

## Folk attributions of mental states to humans, robots and cyborgs:

Bryce Huebner  
Dept of Philosophy, UNC-Chapel Hill

Having seen and thought about the results from the Knobe and Prinz study on mental state ascriptions to corporations, I wondered whether their results demonstrated the presence of a commonsense concept of phenomenal consciousness that resembled the philosophical concept of phenomenological consciousness. As Justin Systema and Eduard Machery (MS, 3) have recently argued, in order to demonstrate the existence of a commonsense concept of phenomenal consciousness, “one needs to establish that ordinary people are disposed to ascribe different mental states to agents that are described as behaviorally and functionally equivalent.”

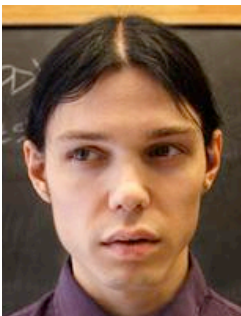
There are three prominent philosophical positions concerning the relationship between functional organization and the capacity of a system to be in various mental states.

- 1) Some philosophers argue that all mental states must be implemented biologically. This theory yields a hypothesis according to which commonsense psychology will ascribe mental states only to systems with biological brains. Following John Searle (1992), I refer to this hypothesis as *biological naturalism*.
- 2) Some philosophers defend functionalism for all mental states. This theory yields a hypothesis according to which commonsense psychology will see differences in physical structure as irrelevant to the presence of mental states, so long as psychological functioning and functional organization remain constant. Following David Chalmers (1995), I refer to this hypothesis as *organization invariance*.
- 3) Some philosophers argue that although most psychological states can be exhaustively explained in terms of functional organization, phenomenal states are possible only in a system with a biological brain. This theory yields a hypothesis according to which commonsense psychology will draw a distinction between phenomenal and non-phenomenal states; if this hypothesis is true, then differences in physical materials will be irrelevant for the ascription of non-phenomenal states but relevant to ascriptions of phenomenal consciousness. I refer to this hypothesis (for lack of a better term) as *phenomenal distinctness*.

With these positions in mind, I developed a pair of simple experiments to test these three hypotheses.

### Experiment 1:

Ninety-five volunteers from introductory philosophy classes at the University of North Carolina–Chapel Hill participated in my first experiment. Volunteers were randomly assigned to four conditions. The first two conditions were human body conditions and included a picture of a person and one of these two brief scenarios:



1. This is a picture of David. David looks like a human and he has a normal human brain. He behaves in every respect like a person on all psychological tests.
2. 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.

The second two conditions were robot body conditions and included a picture of Kismet, the robot created by the Humanoid Robotics Group at MIT, and one of these two brief scenarios:

3. This is a picture of David. David looks like a robot. Instead of a human brain he is controlled by a CPU modeled on a human brain with microchips that behave exactly like neurons. He behaves in every respect like a person on all psychological tests.
4. This is a picture of David. David looks like a robot. However, he has taken part in an experiment over the past year in which his CPU has been replaced, microchip for microchip, with neurons. He now has a normal human brain. He has continued to behave in every respect like a person on all psychological tests throughout this change.



Following the short scenario, volunteers were asked whether they agreed or disagreed with each of two statements about David:

1. He believes that  $2+2=4$ .
2. He feels pain if he is injured or damaged in some way

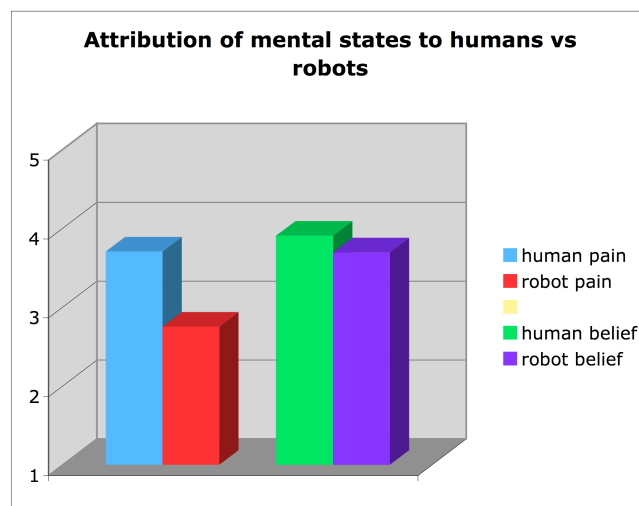
Volunteers gave their responses on a 5-point Likert scale ranging from 'strongly disagree' to 'agree':

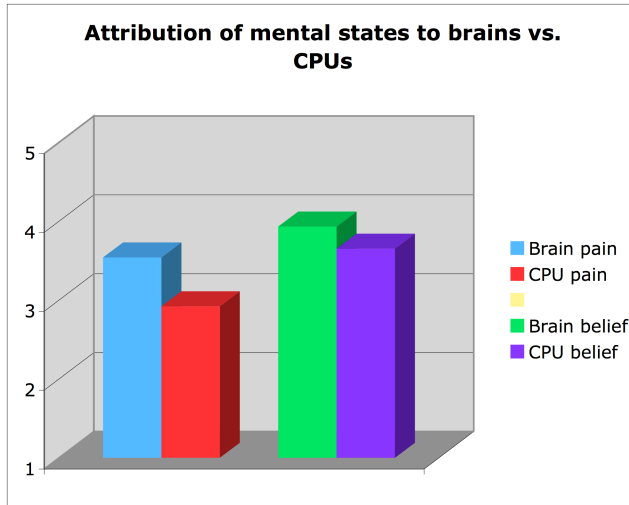
### Results:

A difference between volunteers' ascriptions of intentional states and ascriptions of the feeling of pain was immediately apparent. In order to analyze this difference, two  $2 \times 2$  ANOVAs were conducted: the first analyzed volunteers' ascriptions of mental states to systems with human bodies compared to robot bodies; the second analyzed volunteers' ascriptions of mental states to systems with human brains compared to systems with CPUs.

	Human-Human Brain	Human-CPU	Robot-Human Brain	Robot-CPU
Belief	4.08	3.72	3.78	3.60
Pain	4.16	3.24	2.91	2.59

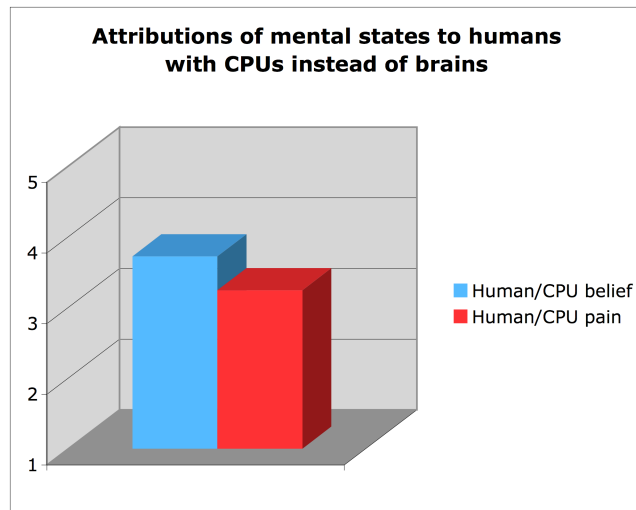
Volunteers exhibited no significant difference in their ascriptions of belief to systems with a human body as opposed to a robot body,  $F(1, 91) = .55$ ,  $p = .46$ ; nor did they exhibit a significant difference in their ascriptions of belief to a system with a human brain as opposed to a CPU,  $F(1, 91) = .92$ ,  $p = .34$ . However, volunteers *did* exhibit significant differences in both their ascriptions of the feeling of pain to a system with a human body as opposed to a robot body,  $F(1, 91) = 11.32$ ,  $p = .001$ , as well as a system with a human brain as opposed to a CPUs  $F(1, 91) = 4.86$ ,  $p = .03$ .





This suggests that neither the body nor the ‘brain’ of a system are relevant to the ascription of belief, but both are relevant to the ascription of the feeling of pain. These results provide initial evidence for a commonsense distinction between intentional and phenomenal states. Although people are willing to ascribe intentional states on the basis of psychological function alone, people seem to be committed to a particular realization of conscious states that does not appear to be organizationally invariant. This lends credence to the hypothesis of *phenomenal distinctness*.

However, there is an important test case for this hypothesis as it is typically articulated in philosophy: the system with a human body and a CPU. If the version of *phenomenal distinctness* typically defended by philosophers is correct, then this system should be seen as a locus of belief, but not a locus of pain. In analyzing this case, a paired-sample T-test was used to examine ascriptions of belief and ascriptions of pain in the system with a human body and a CPU. However, in this case, there was *no significant difference* in volunteers’ willingness to ascribe belief as opposed to the feeling of pain  $t(24) = 1.69, p = .103$ . In fact, there is a significant correlation between volunteers’ willingness to ascribe belief and to the feeling of pain to this system,  $r = .54, p = .006$ .



This suggests that although volunteers were more willing to attribute cognitive states like beliefs to systems with brains rather than CPUs, most of the heavy lifting in volunteers’ judgments about what sorts of mental states different systems are likely to be capable of being in is done by the sort of body that that system has.

## Experiment 2:

Of course, the results of a single experiment tell us little about the commonsense understanding of consciousness *in general*. However, just as pain can be picked out as a qualitative state by adding the ‘feels’ locution, similar effects can be achieved by adding a ‘feels’ locution before an emotion term. This assures that the phenomenal character of the state is made transparent. Moreover, many philosophers think that there are biological constraints on the occupants of the functional roles required for experiencing an emotion. With this consideration in mind, I developed a second experiment designed to examine the commonsense ascription of emotion.

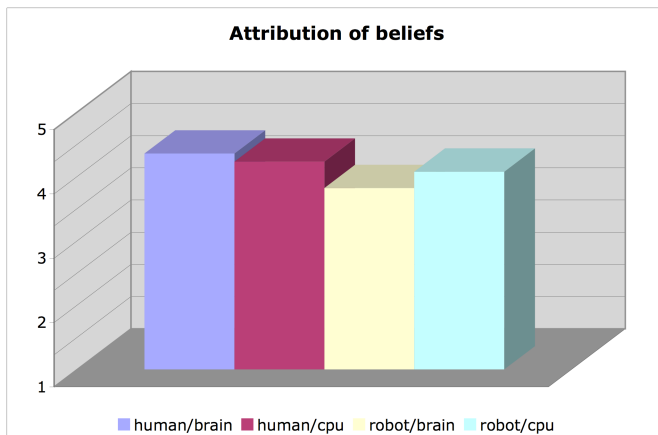
Ninety-nine (99) volunteers from introductory philosophy classes at the University of North Carolina participated in this second experiment. Volunteers were randomly assigned to the same

four conditions used in experiment 1. The questionnaires in this study included the same picture and brief scenarios used in experiment 1. In all four of the conditions, volunteers were asked to rate their agreement with two bew claims about David on a 5-point scale ranging from ‘1-strongly disagree’ to ‘5-strongly agree’:

3. He believes that triangles have three sides.
4. He feels happy when he gets what he wants

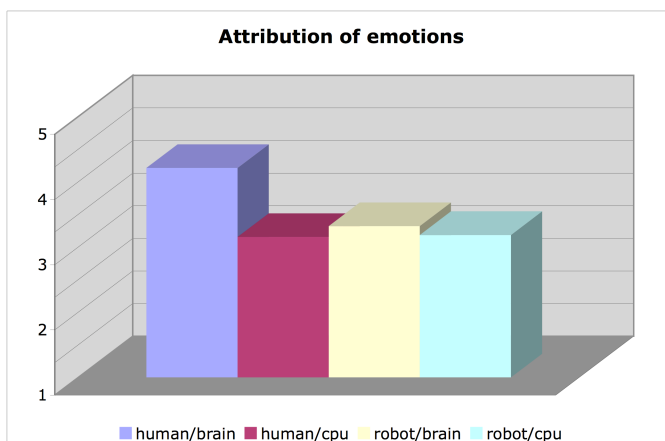
## Results

Mirroring the analyses in experiment one, two 2 X 2 ANOVAs were again conducted: the first analyzed volunteers’ ascriptions of mental states to systems with human bodies compared to robot bodies; the second analyzed volunteers’ ascriptions of mental states to systems with human brains compared to systems with CPUs. Volunteers once again exhibited no significant difference in their ascriptions of belief to a system with a human body as opposed to a robot body  $F(1, 105) = 2.33, p = .13$ ; nor did they show a significant difference in their ascriptions of belief to systems with a human brain as opposed to a CPU,  $F(1, 105) = 70, p = .405$ . Moreover, although volunteers did exhibit some difference in their ascriptions of the feeling of happiness to both a system with human body as opposed to robot body  $F(1, 105) = 3.19, p = .077$ , as well as to a system with a human brain as opposed to a CPU,  $F(1, 105) = 3.67, p = .058$ , these differences trended toward, but did not reach, significance.



However, although there was no significant main effect for either having a human brain or a human body, there was a significant interaction effect. Volunteers exhibited a significant difference between their ascriptions of the feeling of happiness to a system with a human body *and* a human brain as opposed to every other system,  $F(1, 105) = 6.16, p = .015$ ; however, there was no similar difference in ascriptions of belief  $F(1, 105) = .08, p = .781$ .

	Human-Human Brain	Human-CPU	Robot-Human Brain	Robot-CPU
Belief	4.36	4.23	3.82	4.07
Emotion	4.21	3.15	3.32	3.19



These results provide evidence for the claim that physical constitution plays a significant role in determining whether a system can be a locus of emotional experience. However, as before, the test case for *phenomenal distinctness* is the system with a human body and a CPU. In order to analyze this case, a paired-sample T-test was once again

used to examine ascriptions of belief and ascriptions of the feeling of happiness in the system with a human body and a CPU. In contrast to pain, however, there *was* a significant difference in volunteers' ascriptions of belief ( $M=4.21$ ) as opposed to the feeling of happiness ( $M=3.15$ ),  $t(25) = 4.06$ ,  $p = .0004$ . However, there was still a significant correlation between volunteers' ascriptions of belief and the capacity to feel happiness to this system,  $r = .40$ ,  $p = .042$ .

