an entity has moral status, but also whether it can be a locus of various mental states.

In line with this hypothesis, Gray et al. (2007) report that people's judgments about the amount of pain that they would feel if they had to harm an entity were highly correlated with the capacity of that entity to have subjective experiences.[23] More intriguingly, using a narrative method, Anthony Jack and Philip Robbins (reported in Robbins 2008) asked participants to report on the degree of moral concern that they felt for lobsters after reading a brief story about lobsters starving to death in traps. These participants were then asked to recalculate the degree of moral concern that they felt for lobsters after being told that lobsters were either intelligent but insentient, or sentient but not intelligent. Jack & Robbins found that "ratings of moral patiency by the participants significantly decreased relative to baseline in the first scenario (intelligence without sentience) and increased in the second (sentience without intelligence)" (Robbins 2008).

Again, it might seem unsurprising that judgments about whether an entity should be seen as a locus of moral concern are intimately bound up with our judgments about the capacity of that entity to be concerned with how things go for it. However, the key result in this study by Jack & Robbins is that the boundaries between persons and objects are fuzzy, and thereby susceptible to individual variation and perhaps susceptible to modulation by experience with a particular sort of entity. As 'natural psychologists' we are forced to rely on a heterogeneous collection of tools, heuristic strategies, and biased background assumptions—many of which are poorly

[23] The evidence comes from comparative judgments offered in response to the question, "If you were forced to harm one of these characters, which one would it be more painful for you to harm?" Gray and her colleagues found that these responses were highly correlated with their experience dimension (r= 0.85), but only weakly correlated with agency (r=0.26).

articulated—in attempting to make sense of the entities that pervade our social worlds. More importantly, although some of these strategies (perhaps the strategies that are adopted in ascribing beliefs) are deeply entrenched in our psychology, and are perhaps quite difficult to revise, others—particularly our intuitions about the features of an entity that underwrite the capacity for subjective experience—seem to be open to a substantial degree of individual variation as well as being open to revision. We can come to treat out-group members as animals, lacking in the capacities for higher cognition and complex emotions; and we can come to treat them as automatons, lacking in 'feeling' all together (Haslam 2006). But, we can also adopt anthropomorphizing strategies that allow us to treat our dogs, cats, and lizards as possessing human-like characteristics. So, we should not expect there to be a single uniform strategy that is adopted by all people in ascribing mental states. This claim, I argue, has important consequence for understanding the relationship between our ascriptions of mental states and the structure of our moral psychology.

In his recent book on the psychology of commonsense dualism, Paul Bloom (2005) has argued that by treating something as a merely physical object we automatically exclude it from realm of moral concern. One way in which this can occur is by mobilizing emotions like disgust, contempt, or fear against some identified population of humans. Such a strategy has often been used to dehumanize, and, thereby, to remove some humans from the category of persons. However, it is important also to note that this process of dehumanization not only removes an entity from the realm of moral considerability, it also removes them from the category of entities that are treated as having the mental states that we should expect to find in an ordinary human.

In support of this claim, a recent study by Nick Haslam et al. (2008) examined the perceived differences between ascriptions of mental states to humans, animals, robots, and superhuman entities. For a variety of mental states, participants rated the extent to which each of the various entities compared to humans in their ability to experience that state. This experiment, I contend, offers the key insight that is required for understanding the results of the studies that I have reported above. Haslam et al. (2008) found that animals were reliably treated as lacking higher emotions and higher cognition; robots were reliably seen as lacking emotion all together, and as having impoverished capacities for higher cognition; and, superhuman entities were no different from humans in their emotional capacities, but had enhanced capacities for higher cognition and perception. On the basis of these data, Haslam and his colleagues argue that the capacities for emotion and desire (core components of what they term 'human nature') differentiate robots from humans, but that higher cognition and more refined emotions (core concepts of 'human uniqueness') are what distinguish us from non-human animals.

Following Haslam (Haslam et al. 2008), I suggest that we distinguish two sorts of strategies that we adopt in evaluating the ascription mental states to various entities. The first is a strategy that is sensitive to considerations of agency; the second is sensitive to considerations of personhood.[24] When we focus on considerations of

[24] An anonymous reviewer worries that the term 'personhood' is ill suited for the work I put it to. After all, some of these mental states may be shared, at least to some degree, by non-human animals. I retain the term because of its prominence in moral philosophy, and because of the distinction between the intentional stance and the personal stance (Dennett 1978b). However, I concede that 'personhood' is less than ideal.

agency we attend to the capacity of an entity to engage in flexible, goal-directed behavior. Specifically, we assume that goal-directed behaviors are predicated on the capacity for rational decision-making, and agency yields hypotheses about the sorts of mental states that would rationally explain purposive behavior. The agency strategy, I suggest, provides a first, commonsense, approximation of what Dennett calls the intentional stance (Dennett 1987). When we rely on considerations of personhood, by contrast, we focus on the states that allow an entity to be concerned with how things go for her. The personhood strategy, I suggest, provides a first, commonsense, approximation of what Dennett (1978b) once called the personal stance. In the most familiar cases, considerations of personhood and agency tend to converge—that is, most of the agents that we are familiar with also happen to be persons. However, as the results reported by Haslam and his colleagues suggest, these capacities can become dissociated for some sorts of ascriptions. My suggestion is that when we turn to the most interesting cases from the standpoint of the philosophy of mind, the intuitions that are produced by relying on each of these different strategies for making sense of the inner mental life of another person are likely to diverge.

Do androids feel happy when they get what they want?

The relationship between the perception of an entity as capable of being in affective states and perceived personhood is likely to be incredibly complex. Moreover, because of the malleability of these judgments, the structure of this relation is unlikely to be recovered from the analysis of judgments that have been offered in response to thought experiments. With this in mind, I suggest that there are two problems that we face in drawing inferences about the commonsense understanding of the mind in focusing exclusively on the survey methods that have been used to examine the structure of the commonsense understanding of consciousness. First, it is often pointed out that the cases with which people are presented are radically unfamiliar cases. Robots with human brains, robots that are smelling bananas, and emotional corporations lie outside of the range of entities we typically encounter. If it is true that the strategies that we adopt in making sense of other minds are sensitive to our interactions with an entity, this is likely to yield wide variation in judgments as a result of the features of an entity that a person happens to find most salient. As with the data that I have reported above, this is likely to introduce ambiguity into the results (indicated by wide standard deviations) because it might be unclear which of the heuristic strategies that we tend to use in making sense of other entities apply to these cases.

Second, and relatedly, presenting these cases in the form of surveys removes participants from the social environments in which they can receive feedback from an entity. I do not want to deny that we have initial intuitions about the capacities of various entities to be in various mental states. However, when we are faced with an entity that is unfamiliar, these intuitions are typically submitted to feedback from our social environments that allow us to evaluate the normative acceptability of our ascriptions. If I judge that an entity is not capable of feeling pain, for example, but then I come to interact with that entity and it displays a lot of pain behavior, I am

likely to revise my judgment. While survey results can tell us that people have a difficult time knowing what to say about philosophically difficult cases, they may not be the sorts of experiments that will provide us with rich insights into the reasons why the world seems to us as it does.

As I argued above, if an entity has the capacity for subjective experience, it can fear its own death, hope that tomorrow will be a better day, and even be happy when it has an exciting philosophical interaction with a new colleague. The ascription of experiential states, then, immediately provides us with the information that we need to treat that system as a locus of moral concern. For most of us, whether we can empathize with a system might be specified in terms of having a human body or having a biological brain. However, perhaps the factors that are relevant to our capacity to empathize are, at least to some extent, malleable and determined by our interactions with that system.

Consider the depiction of the replicants in the 1982 film *Blade Runner*.[25] Replicants are bioengineered cyborgs that are *physically indistinguishable* from ordinary humans. They are cognitively adept and able to flexibly adapt to their environments in ways that are not fully designed into their program. However, they are affectless entities: they don't feel lonely, they aren't bothered by physical exhaustion, the don't love, and they don't fear their own deaths. This lack of affect leads ordinary people to systematically dehumanize the replicants: they are referred to as skin-jobs, treated as mere objects (with some being viewed as 'basic pleasure-models' for the use of military personnel), and there is nothing wrong with 'retiring' a replicant that behaves in problematic ways (after all, they are just robots). From the beginning of the film, it is clear that the key difference between a human and a replicant is the capacity for emotional experience.[26] However, as the film progresses, viewers gain the impression that some replicants do have the capacity for emotional experience. Having learned when they will be retired, four replicants return to earth—apparently fearing death.[27] A replicant named Rachel shows a deep sadness when she learns that all of her memories have been fabricated. This evokes a feeling of compassion in the Blade Runner, Deckard, eventually leading Deckard to fall in love with Rachel. Finally, in the culminating scene, the leader of the escaped band of replicants destroys the impression that he is a 'mere thing' by showing compassion and saving the life of Deckard. In short, viewers are offered the chance to empathize with plight of the replicants, to imagine what it is like to be one of them, and to reflect on whether it is wrong to 'retire' a replicant.[28]

[25] Thanks to Robert Briscoe for this intriguing analogy.

[26] Replicants are not easily distinguished from ordinary humans. So, the Blade Runners use a machine that measures low-level physiological responses to various scenarios to determine whether a entity is human or mere machine.

[27] This is clearest in a conversation between the Blade Runner, Rick Deckard, and a replicant, named Leon that he is trying to retire (Mulhal 1994): Leon: 'My birthday is April 10th, 2017. How long do I live?'; Deckard: 'Four years.'; Leon: 'More than you. Painful to live in fear, isn't it? Nothing is worse than having an itch you can't scratch.'

[28] The closing scene in the Final Cut version of the film (released in 2007) provides evidence that Deckard is a replicant, complete with fabricated memories and implanted daydreams about unicorns. If Deckard is a replicant, he must be retired. But, having gone through the emotional highs and lows in this film, having felt genuine compassion for Deckard, the thought that he too could be 'retired' seems beyond the pale of moral acceptability.

This discussion of Blade Runner is admittedly impressionistic, and in itself is nothing more than an intuition pump. However, emerging research on "relational artifacts" and human-robot interaction suggests that something like this is true of repeated interactions with robotic entities. Sherry Turkle, for example, has argued on the basis of both laboratory experiments and ethnographical data that people who spend some time interacting with Kismet, or even with a less sophisticated robot, rapidly come to see even these robotic entities as having feelings. Specifically, Turkle (2006) argues Kismet's behavior demands interaction, and that 'she' thereby invites children to see 'her' as a playmate, as someone with whom a mutual connection can be generated. Perhaps more interestingly, Turkle (2006) reports a wide range of individual differences both in children and in elderly people in nursing homes—suggesting that some of them are willing to ascribe emotional states to the robots with whom they interact and others see robots as capable of nothing more than artificial love. A more complete heterophenomenological analysis of how we understand the apparent mental lives of robots and cyborgs requires richer anthropological investigations into the ways in which experience and interactions with an entity can mitigate dehumanization and foster a view of a non-human entity as a locus of subjective experience.

The fact of the matter is that we do not encounter cyborgs and their functional zombie ilk in our daily lives. Thus, when we are presented with the possibility of a functional zombie, we have to rely on other sorts of heuristic strategies to make sense of the type of entity that it is. In this task, we have only two options: focus on similarities between humans and cyborgs (e.g., goal-directed behavior, animacy cues, expressions of emotions, etc), or focus on the differences. Unfortunately, if the entity is really functionally equivalent to an ordinary human, it will be sensitive to whatever normative constraints happen to operate over our mental state ascriptions. She will update her beliefs in light of new evidence; she will express trepidation at the prospect of moving to a new city; she will worry about the future; and she might even try to seduce you with a subtle wink from across the bar—if, that is, she happens to find you attractive (cf., Levy 2007). But by this point in time, the robotic 'it' has become a 'she' from the commonsense perspective, and there is little reason to suppose—from that perspective—that she is lacking in any sort of subjective experience. At this point, ex hypothesi, we could have no evidence that she was—and neither could she.

Keeping this in mind, I suggest that the assumptions about emotions implicit in Blade Runner are likely to get straight to the heart of an important aspect of commonsense psychology—though this hypothesis calls for further empirical investigation. While we may begin by thinking of only the paradigmatic person as a locus of subjective experience, the modulation of empathetic responses might be able to evoke categorical changes in the perceived status of an entity. Ridley Scott, (the director of Blade Runner) effectively transforms the replicants from *mere things* into *persons* by invoking empathetic responses of some sort; the question, however, is whether this can occur in our social world.

This suggests that questions about the value of commonsense psychology call for an approach that differs from the one typically adopted in contemporary philosophy of mind. It is typically assumed that commonsense psychology is a predictive enterprise that is grounded on the attempt to accurately represent the internal states

of other entities. As such, it is assumed that commonsense psychology is concerned with the same properties of the mind with which philosophical and scientific theories are concerned. As Jerry Fodor (1981) puts the point:

> Beliefs and desires, just like pains, thoughts, sensations, and other episodes, are taken by folk psychology to be real, intervening, internal states or events, in causal interaction, subsumed under covering laws of a causal stripe.

Whatever evidence Fodor may have about the commonsense understanding of the mind, his intuitions cannot license such a strong claim about how things appear to people in general. His claim does express a belief that lies at the heart of a great deal of recent work in philosophy and psychology on the commonsense understanding of minds. I propose, however, that we withhold judgment about how people understand the mind until we have cataloged the relevant data. This, I take it, is what is required in treating the results of experimental studies, as well as our on intuitions, as heterophenomenological data about how the world seems to us.

It is typically assumed, if only tacitly, that functional zombies are possible because we can imagine an entity that is functionally organized like us but that lacks experiential states. However, in coming to recognize that the commonsense understanding of subjective experience is intimately tied to the interpretive strategies that we use in structuring the realm or moral consideration, we should question even the most philosophically informed intuitions about the mental lives of such homunculi-headed systems. Perhaps the reason that functional zombies seem possible tells us more about our capacity for dehumanizing than it tells us about the truth of computational theories of the mind. Perhaps, that is, the philosophical debates about functional zombies reveal little more than a set of highly malleable intuitions, molded throughout the course of our philosophical and scientific educations, which help us to navigate our peculiar social environments.

Acknowledgements This paper has benefited greatly from conversations with Robert Briscoe, Jacek Brozozwski, Joshua Knobe, Jesse Prinz, Dylan Sabo, and Daniel Star. I am also grateful for helpful suggestions from Mike Bruno, Dan Dennett, Justin Junge, Bill Lycan, Eric Mandelbaum, Susanne Sreedhar, Daniel Stoljar, Justin Sytsma, Neil Van Leeuwen, and my audiences at the Boston University Colloquium in the Philosophy of Science and the Society for Philosophy and Psychology (2008).

References

Arico, A. (2007). *Should corporate consciousness be regulated? (Poster)*. Toronto, ON: Society for Philosophy and Psychology.

Arico, A., Fiala, B., Goldberg, R., & Nichols, S. (submitted). The folk psychology of consciousness.

Aydede, M. (2001). Naturalism, introspection, and direct realism about pain. *Consciousness and emotion, 2*(1), 29–73.

Block, N. (1978). Troubles with functionalism. *Minnesota Studies in the Philosophy of Science, 9*, 261–325.

Block, N. (1995). On a confusion about a function of consciousness. *Behavioral and brain sciences, 18*, 227–246.

Bloom, P. (2005). *Descartes baby*. New York: Basic Books.

Chalmers, D. (1996). *The conscious mind*. Oxford: Oxford University Press.

Cohen, J. (1994). The earth is round ($p<.05$). *American psychologist, 49*, 997–1003.

Cummins, R. (1983). *The nature of psychological explanation*. Cambridge, Ma.: MIT.

Dennett, D. (1978a). *Brainstorms*. Cambridge, MA: Bradford Books.

Dennett, D. (1978b). *Mechanism and responsibility, Brainstorms*. Cambridge, MA: MIT.
Dennett, D. (1987). *The intentional stance*. Cambridge, MA: Bradford Books.
Dennett, D. (1988). *The unimagined preposterousness of zombies, Brainchildren*, pp. 171–179. Cambridge, MA: MIT.
Dennett, D. (1992). *Consciousness explained*. New York: Penguin.
Dixon, P. (2003). The p-value fallacy and how to avoid it. *Canadian Journal of Experimental psychology, 57*, 189–202.
Fodor, J. (1981). *RePresentations*. Cambridge, MA: MIT.
Gergely, G., & Csibra, G. (2003). Teleological reasoning in infancy: the naive theory of rational action. *Trends in cognitive science, 7*, 287–292.
Gigerenzer, G. (2002). *Adaptive thinking*. Oxford: Oxford University Press.
Gigerenzer, G. (2004). Mindless statistics. *Journal of socio-economics, 33*, 587–606.
Goodman, N. (1955). *Fact, fiction, and forecast*. Cambridge, MA: Harvard University Press.
Goodman, N. (1972). *Problems and projects*. New York: Bobbs-Merrill.
Gray, H., Gray, K., & Wegner, D. (2007). Dimensions of mind perception. *Science, 619*, 315.
Griffin, R., & Baron-Cohen, S. (2002). The intentional stance: Developmental and neurocognitive perspectives. In A. Brook & D. Ross (Eds.), *Daniel Dennett*, pp. 83–116. Cambridge: Cambridge University Press.
Haslam, N. (2006). Dehumanization: an integrative review. *Personality and Social Psychology, 10*(3), 252–264.
Haslam, N., Kashima, Y., Loughnan, S., Shi, J., & Suitner, C. (2008). Subhuman, inhuman, and superhuman: contrasting humans and nonhumans in three cultures. *Social Cognition, 26*, 248–258.
Heider, F., & Simmel, M. (1944). An experimental study of apparent behavior. *American Journal of Psychology, 57*, 243–259.
Huebner, B. (2009). Commonsense concepts of phenomenal concepts: Does anyone care about functional zombies? *European reivew of philosophy.*
Huebner, B., Bruno, M., & Sarkissian, H. (2009). What does the nation of China think of phenomenal states? *European reivew of philosophy, XX*(YY), ZZ.
Jackson, F. (1982). Epiphenomenal qualia. *Philosophical quarterly, 32*, 127–136.
Johnson, S., Slaughter, V., & Carey, S. (1998). Whose gaze would infants follow? The elicitation of gaze following in 12-month-olds. *Developmental Science, 1*, 233–238.
Kauppinen, A. (2007). The rise and fall of experimental philosophy. *Philosophical explorations, 10*, 119–122.
Knobe, J., & Prinz, J. (2008). Intuitions about consciousness: experimental studies. *Phenomenology and the cognitive sciences, 7*, 67–85.
Levy, D. (2007). Love and sex with robots. New York: Harper Collins.
Lewis, D. (1980). Mad pain, martian pain. In N. Block (Ed.), *Readings in the philosophy of psychology*, Vol. 1, pp. 216–222. Cambridge, MA: Harvard University Press.
Lycan, W. (1987). *Consciousness*. Cambridge, MA: Bradford Books.
Mulhal, S. (1994). Picturing the human (body and soul). *Film and philosophy, 1*, 87–100.
Rey, G. (1980). Functionalism and the emotions. In A. Rorty (Ed.), *Explaining emotion*, pp. 163–195. Berkeley: University of California Press.
Robbins, P. (2008). Consciousness and the social mind. *Cognitive Systems Research, 9*, 15–23.
Robbins, P., & Jack, A. (2006). The phenomenal stance. *Philosophical studies, 127*, 59–85.
Searle, J. (1980). Minds, brains, and programs. *Behavioral and brain sciences, 3*(3), 417–457.
Searle, J. (1992). *The rediscovery of the mind*. Cambridge, MA: MIT.
Shimizu, Y., & Johnson, S. (2004). Infants' attribution of a goal to a morphologically novel agent. *Developmental Science, 7*, 425–430.
Sytsma, J., & Machery, E. (submitted). Two conceptions of subjective experience.
Sytsma, J., & Machery, E. (2009). How to study folk intuitions about phenomenal consciousness. *Philosophical psychology.*
Turkle, S. (2006). Relational artifacts with children and elders: The complexities of cybercompanionship. *Connection Science*, *18*, 347–361.
Turkle, S., Breazeal, C., Dasté, O., & Scassellati, B. (2006a). First encounters with Kismet and Cog: Children respond to relational artifacts. In P. Messaris & L. Humphreys (Eds.), *Digital media: Transformations in human communication*. New York: Peter Lang.
Turkle, S., Taggart, W., Kidd, C., & Dasté, O. (2006b). Relational artifacts with children and elders: the complexities of cybercompanionship. *Connection Science*, *18*.