Intuitions about Consciousness: Experimental Studies[1]

Joshua Knobe Jesse Prinz
University of North Carolina – Chapel Hill

(Forthcoming in *Phenomenology and Cognitive Science*)

Philosophers have long been concerned with intuitions about consciousness, but this interest usually takes a peculiar form. The fundamental goal is typically not to understand the intuitions themselves, with all the psychological intricacies. Instead, what philosophers really want to understand is the true nature of consciousness, and they turn to intuitions as a way of getting indirect evidence about this other topic.

This emphasis strikes us as unfortunate. Intuitions about consciousness are fascinating phenomena, amply worthy of study in their own right. The fact that people have the intuitions they do can teach us something valuable about the way people ascribe mental states, the way they think about non-human animals, perhaps even the way they make moral judgments.

Our aim here, then, is to conduct a straightforward investigation into people's intuitions about consciousness. In pursuing this line of inquiry, we truly have no ulterior motives. It is not as though we are trying to present a theory about the true nature of consciousness and have simply chosen to argue for it in a roundabout way. Rather, we are genuinely intrigued by the intuitions themselves, and we want to get a better understanding of the psychological mechanisms that generate them. Our paper therefore draws on a number of different lines of existing research, including research in 'theory of mind' (e.g., Gopnik & Meltzoff; Scholl & Leslie 1999), research in consciousness studies (e.g., Block 1978; 1995), and research about how people determine which sorts of entities are capable of having mental states (Inagaki & Hatano 1991; Johnson 2000).

Because our aims are somewhat unusual, we will be making use of a somewhat unusual method. First we introduce hypotheses about the psychological mechanisms underlying people's intuitions; then we put these hypotheses to the test using systematic experiments.

I

We begin by setting out two initial hypotheses. These hypotheses will not be concerned directly with the actual patterns of people's intuitions. Instead, they will be concerned with certain underlying psychological processes. But when the two hypotheses are put together (and combined with a few plausible assumptions), they yield definite testable predictions.

---

[1] The second author wishes to make it known that the first author actually did the majority of the work on this paper. (However, the first author wishes to make it known that the second author is just being silly and really ought to stop denigrating his important contributions. [However, the second author wishes to make it known that the first author suffers from occasional delusions about authorship.])

For helpful comments and suggestions, we are grateful to Ned Block, Paul Bloom, Bryce Huebner, Tony Jack, Uriah Kriegel, Tania Lombrozo, Bill Lycan, Bertram Malle, Ram Neta, Shaun Nichols, Philip Pettit, David Velleman, and the anonymous author of Mixing Memory. We are especially grateful to Kriegel, whose incisive comments led to major changes in a number of aspects of our paper.

1. Our first hypothesis is that ordinary people – people who have never studied philosophy or cognitive science – actually have a concept of phenomenal consciousness. In particular, we hypothesize that people often make use of the concept of phenomenal consciousness when they are ascribing mental states. Thus, suppose that a person is wondering whether or not to make the ascription:

(1) Sasha is now experiencing great pain.

The person will recognize that mental state ascriptions like (1) require phenomenal consciousness. Hence, she will ascribe the state only if she believes that the agent under discussion (in this case, a person named 'Sasha') is capable of having phenomenally conscious states.

To a first glance, this first hypothesis may seem a bit absurd. After all, it is clear that most people would not understand the *words* 'phenomenal consciousness,' and when one tries to explain the concept in a classroom, students often have trouble understanding what it amounts to. It would certainly be foolish, then, for us to suggest that people ordinarily have explicit beliefs about whether particular mental state types do or do not require phenomenal consciousness. But that is not at all what we have in mind. What we mean to suggest is rather that people grasp this concept at a purely tacit level. In other words, the suggestion is that people are actually applying the concept all the time; it's just that they normally have no awareness of doing so.

To a get a sense for what we mean here, it might be helpful to consider the way research typically proceeds in linguistics. Linguists quite often suggest that people categorize words or phrases using a particular concept even when they have no awareness at all of doing so. For a simple example, consider the following four phrases:

(2) a. professorial gentleman
b. regal style
c. financial planner
d. criminal investigation

A linguist might say that we are actually categorizing the adjectives in these phrases using various complex concepts – e.g., that we are classifying the adjectives in the first two phrases (2a-b) as 'qualitative adjectives' and those in the latter two phrases (2c-d) as 'relational adjectives.' Of course, we are not normally aware of making any such distinction, but one can see that some such thing must be going on when one tries to change the phrases around as follows:

(3) a. The gentleman was professorial.
b. The style was regal.
* c. The planner was financial.
* d. The investigation was criminal.

What we see here is that people find it acceptable to use the adjectives *professorial* and *regal* in contexts where they don't find it acceptable to use *financial* or *mayoral*. Evidence like this can offer us insight into the ways in which people ordinarily classify these words.

Our aim here is to use a similar strategy to study ascriptions of mental states. Thus, consider the following four mental state ascriptions:

(4) a. Sasha is vividly imagining a purple square.
b. Sasha is experiencing intense joy.
c. Sasha is wondering what to do.
d. Sasha is considering his options.

Our hypothesis is that people have a concept of phenomenal consciousness and that they use this concept to distinguish between different types of mental state ascriptions. Specifically, we would suggest that people are classifying the first two ascriptions (4a-b) as ascriptions that require phenomenal consciousness and that they are classifying the latter two (4c-d) as ascriptions that do not require phenomenal consciousness. We propose to provide evidence for that claim by showing that it is possible to change around all four ascriptions in the same way, such that people will find the revised versions of (4c-d) acceptable, but they will regard the revised versions of (4a-b) as completely incorrect.

2. Before we can propose our second hypothesis, we need to introduce a few technical terms:

- The *functional role* of a state is the profile of its typical causes and effects. If we wanted to characterize the functional role of anger, we might mention that people often get angry when they perceive themselves to be victims of a slight, that angry people often seek some form of revenge, and so on.

- The *physical constitution* of an entity is its actual physical make-up. If we wanted to characterize the physical constitution of a human being, we might say that human beings have four limbs and a head, that this head contains a brain, that the brain contains neurons, and so forth.

The key point for present purposes is that, even when two entities are extremely different in their physical constitutions, there can sometimes be a certain sort of isomorphism between their states. Specifically, it can sometimes be possible to map the states of the first entity onto those of the second in such a way that all of the causal generalizations that apply to types of states in the first entity also end up applying to types of states in the second. In such cases, we will say that the states of the two entities have *similar functional roles*.[2]

Now, suppose we encountered an entity whose physical constitution was very different from our own but whose states were extremely similar to our mental states in their functional roles. It would be possible to predict and explain this entity's behaviors using all of the same causal generalizations we normally apply to ourselves. We could ascribe to the entity states of belief, intention, anger, etc. and then use these ascriptions to generate predictions that would be just as accurate as those we would make of an ordinary human being. Still, it seems that an important question remains. We might know that we could predict the entity's behavior extremely well by *saying* that it was angry… but there remains an open question as to whether or not the entity truly would *be* angry.

This question has been discussed at great length in recent work in the philosophy of mind. In trying to answer it, philosophers have drawn on arguments from

[2] We thank Uriah Kriegel for his extremely helpful comments on the formulation of these distinctions.

metaphysics, philosophy of science, logic and, above all, people's intuitions about particular cases.[3] But a funny thing has happened as research in this tradition proceeds. Although philosophers first became interested in people's intuitions because they wanted to solve a more general problem in the philosophy of mind, it has gradually become clear that people's intuitions show surprisingly intricate patterns and are worthy of study in and of themselves. It is this sort of study that we take up here.

We come then to our second hypothesis. This hypothesis is that information about physical constitution plays different roles in different kinds of mental state ascriptions. Specifically, information about physical constitution plays a special role in those ascriptions that require phenomenal consciousness – a role that it does not play in other kinds of mental state ascription.

Before going any further, we need to get clear about what this hypothesis does and does not entail. It certainly does not say anything very specific either about the process underlying ascriptions of states requiring phenomenal consciousness or about ascriptions of other sorts of states. All it says is that there is a particular type of *difference* between these different kinds of ascriptions. Still, the claim it makes is a fairly surprising one. One might have thought that there was some general truth about how people ascribe mental states and that, whatever people turned out to be doing with information about physical constitution, they would at least do the very same thing for all kinds of mental state ascriptions. Our second hypothesis denies this. It asserts that ascriptions of states that require phenomenal consciousness are governed by certain special criteria that do not apply to any other mental state ascriptions.

3. Putting these two hypotheses together, we arrive at a new prediction. From the first hypothesis we learn that people classify mental states like those ascribed in (4a) as requiring phenomenal consciousness.

(4) a. Sasha is vividly imagining a purple square.

and that they regard mental states like those ascribed in (4c) as not requiring phenomenal consciousness.

(4) c. Sasha is wondering what to do.

But from the second hypothesis, we learn that the process underlying ascriptions of states that require phenomenal consciousness makes use of information about physical constitution in a way that other mental ascriptions do not. We now arrive at a somewhat surprising prediction. Suppose that we switch around both of these sentences by getting rid of the word 'Sasha' in both and replacing it with a description of a very different sort of entity. This entity would be capable of having states with functional roles that resembled those of human mental states, but it would be radically different from a human being from a physical perspective. If we chose just the right sort of entity, we should find that people regard ascriptions to it of states that do not require phenomenal consciousness

[3] Here we want to single out for special praise the work of Block (1978; 1995). His pioneering research in consciousness studies has deeply influenced the experimental studies presented below. In fact, the very term 'phenomenal consciousness' is borrowed from Block (1995), where he argues explicitly that people grasp the distinction between phenomenal and non-phenomenal states. Of course, Block's primary aim in those papers is somewhat different from our own, in that he is trying to use facts about people's intuitions as part of an inquiry into the true nature of consciousness.

as perfectly acceptable but that they should regard ascriptions to it of states that do require phenomenal consciousness as completely wrong.

II

In thinking about these issues, philosophers often resort to bizarre science-fictional entities like giant computers made of strung-together soda cans or robots controlled by troops of miniature girl scouts. Our focus here will be on examples of a more pedestrian variety. We will be concerned with entities like corporations, clubs and nations. In other words, we will be concerned with the sorts of entities usually referred to as *group agents***.**

From the standpoint of physical constitution, group agents are radically different from individual human beings. In individual humans, decision-making is realized by neurons, synapses and firing rates. In a group agent, decision-making might be realized by committees, memos and emails. Clearly, the decision making of group agents can be realized by physical objects that have no parallel in individual humans.

And yet, we do often ascribe mental states to group agents. It seems quite natural to say that Microsoft 'intends' to release a new product or that it 'believes' that Netscape is one of its main competitors. Presumably, our willingness to ascribe these mental states stems not from a similarity in physical constitution but from a similarity in functional roles.

In saying all this about group agents, we are simply echoing the view that has become standard among researchers in the field. But someone might object to that view. He or she might say:

> When people say that a corporation 'intends' to do something, they aren't really ascribing a mental to a group as such. What they mean is that certain *members* of the group have the mental state in question. Sometimes they have a funny way of expressing themselves, but that is just some kind of metaphor or shorthand.

Most researchers who have thought seriously about these questions would reject objections like this one. They think that the expressions under discussion here are not just shorthand and that people really are ascribing mental states to groups. (For discussion, see Bloom & Veres 1999; Gilbert 1992; Huebner in prep.; Kashima et al. 2005; Morris et al. 2001; Pettit 2003; Searle 1995; Solan 2005; Tollefsen 2002; Tuomela 1995; Velleman 1997.)

Before we go on to describe our own experiments, it might be helpful to review some of the arguments that have convinced researchers of this view.

> *Huebner's argument*: Consider the way we ordinarily go about ascribing mental states to a person. We might think that the person's behavior is ultimately determined by certain patterns of neural activity in his or her brain, but we do not typically try to explain the person's behavior in terms of the states of individual neurons. Instead, we use more abstract psychological generalizations. These generalizations can be considered 'robust' in the sense that they would continue to hold even if the properties of the individual neurons had been somewhat different.
>
> Now consider the way we ordinarily ascribe mental states to groups. We might believe that the behavior of the group as a whole ultimately depends on the activities of the individual members, but we often explain group behavior in a

way that does not make explicit reference to any specific members. Instead, we rely on more abstract generalizations about the nature of group behavior. These generalizations can also be considered 'robust,' since they would continue to hold even if the properties of the individual members changed in various ways.

In short, the relationship between ascriptions of states to a group and ascriptions of states to individual members is more or less the same as the relationship between ascriptions of states to a person and ascriptions of states to his or her individual neurons. (Huebner in prep.)

Velleman's argument: A philosophy department can intend to hire a new professor. But it seems that (a) one can't intend to perform a behavior if one doesn't think oneself capable of performing that behavior and (b) no individual member of the department believes him or herself capable of hiring a new professor. Therefore, the only entity that can have the intention is the department itself. (Velleman 1997)

Pettit's argument: If we ascribe beliefs to groups by simply summing up the beliefs of the members, the sets of beliefs we end up ascribing will sometimes turn out to be wildly inconsistent. For example, suppose that we ascribe a belief to the group whenever that belief is held by the majority of the members. It can then turn out that the group believes each of two propositions to be true (because each of these propositions has won the assent of the majority of members) but that the group also believes the conjunction of those propositions to be false (because the majority of the members believe at least one of the propositions to be false). If we want to ascribe coherent sets of beliefs, we will need to find some way of ascribing them that does not just amount to tallying up the beliefs of the members. (Pettit 2003)

The Arizona Experimental Philosophy Lab's argument: The Arizona EPL ran a study in which they explicitly asked subjects whether ascriptions of mental states to corporations were literal or figurative. Subjects were given a series of sentences and asked to rate them on a scale from 1 ('figurative') to 7 ('literal'). The key sentence was: 'Some corporations want lower taxes.' Subjects gave this sentence a rating of 6.2 – a truly resounding vote against the view that such sentences are purely figurative. (Arico, et al. 2006)

In what follows, we build on the work of these earlier researchers. We assume that people truly are ascribing mental states to groups and then use these ascriptions as a way of getting a handle on the structure and function of people's concept of phenomenal consciousness.

III

It is a striking fact about group agents that we ascribe to them some types of mental states but not others. We might say that Microsoft intends something or wants something or believes something… but there are other kinds of ascriptions that we would never make to Microsoft. For example, we would never say that Microsoft was feeling depressed. This is a puzzling phenomenon, and one can learn a lot about ordinary mental state ascription by trying to understand how it arises.

First of all, it should be emphasized that a state of a corporation easily could have a functional role similar to the one that people ordinarily associate with feeling depressed. So, for example, suppose that Microsoft had a department in charge of monitoring net cash flow. When cash flow becomes too low, it sends out a warning to all other departments of the corporation. Those other departments then stop moving forward on the projects they had previously been pursuing and instead take time to reflect on any mistakes they might have been making in their overall approach. This state, or something very much like it, would show the profile of causes and effects normally associated with feelings of depression. Or, at a very minimum, it would be just as similar in function role to depression as the 'intentions' of a corporation are to those of a human individual.

Yet the fact remains that people do not normally ascribe feelings of depression to group agents. We suspect that their unwillingness to make such ascriptions has nothing to do with dissimilarities in functional roles. Instead, we propose to explain it in terms of our two hypotheses:

> (1) People tacitly classify *feeling depressed* as a state requiring phenomenal consciousness
>
> (2) Ascriptions of states requiring phenomenal consciousness are sensitive in a special way to information about physical constitution.

To test this explanation, we conducted a series of experiments.

Study 1

The first thing we need to show is that people do ascribe various mental states to group agents but that they do not ascribe states that require phenomenal consciousness. Here we do not simply want to know whether people will be willing to ascribe certain states when pressed; we also want to know whether people are naturally inclined to ascribe those states in ordinary life.

The first author's beloved wife Alina Simone came up with the perfect solution. She entered into Google a series of phrases that ascribed mental states to group agents. Some ascribed phenomenal states; some ascribed non-phenomenal states. By comparing the number of hits that each type of ascription received, we can see whether people are more inclined to ascribe certain types of states than they are to ascribe others.

Here are the phrases ascribing non-phenomenal states, along with the number of hits that each phrase received:

'Microsoft intends'	25,700
'Microsoft decides'	11,400
'Microsoft tries'	52,600
'Microsoft wants'	135,000
'Microsoft believes'	31,100
'Microsoft hopes'	56,600
'Microsoft loves'	4,100
'Microsoft hates'	970

And here are the phrases ascribing phenomenal states:

'Microsoft feels depressed'	0

'Microsoft experiences joy'	0
'Microsoft feels happy'	0
'Microsoft feels pain'	2
'Microsoft feels angry'	0
'Microsoft feels scared'	0

The difference between the number of hits received by phrases in each of these two groups was so dramatic that, even with such a small sample size, one actually obtains a statistically significant effect.[4]

But now we face a problem. We know that people use certain English expressions more frequently than others, but we do not know precisely *why* they do this. It could be that the whole effect is due to some trivial difference like the number of words contained in each expression or the frequency with which people generally ascribe different types of states. What we want to know now is whether people are refraining from ascribing certain states to group agents because they actually regard those ascriptions as unacceptable. To address this question, we ran a series of studies on human subjects.

Study 2

We began with a study in which subjects were given a list of sentences that ascribed mental states to group agents and then asked whether each of these sentences 'sounds natural' or 'sounds weird.' Some of the sentences ascribed non-phenomenal states; others ascribed phenomenal states. The two types of sentences were mixed together, and the order of presentation was randomized.

The sentences ascribing non-phenomenal states were:

- Acme Corp. believes that its profit margin will soon increase.
- Acme Corp. intends to release a new product this January.
- Acme Corp. wants to change its corporate image.
- Acme Corp. knows that it can never compete with GenCorp in the pharmaceuticals market.
- Acme Corp. has just decided to adopt a new marketing plan.

The sentences ascribing phenomenal states were:

- Acme Corp. is now experiencing great joy.
- Acme Corp. is getting depressed.
- Acme Corp. is feeling excruciating pain.
- Acme Corp. is experiencing a sudden urge to pursue internet advertising.
- Acme Corp. is now vividly imagining a purple square.

For each sentence, subjects were asked to provide a rating on a scale from 1 ('sounds weird') to 7 ('sounds natural'). The mean ratings were as follows:

[4] For detailed methodological and statistical information, see Knobe and Prinz (2006).

Non-phenomenal states:

- 6.6: Deciding
- 6.6: Wanting
- 6.3: Intending
- 6.1: Believing
- 5.2: Knowing

Phenomenal states:

- 4.7: Experiencing a sudden urge[5]
- 3.7: Experiencing great joy
- 2.7: Vividly imagining
- 2.5: Getting depressed
- 2.1: Feeling excruciating pain

As the table shows, even the most acceptable phenomenal state was still deemed less acceptable than the least acceptable non-phenomenal state. More generally, there was a statistically significant effect such that people gave lower ratings for the phenomenal states than for the non-phenomenal states.[6]

Study 3

The results reported thus far seem to indicate that people are unwilling to ascribe to group agents states that require phenomenal consciousness. We now turn to questions about precisely what sort of criteria people are applying here. What exactly is it about group agents that makes people regard them as unable to have certain kinds of states?

One possibility would be that subjects' judgments are based on *similarity to humans*. Subjects start out with the premise that human beings have phenomenal consciousness. Then, when they are wondering whether some other sort of agent has phenomenal consciousness, they simply ask whether its physical constitution is sufficiently similar to that of human beings. Since the physical constitution of a

[5] We were surprised that subjects gave such high ratings for 'experiencing a sudden urge,' and we therefore ran a quick follow-up study to get a better handle on the phenomenon. Some subjects received the sentence 'Acme Corp. is experiencing a sudden urge to pursue internet advertising.' Others received a sentence that was exactly the same except that the word 'experiencing' was replaced with 'feeling.' The mean rating for subjects who received the version with 'experiencing' was 2.9; the mean for subjects who received the version with 'feeling' was 3.9. The overall mean was 3.5. These results suggest that the original ratings may have been artificially high as a result of sheer chance.

[6] In a striking development, Michael Bruno, Bryce Huebner and Hagop Sarkissian (unpublished data) have conducted a cross-cultural study demonstrating that this effect also arises among subjects in Hong Kong. (Even more interestingly, the study showed a significant effect such that the difference between ascriptions to groups and ascriptions to individuals is smaller for Hong Kong subjects than it is for American subjects.)

corporation is extremely unlike that of a human being in numerous respects, subjects conclude that corporations do not have phenomenal consciousness.

But there is also another possibility. Perhaps subjects are not thinking at all about similarity to human beings. Perhaps they are applying a far more specific restriction on constitution (say, a restriction against agents that are composed of other agents). On this latter view, people might be willing to ascribe phenomenal states to agents that are very, very different from us – just as long as those agents do not violate the specific restriction.

To decide between these conflicting hypotheses, we ran a follow-up experiment. All subjects were given a description of an agent that is not in any sense made up of smaller agents but which nonetheless has a physical constitution radically different from our own:

> Once there was a powerful sorceress. She came upon an ordinary chair and cast a spell on it that endowed it with a mind. The chair was still just made of wood, but because of the magic spell, it could now think complex thoughts and form elaborate plans. It would make detailed requests to the people around it, and if they didn't do everything just as it wanted, it would start complaining. People used to call it the Enchanted Chair.

Note that this passage ascribes to the chair only states that do not require phenomenal consciousness. (Indeed, it only ascribes states that people would be perfectly happy to ascribe to a corporation.) The key question now is whether people will automatically conclude that the chair is also capable of having states that require phenomenal consciousness. Subjects were therefore asked the question: 'Can the Enchanted Chair *feel happy or sad*?'

In addition, all subjects were also given a brief description of the Acme Corporation. They were then asked a question designed to see whether they would ascribe phenomenal states to that corporation, namely: 'Can Acme Corp. *feel happy or sad*?'

Both answers were given on a scale from 1 to 7. Subjects once again refused to ascribe phenomenal states to the corporation (average rating: 1.8), but they were happy to ascribe phenomenal states to the chair (average rating: 5.6). This difference was statistically significant.[7]

The moral here is clear. From the standpoint of physical constitution, a wooden chair is extremely different from a human being. Yet people were perfectly willing to ascribe phenomenal states to the chair. It therefore appears that people do not simply refuse to ascribe phenomenal states to any agent that differs from human beings in its physical constitution. They must be making use of some more specific restriction that rules out group agents on independent grounds. In philosophical jargon, our respondents were committed to the multiple realizability of phenomenal states, but they were also willing to impose certain specific restrictions on physical constitution.

Study 4

[7] Here it is natural to wonder whether people would also be willing to ascribe phenomenal states to the corporation if it had been enchanted by a sorceress. We do not yet have any experimental data on this question, but Adam Arico and Shaun Nichols are designing a study to address the issue.

The results of these first three experiments show that people are not willing to apply certain kinds of sentences to group agents. It seems that a full explanation of this effect would consist of two basic parts. First it would provide an account of the way in which people map the actual words in the sentences onto various underlying concepts; then it would provide an account of why people are unwilling to apply these concepts to group agents.

For a simple example, consider the fact that people seem unwilling to say that a group agent can be 'feeling upset.' What we want now is a step-by-step explanation of the process that leads up to this intuition.

Here is one possible view. First people map the phrase 'feeling upset' onto the concept *upsetness*; then they determine that no group agent can satisfy the criteria associated with the concept of upsetness. This view is a plausible one, but we suspect that it is actually incorrect.

Instead, we want to propose a slightly more complex account. When people hear the phrase 'feeling upset,' they recognize that this phrase cannot correctly be applied to an agent unless that agent fulfills both the criteria associated with the concept *upsetness* and the criteria associated with the concept *phenomenal consciousness*. There is actually no obstacle to a group agent fulfilling the criteria associated with the concept of upsetness. People's reluctance to apply phrases like this one to group agents derives entirely from the criteria associated with the concept of phenomenal consciousness.

In other words, people should be perfectly willing to ascribe upsetness to a corporation. The problem is simply that they don't think corporations are capable of genuinely *feeling* anything. If they had some way of saying that a corporation was in a state of being upset without implying that the corporation actually felt anything, they should be perfectly happy to do so.

To test this hypothesis, it would be helpful to find a way of holding fixed the degree to which people ascribe upsetness and varying only the degree to which they ascribe phenomenal consciousness. Consider, in this light, the following pair of sentences:

(5) a. Acme Corp. is feeling upset.
b. Acme Corp. is upset about the court's recent ruling.

Here it seems that both sentences ascribe upsetness to a corporation. The chief difference between them is just that only the first sentence ascribes phenomenal consciousness. The second sentence seems to indicate that the corporation is in a state of upsetness without also indicating that the corporation was genuinely capable of having feelings.

Similarly, consider the pair:

(6) a. Acme Corp. is feeling regret.
b. Acme Corp. regrets its recent decision.

Here is seems that both sentences ascribe regret to a corporation, but only the first also ascribes phenomenal consciousness.

In our fourth study, we presented these two pairs of sentences to subjects and asked them to rate the sentences in each pair on a scale from 1 ('sounds weird') to 7 ('sounds natural'). The mean responses were as follows:

	With 'Feeling'	Without 'Feeling'

Upset	1.9	5.3
Regret	2.8	6.1

Note that subjects gave far higher ratings to the sentences that did not include the word 'feeling.' This difference is statistically significant.

Looking at these results, it seems clear that people are not showing an across-the-board tendency to reject ascriptions of upsetness and regret to group agents. On the contrary, it seems that people are perfectly willing to say that a group agent can be in a state of upsetness or regret. The problem is simply that it cannot *feel* upset or *feel* regret. In short, it seems that people's reluctance to say that a group agent 'feels upset' stems not from the criteria associated with their concept of upsetness but rather from the criteria associated with their concept of phenomenal consciousness.

Discussion

The experiments reported here have focused on ascriptions of mental states to group agents, but the ultimate aim has been to reach a better understanding of precisely what is going on in ordinary cases of mental state ascriptions. Thus, suppose someone asserts (in the course of an ordinary conversation) that 'George has been feeling upset.' Our hypotheses say that:

1. People will tacitly classify this ascription as one that requires phenomenal consciousness.
2. People will regard it as acceptable only if they feel that the mental state in question has the right sorts of physical constitution.

The significance of our study of group agents is just that it provides *evidence* for these hypotheses about what goes on in ordinary cases. One can figure out whether or not ordinary ascriptions of 'feeling upset' are sensitive to information about physical constitution by checking to see how people react in the unusual cases where they are asked about agents that do not have the right sorts of constitution.

IV

We now attempt to embed our findings within a richer theoretical framework. The framework to be offered does not follow directly from the experimental results presented thus far, but it might be tested in further experimental research.

We begin with the fact that people only ascribe mental states to entities that fall within a special class. Entities that fall within the class are usually called *agents*, and a great deal of research has gone into determining precisely how particular entities end up getting classified as agents in particular circumstances (Baron-Cohen 1997; Barrett 2000; Johnson 2000). Research in this tradition is not concerned with the mechanisms underlying ascriptions of specific mental states. It is concerned instead with the processes people use to determine whether a given entity is the sort of thing that has any mental states at all. In other words, it is concerned with questions about why, e.g., we ascribe mental states to hamsters but not to toasters.

When research in this tradition has turned its attention to the topic of groups, it has invariably concluded that people truly do classify some groups as agents (Bloom & Veres 1999; Kashima et al. 2005; Morris et al. 2001; O'Laughlin & Malle 2002).

Experimental results indicate that people ascribe mental states to a group of moving shapes just as automatically as they would to a single moving shape (Bloom & Veres 1999) and that Asian subjects actually regard the mental states of groups as just as salient as those of individuals (Kashima et al. 2005). Whatever features turn out to be necessary for an entity to be classified as an agent, it seems that some groups definitely do have those features.

But it also appears that the results presented above cannot easily be explained using a simple distinction between agents and non-agents. We therefore join the growing chorus of researchers calling for a more complex framework (Gray et al. forthcoming; Robbins & Jack 2006). In addition to the concept of an *agent*, it seems that we need to posit a second concept – the concept of an *experiencer*. Then we can say that people ascribe beliefs, desires, intentions, etc. to anything they have classified as an agent but that they only ascribe phenomenal states to those entities that they have specifically classified as experiencers. Since group agents are classified as agents but not as experiencers, people are willing to say that they have beliefs, desires and intentions but are not willing to say that they have phenomenal states.

A question now arises about precisely how people determine whether or not an entity counts as an experiencer. We are not at all sure how to answer that question. As people observe an organism's behavior and physical make-up, it seems that they thereby obtain information that is in some way relevant. (For example, when people spend a lot of time working with fish, it seems that they acquire information that in some way affects their intuitions as to whether fish have phenomenal consciousness.) Still, it is hard to say exactly what people are looking for in cases like these. All we can know from the present studies is that, whatever criteria people use, those criteria are not satisfied by group agents. Clearly, more research is needed.

Nonetheless, it does seem that our results provide strong evidence against one possible view. Suppose that someone said:

> All types of mental states are ascribed in more or less the same way. In all cases, people simply check to see whether there is a state with the right sort of functional role, and if there is, they apply the corresponding mental state concept.

We see little hope of reconciling this hypothesis with our data. It seems that a group is no less capable of having a state with the functional role of depression than it is of having a state with the functional role of intention, and yet people are willing to ascribe one of these states but not the other. Something more complex appears to be afoot. We propose that the data are best explained by positing a distinct process that is not just a matter of checking functional roles.

V

Thus far, we have been engaged in a detailed study of the application of certain particular concepts. We now ask whether our findings might have any implications for broader questions about the nature of folk psychology.

To begin with, we can ask why anyone would have thought that folk psychology was functionalist in the first place. Clearly, the answer is not that researchers derived this conclusion from empirical studies of mental state ascriptions. That is, it is not as though

cognitive scientists just went out and studied a lot of different kinds of mental state ascriptions, found that all of them were best understood functionally, and then concluded that folk psychology as a whole was probably functionalist. Instead, it seems that the idea that folk psychology is fundamentally functionalist was derived from a far broader view – a kind of *grand vision* of the nature of folk psychology.

This grand vision says that folk psychology should be understood, most fundamentally, as a tool for predicting and explaining behavior. Researchers who subscribe to this vision often suggest that folk psychology is in many ways similar to a scientific theory. Just as a scientist might posit unobservable entities in order to predict and explain the behavior of the observables, so too the folk psychologist posits unobservable mental states as a way of predicting and explaining human behavior. The key claim here is that we will be able to understand why people ascribe mental states in precisely the way they do if we reflect on the ways in which these ascriptions facilitate the activities of prediction and explanation.

Starting from this grand vision, it is only a short step to the view that folk psychology must be functionalist. After all, if the vision is correct, it seems that the only properties of mental states that could play a role in folk psychology are those properties that might contribute to prediction and explanation – and the only properties that could be helpful in prediction and explanation are those that have something to do with the state's causes and effects. This chain of reasoning strikes us as a powerful and compelling one.

Yet the results reported here have moved us to accept a theory that does not fit well with the functionalist view. On this theory, certain mental state ascriptions are based on the classification of particular entities as 'experiencers,' and this classification is based in turn on a complex system of non-functionalist principles. It is hard to see how a psychological mechanism like this one could be best understood as a tool for predicting and explaining behavior.

To bring out the problem here, it might be helpful to emphasize that we seem to be uncovering a mechanism that specifically blocks the ascription of certain mental states even in cases where ascriptions of those mental states would facilitate prediction and explanation. Thus, suppose we find that we can do a better job of predicting and explaining the behavior of a given entity if we sometimes ascribe to it feelings of depression. If the entity in question has the wrong type of physical make-up – e.g., if it is a group agent – a special type of psychological mechanism will kick in and block the ascription of depression to the entity. How exactly could such a mechanism be understood as a tool for facilitating prediction and explanation?

If we had overwhelming evidence for the thesis that all aspects of folk psychology were best understood in terms of prediction and explanation, the right thing to do now would be to introduce some ad hoc assumption that allowed us to reconcile the thesis with our data. But the truth is that this thesis has been under attack in a number of other domains. In fact, a growing body of experimental research now points to a rather different picture of the nature of folk psychology. This research suggests that, although some aspects of folk psychology may indeed be best understood as tools for prediction and explanation, others are best understood in terms of their role in facilitating *moral* judgment (e.g., Cushman 2006; Knobe 2005, forthcoming; Leslie, et al. 2006; McCann 2005; Mele 2003). In other words, the suggestion is that we won't be able to make sense of every last aspect of folk psychology just by thinking about the importance of

prediction and explanation. Some aspects of folk psychology will only begin to make sense when treat them as tools for facilitating judgments about what is right or wrong, praiseworthy or blameworthy.

In light of this new wave of research, we think it would be a mistake to suppose that there must be some way to understand our findings in terms of the use of folk psychology in prediction and explanation. Instead, we suggest that the best approach would be to consider *all* of the various uses to which phenomenal state ascriptions are put. Then we can ask whether the findings can actually be understood more simply or elegantly in terms of some other use.

Let us try, then, to put aside our theoretical preconceptions and take a fresh look at the phenomena. We can consider a prototypical case in which someone might wonder whether an entity is capable of having phenomenal states and then ask how this sort of question is best understood. Suppose, for example, that we are observing a fish that has been injured and is squirming about helplessly. If we had to say what mental state this fish was in, we might be tempted to say that it is in 'pain.' Yet it does seem that there is still a legitimate question as to whether or not the fish is truly capable of phenomenal consciousness. Thus, a person might well think to herself: 'I see that the fish is squirming, but is it truly capable of feeling pain? Can a fish truly *feel* anything at all?' The experiments reported above suggest that people actually engage in certain highly complex psychological processing when trying to address questions like this one. What we want to understand now is what role all of this psychological processing actually serves in their lives.

It is certainly a bit difficult to see how all of this processing could be justified in terms of its potential to facilitate future behavioral predictions. (In fact, our bet would be that this processing doesn't end up facilitating behavioral prediction at all.) Yet, it isn't at all difficult to see how the answer to this question might play a role in a person's future decisions. Regardless of whether it in any way facilitates behavioral predictions, it can certainly influence a person's subsequent *moral* judgments. The more certain we are that an entity is capable of having phenomenal states, the more certain we will be that it is important to treat that entity with moral concern.

This point comes out especially clearly in recent work by Gray, Gray and Wegner (forthcoming) and Jack, Roepstorff and Robbins (2006). These researchers conducted experiments to see what sorts of mental state ascriptions most affected people's moral judgments. The results showed that people's judgments that a given entity was worthy of moral concern were affected far more by ascriptions of phenomenal states than by ascriptions of other sorts of mental states. In other words, when we are wondering whether to treat an entity with moral concern, we are not principally concerned with questions about whether this entity is capable of complex reasoning, planning or comprehension – what we really want to know is whether or not the entity is capable of having genuine feelings.[8]

[8] As the researchers rightly emphasize, phenomenal consciousness is specifically relevant to judgments of *moral patiency* (judgments about whether it would be wrong to do certain things to a given entity) rather than to judgments of *moral agency* (judgments about whether it would be wrong for the entity itself to do certain things). Judgments of moral agency appear to depend more on ascriptions of non-phenomenal states, such as beliefs, desires and intentions.

Now, if we focus on these moral concerns, it becomes easy to see why people might find it important to determine whether or not a given agent has phenomenal consciousness. There is no need to construct some complex story about how ascriptions of phenomenal consciousness might actually be able to facilitate behavioral prediction. Instead, we can simply rely on the straightforward idea that ascriptions of phenomenal consciousness have an impact on subsequent moral judgments. Take the person who is staring at the fish and wondering whether it is genuinely capable of phenomenal consciousness. There is a clear and rather obvious sense in which the question she is asking might be relevant to her future behaviors. She needs to know whether fish can truly feel pain because she needs to know what sorts of moral obligations she has toward the fish.

VI

The experiments presented above suggest that people make use of surprisingly complex criteria when trying to decide whether a given entity counts as an experiencer. These criteria enable people to determine, e.g., that an Enchanted Chair can have phenomenal consciousness but a corporation cannot. A question arises as to why people go through all of the cognitive effort necessary to make such determinations. The answer, we suggested, is that ascriptions of phenomenal consciousness prove important in certain types of moral judgments.

In light of this suggestion, we can revisit the question as to why people perform the psychological processing necessary to determine that group agents do not have phenomenal consciousness. We noted above that this processing does not appear to offer people much help in actually predicting the behaviors of group agents. Still, it seems that this processing may serve an important function in people's lives. Specifically, it may help people to make certain kinds of *moral* judgments about actions that affect group agents.

To test this hypothesis, we conducted two additional studies.

Study 5

Subjects were randomly assigned to receive either a story about an agent who dismantles an Enchanted Chair or a story about an agent who dismantles a corporation. In each case, subjects were asked to say whether or not the agent's behavior was morally wrong.

The story about the Enchanted Chair went as follows:

> Once there was a powerful sorcerer. He took a perfectly ordinary chair and cast a spell on it that endowed it with a mind. From then on, it was called the Enchanted Chair.
>
> The sorcerer had been hoping that the Enchanted Chair would do his bidding, but instead it quickly developed goals of its own. It had its own unique interests and was always pursuing some unusual project that no one would have expected. It would strive to accomplish these projects in every way it could.
>
> In the end, the sorcerer decided to dismantle the chair. But he also made sure that each of the individual parts of the chair were kept intact and used to make good pieces of furniture after the chair was destroyed.

After receiving this story, subjects were asked: 'Was it *wrong* of the sorcerer to dismantle the chair?'

The story about the corporation then went as follows:

> Once there was a powerful businessman. He took out a big loan and created a new corporation. From then on, the corporation was called *EnChair*.
>
> The businessman had been hoping that EnChair would do his bidding, but instead it quickly developed goals of its own. It had its own unique interests and was always pursuing some unusual project that no one would have expected. It would strive to accomplish these projects in every way it could.
>
> In the end, the businessman decided to dismantle the corporation. But he also made sure that each of the individual employees got very good jobs after the corporation was disbanded.

After receiving this story, subjects were asked: 'Was it *wrong* of the businessman to dismantle the corporation?'

All answers were recorded on a scale from 1 ('not wrong at all') to 7 ('very wrong'). The mean response for the chair was 4.2; the mean response for the corporation was 2.4. This difference is statistically significant.

We argued above that ascriptions of phenomenal consciousness have a substantial impact on people's moral judgments. This general claim offers us a simple explanation of the data obtained in the present experiment. People show less moral concern for the corporation than they do for the Enchanted Chair because they are more inclined to ascribe phenomenal consciousness to the chair than they are to the corporation.

It now becomes possible to offer a hypothesis about why people go through all of the psychological processing necessary to determine that corporations do not have phenomenal consciousness. There is no need to suppose that this processing somehow enables them to do a better job of predicting corporate behavior. Instead, one can simply say that people need to know whether or not corporations have phenomenal consciousness because they need to know whether or not it would be morally wrong to perform actions that harm corporations.

Study 6

In discussing this question with other researchers, we find that many of them see us as proposing a radical and counterintuitive doctrine. We think that this reaction gets things exactly backwards. While it is true that many cognitive scientists regard all aspects of folk psychology as tools for behavioral prediction, we think that *they* are the ones upholding a radical and counterintuitive doctrine. Meanwhile, we see ourselves as simply standing up for the commonsense view.

To address this issue, we ran one final experiment. Subjects were asked to frame their own hypotheses about why people might be interested in ascribing certain kinds of mental states. Some subjects were asked why people might be interested in ascribing a capacity for memory; others were asked why people might be interested in ascribing a capacity for consciousness. We thought that subjects would offer different hypotheses for these different kinds of ascriptions. Specifically, we thought that they would naturally tend to explain ascriptions of memory in terms of the aim of prediction and explanation but that they would spontaneously explain ascriptions of consciousness in terms of a need to form moral judgments.

Subjects in the *memory condition* received the following question:

> Imagine a person who has a job working with fish. He finds himself wanting to know the answer to a particular question about them. Specifically, he wants to know whether fish are capable of *remembering* which part of a lake has the most food.
>
> Why do you think he might want to know this? Why might the question be important to him?

Subjects in the consciousness condition received a question that was almost exactly the same, except that the person was described as wondering whether fish were 'genuinely capable of feeling anything':

> Imagine a person who has a job working with fish. He finds himself wanting to know the answer to a particular question about them. Specifically, he wants to know whether fish are genuinely capable of *feeling* anything.
>
> Why do you think he might want to know this? Why might the question be important to him?

After reading each question, subjects provided a free-response answer in the space below. These answers could then be coded into categories for statistical analysis.

First, we went through each of the responses and determined whether or not it said that the man would be interested in ascribing the relevant capacity for reasons having to do with *prediction, explanation or control*. Here is an example of a response that was classified in this first category:

> So it will be easier to feed them, b/c he only has to distribute food in one place or so he'll know where to go in order to give bait, if they are capable of remembering such things.

Second, we went through each response and determined whether or not it said that the man would be interested in ascribing the relevant capacity for reasons that had to do with making *moral judgments*. Here is a response that was classified in this second category:

> He might want to know whether fish genuinely feel things because in doing his job, he does lots of things to the fish that might possibly hurt them if they can really feel things. It might be important to him to find out if he causes them pain because he might feel it is unethical or immoral to cause harm to other things. He could hold this belief for several reasons such as religion.

These two kinds of categorization were performed independently, so that any given response could be coded into one category, into both, or into neither.

Overall, responses in the memory condition fit well with the traditional 'grand vision' about the function of folk psychology. Subjects overwhelmingly responded that the man would be interested in ascribing a capacity for memory because this ascription could enable him to predict, explain or control behavior. (100% of responses referred to prediction, explanation or control; 9% referred to moral judgment.)

But responses in the consciousness condition were very different. In that condition, subjects *did not* refer to an interest in prediction, explanation or control. Instead, the overwhelming tendency was to explain these ascriptions in terms of an interest in moral judgment. (0% of responses referred to prediction, explanation or control; 100% referred to moral judgment.)

Of course, it is possible that people will turn out to be mistaken here. That is, it is possible that people *believe* that they are interested in these questions primarily for moral reasons but that they are *really* interested in these questions primarily as a way of facilitating subsequent prediction, explanation and control. Yet, although this sort of mistake is possible, we see no specific reason to believe that it is taking place here. Indeed, we see no reason at all to think that ascriptions of consciousness are best understood as tools for the prediction, explanation and control of behavior.

VII

We therefore tentatively offer a more complex account of the function of these ascriptions. Suppose, e.g., that a person concludes

(7) George is feeling upset.

The evidence provided above suggests that it would be a mistake to consider this judgment as a whole and ask what role it might serve in people's lives. Instead, we need to break it down into two parts – an ascription of upsetness and an ascription of phenomenal consciousness – and consider each of them separately. It may very well turn out that ascriptions of upsetness serve primarily to facilitate behavioral prediction, but it does not appear that this same approach can be helpfully applied to ascriptions of phenomenal consciousness. On the contrary, it seems that ascriptions of phenomenal consciousness are best understood in terms of their role in facilitating moral judgment.

References

Arico, A., Fiala, B. & Nichols, S. 2006. The Folk Psychology of Consciousness. Unpublished manuscript. University of Arizona.

Baron-Cohen, S. 1997. *Mindblindness: An Essay on Autism and Theory of Mind*. MIT Press, Boston, MA.

Barrett, J. 2000. "Exploring the Natural Foundations of Religion." *Trends in Cognitive Sciences* 4: 29-34.

Block, N. 1978. "Troubles with Functionalism." *Minnesota Studies in the Philosophy of Science, Volume IX*. C. Wade Savage ed. Minneapolis: University of Minneapolis Press, 261-325.

Block, N. 1995. "On a Confusion about the Function of Consciousness." *Behavioral and Brain Sciences* 18: 227-247.

Bloom, P. & Veres, C. 1999. "The Perceived Intentionality of Groups." *Cognition* 71: B1-B9.

Cushman, F. 2006. "Judgments of Morality, Causation and Intention: Assessing the Connections." Unpublished manuscript. Harvard University.

Gilbert, M. 1992. *On Social Facts*. Princeton: Princeton University Press.

Gopnik, A. & Meltzoff, A.N. 1997. *Words, Thoughts, and Theories*. Cambridge, MA: MIT Press.

Gray, H., Gray, K. & Wegner, D. forthcoming. Dimensions of Mind Perception. *Science*.

Huebner, B. in prep. Ph.D. dissertation. University of North Carolina – Chapel Hill.

Inagaki, K & Hatano, G. 1991. "Constrained Person Analogy in Young Children's Biological Inference." *Cognitive Development* 6: 219-231.

Jack, A. I.; Roepstorff, A; Robbins, P. 2006. "The Genuine Problem of Consciousness: Trusting the Subject" Manuscript under revision.

Johnson, S. 2000. "The Recognition of Mentalistic Agents in Infancy." Trends in Cognitive Sciences 4: 22-28.

Kashima, Y., Kashima, E., Chiu, C-Y., Farsides, T., Gelfand, M., Hong, Y-Y., Kim, U., Strack, F., Werth, L., Yuki, M., & Yzerbyt, V. 2005. "Culture, Essentialism, and Agency: Are Individuals Universally Believed to be More Real Entities than Groups?" *European Journal of Social Psychology* 35: 147-169.

Knobe, J. 2005. "Theory of Mind and Moral Cognition: Exploring the Connections." *Trends in Cognitive Sciences* 9: 357-359.

Knobe, J. forthcoming. "Folk Psychology: Science and Morals." In Hutto, D. & Ratcliffe, M. ed. *Folk Psychology Reassessed*. Kluwer/Springer Press.

Knobe, J. & Prinz, J. (2006). "Experimental Studies of Intuitions about Consciousness: Methodological and Statistical Details." <http://www.unc.edu/~knobe/ConscDetails.pdf>

Leslie, A., Knobe, J. & Cohen, A. 2006. "Acting Intentionally and the Side-Effect Effect: 'Theory of Mind' and Moral Judgment." *Psychological Science* 17: 421-427.

McCann, H. 2005. "Intentional Action and Intending: Recent Empirical Studies." *Philosophical Psychology* 18: 737-748.

Mele, A. 2003. "Intentional Action: Controversies, Data, and Core Hypotheses." *Philosophical Psychology* 16: 325-340.

Morris, M., Menon, T. & Ames D. 2001. “Culturally Conferred Conceptions of Agency: A Key to Social Perception of Persons, Groups, and Other Actors.” *Personality and Social Psychology Review* 5: 169-182.

O’Laughlin, M., &. Malle, B. F. 2002. “How People Explain Actions Performed by Groups and Individuals.” *Journal of Personality and Social Psychology* 82, 33-48.

Pettit, P. 2003. “Groups with Minds of their Own.” In F. Schmitt ed., *Socializing Metaphysics*. New York: Rowan and Littlefield: 167-193.

Robbins, P. & Jack, A. 2006. “The Phenomenal Stance.” *Philosophical Studies* 127: 59-85.

Searle, J. 1995. *The Construction of Social Reality*. Free Press, New York.

Scholl, B. J, & Leslie, A. M. 1999. Modularity, Development and ‘Theory of Mind.’ *Mind & Language* 14: 131-153.

Solan, L. 2005. "Private Language, Public Laws: The Central Role of Legislative Intent in Statutory Interpretation," *Geo. L.J.* 93: 427.

Tollefsen, D. 2002. “Organizations as True Believers,” *Journal of Social Philosophy* 33: 395-410.

Tuomela, R. 1995. *The Importance of Us: A Philosophical Study of Basic Social Notions.* Stanford: Stanford University Press.

Velleman, D. 1997. “How to Share an Intention”, *Philosophy and Phenomenological Research* 57: 29-50.