Empathy as the Moral Sense?

Antti Kauppinen

University of Tampere / Academy of Finland

Final draft, January 5, 2017

Forthcoming in *Philosophia*

Let's suppose that it is possible for us to come to know by empathy what other people are thinking and feeling, where empathy is some kind of direct perception of what goes on in their minds that isn't based on inference of any kind. Given that the thoughts of other people, in turn, may reveal further facts about them and the external world, empathy may be a way of coming to know non-psychological facts as well. This informational role of empathy is the main focus of Michael Slote's 'The Many Faces of Empathy'. Slote argues that empathy is a way of acquiring knowledge of the world and moral facts, and is essential to the transmission of knowledge by testimony.

In this comment, I'll focus on Slote's claims about empathy's role in moral epistemology. He holds an unusual combination of views, according to which we can know basic moral truths a priori, even though they are not analytic, and moral properties are natural properties. We're then capable of detecting a posteriori by natural means – empathy – whether these properties are instantiated in particular cases. My challenge will focus on whether it is possible to combine a priori moral knowledge with a kind of non-analytic naturalism. Since Slote's main argument for the view is based on a thesis in the philosophy of language, this will require close examination of the notion of reference-fixing. I'll argue that Slote's account isn't as Kripkean as he thinks, and consequently inherits some familiar challenges that analytic naturalism faces.

1. Empathy as a Source of Moral Knowledge

Slote's earlier work on sentimentalism (Slote 2007, Slote 2010) has emphasized the role of empathy in ethics, so it is not surprising that he highlights its importance for moral knowledge. Indeed, he says that "empathy is the special moral sense that Hutcheson was talking about". The structure of his argument for this claim is straightforward, though the premises are contentious:

Empathy Is the Moral Sense

1. We can come to know the warmth or coldness of an agent's motives by empathy.
2. Moral goodness or virtue consists in being a warm-hearted person, and badness or vice in being cold-hearted toward others.
3. So, we can come to know the virtue or vice of others by empathy.

In this paper, I'm going to grant Slote the first premise of the argument. It is in any case plausible that if we can empathically *take on* the emotions of others, we're also in a position to *know* what their emotional state is. What I'll challenge is the second premise. I'll spend most of my time discussing Slote's metaethical argument for it, but I'll start by saying a few things at a more intuitive level. To begin with, there is something appealing about the idea that virtue consists in tending to be warm and caring. It is no doubt *part* of being virtuous. Yet on the face of it, it doesn't look like warm-heartedness exhausts virtue. Some core virtues, such as honesty and courage, don't seem to have much to do with warmth, although they don't necessarily conflict with it. You don't need to be particularly empathic to be honest, just principled or conscientious. More problematically for Slote, some virtues seem to *conflict* with warmth, at least on particular occasions. Here I'm thinking of what Hume called, in part for just this reason, artificial virtues, which include justice and promise-keeping. As he famously put it, "Judges take from a poor man to give to a rich; they bestow on the dissolute the labour of the industrious; and put into the hands of the vicious the means of harming both themselves and others." (*Treatise*, 3.3.1) In such cases, justice requires reining in empathy. Worse yet, those of us with any Kantian leanings will think it is morally bad if someone keeps a promise just because they empathize with the promisee – it is, so to speak, a wrong kind of motivating reason in this instance, when simply having promised should suffice. We can agree with this even if we're dubious of Kant's more counterintuitive claim that even helping others only has moral worth when the agent "does the action without any inclination, for the sake of duty alone".

Hume, of course, thinks that we ultimately approve of justice and promise-keeping, because we take these institutions in general to have beneficial consequences, and empathize with those who benefit from them. Slote's own brief account gestures in the same direction as Hume's, I think, when he says that "institutions and laws, as well as social customs and practices, are just if they reflect empathically caring motivation on the part of (enough of) those responsible for originating and maintaining them" (Slote 2010, 125). The similarity with Hume is that empathy is supposed to lie behind institutions and practices as a whole, rather than individual actions (or so it seems). Slote's proposal faces the obvious problem that even good people are fallible, and it is possible for even the most empathic law-makers, say, to come up with unjust laws. What is more, criminal law involves holding people responsible for wrongdoing. Rather than caring, such laws reflect or express reactive attitudes like resentment and indignation.[1] As a matter of fact, I believe that regulated empathy has an essential role in directing such reactive attitudes, when it comes to justice. But in any case, as long as individual just acts are motivated by something other than empathic caring, it seems there must be more than warm-heartedness to virtue.

[1] See Kauppinen 2015 for my take on the expressive function of the law.

Perhaps to sidestep such intuitions, Slote offers a more theoretical supplementary argument for Premise 2 in my reconstruction. It goes like this (and here I can quote him directly):

Virtue Is Caring

1. "Moral virtue … is that property of other individuals or their attitudes or actions that is empathically registered in us as a feeling of being warmed."
2. "The only property of others that can be empathically registered as warmth is warm caringness on their part."
3. So, moral virtue consists in warm caringness.

Here the second premise seems to be unquestionably true in the relevant sense of empathy. So the main issue is whether the first premise is true.

2. Slote on Empathy and Reference-Fixing

The first premise of the Virtue Is Caring argument is a metaethical thesis, more precisely a metaphysical claim. Slote defends it most thoroughly in his 2010 book. Here's a more explicit formulation of the assumption:

> As a way of indicating how the reference of a moral term like "morally good" or "right" is fixed, we can also say that it is a priori that moral goodness (or rightness) is whatever feelings of warmth directed at agents and delivered by mechanisms of empathy are caused by. (2010, 61)

To assess this claim, we need to first get clear on the relationships among reference-fixing, meaning, and the a priori. Slote draws on the work of Saul Kripke, who introduces the notion of reference-fixing in the context of discussing proper names. I'm going to take a little time to review what Kripke actually says, since the details will matter for assessing Slote's argument.

In *Naming and Necessity*, Kripke's initial target is the descriptivist theory of names, according to which names either mean the same as an associated description or cluster of descriptions, or at least refer to whatever satisfies it. Views of this kind can be found in Frege, Russell, and Searle, for example. 'Aristotle', for example, might be associated with the description 'Plato's most famous student' (and perhaps a bunch of others). Kripke observes that in the actual world, a description like "Plato's most famous student" indeed picks out Aristotle. But it is, of course, possible that someone else would have been Plato's most famous student – Aristotle might never have left Stageira, for example. Since the description picks out a different individual in different scenarios, it is non-rigid. The proper name 'Aristotle', however, refers to Aristotle, the very same individual, even when we're talking about different possibilities – it is a rigid designator. For example, for each possible scenario, the truth of "Aristotle was fond of dogs" depends on whether a particular person, Aristotle, was fond of dogs. But Russell, for example, thinks that "Aristotle was fond of dogs" is (or may be) equivalent to "Plato's most famous student was fond of dogs". If that were the case, in the possible situation in which Isokrates was Plato's most famous student, the truth of "Aristotle was fond of dogs" would depend on whether Isokrates liked dogs. The same would follow if we didn't take the name to be synonymous with the description, as long as the associated description determined its reference. It also follows that some sentences that are intuitively true wouldn't be true, and some sentences that are not necessarily true would be necessarily true. For example, the sentence "Aristotle might have never studied with Plato" is true, and the sentence "Aristotle didn't study with Plato" could be true. In contrast, "Plato's most famous student might never have studied with Plato" is false (on a plausible reading), and necessarily so. And "Aristotle was Plato's most famous student" isn't necessary, or a priori for that matter. So clearly, “Aristotle” can’t *mean* the same as “Plato’s most famous student”. Nor does the description determine its reference, although they coincide in the actual world.

Nevertheless, we can make use of the description in a different way to work with names. We can make use of it to *fix the reference* of ‘Aristotle’, just as we might point to Aristotle, if he were here, and say “Aristotle is *that guy*”. Kripke calls this the *initial baptism* (1980, 97). Here the description “fixes the reference by some contingent marks of the object” (106). The description need not even be true of the referent. I think it’s quite likely that after Clyde Tombaugh discovered an object in the solar system beyond Neptune in 1930, astronomers debated what to call what they took to be the ninth planet. Using the description ‘the ninth planet’ to pick out what they were talking about, they settled on ‘Pluto’ – indeed, they could have decided in advance that if a ninth planet is going to be found, it’ll be called ‘Pluto’. But it’s no part of the meaning of ‘Pluto’ that it’s the ninth planet, as evidenced, for example, by the fact that it is no contradiction to say that Pluto isn’t a planet. In any case, once reference is fixed, people who subsequently pick up the name from the original users need not associate the description with the referent in order to successfully talk about it.

Kripke then extends this account to natural kind terms, such as ‘tiger’, ‘gold’, and, as he says, “suitably elaborated”, colour terms like ‘red’. The idea, again, is that we may identify certain kinds or properties by their effects on us, and thus to make use of these effects or the disposition to produce them in us to fix the reference of our terms or concepts, without it being the case that producing these effects in us part of their meaning or content. For example, as things actually are, heat produces certain sensations in us. We pick out heat as whatever property of objects it is that produces these sensations. This property turns out to be mean molecular kinetic energy. Since it is the referent of ‘heat’, the identity claim that heat is mean molecular kinetic energy is an *a posteriori* necessary truth. This property is independent of our thinking or experiencing it in any way. Objects would be just as hot even if we didn't have heat-sensations, and even if we didn't exist at all.

Most importantly for our purposes, Kripke briefly argues that the same goes for colours. For example, yellowness is not a dispositional property (140n71). Again, the idea is that "the reference of 'yellowness' is fixed by the description 'that manifest property of objects which causes them, under normal circumstances, to be seen as yellow'" (ibid.). It is up to scientists to determine what this property of objects is, or whether there is a single such property in the first place. Whatever it turns out to be is what yellowness is. Anything that has it is yellow, and would be yellow, even if there were no creatures capable of seeing colour, or never had been. It is not that 'yellow' *means* something like 'tends to produce a yellow sensation' – one doesn't need to associate such a description with the term to be able to talk about yellow things, and the description need not be true of yellow objects. To be sure, Kripke allows that it can *seem* as if fixing reference in a particular way matters in these cases, but he denies it: "the way reference is fixed seems overwhelmingly important to us in the case of sensed phenomena … the fact that we identify light in a certain way seems to us to be *crucial*, even though it is not necessary" (139, emphasis in the original).

Here's a summary of Kripke on reference-fixing:

a) Some subject(s) S can fix object/kind/property *x* as the reference of term T by a description D (where D may contain reference to S's responses to *x*).

b) Subsequently, S, or any S' at the end of a causal-historical chain of language use, can refer to *x* by T without associating D with *x* – in other words, D is not part of the meaning of T, nor does it determine its reference.

c) As a consequence of b, it is not *a priori* that *x* is D.

d) T can refer to *x* even if it is false that *x* is D.

e) It may be possible to establish the essence of *x* by an *a posteriori* investigation.

I've spent a bit of time on Kripke exegesis, because I want to highlight how Slote's proposal differs from it. Slote's claim, recall, is that empathic feelings fix the reference of moral terms, so that when it comes to 'morally good', D is something like "the thing that causes warm empathic feelings towards agents in us". What this would amount to on the original Kripkean picture is that if *g* is the property that, at the time of the initial baptism, causes (or caused) warm empathic feelings towards agents in us (or those from whom we've inherited the term), then *g* constitutes moral goodness, even if it no longer causes warm empathic feelings in us. Later users who borrow the reference need not have any empathy, or know that empathy played a role in reference-fixing, to refer to the same thing with 'morally good'. But of course, there's no reason to believe that when we use moral terms, we're referring to whatever someone at some unknown point in history picked out. If we did, it would be very difficult to explain why moral judgments tend to have an impact on our motivation, on either motivational internalist or externalist accounts. And it's not as if my utterance of "Slavery is wrong" is false if it turns out that the people from whom I've inherited the use of moral terms empathized more with slave-owners than with slaves.

It is thus unsurprising that Slote departs from the strictly Kripkean model of reference-fixing. According to him, it is *a priori* that the extension is fixed in a certain way. I want to emphasize that this is a radical and unacknowledged difference from Kripke, for whom a description can *either* fix the reference *or* be part of the meaning of a term. (Slote does note that his view is different from Kripke's, but he's talking about a different difference.) For Kripke, competent speakers need not know how reference is fixed. But Slote wants to have both. He says that "empathy and approval enter into the *meaning* of 'right'", because "it is part of the very *meaning* of "right" that its reference is fixed in relation to empathy and (what our theory says constitutes) moral approval, that it *is understood as* *referring to* a property that causes or tends to cause approval, that is, warm feelings toward agents resulting (in the ways described previously) from the operation of empathy" (2010, 67, my emphases).

I've observed that combining a priority and reference-fixing isn't a Kripkean view. But is it nevertheless a coherent picture of how language works? It may be, if some suitable form of generalized two-dimensional semantics is correct. Let me give a very brief outline of this complex view here. In standard possible worlds semantics, each expression has an intension and an extension. For singular terms, the extension is an individual, and the intension is a function from possible worlds to individuals. In the case of a definite description, the intension picks out a different individual in different possible worlds – for example, in some possible worlds, 'the strongest man in the world' picks out me (or my counterpart). But as Kripke argued, when it comes to names and natural kind terms, they pick out the same individual or kind in each possible world (in which they exist). So 'water' always picks out H_2O, given that water is actually H_2O. So in one sense, 'water' and 'H_20' have the same intension as well as extension. Yet it seems obvious that they nevertheless have a different meaning – you can think that there's water in your bottle without thinking that there's H_20 in your bottle. As David Chalmers (2006) puts it, even if it isn't *metaphysically* possible that water is H_2O, it is *epistemically* possible – in one sense, we could discover that the colourless liquid in seas and rivers, or the watery stuff for short, isn't H_2O after all. The sense of 'could' here is that this isn't ruled out *a priori*. To capture this possibility, two-dimensional semanticists propose that each expression has a second kind of intension as well, which is a function from possible worlds to the first kind of intension. The idea is that *if* the watery stuff around us is H_2O, *then* 'water' has the intension that picks out H_2O in every possible world. This has the consequence that on Putnam's Twin Earth, on which the watery stuff is XYZ, there is no water. However, if the watery stuff in the actual world had turned out to be XYZ (which is not ruled out a priori), 'water' would pick out XYZ in every possible world.

This second kind of intension, which Chalmers calls the *primary intension*, is something like a reference-fixing criterion or rule, which we can know without knowing what the world is like, and consequently without knowing what the referent is. In this vein, Frank Jackson (1998) claims that in one sense of 'meaning', what it is to know the meaning of 'water' is to know that it refers to the clear and potable liquid that actually rains down from the sky and fills lakes and oceans we're acquainted with, or watery stuff, for short. (Equivalently and less ambiguously, as a matter of contingent fact, the term 'water' refers to the watery stuff we're acquainted with, given the way we have fixed its reference.) This is something you can know *a priori*, without knowing how things actually are (Jackson 1998, 51). In contrast, it is an a posteriori discovery that 'water' refers to H_2O, because the primary intension of H_2O is different from that of 'water'. But this discovery is only possible, because we know *a priori* that *if* X is the watery stuff (specified in qualitative terms), X is water. As Chalmers sees it, the primary intension of an expression is grounded in its inferential role in this kind of fashion – roughly, in our dispositions to apply it to scenarios when they are considered as actual. This means that primary intensions need not correspond to descriptions, though it may be possible to roughly characterize them using them, as I have done in the case of 'water'.

If Slote adopts something like this semantic model, he can coherently claim that it is a priori that moral goodness is whatever actually tends to cause warm empathic feelings in us, and that this is something any competent speaker is in a position to know, because they know the relevant reference-fixing rule. The question, then, is why we should think that the primary intension of 'morally good' is something like "whatever tends to give rise to warm empathic feeling toward agents in us". (Call this thesis Empathy Fixation.) I think Slote's argument is along the following lines. First, it seems he takes it to be a priori that we fix the reference of "morally good" to those things that tend to give rise to moral approval in us. (Call this Approval Fixation.) Second, according to Slote moral approval consists in empathic warmth derived from and directed toward an agent's motives (Call this Empathic Approval). He clearly believes that this claim in moral psychology is a priori knowable. In effect, Slote's argument goes from Approval Fixation via Empathic Approval to Empathy Fixation. And taking into account the fact that we can only come to have warm feelings towards an agent *via empathy* if the agent herself has warm empathic feelings toward someone else (and that we can also know *this* a priori), it follows that moral goodness consists, necessarily and a priori, in warm empathic feelings. (This last step is the difference from standard reference-fixing accounts that Slote acknowledges.) In other words, the second premise of Empathy As Moral Sense is true, in spite of its counterintuitiveness.

3. Slote's Burdens

This is an ambitious chain of argument, to say the least. Let's begin assessing it by observing that departing from a pure reference-fixing account to a hybrid two-dimensionalist view, or something like it, means that Slote incurs the same explanatory burdens as the analytic naturalist accounts, which he rejects. In particular, because so much is loaded in the *meaning* of 'morally good', the following questions are only seemingly open:

> A's motives tend to cause moral approval in us, but is A morally good?
>
> A tends to act out of empathic motives, but is A morally good?

I submit that these are genuinely open questions – indeed, the first is very close to the kind of example that Moore (1903) considered. But if it were part of the meaning of 'morally good' that whatever tends to cause approval in us is good, there would be no open question here.

Any competent speaker, someone who fully grasps the meaning of the term or content of the concept, would be in a position to know that what tends to cause approval in us is morally good, and indeed that nothing else is. So anyone who thinks the question is open would be conceptually confused or at least failing to exercise their competence. I do not see any reason to think so. It seems blatantly obvious that we might morally approve of things that are not morally good. Slote might, to be sure, claim that this is because of a mistaken conception of what it is to morally approve of something, but as I'll soon argue, his view of moral approval is equally questionable. So Approval Fixation is probably false, and certainly not a priori. It is a more clearly first-order moral question whether acting out of empathic motives is good, but it doesn't look like one of 'moral fixed points', to use Terence Cuneo and Russ Shafer-Landau's (2014) expression, even if such things exist. For Cuneo and Shafer-Landau, moral fixed points are conceptual truths that must be included in any system of norms that qualifies as moral (for people like us in a world like ours), such as "It is pro tanto wrong to break a promise on which another is relying simply for convenience's sake" and "If acting justly is costless, then, *ceteris* paribus, one should act justly". When people like Jesse Prinz (2009) and Paul Bloom (2016) criticize acting from empathy, they may be making a normative mistake, but there's no reason to think they're making a *conceptual* error as well. So this, too, is an open question and cannot be settled a priori. Consequently, Empathy Fixation isn't a priori knowable (and may be false).

What is more, Slote's related necessity claims regarding moral approval rule out some genuine possibilities, so they can't be true either. Consider the following scenario:

> *Immanuel Can't*
>
> Joachim is a shopkeeper who always handles the customers' money honestly, simply out of sense of duty and without any warm feeling toward the customers, since he happens to have a dour and cold personality. Immanuel morally approves of Joachim

> and his motives. He recommends Joachim's shop to everyone and defends his honesty if others criticize him.

This seems to be to be a perfectly possible situation. But for Slote it isn't. If moral approval consists in warm empathic feelings towards an agent, and we can have such feelings only toward an agent who herself has warm feelings towards others, it is *impossible* for us to morally approve of anything other than empathic warmth. Note that the claim is *not* a *normative* one – it isn't that someone like Immanuel is mistaken. It is that someone like Immanuel *cannot exist*. And this is very hard to believe. Empathic Approval must be false.

And indeed, we have independent reason to think that Empathic Approval is not true. I think the following kind of platitudes we have regarding moral approval yield desiderata for any theory of (non-judgmental) moral approval:

a) It is possible to morally approve/disapprove of something without believing it is morally good or right/bad or wrong.

b) Moral approval is an attitude with intentional content – it involves construing some X as good or right (or, possibly, in some related non-conceptual way).

c) There is a difference between thinking that something is good and thinking that it is right (permissible).

d) If you think something is *morally* right or wrong, you care about other people's response to it.

e) Normally, morally disapproving of something motivates you to avoid doing it.

f) At least some forms of moral disapproval address an RSVP to the target.

I think most of these are fairly self-explanatory. I'll just note that (d) is one feature of moral approval or disapproval that distinguishes it from other kinds of approval, for example mere liking. It is logically possible to like something without caring whether others do likewise. But if you morally approve of something, you're not indifferent to others disapproving of it (Blackburn 1998) – though you might of course fail to do anything about such third parties. The platitude (f) is based on Stephen Darwall's (2007) insights regarding the second-personal nature of morality. He highlights the importance of especially negative moral attitudes addressing a claim or demand to their target, where the content of the demand is acknowledging past wrongdoing in attitude (by repenting) and action (by making amends).

Does Slote's account meet these desiderata better than competing views? I don't think so. Here is what he says in some more detail about approval and disapproval:

> If agents' actions reflect empathic concern for (the well being or wishes of) others, empathic beings will feel warmly or tenderly toward them, and such warmth and tenderness empathically reflect the empathic warmth or tenderness of the agents. (2010, 34-35)
>
> If a person's actions toward others exhibit a basic lack of empathy, then empathic people will tend to be chilled (or at least "left cold") by those actions, and I want to say that those (reflective) feelings toward the agent constitute moral disapproval. (2010, 35)

Here approval and disapproval consist of feelings with a certain presumably metaphorical temperature. For Slote, the causal history of these feelings is essential to what they are, and in particular their intentional content:

> Even if the empathic warmth the approver feels for the warmth of someone concerned about others takes in that very concern and is to that extent focused on those others, its immediate source is the agent, not those the agent is concerned about, and what approval is approval of depends to a substantial extent, I think, on such

> causal matters rather than on pure phenomenology or seeming intentionality. (2010, 39)

While this account satisfies the first desideratum, I believe it fails to meet the desiderata (b) to (f). Let me start with intentionality, which Slote himself explicitly treats as a desideratum. I believe that the empathic feelings he talks about either lack aboutness altogether or have the wrong kind of content. What, after all, are the ‘warmth’ and ‘chill’ that he talks about? The most natural take is that they are *sensations* of some kind, just like literal warmth and chill. When I think about the kind actions of some people, I do get a warm feeling, which is a form of pleasure. Similarly, chill seems to be some kind of displeasure we get when we put ourselves in some people’s shoes – although, to be honest, it is hard to see this as a chill caught from the agent, as Slote does. (The wrongdoer might be quite pleased with himself.) In any case, such sensations or hedonic feelings are not *about* anything. They can’t constitute approval *of* something, because they have no intentional content. They don’t represent or present the world as being in one way or another, so they can’t be accurate or inaccurate, not to mention true or false. To be sure, there are people like Fred Feldman (2004) who argue that there are attitudinal pleasures, which we express by saying that we are pleased about something. But as I’ve argued elsewhere, the so-called ‘attitudinal pleasures’ are in fact just various positive emotions rather than kinds of pleasure (Kauppinen 2013b).

What else could warmth and chill be? Perhaps desires, which are mental states that do have a direction of fit. But they lack the inherent phenomenal quality that is essential to Slote’s view, and in any case we don’t necessarily take on a helpful person’s desires when we morally approve of her. Could they be emotions or some other kind of non-cognitive attitude, then? Here again we must bear in mind that warmth or chill must be something we can catch empathically from the agent – whatever our state is, it must be identical with (or at least similar to) that of the agent. It is not possible on Slote's view that disapproval would consist of, say, resentment, since the agent we disapprove of might not act out of resentment toward others. They might, for example, act out of greed. But empathizing with greed doesn't amount to disapproval. So it's very difficult to say what warmth and chill actually are beyond metaphorical terms.

Could Slote say that they are *sui generis* states that get their intentional content from their causal origin? I guess something like this is suggested by what he says. But as I've argued elsewhere, the general principle that feelings are directed toward their empathic origin results in absurdities. Suppose you're angry with Joe, and I empathize with you. If Slote's principle were true, I would consequently be angry with you. So his principle can't be true. It is possible, to be sure, to maintain that it's only *sui generis* states of warmth and chill that have intentional content in virtue of their causal origin. But this is difficult to see as anything but an *ad hoc* move.

I will discuss the rest of the desiderata more briefly, since problems relating to them have been raised by other commentators before (e.g. D'Arms 2011, Prinz 2011, Stueber 2011). Regarding (c), it seems Slote's account can't distinguish between approval of something as good and approval of something as right. After all, there's just one feeling of empathic warmth involved in each case, and it is directed towards the same object, the agent. Regarding d, if moral approval or approval distinctively involves attitudes toward third parties as well, it cannot be second-hand warmth, since the original agent's empathic response to, say, the person they're helping does not in any way refer to third parties. Next, as Karsten Stueber (2011), for example, has pointed out, feelings of warmth towards others, and perhaps even more clearly cold feelings towards wrongdoers, do not seem to play the same motivational roles as moral approval and disapproval do. If I'm chilled by someone's motives in pursuing their own profit, it doesn't, sadly, follow that I'm necessarily in any way motivated to avoid pursuing my own profit. And finally, as Neil Roughley (2015), among others, has emphasized, feelings of chill are not second-personal reactions, and consequently do not address a demand to the agent.[2]

In sum, like many others, I am very skeptical of understanding moral approval and disapproval in terms of empathic warmth or chill, in part because it seems no emotional response we can take on from an agent is going to be able to play the roles that moral attitudes do. I'm not going to defend a positive account of them here, but to give a rough idea, I believe that moral sentiments consist in dispositions to praise or blame combined with an authority-independent normative expectation that others share the same disposition (Kauppinen 2013a). These dispositions will be manifest in different emotions, such as admiration, pride, anger, or guilt, in different contexts. Occurrent emotions constitute moral approval or disapproval if and only if they manifest such sentiments. For example, anger as such doesn't involve seeing the target as having done something wrong. But anger that involves a normative expectation that others share it, and that results from a disposition that would manifest itself as pleasure were the agent punished for her action, does constitute an appearance of wrongness. On this view, empathy does not play a constitutive role in moral attitudes – it is possible for someone who lacks empathy to morally approve or disapprove of others, and possible for anyone to do so on occasions on which they do not empathize.

4. Conclusion: Empathy's Roles in Moral Epistemology

Let's assume that what I've said suffices to show that moral virtue doesn't consist in empathic warmth or disposition for it, or at any rate that Slote's metaethical arguments fail to establish

[2] In fairness to Slote, he has responded, or attempted to respond, to at least some of these criticisms, for example in Slote 2013. I am not convinced by the responses he sketches. For example, he thinks anger cannot amount to moral disapproval, because it is a 'hot' emotion. I disagree with the underlying assumption here, but also want to note that indignation or righteous anger need not be as hot as non-moral anger, which seems to be what Slote has in mind.

this against the intuitive counterevidence. If this is the case, then even if we grant that empathy is a kind of perception of an agent's feelings, it won't amount to moral perception or moral sense of any sort, because knowing an agent's warmth or coldness is not the same thing as knowing her virtue or vice. It doesn't follow, of course, that knowledge of the agent's motives isn't relevant to determining her virtue or vice, or the moral status of her actions. It is just that something more than empathy with the actual motives is needed.

Might empathy in the sense of taking on another's emotional responses play a different role in such process of acquiring a posteriori moral knowledge? Hume and Smith certainly thought so. Both, in different ways, believed that to know whether an agent is virtuous, we need to occupy a perspective that is not only informed but also impartial or general when we have emotional responses that constitute appearances of virtue or vice. (Indeed, it might be that our emotions only constitute moral appearances when they are felt from such a common or general point of view, as Hume suggests.[3]) And both thought that to achieve such a perspective, we need to depart from (or enlarge) our first-personal point of view by way of empathically seeing things as certain kinds of other people see or would see them.

Consider the case of the honest shopkeeper Joachim, for example. I've already observed that we can't come to have feelings of approbation towards Joachim by empathizing with his feelings. But it might be that his honesty, if known, gives rise to appreciation or trust or gratitude in those who buy things from him. If we empathize with these feelings, we will also come to feel positively towards him, and this positive feeling might constitute moral approval. But of course, it might be that Joachim's motives remain unknown to his customers and everyone else, or it might be that the customers are prejudiced toward his kind, or selfish

[3] "Tis only when a character is considered in general, without reference to our particular interest, that causes such a feeling or sentiment, as denominates it morally good or evil." (Hume, *Treatise*, 472) I used to agree with this view (see Kauppinen 2010), but have come to think that we can have genuine (if often inappropriate) moral feelings without empathizing.

and ungrateful. So empathy with the actual feelings of patients or observers of an action will not be a source of moral knowledge either, because the actual feelings of people are not a reliable guide to what is good or bad. But we can also imaginatively place ourselves in the shoes of a patient or observer who lacks such defects, and come to feel some praise-emotion, and form a normative expectation that others feel the same way if acquainted with the facts of the case. This is roughly what Hume says when he emphasizes that moral approbation is properly felt from a common point of view, and what Smith means when he says that we must adopt an impartial observer's perspective when judging actions.[4] Such feelings of moral approval are broadly empathic – empathic because they result from putting ourselves in the shoes of another and taking on their feeling, but only broadly so, because they involve taking on the hypothetical rather than actual emotional responses. Drawing on psychological research on emotion regulation, I've used the term *ideal-regulated empathy* for this kind of process (Kauppinen 2014). My contention is that if we form the true belief that Joachim is at least somewhat virtuous on the basis of informed and ideal-regulated empathic approval, the quasi-perceptual emotional appearance or intuition that he is at least somewhat virtuous provides defeasible justification for the belief, and it is not a lucky accident that the belief is true.

So I believe this kind of broadly empathic process is potentially a source of a posteriori moral knowledge, and not only about individual agents or actions, but also about the rightness or wrongness of act-types. If this is along the right lines – and I do not pretend that I've offered an adequate defense here – our capacity for empathy will have an important

[4] Beyond this broad similarity, the details of Hume's, Smith's, and my own view differ considerably. I discuss the relationship among these accounts in Section 2 of Kauppinen 2014.

and perhaps irreplaceable role in moral knowledge, even if it does not constitute a moral sense.[5]

References

Blackburn, Simon 1998. *Ruling Passions*. Oxford: Clarendon Press.

Bloom, Paul 2016. *Against Empathy: The Case for Rational Compassion*. New York: HarperCollins.

Chalmers, David 2006. Two-Dimensional Semantics. In E. Lepore and B. Smith (eds.), *Oxford Handbook of the Philosophy of Language*. New York: Oxford University Press.

Cuneo, Terence and Shafer-Landau, Russ 2014. Moral Fixed Points: New Directions for Moral Non-Naturalism. *Philosophical Studies* 171(3): 399–443.

D'Arms, Justin 2011. Empathy, Approval, and Disapproval in Moral Sentimentalism. *Southern Journal of Philosophy*, 49 Supplement 1: 134–141.

Darwall, Stephen 2007. *The Second-Person Standpoint*. Cambridge, MA: Harvard University Press.

Horgan, Terence and Timmons, Mark 1992. Troubles on Moral Twin Earth: Moral Queerness Revisited. *Synthese* 92: 221–260.

Jackson, Frank 1998. *From Metaphysics to Ethics*. New York: Oxford University Press.

Kant, Immanuel 1785/1999. *Groundwork of the Metaphysics of Morals*. In Mary J. Gregor (ed.) *Practical Philosophy*. Cambridge: Cambridge University Press, 37–108.

Kauppinen, Antti 2010. What Makes a Sentiment Moral? *Oxford Studies in Metaethics* 5_ 225–256.

[5] Thanks to Lilian O'Brien, Neil Roughley, Tuukka Tanninen, and participants at a moral psychology seminar at the University of Duisburg-Essen in December 2016 for helpful comments and discussions on the issues discussed in this paper.

Kauppinen, Antti 2013a. A Humean Theory of Moral Intuition. *Canadian Journal of Philosophy* 43 (3): 360–381.

Kauppinen, Antti 2013b. Meaning and Happiness. *Philosophical Topics* 41 (1): 161–185.

Kauppinen, Antti 2014. Empathy, Emotion Regulation, and Moral Judgment. In Heidi Maibom (ed.), *Empathy and Morality*. New York: Oxford University Press.

Kauppinen, Antti 2015. Hate and Punishment. *Journal of Interpersonal Violence* 30 (10): 1719–1737.

Kripke, Saul 1980. *Naming and Necessity*. Cambridge, MA: Harvard University Press.

Moore, G. E. 1903. *Principia Ethica*. Cambridge: Cambridge University Press.

Prinz, Jesse 2011. Against Empathy. *Southern Journal of Philosophy* 49 Supplement 1: 214–233.

Roughley, Neil 2015. On the Objects and Mechanisms of Moral Approval and Disapproval. In *On* Moral Sentimentalism. Newcastle Upon Tyne: Cambridge Scholars Publishing, 28–40.

Slote, Michael 2007. *The Ethics of Care and Empathy*. New York: Oxford University Press.

Slote, Michael 2010. *Moral Sentimentalism*. New York: Oxford University Press.

Slote, Michael 2013. *From Enlightenment to Receptivity: Rethinking Our Values*. New York: Oxford University Press.

Stueber, Karsten 2011. Moral Approval and Dimensions of Empathy: Comments on Michael Slote's Moral Sentimentalism. *Analytic Philosophy* 52 (4): 328–336.