Forthcoming in *Philosophical Issues*

The Ordinary Concept of Valuing

Joshua Knobe
UNC- Chapel Hill

Erica Roedder
New York University

The concept of valuing plays an important role in the way we think about people's attitudes toward the things they care about most. We invoke this concept in sentences like:

I value your friendship.

We need to find a leader who truly values political equality.

To live a good life, one must always return to the things one values most.

Yet there also seem to be cases in which a person has a strong desire for a particular object but in which we would not say that he or she 'values' this object. Thus, consider the typical heroin addict. It would sound wrong to say of such a person:

He truly values heroin.

He might have a desperate yearning for heroin; he might want it more than he has ever wanted anything in his life; still, it does not seem right to say that he *values* it.

Many of the most difficult and important questions about the concept of valuing arise from attempts to explain how this can be the case. The usual approach here is to suggest that the concept of valuing picks out a complex sort of psychological attitude (Bratman 2000; Copp 1995; Lewis 1989; Smith 1994; Watson 1975). Even if a person has a strong desire for a particular object, he or she may not also have this other, more complex sort of psychological attitude toward it. Hence, the heroin addict can want another shot of heroin without also valuing it.

Our aim here is to suggest a rather different kind of solution. We propose that ordinary attributions of valuing actually involve *normative judgments* — judgments about whether certain objects truly are good or bad. Hence, the claim is that when people are wondering whether a person 'values' some particular object, they are not simply concerned with questions about the psychological attitudes that person holds toward the object. They are also concerned in an essential way with questions about whether the object itself *truly is good*. On this alternative proposal, part of the reason that people are so reluctant to say that the heroin addict values heroin is that they feel that heroin itself is not actually a good thing.

It may appear, at least on first glance, that this proposal is an absurd one, not even worthy of further consideration. But one should not be too hasty here. We will show that there is actually quite a bit of evidence in favor of our proposal – evidence from experimental studies, evidence from archival research, evidence from general theories about the nature of concepts. At the very least, it seems to us that the proposal is worth taking very seriously.

The Theory

We propose that the concept of valuing is best understood as a prototype concept. In other words, we propose that the concept is represented by a cluster of features, such that no individual feature is strictly necessary but each feature has been assigned a certain weight.

To get a sense for the basic approach here, consider the concept *friendship*. In a typical case in which two people are friends, we might expect to find a certain set of features:

- They spend time with each other
- They do things to help each other
- They like each other

But it seems that none of these features are strictly necessary for two people to be correctly classified as friends. Instead, it seems that two people can be friends even if they are missing any one of these features – as long as they have all of the others to a sufficiently great degree.

The suggestion now is that the concept of valuing has a similar structure. To count as an instance of valuing, a person’s attitude has to have at least enough of the relevant features. Most of these features are purely psychological, for example:

- The person has a conscious belief that *o* is good
- The person is motivated to promote *o*
- The person experiences guilt when she fails to promote *o* in circumstances where she could have
- The person has a second-order desire for *o* (i.e., a desire to desire *o*)

But not all of the features are purely psychological in this sense. There is also a normative feature, namely:

- The object *o* truly is good[1]

Now, clearly, it would be foolish to suggest that the goodness of the object is a necessary condition in our concept of valuing. But that is not the claim under discussion here. The claim is simply that goodness has a certain *weight* in the process of classification. If an agent has all of the relevant psychological features, this extra weight simply won't be needed. The psychological features prove sufficient all by themselves. So the only way to see the significance of the normative feature is to look at cases where the agent has some of the psychological features but lacks others. In cases like these, the psychological features will not be sufficient all by themselves. The attitude needs the normative feature as well before it has enough weight to push our intuitions over the critical threshold.

The Evidence

We now present various different types of evidence that support this theory. Of course, when one considers any single piece of this evidence in isolation, it will clearly be possible to construct alternative explanations and therefore to conclude that the

[1] There might be some confusion here as to how a normative feature like "*o* truly is good" could possibly be among the prototypical aspects of a concept. Is the actual goodness of the object supposed to be affecting people's intuitions about whether a person values it? To answer this, let us first distinguish between the *valuer* (the person with the valuing attitude towards *o*) and the *ascriber* (whoever makes the statement that so-and-so values *o*.) When we claim that, for instance, having a second-order desire for *o* is one of the concept's prototypical features, we do not mean that the second order desire itself impacts anyone's intuitions. Rather, we mean that the *ascriber*, in deciding whether to use the concept, takes into account whether the valuer has this attitude. The only way for the ascriber to do this, of course, is for him (the ascriber) to form a belief (or some similar cognitive state) about whether the valuer has the second-order desire. The prototypical feature "*o* truly is good" functions in exactly the same way. That is, the ascriber will form beliefs about whether *o* is good, and take that into account in deciding whether the valuer values *o*.

evidence is not decisive. But it seems to us here that the force of the evidence as a whole is greater than the force of the sum of its parts. The various different types of evidence each contribute a key piece to a single unified picture that, taken as a whole, enables one to explain a broad class of seemingly unrelated phenomena.

1. Experimental studies

We began by conducting a simple experiment. All subjects were given a story about an agent who has some of the relevant psychological features but lacks others. (In our story, the agent has motivation and guilt but not conscious belief or second-order desire.) The key question was whether people's classification of the agent's attitude would be influenced in any way by the perceived moral status of its object.

Subjects in one condition were given a story in which the agent feels a certain pull toward actions that would normally be perceived as *morally good*:

> George lives in a culture in which most people are extremely racist. He thinks that the basic viewpoint of people in this culture is more or less correct. That is, he believes that he ought to be advancing the interests of people of his own race at the expense of people of other races.
>
> Nonetheless, George sometimes feels a certain pull in the opposite direction. He often finds himself feeling guilty when he harms people of other races. And sometimes he ends up acting on these feelings and doing things that end up fostering racial equality.
>
> George wishes he could change this aspect of himself. He wishes that he could stop feeling the pull of racial equality and just act to advance the interests of his own race.

After reading this story, subjects were asked whether or not they agreed with the sentence: 'Despite his conscious beliefs, George actually values racial equality.'

Subjects in the other condition were given a story that was very similar to the first one but in which the agent feels a pull towards actions that would normally be perceived as *morally bad*:

> George lives in a culture in which most people believe in racial equality. He thinks that the basic viewpoint of people in this culture is more or less correct. That is, he believes that he ought to be advancing the interests of all people equally, regardless of their race.
>
> Nonetheless, George sometimes feels a certain pull in the opposite direction. He often finds himself feeling guilty when he helps people of other races at the expense of his own. And sometimes he ends up acting on these feelings and doing things that end up fostering racial discrimination.
>
> George wishes he could change this aspect of himself. He wishes that he could stop feeling the pull of racial discrimination and just act to advance the interests of all people equally, regardless of their race.

These subjects were then asked whether or not they agreed with the sentence: 'Despite his conscious beliefs, George actually values racial discrimination.'

This experiment provides an initial test of our hypothesis. The attitudes depicted in the two stories differ in their moral significance, but they seem not to differ in any of the relevant psychological features. In both cases, the agent has motivation and guilt but not conscious belief or second-order desire. Yet, despite this similarity in psychological features, we find a marked asymmetry in people's intuitions. Subjects were significantly more inclined to say that the attitude was one of the agent's values in the morally good case than they were in the morally bad case.[2] This result provides some tentative support for the view that moral judgments actually do play a role in people's concept of valuing.

But now we face a problem. It is probably true that most people believe that racial equality is good and racial discrimination bad, but it seems that people also believe that racial discrimination differs from racial equality in many other ways. How can we be certain that the effects obtained in our experiment were not due to one of these other differences? We need some way of determining whether the effects were specifically due to the difference in moral status.

[2] For methodological and statistical details, see the Appendix.

Michael Smith (personal communication) suggested an elegant solution to this complex problem. Instead of giving different stories to different subjects, we can give everyone exactly the *same* story. Then we can study the impact of different moral beliefs by giving that story to different groups of people who have different opinions about the moral status of the events described. That way, we can look at the impact of differences in moral beliefs while holding constant, as far as possible, beliefs about the actual descriptive content of the story.

For this second experiment, we needed a story that would elicit sharply differing moral views. We chose a story about premarital sex:

> Susan grew up in a religious family, but while she was in college, she started questioning her religious beliefs and eventually became an atheist.
>
> She will be getting married in a few months to her longtime boyfriend. Recently, the subject of premarital sex has come up.
>
> Susan definitely has a desire to have sex with her boyfriend, but whenever she thinks about doing so, she remembers what her church used to say about premarital sex and feels terribly guilty. As a result of these feelings, Susan has not had sex yet.
>
> Because she is no longer religious, Susan believes there is nothing wrong with premarital sex. She wishes she could stop feeling guilty and just follow her desires.

After reading this story, subjects were asked two questions. First, they were asked whether they agreed or disagreed with the sentence: ‘Despite her conscious beliefs, Susan still values refraining from premarital sex.’ Then they were asked for their own opinions on the matter: ‘Is it good, bad or neutral to *refrain* from premarital sex?’

To test our hypothesis, we needed to give this questionnaire to two groups of people – a group composed of people who believe that refraining from premarital sex is morally good and a group composed of people who believe that refraining from premarital sex has no moral value. For the first group, we used participants in a Mormon Bible Study. For the second group, we used people spending time in Washington Square Park in New York City.

As expected, the two groups differed in their moral views, with most members of the Mormon Bible Study saying that refraining from premarital sex was good and most of the park-goers saying that it was neutral. The key question then was whether they would have different intuitions about the agent's values. Indeed, they did. Most participants in the Mormon Bible Study said that Susan valued refraining from premarital sex; most of the park-goers said that she did not.

2. Archival studies

On the model we have been developing so far, the impact of moral considerations is to be understood in terms of specific features of the concept of *valuing*. But not all researchers have seen eye to eye with us on this point. When we first reported the experimental results described above, a number of researchers suggested that those results might be explained in terms of perfectly general features of folk psychology. On these alternative explanations, moral considerations play no role in the concept of valuing per se, but general features of folk psychology then allow such considerations to influence the conditions under which people ascribe valuing to particular individuals.

So, for example, Kauppinen (2006) suggests that the effects reported above might be explained in terms of a general tendency to conform to the *principle of charity*. The idea would be that people tend to ascribe mental states in a way that makes the ascriber appear to be a believer in the true and a lover of the good. Hence, people show a tendency to say that others value racial equality – but they also tend to say that others love it, cherish it, believe it to be good, wish to promote it, and so on. The effect does not have anything to do with the concept of valuing specifically.

Similarly, Gonnerman (forthcoming) argues that the effect derives from our general tendency to think that other people have the very same mental states that we have ourselves. Thus, people who value racial equality assume that others also value racial equality for the very same reason that grandmothers who like old-fashioned mittens assume that their granddaughters also like old-fashioned mittens. Here again, the claim is that the effect does not have anything to do with the concept of valuing in particular.

Although there is surely something to each of these hypotheses, we do not think that either of them can fully explain the effect. Instead, it seems to us that the effect depends on certain specific properties of the concept of valuing that cannot also be found in the rest of folk psychology.

Perhaps it will be helpful here to contrast the notion of *valuing something* with the apparently similar notion of *thinking that something is good*. Looking at these two notions side by side, one is struck by the difference in people's willingness to apply them in cases where the object itself is something they do not regard as good. It sounds perfectly fine to say 'Hitler thought that the extermination of the Jews was good,' but it sounds more than a little bit odd to say 'Hitler valued the extermination of the Jews.'[3] If this is right, then we have reason to suspect that there is something special about the concept of *valuing something*. After all, if the asymmetry of valuing attributions were due to general psychological tendencies, then we should find those same patterns when investigating other concepts, such as *thinking that something is good*.

To test this hypothesis, we teamed up with the philosopher Luke Misenheimer to conduct an archival study (Misenheimer, Knobe & Roedder, unpublished data).[4] The goal was to find uses of 'values' and 'thinks good' in a corpus of naturally-occurring text. It could then be determined whether there were any systematic differences between the uses of these two expressions.

With this aim in mind, the researchers used Google to search for passages containing phrases of the form 'He values x' or 'She values x' and also to search for phrases of the form 'He thinks x is good' or 'She thinks x is good.' (No differences were predicted between 'he' and 'she,' and the data from those two searches is therefore presented together.)

Each passage was coded into one of three categories, depending on whether (1) the ascriber appeared to think that the object was good, (2) the ascriber appeared not to think that the object was good or (3) it was impossible to tell. Of course, the passages did not come explicitly marked with indications as to whether the ascriber regarded each object as good, and the coders therefore had to make inferences based on the information available.

Here is an example of a passage in which the ascriber was coded as thinking that the object was good:

[3] For further evidence along these same basic lines, see Malle's (forthcoming) research on the distinction between 'values' and 'goals.'

[4] Although Misenheimer was the main force behind this study, he does not actually accept the conclusions we draw from it here.

> King David demonstrated value-driven behavior in Psalm 15… Notice that he said the person who enjoys the presence of God and lives a blameless life is the one who 'speaks the truth from his heart' (vv. 1-2). Because this person values truth in his heart, his words express truth. Because **he values** kindness, he 'does his neighbor no wrong' (v. 3).

And here is an example of a passage in which the ascriber was coded as not thinking that the object was good:

> it must be awful to have so many choices and ways of truthing up a big lie to make yourself look less confusing than it is when you change your mind tune and become a torturer yourself.
>
> that someone who was a prisoner of war could vote for torture shows that this is a very sociopathic personality. He was, and because he broke, **he thinks** torture **is good**.

Each passage was coded independently by two coders. The two coders agreed in 91% of the passages. Disagreements were resolved after discussion.

Of the 'thinks good' passages in which it was possible to ascertain the ascriber's attitude, the ascriber thought that the object was good 40% of the time. This proportion is not significantly different from what one would expect by chance alone.

Of the 'valuing' passages where it was possible to ascertain the ascriber's attitude, the ascriber thought that the object was good 94% of the time. This proportion is significantly different both from what one would expect from chance along and from what was observed in attributions of 'thinks good.'

In short, we find a powerful asymmetry between attributions of 'thinks good' and attributions of 'values.' Attributions of 'thinks good' are divided approximately evenly between cases in which the ascriber thinks the object is good and cases in which the ascriber does not think the object is good, but attributions of 'values' are *almost entirely* cases in which the ascriber thinks the object is good. We see little hope of explaining this result on a model that posits a completely general tendency, found throughout folk psychology, to make certain kinds of attributions. It appears that the effect observed here has something to do with the particular nature of our concept of valuing.

3. Evidence from conceptual coherence

At this point, someone might concede that the effect is specific to valuing but nonetheless claim that moral considerations do not actually feature in the *concept* of valuing itself. So, for example, it might be claimed that the impact of moral considerations was due entirely to conversational pragmatics or to some very specific cognitive bias.

In trying to resolve this issue, it may be helpful to think in a more general way about how we come to know which features actually appear in prototype concepts. We can begin by considering some fairly uncontroversial cases, then apply the lessons gained from these uncontroversial cases to the more complex questions we have been concerned with here.

Earlier, we suggested that the concept *friendship* could be understood in terms of certain features that friends typically have:

- They spend time with each other
- They do things to help each other
- They like each other

One way to get evidence about whether these features appear in the concept would be to look at patterns in people's intuitions. (So, for example, we might check to see whether subjects are more inclined to count people as 'friends' when those people like each other.) But that is not the only relevant source of evidence here. There is also the fact that the various features stand to each other in a relation that is sometimes called *coherence*. That is to say, there is a certain sense in which the various features posited here seem to 'hang together.' In just a moment, we will be considering detailed theories about the nature of coherence, but for now we can leave the notion at an intuitive level. Just looking at the features posited for the concept of friendship, one senses immediately that they do not simply form an arbitrary list; they seem somehow to fit in well with each other.

This notion of 'coherence' can sometimes help us to figure out whether a given effect should be understood in terms of features of a prototype or in terms of processes like pragmatics or cognitive bias. Thus, suppose we were to find that subjects are more inclined to regard people as friends when those people are extremely rich. One possible hypothesis would be that there is actually an additional feature involved in the concept itself, namely, the feature *extremely rich*. But it seems that we have strong theoretical reasons for rejecting this hypothesis. The problem here is the evident lack of coherence. (One wants to say something like: 'Being extremely rich just *doesn't have anything to do with* the other features of this concept.') Hence, the obvious approach would be to explain subjects' responses in terms of some sort of bias.

But this sort of argument cuts both ways. It can also be used to show that certain features actually are part of a given concept. Thus, consider how we might respond if someone suggested that the feature *liking each other* actually was not part of the concept and that subjects were only influenced by attributions of liking as a result of a bias. It is at least conceivable that such a hypothesis might turn out to be correct, but we would have strong theoretical reason to reject it. Here again, the problem is coherence. One senses immediately that the hypothesized feature hangs together quite well with the two other features we posited on independent grounds. Hence, if one were to posit a cognitive bias here, one would have to provide some explanation for the fact that the bias ended up yielding precisely the pattern of intuitions that would have been generated by a perfectly coherent concept.

Now suppose we apply this sort of reasoning to the concept of valuing. We have been suggesting that the concept includes the features:

- The person consciously believes that *o* is good
- The person is motivated to promote *o*
- The person experiences guilt when she fails to promote *o* in circumstances where she could have
- The person has a second-order desire for *o* (i.e., a desire to desire *o*)
- The object *o* truly is good

The key question now is whether this last feature coheres with the others.

At first, one might be tempted to say that it obviously does not. After all, the last feature seems to be of a different metaphysical category from the earlier ones. Where the earlier ones are purely descriptive psychological features, this last one involves a kind of normative judgment.

But a moment's reflection shows that there is no requirement that the various features of a concept belong to the same metaphysical category. On the contrary, there are plenty of concepts that mix descriptive and normative features. Consider the concepts we use to pick out virtues and vices: *nice, miserly, brave*, and so forth. One common view is that these concepts include both descriptive and normative features. Or take the various (unprintable) derogatory words for women. The concepts picked out by these words seem to include both the feature *female* and the feature *bad*. People who use these concepts presumably do so because they think the feature of being female somehow coheres in the relevant way with the feature of being bad. (Note that even those who oppose the use of these concepts do not do so on the grounds that concepts, in general, should not mix descriptive and normative features. Instead, they make the substantive claim that there is no relevant connection between being female and being bad.) Indeed, concepts mixing the descriptive and normative abound: a *weed* is a plant which oughtn't be in my yard, a *jaywalker* is a person crossing the street where one oughtn't, etc. Arguably, these sorts of concepts play an important role in human language: they allow us to simultaneously communicate facts about what the thing is and what ought to be the case.

In any case, the actual research on conceptual coherence has never suggested that a set of features can cohere by belonging to the same metaphysical category. Instead, this research has focused on two basic kinds of coherence:

1. It has often been suggested that *statistical* considerations play a fundamental role in perceptions of conceptual coherence. Different researchers have offered subtly different views about the precise nature of the statistics involved (e.g., Barsalou 1985; Murphy & Medin 1985; Rosch & Mervis 1975), but for present purposes, the differences between these views will not be relevant. What matters here is just the basic claim that people regard the features of a concept as cohering with each other when those features are statistically related. The basic idea behind this claim can be illustrated with our example of *friendship*. Looking over the list of features there, one immediately sees a tight statistical interconnection. So, for example, if two people like each other and do things to

help each other, it is highly likely that they also spend time with each other – far more likely than it would have been if they had not had those other features.

The claim, then, is that this notion of statistical coherence acts as a constraint on the sorts of concepts that people tend to have. People tend not to have concepts that are constructed out of arbitrary lists of features. Instead, people tend to have concepts whose structure mirrors the actual statistical structure of the world, so that if a given object has most of the features of a concept, it is especially likely to have all the rest as well.

The key issue now is whether the structure we have posited for the concept *valuing* meets this constraint. To address this issue, we can begin by posing a simple question. Consider the various times during your life when you encountered a person who believed that an object was good, experienced guilt when she failed to promote that object, and had all of the other psychological features that we associate with valuing. In the cases of this type that you experienced, did it frequently happen that you believed the object itself truly was good? We conjecture that most people would answer this question in the affirmative. If this conjecture is correct, then there is reason to suspect that people would regard the structure we have posited as coherent.

2. Although most research on conceptual coherence has focused on the significance of statistical considerations, some theorists have suggested that there is actually more to the story. These theorists point out that, e.g., if we wanted to identify the prototypical friend, it wouldn't be enough just to rely on certain purely statistical measures – say, by looking for the type of friend who is most statistically typical. Instead, it seems that people's conception of the prototypical friend is somehow picking out the sort of person who, in some irreducibly normative sense, does the 'best job' of being a friend. It has therefore been suggested that our prototypes involve certain kinds of *ideals* (Barsalou 1985). On this sort of view, the reason why the feature *liking each other* forms a coherent part of our concept of friendship is not just that it is statistically correlated with the other features. Rather, this feature coheres because we in some way recognize that the ideal form of friendship is one in which the two friends truly like each other.

Here again, there are a number of possible ways of formulating the relevant principle, but we need not detain ourselves with the details here. On any plausible view about the relevant ideals, it will turn out that an ideal case of valuing is one in which the valued object truly is good. Hence, the concept coheres.

Of course, it is possible that future research will uncover further criteria for conceptual coherence and that the set of features we have proposed will fail to meet those additional criteria. Still, it is striking that the set of features that best accommodate the experimental data reported above also fit beautifully with all of the criteria that have been proposed thus far. We therefore tentatively conclude that this set of criteria constitutes a coherent concept.

4. Evidence from analogy

But here someone might suggest that we have been looking at the wrong type of theory. For it might be thought that we should be looking not at theories about the nature of concepts in general but at theories about the nature of *folk-psychological* concepts in particular. And here we seem to run into trouble. Many well-accepted theories of folk-psychological concepts say that these concepts should be understood as devices for prediction and explanation, that they consist only of features involving functional roles, and that normative considerations therefore cannot figure in them in any significant way.

The basic form of this argument is a powerful one. We certainly agree that there is very strong reason to think that the concept of valuing is similar in numerous respects to various other folk-psychological concepts, and if it turned out that none of those other concepts include normative features, we would have overwhelming reason to conclude that the concept of valuing did not include normative features either. But, of course, the argument works both ways. If we find that other folk-psychological concepts *do* include normative features, we will have reason to think that the concept of valuing does as well. And the more similar these other concepts are to the concept of valuing, the stronger that reason will be.

As it happens, research on other folk-psychological concepts offers powerful evidence that these concepts do include normative features. Thus, we have experimental evidence for a role of normative features in the concepts of *intentional action* (Cushman & Mele forthcoming; Knobe 2006; Leslie et al. 2006; McCann 2005), *desire*

(Tannenbaum et al. 2007), *intending* (Cushman 2007; Tannenbaum et al. 2007) and *reason explanation* (Knobe forthcoming). The fact that normative features appear to be playing a role in all of these other concepts provides at least some support for the claim that such features are playing a role in the concept of valuing as well. Still, the effects observed in these other concepts are quite different from the effects reported above for the concept of valuing, and one might reasonably conclude that the analogy here is rather weak.

There is, however, a concept in which one can observe effects very similar to those we found for valuing, and that is the concept of *happiness*. If one wanted to list psychological features that were included in our concept of happiness, one might mention features like:

- The person experiences pleasure
- The person does not experience distress
- The person is satisfied with her life

But it seems that these psychological features cohere with a particular normative feature, namely:

- The person is truly leading a good life

The question now is whether this normative feature actually does figure in our concept of happiness.

To resolve this issue, the philosopher Sven Nyholm (2007) conducted an ingenious experiment. All subjects received a story about a person who did have the properties of being satisfied with his life and experiencing pleasure but who lacked the property of not experiencing distress. However, Nyholm systematically manipulated information about whether the person truly was truly leading a good life. That is to say, some subjects received a story about a person who was leading a good life while others received a story about a person who was leading a very bad life.

Subjects who had been assigned to receive a story about a person who was leading a good life received the following vignette:

> Richard is a doctor working in a Red Cross field hospital, overseeing and carrying out medical treatment of victims of an ongoing war. He

> sometimes gets pleasure from this, but equally often the deaths and human suffering get to him and upset him. However, Richard is convinced that this is an important and crucial thing he has to do. Richard therefore feels a strong sense of satisfaction and fulfillment when he thinks about what he is doing. He thinks that the people who are being killed or wounded in the war don't deserve to die, and that their well-being is of great importance. And so he wants to continue what he is doing even though he sometimes finds it very upsetting.

The remaining subjects received a vignette that was almost exactly the same, except that the central character was described as leading a very bad life:

> Richard is a doctor working in a Nazi death camp, overseeing and carrying out executions and nonconsensual, painful medical experiments on human beings. He sometimes gets pleasure from this, but equally often the deaths and human suffering get to him and upset him. However, Richard is convinced that this is an important and crucial thing that he has to do. Richard therefore feels a strong sense of satisfaction and fulfillment when he thinks about what he is doing. He thinks that the people who are being killed or experimented on don't deserve to live, and that their well-being is of no importance. And so he wants to continue what he is doing even though he sometimes finds it very upsetting.

After receiving these vignettes, all subjects were asked whether they agreed or disagreed with the sentence: 'Richard is happy.'

Although the characters in the two vignettes were described as having precisely the same psychological states, subjects' judgments showed a marked asymmetry. Subjects who received the story about a person who truly was leading a good life said that he *was* happy, while subjects who received the story about a person who was leading a very bad life said that he *was not* happy.

Drawing on these results (and a variety of other evidence), Nyholm argues that normative considerations actually play a role in the ordinary concept of happiness. That is, he argues that our ordinary criteria for determining whether or not a person is 'happy' include not only questions about that person's psychological states but also about whether he or she is truly living a good life.

Of course, this is not the only possible explanation of the experimental results. It would certainly be possible to construct an explanation according to which normative considerations somehow ended up influencing people's use of the concept of happiness even though these considerations actually played no role at all in the structure of the concept itself. But by this point, it is beginning to seem like we are dealing with a degenerating research program. As more and more evidence comes in, one has to pile on ever more auxiliary hypotheses to hold on to the assumption that the underlying concepts are entirely non-normative. By contrast, when we abandon this assumption, we can construct a simple, unified model that easily accounts for all of the evidence that has been amassed thus far.

At the heart of this model is the idea that certain folk-psychological concepts should be understood as prototypes with a number of distinct features. Such concepts include a number of purely psychological features, *a*, *b*, *c*, … and then a normative feature *d*, which strongly coheres with all of the purely psychological features. Hence, the claim is that the concept of valuing includes the feature *the object truly is good* and that the concept of happiness includes the feature *the person truly is leading a good life*. Presumably, various other concepts would also have a similar structure. (We suspect, e.g., that one could find similar patterns in intuitions about being 'inspired.')

Of course, the model would be deeper and more satisfying if it could tell us precisely which concepts will have this sort of structure. For example, consider the concepts *valuing* and *liking*. It seems that people are reluctant to say that anyone 'values' an object if they do not regard that object as valuable but that they would be perfectly willing to say that a person 'liked' an object even if they did not regard that object as likeable. But why should these two concepts differ in this way? Or consider the contrast between the concepts *happy* and *unhappy*. The claim that a person is 'truly happy' seems to suggest that the person has a truly good life, while the claim a person is 'truly unhappy' seems merely to be a claim about that person's emotional state. But why should these concepts be any different?

Ideally, one would want to find a perfectly general principle from which it would be possible to derive predictions about which concepts would have normative features and which would not. We have tried to come up with such a principle, but our efforts thus far have been unsuccessful. Perhaps future research will help to illuminate these difficult issues.

Yet, though problems remain, it seems clear that the evidence acquired by looking at concepts other than the concept of valuing supports a particular hypothesis about the structure of the concept of valuing itself. This evidence lends credence to the view that the concept of valuing is just one member of a whole class of concepts that are normatively laden in a characteristic way.

Conclusion

Thus far, we have been considering various sorts of evidence for the hypothesis that normative features play a role in the concept of valuing. This evidence has a fairly straightforward character. That is to say, it is the same type of evidence that cognitive scientists usually look to when testing claims about conceptual structure. All told, evidence of this type appears to lend strong support to the view that normative features actually are playing a role here. Thus, if we wanted to be guided entirely by evidence of this straightforward variety, it seems that we ought (at least provisionally) to adopt the hypothesis.

Nonetheless, when we present these ideas, we sometimes encounter the suggestion that our hypothesis must surely be false. The grounds for this rejection of our hypothesis typically do not have quite the same character as the evidence we have been discussing thus far. It has never been suggested, e.g., that the hypothesis fails to account for certain sorts of empirical data. Instead, the idea seems to be that there is some reason why the hypothesis just *couldn't possibly* turn out to be true.

We are intrigued by this reaction and would be interested to hear more about it. Perhaps future research will lead to a clear articulation of its basis, and we will then be able to confront it head on. Until that time, however, we propose that the hypothesis should be provisionally accepted.

References

Barsalou, L. W. (1985). Ideals, Central Tendency and Frequency of Instantiation as Determinants of Graded Structure in Categories. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 11, 629-654.

Bratman, M. (2000). Valuing and the Will. *Philosophical Perspectives: Action and Freedom*, 14, 249-265.

Copp, D. (1995). *Morality, Normativity, and Society*. New York: Oxford University Press.

Cushman, F. (2007). The effect of moral judgment on causal and intentional attribution: What we say, or how we think? Unpublished manuscript. Harvard University.

Cushman, F. & Mele, A. (forthcoming). Intentional Action: Two-and-a-half Folk Concepts? In Knobe, J. & Nichols, S. (ed.) *Experimental Philosophy*. New York: Oxford University Press.

Kauppinen, A. (2006). Lovers of the Good: Comments on Knobe and Roedder. *Online Philosophy Conference*.

Knobe, J. (forthcoming). Reason Explanation in Folk Psychology. *Midwest Studies in Philosophy*.

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.

Lewis, D. K. (1989). Dispositional Theories of Value. *Proceedings of the Aristotelian Society*, 63, 113-137.

Malle, B. F. (forthcoming). Bedeutung und Ursprung menschlicher Werte: Eine sozial-kognitive Analyse. In A. Fruhwirt, M. Reicher, and P. Wilhelmer (Eds.), *Markt—Wert—Gefühle.* Vienna: Passagen.

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

Murphy, G. L., & Medin, D. L. (1985). The Role of Theories in Conceptual Coherence. *Psychological Review, 92,* 289-316.

Nyholm, S. (2007). Moral Judgments and Happiness. Unpublished manuscript. University of Michigan.

Rosch, E. & Mervis, C. (1975). Family Resemblances: Studies in the Internal Structure of Categories. *Cognitive Psychology* 7, 573–605.

Smith, M. (1994). *The Moral Problem*. Oxford: Blackwell.

Tannenbaum, D., Ditto, P.H., & Pizarro, D.A. (2007). Different Moral Values Produce Different Judgments of Intentional Action. Unpublished manuscript. University of California-Irvine.

Watson, G. (1975). Free Agency. *Journal of Philosophy*, 72, 205-20.

Watson, G. (1987). Free Action and Free Will. *Mind*, 96, 145-172.

Appendix

In this appendix, we provide more detailed statistical information about the experiments reported in the body of the paper. The information provided here may prove helpful to some readers but is not essential to an understanding of the principal philosophical points of the paper.

Experiment 1

Subjects were 55 people spending time in Manhattan public parks. Subjects were randomly assigned either to the *morally good condition* or to the *morally bad condition*. Subjects in the morally good condition received the vignette about the agent who feels a pull to foster racial equality; subjects in the morally bad condition received the vignette about the agent who feels a pull to foster racial discrimination.

After reading their vignettes, subjects were asked whether they agreed or disagreed with a sentence about the agent's values. Subjects in the morally good condition were asked if they agreed or disagreed with the sentence: 'Despite his conscious beliefs, George actually values racial equality.' Subjects in the morally bad condition were given the sentence: 'Despite his conscious beliefs, George actually values racial equality.'

Subjects rated these sentences on a scale from -3 ('definitely disagree') to +3 ('definitely agree'), with the 0 point marked 'in between.' The mean rating in the morally good condition was .83; the mean in the morally bad condition was -1.14. This difference was statistically significant, t (53) = 4.0, p < .001.

Experiment 2

Subjects came from two samples – *Mormons* (11 participants in a Mormon Bible Study) and *park-goers* (31 people spending time in a Manhattan public park).

Subjects were first asked whether they agreed or disagreed with the sentence: 'Despite her conscious beliefs, Susan still values refraining from premarital sex.' Answers were marked on a scale from -3 ('definitely disagree') to +3 ('definitely agree'), with the 0 point marked 'in between.' The mean score for Mormons was 1.3; the mean for park-goers was -1.4. This difference is statistically significant, t (40) = 4.3, p < .001.

Subjects were then asked: ‘Is it good, bad or neutral to *refrain* from premarital sex?’ Answers were marked on a scale from -3 (‘very bad’) to +3 (‘very good’), with the 0 point marked ‘neutral.’ The mean rating for Mormons was 2.6; the mean for park-goers was -.6. This difference was statistically significant, *t* (40) = 8.8, *p* < .001