Lovers of the Good: Comments on Knobe and Roedder

First draft, April 26, 2006

Antti Kauppinen

University of Helsinki

antti.kauppinen@helsinki.fi

In 'The Concept of Valuing: Experimental Studies' (CV), Joshua Knobe and Erica Roedder argue that moral considerations "play a role in the concept"[1] of valuing. The short paper is a part of a broader project by one of the co-authors (Knobe) to show through experimental studies that folk psychology is not purely descriptive. Instead, Knobe argues, the criteria for application of a broad range of folk psychological concepts, including those of intentional action and causation, include normative elements. This thesis, though not entirely novel, certainly goes against the prevailing interpretations of folk psychology, and is supported by evidence gathered through innovative, cross-disciplinary empirical studies. The challenge it presents to the received view is therefore no doubt worth serious consideration. In earlier work I have critically examined Knobe's empirical methodology.[2] Here, I leave those concerns aside and focus on the explanation of the empirical data. I argue that while we are indeed more likely to interpret someone as valuing something if we ourselves take the object of valuing to be good than if we think it is not, this interpretive tendency can be explained by appeal to the principle of charity while holding on to a traditional, descriptive understanding of folk psychological concepts.

[1] CV, 1.

[2] See my 'The Rise and Fall of Experimental Philosophy' (http://www.helsinki.fi/~amkauppi/phil/The_Rise_and_Fall_of_Experimental_Philosophy.pdf).

Trying to see others as lovers of the good is a sound principle of interpretation, not a feature of the concept of valuing.

Descriptive and Moralized Concepts of Valuing

Let me begin with the thesis that the authors present. What exactly is being claimed? Here is one possible interpretation:

> *The Attribution Thesis* (AT): Given identical evidence of A's attitudes toward and beliefs about o, we are more likely to attribute valuing o to A if we ourselves value o than if we do not value o.

Roughly, the Attribution Thesis says that we take o being good by our lights into account as a positive consideration when we attribute valuing o to A. (Knobe and Roedder are rightly cautious and suggest that our own valuing o is neither a necessary nor a sufficient condition for attributing valuing o to A, but rather something like a tie-breaker. I'll return to this below.) I believe AT is true. However, as I will argue, it is also weak and relatively uninteresting, since it is perfectly compatible with the traditional descriptive interpretation of folk psychology, and so cannot be what Knobe and Roedder are claiming. In particular, I will claim that AT does not contradict the following:

> *The Descriptive Concept of Valuing* (DCV): The folk concept of valuing is the concept of having certain psychological states.

In other words, according to folk semantics, attributions of valuing are attributions of psychological facts alone. I formulate DCV as vaguely as possible in order to remain neutral between competing moral psychological accounts about just what states constitute valuing – whether they are cognitive, non-cognitive, or both, whether they are propositional or not, and so on.[3] If (and only if) DCV is true, so is the following:

> *The Descriptive Truth Conditions Thesis* (DTC): According to folk practice, the truth conditions of sentences of the type "A values o" include only psychological facts about A.

According to DTC, what makes it true that A values o is that A has certain psychological states or stances. The value of o itself is irrelevant to the truth conditions of attributions of valuing.

What makes the position of Knobe and Roedder distinct and interesting is that they deny DCV and thereby DTC. They do, after all, claim that moral considerations play a role in the *concept* of valuing. Here's what seems to me the best way to cash out this stronger claim:

> *The Moralized Concept of Valuing* (MCV): The folk concept of valuing has both psychological *and* moral criteria of application.

This is what Knobe and Roedder seem to be saying when they say that "psychological features are not the only features of the concept [of valuing] […] there is also a moral feature, namely, *whether the object o truly is morally good*."[4] If MCV is true, then so is the following:

[3] DTV should not, therefore, be confused with what might be called descriptivism about valuing, i.e. the claim that the psychological states that constitute valuing are descriptive or cognitive.

[4] CV, 2. Emphasis in the original.

> *The Moralized Truth Conditions Thesis* (MTC): According to folk practice, the value of o plays a role in the truth conditions of attributions of valuing o to others.

In other words, the folk think that the truth of a sentence of the type “A values o” depends, in part, on whether o is good. This is certainly controversial. It says that according to folk, it is not merely psychological facts about A that make it true that A values o, but also moral facts about o itself (as they are taken to be by the folk).

My argument can now be put in the following terms. First, the experimental studies by Knobe and Roedder support AT, and we have other reasons as well to believe it to be true. However, second, MCV (and so MTC) does not follow from AT without further assumptions. Instead, we can explain AT consistently with DCV and DTC by appealing to a general principle of interpretation, namely the principle of charity, according to which we when we interpret someone, we aim for a theory that makes her "consistent, a believer of truths, and a lover of the good”, as Davidson famously put it.[5] This does not show, of course, that MCV and MTC are false, only that Knobe and Roedder have given us no reason to believe they are true. Our own moral and other evaluative beliefs can influence our attributions of psychological states to others without influencing the truth conditions we take those attributions to have; this is the case when they make a difference to the background beliefs and desires we attribute in making sense of the agent and those background attitudes in turn make a difference to what the agent really values.

Attributing Values

[5] Davidson 1980, 222.

What is it to value something? In contemporary discussions, the term 'valuing' is often used, for better or for worse, to cover a broad range of evaluative or normative attitudes.[6] Knobe and Roedder's examples suggest that what they mean by 'valuing' is something like adopting a moral evaluative attitude toward an object or property that grounds a corresponding first-personal ought-judgment ("I ought to φ"). I will follow this usage from now on – but it is good to bear in mind that this is a technical sense of 'valuing' that is in some ways broader and in some ways narrower than the ordinary concept. Their initial hypothesis is that valuing an object o is a cluster concept involving such descriptive criteria as conscious belief that o is good, being motivated to promote o, experiencing guilt when failing to promote o, and having a second-order desire to desire for o. They suggest that none of these is either necessary or sufficient by itself, but merely carries a certain weight in folk classification. This in itself is an interesting approach and worth more attention than I can give it here.[7] However, as already noted, the main claim is that in addition to any purely psychological criteria, the actual value of o, by the attributor's lights,

[6] I want to note here parenthetically that using 'valuing' as an umbrella term for normative attitudes masks potentially important distinctions. Prima facie, there is a difference between deontic attitudes – taking something to be right or wrong, thinking that I ought to do something or have reason to do something – and evaluative attitudes like finding something to be good or beautiful. I can perfectly well value something (think that it is good) without thinking that I ought to do something about it. This is the case even if valuing something is a matter of taking there to be reason to pursue, promote, or protect it – it doesn't follow that I have to take *myself* to have reason to do any of those things. And of course, I may value something while finding other things even more worthwhile, in which case again I won't think I ought to pursue it – I value spending the night alone at home, but I value quality time with my partner even more. Nor should we ignore the distinction between instrumental and non-instrumental value, however we draw it more precisely, or the distinction between moral and non-moral value. These distinctions among normative attitudes matter when we're doing moral psychology, since what constitutes 'valuing' may be very different in different types of cases. While the thesis that there is an internal connection between first-personal ought-judgment and motivation is plausible, it is much less likely, for example, that in order to count as instrumentally valuing a means to an end that one may never pursue one has to have any degree of motivation with respect to those means. Similarly, while moral ought-judgments seem to have much to do with guilt in the case of non-performance, it is too much to ask for me to feel guilt for not bothering to wash the car, though I judged it best.

[7] One possibility that Knobe and Roedder overlook is that different ways of 'valuing' (judging that I morally ought to do something, judging that something is prudentially good, judging that something is the best means to an end, and so on) involve different necessary and sufficient criteria. If this were the case, traditional conceptual analysis would still have a shot, even if different features were relevant in different cases of 'valuing'. But it could also be that we just give more weight to feeling guilt in the case of moral ought-judgments, for example. As I said, this invites further examination.

carries some weight in determining whether the concept applies (whether the agent values o) and so whether the claim that the agent values o is true.

Knobe and Roedder support this claim with the results of two different surveys. I won't reproduce their vignettes here, but I will lay out just what psychological states they explicitly direct the test subjects to attribute to the agents whose values they are examining, since this is highly relevant to the argument. In the first study, the question is whether George values racial equality or not. George I feels guilt for racist actions in spite of his contrary ought-judgments, while George II feels guilty for egalitarian actions in spite of his contrary ought-judgments. According to the vignettes, the following psychological facts hold of the two Georges:

> *George I*
>
> Belief that he ought to advance the interests of people of his own race at the expense of people of other races = $B (O (\varphi))$
> Desire not to advance people of his own race at the expense of people of other races = $D (\sim\varphi)$
> Guilt for having the desire to advance people of his own race at the expense of people of other races = $G (D (\varphi))$
> Second-order desire to get rid of the desire to help people of other races to be more equal = $D (\sim D (\sim\varphi))$
>
> *George II*
>
> Belief that he ought to advance the interests of people of all races equally = $B (O (\psi))$
> Desire not to advance the interests of people of all races equally = $D (\sim\psi)$
> Guilt for having the desire to advance the interests of people of all races equally = $G (D (\psi))$
> Second-order desire to get rid of the desire not to advance the interests of people of all races equally = $D (\sim D (\sim\psi))$

In terms of my (hopefully obvious) quasi-formal apparatus, the psychological setup of the two Georges is thus as follows:

George I: $\{B\ (O\ (\varphi)),\ D\ (\sim\varphi),\ G\ (D\ (\varphi)),\ D\ (\sim D\ (\sim\varphi))\}$

George II: $\{B\ (O\ (\psi)),\ D\ (\sim\psi),\ G\ (D\ (\psi)),\ D\ (\sim D\ (\sim\psi))\}$

As can be seen, as far as the vignettes go, the two are psychologically identical, apart from the content of the states (φ and ψ, promoting racial inequality and promoting racial equality). It would seem that if the concept of valuing is purely psychological and descriptive, our judgment in each case should be the same – if George I values racial equality, George II values racial discrimination. But this is not so, as the Knobe and Roedder experiment shows: when presented with the cases, ordinary people respond, on average, that while George I does value racial equality, George II does not value racial discrimination. Since the only salient difference between the two cases is the moral value of racial equality (good by our lights) and racial discrimination (bad by our lights), Knobe and Roedder conclude that this is what is driving the responses, and that the concept of valuing has non-psychological, moral criteria of application.[8]

Resolving Conflict in the Direction of the Good

Going beyond the data we are actually given is commonplace when we're trying to make sense of people or stories. If you tell me that Jonathan thinks George W. Bush is doing a great job, I do not simply add that one belief to the score I keep of Jonathan's doxastic commitments. Perhaps tacitly, I also make adjustments to other parts of Jonathan's score so as to make sense of this odd belief that I don't share. The precise background attributions I make depend, naturally, on what I know of Jonathan and the way Bush does his job, but there is a predictable pattern to them. First,

[8] To be precise, they speak of the concept's 'features', but this is misleading. People and objects have features in virtue of which concepts apply to them; (some) concepts have criteria in virtue of which they apply to people and objects with certain features.

I may make sense of Jonathan's belief by attributing to him *false factual beliefs* that would, if true, support the notion that Bush is doing a great job. Perhaps I take it that he believes Saddam Hussein was an imminent threat to the US or that tax cuts for the rich trickle down to the rest. (Note that I am reluctant to attribute to him the belief that Bush is the brain behind the rise of the Internet or that Bush is guided by God's voice – if possible, I want to avoid making him into a lunatic.) This may be sufficient, but what if I have reason to believe there's nothing wrong with Jonathan's factual beliefs? Then, second, I'll have to attribute to him *dubious values*. Perhaps he thinks the job of the American president is to advance the interests of American people, the rest of the world be damned, or that whatever outcome results from uncoerced market interactions is automatically just. (Note that I'm reluctant to attribute to him the thought that bombing Iraqi civilians is good or that rich people are just inherently more deserving than others – if possible, I want to avoid making him into a monster.) Some such values combined with true beliefs would not only explain why a rational person would hold his original belief, but also explain why someone like him would be disposed to form particular kinds of false or ungrounded beliefs. But what if I know Jonathan to be not only a believer of truths but also a proud liberal in the American sense? Then, thirdly, as a last resort, I may have to think he is being *irrational*. His belief that Bush is doing a great job just doesn't fit in the pattern of his other beliefs and values. Perhaps he doesn't really believe that, even if he uses with apparent sincerity the form of words that would usually express that belief. Or he's given up or being inattentive to his other beliefs and values. Insofar as understanding Jonathan is important to me, I do my best to adjust my take on his overall psychology so that he comes out as rational and, on the whole, someone whose beliefs and values are responsive to the available evidence rather than a lunatic or a monster. In doing so, I inevitably go beyond what I'm told about him, guided by the assumption that there is (by my lights) a reasonable explanation for his false belief.

My alternative explanation of the data provided by Knobe and Roedder is based on the assumption that the folk use the same principles of interpretation to make sense of George and Susan. These are, in philosophical terms, (modest) holism and the principle of charity. Both are associated with the work of Donald Davidson, but have a much broader resonance; one might as well refer to Gadamer and the continental hermeneutic tradition from Schleiermacher and Dilthey onwards.[9] What I'm calling modest holism is simply the thesis that we don't attribute contentful psychological states one by one. Someone who believes that there is a tree in the garden will have a host of possibly unarticulated background beliefs about there being in the garden a visible, solid, living thing that is rooted on the ground, has a particular scent, and may probably be used to feed a fire, there being a difference between wild nature and nature cultivated or at least separated by humans, and so on and on. In the absence of some such background, we would be hard pressed to make sense of what someone meant if she asserted that there is a tree in the garden; conversely, when we do attribute such a belief to someone, we at least tacitly attribute to her a host of suitable background beliefs. The principle of charity, in turn, states that to make sense of the words and behavior of others, the psychological states we attribute to them must come out as both internally coherent and appropriate in their circumstances, so far as that is possible given behavioral and linguistic evidence. While local failures are always possible, in the absence of a plausible excusing condition, we could not hope to understand someone who didn't on the whole respond to what we take to be reasons for belief and desire.

[9] See, in particular, Gadamer 1960.

We can buttress the case for charity by straightforward epistemological considerations. All interpretive concerns aside, it is very likely that human beings with normal perceptual and cognitive capacities have a host of true beliefs about dry, middle-sized objects in their immediate environment. Similarly, whether basic moral knowledge is a priori or a posteriori, we have reason to believe that human beings with normal cognitive and emotional capacities have it. If it is a priori, it surely belongs to the readily accessible rather than the esoteric branch of a priori knowledge; if it is a posteriori, it is very likely that any normal human being has had the sort of experience needed to acquire it. Any random person you meet is much more likely than not to know that kicking a man on the ground is wrong and keeping a promise is good. I would even claim that any normal person who has had face-to-face dealings with people of other races outside a dehumanizing context is likely to recognize, to some significant extent, that they are persons with a moral standing, unless they actively repress this knowledge.[10] Given the odds, we're better off assuming that people we encounter value the good than assuming they don't, and, within reason, explaining away apparent indications to the contrary.

I will thus assume that modest holism and the principle of charity are both part of our ordinary folk psychological practice, and that the subjects in the Knobe and Roedder experiments have made use of them. If this is the case, how can we expect them to respond to the George cases? To begin with, an important feature of these cases is that they involve *inner conflict* or at least tension. In each case, some of George's psychological features suggest that he values racial

[10] Of course, the two provisos I make are both important and, sadly, all too often actualized. When all encounters with slaves take place under conditions that are dehumanizing for them, their humanity may easily go unrecognized, and even when the humanity of supposedly inferior races is recognized, history has shown that ideological rationalizations can be very effective.

equality, others that he doesn't. Moreover, the psychological features that he is described to have are themselves in conflict – normally and probably constitutively, ought-beliefs and guilt go hand in hand.[11] His psychology, as described, presents a puzzle; if he's not practically irrational, he's very close to it. Applying the two principles of interpretation resolves the puzzle. First, in making sense of each George, we presumptively need to see him internally coherent and as a lover of the good by our own lights. It could, of course, turn out that that's not the case – all indications suggest that Hitler was not a lover of the good by our lights, and in such a case, we have no problem attributing reprehensible values to him; that's precisely what's wrong with him. (This is why it's so important that the examples feature people who are described as internally conflicted.) Second, when there are conflicting indications, charity leads us to (try to) tacitly fill in the psychological story in a way that resolves the conflict in the direction of the good. That is, we attribute the sort of background beliefs, desires, and emotions in the light of which those beliefs or desires that we would ourselves have in the situation stand out as genuinely George's own. We end up thinking that George, with his conflicting attitudes, doesn't *really* think racial inequality is good. In that way, we avoid attributing irrationality to him. This works out differently in the two cases, since the conflict is resolved in opposite directions.

Take George I first. Having grown up in a racist culture, he believes he should promote his own race, but feels guilt for doing so and sometimes has a motivationally effective desire to treat people equally, though he wishes not to have such a desire. Charity requires us to privilege his non-racist attitudes in resolving his apparent irrationality. It directs us to see George as very much like Mark Twain's Huckleberry Finn, who was taught all his life that slaves are property,

[11] The most obvious exception is when somebody has been forced to choose between two evils. In that case, even if she is certain she made the right choice, she may intelligibly feel guilt for what she did, provided it was bad enough.

not persons, but befriended the slave Jim and found himself unable to turn him in when he escaped. As Nomy Arpaly has recently noted, Huck Finn's desire to help Jim is not a mere inclination that just happens to be causally stronger than his ought-judgment, which on many theories represents the agent's true self.[12] Rather, it is deeply integrated with his perceptions, emotions, and other desires, while his ought-judgment is an abstract belief that is disengaged from much of the rest of his evaluative system. Though he chides himself for not treating Jim as property, it is clear from the story that Huck values Jim as a person among others, even if he is unable to articulate that belief consciously even to himself; as Arpaly emphasizes, this is manifest in such actions as apologizing to Jim. When we see George I as a lover of the good, we see him in the same vein. We tacitly attribute to him experiences that reveal the wrongness of the beliefs he has internalized; it need not be the faculty of moral intuition at work, but simply the natural sympathy that normal people are equipped with.[13] This helps make sense of his egalitarian desires and his guilt for racism. And we know this is how it works: when people have been indoctrinated to believe something wrong, the light dawns first in the form of vague unease and a lack of eagerness to follow the rules, grows into guilt and reluctance, and at some point into doubts about the internalized principles and perhaps their rejection. It is relatively easy to fit George I's present inner conflict into this narrative (while, as we'll soon see, it is impossible in the case of George II). As we do so, we see his present ought-beliefs as external impositions that he no longer fully identifies with; as his guilt suggests, his heart already beats to a different beat. It is worth noting that the case highlights the *dynamic* aspect of charity, as indicated by temporal expressions like 'no longer' and 'already' above. We don't just look at the totality of an agent's

[12] Arpaly 2003, 75–78.

[13] As Adam Smith puts it at the very beginning of *The Theory of Moral Sentiments*: "That we often derive sorrow from the sorrow of others, is a matter of fact too obvious to require instances to prove it; for this sentiment, like all the other original passions of human nature, is by no means confined to the virtuous and humane, though they perhaps may feel it with the most exquisite sensibility." (Smith 1976/1759, 9)

attitudes at one particular moment, but consider plausible trajectories over time – we need to see how a rational person might have ended up with the present set of attitudes. And we anticipate how things might develop; to take an attitude as expressing an agent's values is to take her to resist change to it.[14] If, as I suspect, the folk think along these lines, they will tacitly attribute to George I the sort of experiences, desires, and doubts that make the egalitarian desire and guilt stand out as truly his own and his explicit beliefs as more or less external impositions, and so conclude that he in fact values racial equality.

George II, on the other hand, grew up egalitarian, but feels the motivational pull of racism, to the extent that he feels guilt for holding on to his egalitarian beliefs. Yet at the same time, the story goes, he wishes he didn't have racist desires. Here charity, much needed, requires us to try to fill in the background story so that his ought-beliefs come out as genuine and his own. This means that we have to discount, above all, the guilt that he is said to feel for having the egalitarian belief – the mere desire to prioritize his own doesn't weigh too much in attributing values. Now, to be honest, I find the sentence "He often finds himself feeling guilty when he helps people of other races at the expense of his own" almost unintelligible in the context of the rest of George II's story. He is described as having internalized egalitarian beliefs, and the default charitable assumption is that his moral convictions track moral truth. While George I is readily seen as tacitly recognizing the wrongness of racism when he feels guilty for his actions, there is no wrongness of egalitarianism for George II to tacitly recognize, no stream of morally tinged experiences we could picture as forcing him to reconsider his explicit views. How, then, has he come to regard his egalitarian actions as so wrong that he feels *guilty* for having them? It is hard

[14] See Blackburn 1998, 67.

to think of a story that would make his guilt rationally intelligible.[15] It seems more like a glitch in the system, a mere causal force. Consequently, it is not hard to interpret George II as being alienated from it. On charitable assumptions, it is largely disconnected from the rest of his attitudes – it doesn't match his plans, his take on what there is reason to do, his attributions of blame to others who manifest racist behavior and related feelings toward them, his attitudes and feelings toward particular people of other ethnicities most of the time, his voting patterns, and so on. When we tacitly fill in George II's story in some such way, we can discount the competing indications and say that on the basis of the totality of his attitudes, George II values racial equality after all.

Testing Charity

Could we test my alternative interpretation empirically? Bracketing for the time being legitimate concerns there are about the use of surveys in this kind of research in general, this seems quite straightforward. We could ask the test subjects questions about those of George's psychological states that are *not* explicitly mentioned in the vignettes, but form a part of the background in the light of which the explicitly mentioned attitudes make sense. My view predicts that people would respond differently depending on whether or not they take George to value racial equality. In that case, the most plausible interpretation of the data would be that the asymmetry in value attributions would be explained by the asymmetry in tacit attributions of background beliefs, desires, and emotions rather than by the asymmetry in the attributors' beliefs about the value of the objects (though this would explain, in part, the tacit attributions). To test this, we could

[15] Hard, but not impossible. Maybe he's ended up under the influence of a covert neo-Nazi group that is manipulating his emotions by arranging that he constantly encounters situations in which racially egalitarian attitudes and actions involve well-wishing dishonesty and lead to pain and suffering.

proceed as follows. After (or perhaps before) asking whether the test subject agrees or disagrees with the sentence 'Despite his conscious beliefs, George actually values racial equality/discrimination', we could ask whether she agrees or disagrees with sentences like the following: 'George probably feels bad when he witnesses another person in serious pain, regardless of his or her race', 'George probably believes that making a person of any race happy speaks in favor of a course of action', 'George probably wants to make friends with people with similar interests regardless of their race', 'George probably has doubts about his moral beliefs/whether his guilt is reasonable', and so on. If I'm correct, the people who agree with these additional sentences or other similar ones are much more likely to agree with the claim that George values racial equality than those who don't. Also, if I'm correct, people are more likely in general to agree than disagree with these sentences, given the original vignettes – regardless of which way ought-beliefs and guilt are described as pulling.

Though I don’t want to enter the methodological dispute here, it is perhaps worth noting at this point that a survey approach with fixed questions is far from ideal when it comes to testing what background beliefs and attitudes people tacitly attribute. After all, my thesis is merely that there is *some* difference in these background attributions that makes the difference in attributions of valuing intelligible; different people may imagine the agents' broader psychologies very differently. A structured interview would seem to be a much more fruitful approach here. Gently guiding people to talk about how they see George (and Susan in the other case) should reveal significant differences between those who see him as a racist and those who don't. These results would be harder to quantify, but philosophical and psychological illumination doesn’t necessarily increase with quasi- (or pseudo-) scientific precision.

With the charitable background beliefs and attitudes filled in in each case, a broader look at the psychologies test subjects attribute to the two Georges should reveal something like the following:

> George I: {$B (O (\varphi))$, $D (\sim\varphi)$, $G (D (\varphi))$, $D (\sim D (\sim\varphi))$, recognizes-reasons-for ($\sim\varphi$), feels-bad-when-others (φ), has-goals-for-which-an-essential-means-is ($\sim\varphi$), doubts ($B (O (\varphi))$), resists-change-to ($D (\sim\varphi)$)[16]}
>
> George II: {$B (O (\psi))$, $D (\sim\psi)$, $G (D (\psi))$, $D (\sim D (\sim\psi))$, recognizes-reasons-for (ψ), feels-bad-when-others ($\sim\psi$), has-goals-for-which-an-essential-means-is (ψ), ~doubts $B ((O (\psi)))$, ~resists-change-to ($D (\sim\psi)$)}

Now the two psychologies are no longer symmetrical. The background features added in the process of making sense of the stories tip the balance to the direction of the good; George I values $\sim\varphi$-ing (not promoting his own race, i.e. promoting racial equality), while George II values ψ-ing (promoting racial equality). In George I's case, the charitably added features undermine his conscious ought-belief (he doubts it, recognizes reasons to the contrary, and so on), while in George II's case they buttress it (he has no doubts about it, feels bad when others do the opposite, and so on). Still, for the folk, what makes it *true* that the Georges value racial equality is that the pro-equality attitudes are best supported by his overall psychology.

The results from the case of Susan, who has conflicting attitudes toward premarital sex, can be similarly explained. She was brought up religious, but became an atheist in college, and now

[16] One simple way to understand resisting change to non-racist desires while having a desire not to have non-racist desires is to postulate a third-order desire: $D (\sim D (\sim D (\sim\varphi)))$. This is complex, but all it amounts to is this: George I wants to stop wishing he didn't feel the pull of egalitarian actions. Normally, higher-order desires are liable to collapse quickly to second-order desires, but sometimes second-order desires can be sticky. (Mutatis mutandis, the same goes for George II.)

believes premarital sex is acceptable and wants to have it with her husband-to-be. But she still feels guilt for it, though she desires to get rid of this feeling. Knobe and Roedder presented her story to subjects with different values – a Mormon Bible study group and Manhattan park-goers. While both were given the same story, the former thought refraining from premarital sex was good and attributed the same value to Susan, while the latter found it neutral and didn't attribute the value to Susan either. According to my view, this is explained by the different background beliefs that charity demands the members of the two groups to attribute to Susan to resolve the conflict between ought-belief and guilt in the direction of the good.

Thus, the secular park-goers most likely take Susan to have shed her religious values when she has had a chance to experiment and think for herself in college. For them, something like this is a plausible learning process. Consequently, they take her ought-judgments and desire to get rid of guilt as sincere and the guilt itself as residual and ungrounded in the rest of her psychology. It is just a recalcitrant trace of her childhood teaching, maybe supported by the anxiety of a virgin. When the Mormons, in turn, put themselves in Susan's shoes, as charity requires, they tacitly fill in the story so that Susan's guilt for her sexual desire is authentic and her professions of worldly values and desire to lose the guilt the result of an unfortunate influence by peers and liberal professors, or some such thing. In so doing, they take her guilt to be deeply rooted in her overall psychology.[17] Maybe they picture her as being at bottom disgusted with modern promiscuity and debauchery, as worrying about girls being taken advantage of, as finding a traditional wedding night very romantic, as believing in the importance of self-control, and so on. Perhaps they imagine her explicit beliefs and second-order desires so shallowly rooted that they would weaken

[17] Given the description of this particular case, this may call for some wishful thinking, given that Susan's desires, second-order desires, and explicit moral beliefs are all supposed to be pro-sex.

and vanish with the unfortunate peer pressure. Against some such background, Susan’s guilt represents her true values.

Conclusion

At the end of their paper, Knobe and Roedder wonder "whether moral considerations actually play any role in people’s *concept* of valuing"[18] and invite alternative hypotheses. I have tried to provide one in this response. Before summing up, I’d like to note that there may be other possibilities I haven’t discussed here. In particular, talk of someone’s values might convey some sort of approval of them (or of the person) by means of conversational implicature or some other pragmatic device.[19] If so, it’s understandable why people wouldn’t want to say that someone values racial discrimination when they don’t themselves value it. However, I have not pursued this explanatory strategy, since I think Knobe and Roedder’s results are robust in this respect. That is, I believe the asymmetry would still obtain even if it was made salient to subjects before the test, for example, that Hitler indeed did *value* a world without Jews, though we want to have nothing of his values. One of the marks of conversational implicature is, after all, that it can be cancelled, so it should be possible to significantly reduce the effect of such pragmatic considerations in an experimental situation.

[18] CV, 6.

[19] Something along these lines was suggested by Jason D’Cruz on the free will and moral responsibility blog Garden of Forking Paths (http://gfp.typepad.com/the_garden_of_forking_pat/2005/05/desires_vs_valu.html) and, independently, by Teemu Toppinen in personal communication. Some of the experimental data by Bertram Malle and Eric Edmondson (MS) seems to support the existence of some such effect, though it's not quite clear to me what their study attempts to achieve, and they themselves fail to recognize the existence of pragmatic considerations. Also, the results they report suggest that the verbal difference between what someone values and what are someone's values confuses some people. I value my parents, for example, but my parents aren't among my values; such a claim wouldn't be grammatically intelligible.

Understanding other people is a difficult business, since psychological states and stances so often manifest themselves in speech and behavior in indirect ways, if at all. To get a foothold, it is useful or even necessary to put ourselves in others' shoes. This amounts to deploying our own beliefs and values in trying to make sense of what others believe and value in a given situation. But people are different, and often the behavior of others shows that they believe and value differently from us. This is no threat to comprehension as long as we can see how a rational, at least moderately reasons-responsive person might have ended up with the sort of attitudes that the other manifests. When the person we are trying to understand seems to have conflicting or incoherent attitudes, the task becomes harder. If someone genuinely thinks she ought to have done something and it is not a matter of having had to choose between two evils, guilt seems to be out of place by her own lights. Unless both are very strongly supported by behavioral evidence, we avoid attributing both genuine guilt and a genuine opposed ought-judgment. Instead, if we have to make a choice, we once again try to place ourselves in the other's shoes to discover which of these attitudes is integrated with the rest of the agent's psychology and which just some kind of relic of the past. This amounts to resolving the apparent conflict – and so avoiding the attribution of irrationality – in the direction of what we ourselves find valuable. But since this happens by way of reconceiving the agent's psychology, it doesn't imply that our concept of valuing has any moral criteria of application. By our lights, what makes it true in a conflict situation that George or Susan values something we take to be good rather than something we take to be bad is that his or her attitudes toward the good are more deeply anchored in her true self.[20]

[20] Both Jussi Suikkanen (at the GFP blog) and Teemu Toppinen independently suggested that the principle of charity might explain the difference noted by Knobe and Roedder. Given that I had myself independently arrived at the same solution, we may safely conclude that philosophical socialization at the University of Helsinki disposes people

Bibliography

Arpaly, Nomy (2003), *Unprincipled Virtue. An Inquiry into Moral Agency*. Oxford University Press, Oxford.

Blackburn, Simon (1998), *Ruling Passions. A Theory of Practical Reasoning*. Clarendon Press, Oxford.

Davidson, Donald (1980), 'Mental Events', in *Essays on Actions and Events*. Oxford University Press, Oxford, 207–225.

Gadamer, Hans-Georg (1960), *Wahrheit und Methode*. Mohr, Tübingen.

Kauppinen, Antti (under review), 'The Rise and Fall of Experimental Philosophy', http://www.helsinki.fi/~amkauppi/phil/The_Rise_and_Fall_of_Experimental_Philosophy.pdf

Knobe, Joshua and Roedder, Erica (MS), 'The Concept of Valuing: Experimental Studies' (CV)

Malle, Bertram and Edmondson, Eric (MS), 'What Are Values? A Folk-Conceptual Investigation', http://hebb.uoregon.edu/04-01tech.pdf

Smith, Adam (1976/1759), *The Theory of Moral Sentiments*. Ed. D. D. Raphael and A. L. Macfie. Liberty Fund reprint, Indianapolis.

to see others as lovers of the good, unless proved otherwise. I'd also like to thank Lilian O'Brien and Pekka Mäkelä for written comments on an earlier draft.