Sentimentalism, Blameworthiness, and Wrongdoing

Antti Kauppinen

Final draft, August 17, 2015

For Karsten Stueber and Remy Debes (eds.), *Ethical Sentimentalism*. Cambridge University Press.

It is relatively easy for a sentimentalist to make the case that someone is, say, admirable or enviable if and only if it is fitting to admire or envy her. And it's plausible that whether such attitudes are fitting depends on something about those attitudes themselves, or even facts about human tendencies to respond to certain things in certain ways, as sentimentalists say, although this is much more controversial. But for sentimentalism to be a serious rival to competing theories in metaethics, it must also account for central moral properties such as blameworthiness and praiseworthiness of agents and rightness and wrongness of actions in terms of some kind of emotional response. In this paper, I will focus on two tough questions for such *ambitious neo-sentimentalism*. The first concerns the basis on which blaming attitudes are fitting. There's broad agreement that (reactive) blame is merited on the basis of the quality of the agent's will, but it's not trivial to delineate genuinely sentimentalist criteria for a blameworthy quality of will. The second issue concerns the relationship between blame and wrongdoing. While it is *agents* who are to blame, it is *actions* that are right or wrong. The question, then, is how fitting attitudes towards agents can ground the moral status of actions as obligatory, permissible, or impermissible.

On the issue of fitting blame, I begin by sketching a general sentimentalist picture of what makes any attitude fitting. Roughly, on my ideal subjectivist view, fittingness of an attitude is determined by the nature of the attitude together with the related desires of a suitably idealized subject. Blame, I claim, consists of negative reactive attitudes that target an undesirable quality of the will of the agent. I argue that the best motivated characterization of a morally authoritative subject is of a well-informed and self-respecting rational agent who serially occupies the position of each affected party in imagination and gives equal weight to the strength of emotional responses in everyone's shoes without aggregating the numbers – for short, a Nagelian Imp. The consequence of these views for the question of fitting blame is that someone is blameworthy when and because she has a quality of will that a Nagelian Imp wouldn't want her to have – roughly speaking, when she fails to sufficiently constrain or shape her planning by the will of others.

Second, I address a key challenge to ambitious sentimentalism. When it comes to rightness and wrongness, it is fairly clear that an action can be wrong or impermissible even if the agent is not blameworthy. It is, after all, possible for a person to have an *excuse* or be exempt from blame, even if what they do is wrong. But it is tempting to think, as Allan Gibbard (1990) and Stephen Darwall (2006) do, that if someone either is blameworthy or *would be* blameworthy as a result of an action if they lacked an excuse or exemption, then the action is wrong. Call this the Blame-Wrongness Link. Many philosophers have recently challenged it. In particular, they have argued that people can be blameworthy for permissible actions. If so, it seems the link between assessing an agent's quality of the will and the permissibility of her action is decisively broken. How, then, could sentimentalists hope to ground the moral status of actions on attitudes towards agents?

In response to this challenge, I grant that there are cases of blameworthiness without wrongdoing. Nevertheless, I maintain that even in these cases, the permissibility or impermissibility of actions indirectly hangs on fitting attitudes based on a possible quality of will of the agent. Roughly, when it is *possible* to perform an action while meeting all the conditions of accountability without meriting blame, it is permissible. When it is *not possible* to perform an action without meriting blame, unless one has an excuse, it is impermissible. In such cases, there is what I call an *external support* for blame.[1] If this move, call it the *Indirect* Blame-Wrongness Link, is successful, questions about the agent's possible quality of will (and external support for blame) turn out to be central to normative inquiry.

The Indirect Blame-Wrongness Link has interesting normative consequences. In particular, since one's quality of will is centrally a matter of the kind of plans one acts on in the circumstances, and there are often several possible plans for bringing about the same outcome, blameworthiness turns out to depend not only on *what* one intends to bring about, but also on *how* one intends to bring it about. Since permissibility or impermissibility, in turn, is determined by *possible* quality of will, it will depend on how it is *possible* to bring about an outcome; that is, on the causal paths available to the agent and how other agents are involved in the causal paths, not just on the impersonal desirability of the possible outcomes. For example, if the only way to bring about an impersonally desirable outcome requires subordinating the good of one to the good of many, as in the Footbridge trolley case (pushing a heavy person on the tracks to save five others), a Nagelian Imp will predictably resent the agent who chooses the option (insofar as she is accountable), and the act will be impermissible. Roughly, this is because a Nagelian Imp's attitude depends on the *strength* of emotional response in each affected person's shoes, not on the *number* of people who have the response, and that the resentment caused by imagining being used as a means to advance another's good at the expense of a like good of one's own is stronger than the resentment caused by failure to be saved. In short, an independently motivated sentimentalist account of moral metaphysics turns out to provide indirect support for nonconsequentialist normative

[1] This proposal builds on the work of Frances Kamm and others.

[2] In tribute to R.M. Hare, who laid out this kind of approach very clearly (Hare 1981).

[3] The notion of self-respect I'm using here isn't meant to be heavily theoretically loaded – just the ordinary sort of opposite of self-abasement.

[4] This view of blame is defended, among others, by Strawson (1963) and Wallace (1996). It is challenged by Scanlon (2008). I have argued that there's good reason to think that blame comes in both varieties, since forgiveness does, too (Kauppinen forthcoming).

[5] The echoes of Kant's Formula of Humanity are obviously not a coincidence.

[6] For discussion of such cases, see Capes 2012.

[7] As Remy Debes pointed out, sometimes following rules itself displays my attitudes towards other people.

theory – although some might reasonably argue that this shows only that metaethical and normative projects must be pursued hand in hand.

1. Fitting Attitudes and Actual Sentiments

While one can be a sentimentalist about moral judgment or epistemology, in this paper I focus on *metaphysical sentimentalism*. Metaphysical sentimentalists hold that moral properties are grounded in a subject's sentimental response. As Hume puts it in summarizing his view,

> The hypothesis which we embrace is plain. It maintains, that morality is determined by sentiment. It defines virtue to be whatever mental action or quality gives to a spectator the pleasing sentiment of approbation; and vice the contrary. (Hume 1777/2006, 270)

When Hume talks about a spectator, he doesn't mean just anyone. As other passages make clear, he's thinking of someone who adopts a common or general point of view when considering the consequences the agent's motives or traits tend to produce. It is the approbation of such subjects that determines what is a virtue.

While Hume is happy to talk about the relationship between sentimental responses and the properties that give rise to them in causal terms, such reductive approaches are rejected by most contemporary sentimentalists. According to the neo-sentimentalist view, insofar as moral properties are determined by our attitudes, it is *fitting* attitudes that count. We can't simply observe what *causes* approbation to discover what is right or good, but must reflect on what *merits* it. The fact that we're disposed to approve of something doesn't mean it's good or right, since it is always an open question (as Moore said) whether something we're disposed to approve of is good. This is at least in part because moral properties are *normative* for us in a way that powers to cause responses in some subjects are not.

There is a lot to be said for this new orthodoxy. Nevertheless, I want to push back a little bit. There is nothing sentimentalist about Fitting Attitude (FA) accounts as such. After all, non-naturalist intuitionists such as A.C. Ewing (1947) have endorsed FA analyses. For all that FA says, the fact that resentment, say, is fitting towards an agent might be a primitive mind-independent fact that obtains regardless of what sentimental responses people have or could have. The distinction between a generic FA analysis and a *neo-sentimentalist* FA analysis must lie in there being some significant role for sentimental responses or dispositions in making a response fitting. When we ask whether a joke merits amusement, for example, the correct answer will have *something* to do with what actually amuses people. At the extreme, it makes no sense to say that something merits amusement if you can't get anyone to laugh at it.

What, then, makes an attitude fitting for a sentimentalist? I'll set aside nominally sentimentalist views that appeal to the correctness or accuracy of a belief or presentation contained in an emotional response (e.g. Tappolet 2011). Such accounts presuppose that there are attitude-independent facts that make the belief or presentation correct. How, then, should sentimentalists understand fittingness? My own hypothesis, which I can only sketch here, is that evaluative attitudes have both descriptive content and normative content suitably related to the descriptive content. The descriptive content presents things as being in a certain non-normative way, and the normative content refers to the related desires of a subject who is authoritative in one way or another. For example, the content of admiration is something along the lines that the target has done something challenging that is of great worth. It is an ordinary empirical question whether the agent has done something challenging for her, but whether it is of great worth is a different matter. My suggestion is that here we need to turn to the desires of, say, an aesthetically or morally authoritative subject, and ask, roughly, whether such a subject would desire that the target make the challenging effort and desire that others emulate the target. So whether it is morally fitting to admire Mahatma Gandhi (or more specifically his non-violent campaign for independence) depends both on whether what he did was difficult and whether any morally authoritative subject would want him to have done so and want others to do likewise (mutatis mutandis, of course). On this picture, whether the attitude is fitting is jointly determined by non-normative facts and the relevant reactions of one or another kind of authoritative subject. In virtue of the latter feature, evaluative or normative properties are ontologically subjective or attitude-dependent: attitudes come first in the order of explanation.

The strategy of approaching normative issues in terms of the reactions of an idealized subject, or of our responses when we're at our relevant best, was characteristic of traditional sentimentalism of Hume and Smith. Such approaches have not been popular recently. In particular, non-trivial, non-morally specified impartial spectator or ideal observer views in ethics face many serious objections. Even if we leave aside general objections to naturalistic views in metaethics, as I will do here, there are specific challenges that Rawls (1971), for example, nicely pointed out. His objection to impartial spectator accounts can be formulated as a kind of dilemma: Either the impartial spectator is underspecified, so that no normative conclusions follow from the metaethical account, or the account conflates impartiality with impersonality.

Here it is crucial that there are many alternative and competing ways of specifying a morally authoritative subject. What Rawls had in mind was a specification of the impartial spectator that I'll call the Utility Hare.[2] The Utility Hare takes on the preferences of everyone as his own and desires the action or adoption of the rule that maximizes his

[2] In tribute to R.M. Hare, who laid out this kind of approach very clearly (Hare 1981).

expected utility. On such views, famously, the distinction between persons disappears, as it is just the satisfaction of preferences that matters, regardless of who gets the benefit and who pays the cost.

One might reject such an account on the basis of its normative consequences, as Rawls does. But I think the Utility Hare is independently implausible as a specification of a morally authoritative advisor. The point of idealizing, as classical sentimentalists saw it, is to guide our non-ideal sentimental responses so that we can robustly get along without continual strife, on terms that everyone can live with (see Kauppinen 2014). For such guidance, we need to look to or become someone who lacks the features that predictably cause the problems and has features that predictably help human beings avoid them. The fact that we actually defer to or aspire to become certain kinds of agents when making moral judgments reveals that we regard certain perspectives as authoritative. On this view, then, the idealizations involved in an authoritative moral perspective are importantly continuous with our practice. The problem with the Utility Hare is that it isn't a solution to any problem of ours – treating people as distinct individuals whose boundaries matter isn't what gets us into trouble, and aspiring or deferring to a humanly unattainable perspective won't get us out of it. Consequently, it is hard to make the case that we somehow tacitly refer to the Hare's responses when we issue moral verdicts, or that we should accord authority to them while moralizing.

What should we expect a morally authoritative subject to look like, then? Here's what I think is a pretty good model, an impartial spectator I'll call the Nagelian Imp. The Nagelian Imp is an average ordinary everyday self-respecting rational Joe (or Jenny) with some special abilities.[3] He is well-informed and capable of imaginatively occupying the perspectives of each involved in an action and its salient alternatives, and keeping track of

[3] The notion of self-respect I'm using here isn't meant to be heavily theoretically loaded – just the ordinary sort of opposite of self-abasement.

the responses he would have in each position. Those responses depend on two kinds of desire. As a rational and self-respecting agent, the Imp has the kind of basic desires we all do as such. He wants to be able to make up his own mind, and to ensure that, he wants there to be checks and balances against others bypassing or manipulating or incapacitating his decision-making. In short, he cares about his *status* or standing relative to others, as humans generally do. Otherwise, the Imp is chameleon-like, and to some extent takes on the contingent desires of the people whose position he occupies, but they take a backseat to his own desires.

As impartial, the Imp doesn't give any more or less weight to responses from any position. But he doesn't aggregate the positive and negative responses into a big bundle. Rather, he prefers the action (or attitude) that is least unacceptable to the person for whom it is most unacceptable, to use Nagel's (1979) famous phrase. That is, if there are possible actions, x and y, that bear on what the Imp wants in the shoes of A, B, and C, he'll compare the strength of the strongest preference against x to the strength of the strongest preference against y. If, say, he disprefers x in A's position more than y in either B or C's position, he'll prefer y to x. Consequently, if the Imp wouldn't want you to do something, it'll be something you wouldn't want done to yourself in somebody else's position, insofar as you are a rational and self-respecting agent with the desires that come with those features. If you did it, someone else would be worse off than you will be if you don't. So the Imp's responses will be authoritative at least to those who care about living with others on terms that can be expected to have sway with them without external sanctions, as long as they, too, are willing to moderate their demands.

I've argued elsewhere (Kauppinen 2014) that the Imp's responses and fittingness of attitudes are connected as follows: an attitude is morally fitting if and only if any morally authoritative subject would endorse it. By endorsing an attitude I mean roughly wanting the subject to manifest it. The underlying assumption is that part of what makes an attitude what it is can be found in what it motivates or disposes us to do or feel – for example, fear essentially motivates us to flee, admiration to emulate, and intention to act. So, for example, admiration is morally fitting in part if and only if and because the Imp would endorse emulating the target.

2. Blameworthiness and Reactive Attitudes

The reason why being right or wrong is a hard case for a sentimentalist is that such deontic properties are not obviously related to our sentiments, and it is particularly tempting to think they are independent of our attitudes altogether. It makes more sense to think we'll never realize that something is wrong than to think we'll never realize something is disgusting, even if we're acquainted with it. But sentimentalists have attempted to make the connection. For example, Allan Gibbard argues that wrongness can be understood in terms of blameworthiness, and other properties in terms of wrongness. Here is how he links blame, blameworthiness, and wrongness:

> A person is *to blame* for an act if it makes sense for others to be angry at him – from a standpoint of full, impartial engagement […] – and if it makes sense for the person himself to feel guilt for what he has done. Morally wrong act are not always acts a person is to blame for; there are excuses. […] *Wrong acts*, roughly, are acts a person would be to blame for if he chose them in a normal state of mind. (1990, 223)

There are two steps here: first blameworthiness is determined by fittingness of blame, and then wrongness in terms of possibly counterfactual blameworthiness. The second step is what I've called the Blame-Wrongness Link. I'll get to challenges to it in the next section, but first, let's focus on fitting blame.

Let us assume that at least the relevant kind of blame consists of reactive attitudes, most centrally guilt and the form of anger that we call resentment.[4] On the neo-sentimentalist picture I sketched, the fittingness of these sentiments depends on their descriptive content and the related desires of suitable subjects. So let's start with descriptive content. First, I want to reject what seems to me to be a deeply problematic orthodoxy, represented by Jay Wallace's influential account. He maintains that "To be in a state of reactive emotion, one must believe that a person has violated some expectation that one holds the person to" (1996, 19). At the same time, he says that "To hold someone to an expectation, I maintain, is to be susceptible to the reactive attitudes in one's relations with the person" (1996, 18). So reactive attitudes are defined in terms of beliefs about normative expectations, and normative expectations in terms of dispositions to reactive attitudes. This "mutual dependence" of attitudes and expectations may not be strictly circular, but it is at least uninformative and confusing. I believe the way out of this problem is to reject the first claim, and hold on to the plausible thesis that to have a normative expectation of someone is to be disposed to negative reactive attitudes when they fail to perform as desired. Thus, the test for whether I normatively expect you to open the door for me when my hands are full of heavy stuff is whether I'm disposed to blame you if you don't open it.

What, then, is the distinctive descriptive content of reactive attitudes? Strawson (1963/2003, 83) was no doubt right in saying that they are responses to the "quality of the will" of their target. My claim is that they present their target as having either a *desirable* or *undesirable* quality of the will. I'm using the term "will" fairly broadly, to cover both the thought process that led the target to their plans or intentions, and the plans and intentions themselves. Two agents who take the same means to the same end (and thus have the same

[4] This view of blame is defended, among others, by Strawson (1963) and Wallace (1996). It is challenged by Scanlon (2008). I have argued that there's good reason to think that blame comes in both varieties, since forgiveness does, too (Kauppinen forthcoming).

maxim in the Kantian sense) may display a different quality of the will, if they've arrived at the intention by way of different practical reasoning. Your *reactive* attitudes towards me target the way my reasoning and plans take the your *own* will – your own plans and reasons – into account. In that sense, they are essentially intersubjective, and will not be fitting towards beings who lack a capacity to take another's will into account in their planning. Consider resentment. Let me borrow D'Arms and Jacobson's (forthcoming) method and try to interpret resentment in terms of its typical elicitors, motivational tendencies, and attenuators. Among other things, we resent being betrayed or misled, being taken advantage of, and failures to reciprocate. Such actions motivate us to take steps that force the target to acknowledge that we matter more than their behavior suggests. As Adam Smith says,

> The object … which resentment is chiefly intent upon, is not so much to make our enemy feel pain in his turn, as to make him conscious that he feels it upon account of his past conduct, to make him repent of that conduct, and to make him sensible, that the person whom he injured did not deserve to be treated in that manner. (Smith 1759–1790/2002, 112)

I think that the best take on this phenomenon is that resentment presents its target A as having failed to appropriately constrain her planning by the subject B's will – that it doesn't count for the target as it should. It is fitting if this is indeed the case. Here the strictly descriptive content is that A's planning isn't constrained or shaped by B's will in way *w*. The response-dependent normative content is that it is morally desirable that A's planning is constrained or shaped in way *w*, and also that it is morally desirable to impose costs on A to ensure that it is (since resentment manifests itself in behavior that imposes some cost on its target to get her to take the subject's will into account). The old-fashioned sentimentalist way of cashing this out is to say that any suitable impartial spectator would want, indeed insist, A to constrain or shape her planning in way *w*. Since I've briefly argued that the relevant kind of impartial spectator is a Nagelian Imp, my proposal is that blame is fitting if and only if and because the agent fails to constrain or shape her planning and intentions as a well-informed Nagelian Imp would want her to do, so that any such spectator would endorse resentment and guilt on the part of the agent.

What I’ve said above is formal and normatively neutral in the sense that it involves taking no stand on what particular things merit blame. Two people who accept the analysis could still disagree about when blame is fitting – indeed, on the kind of neo-sentimentalist account I’ve sketched, this kind of moral disagreement comes to disagreeing about how a Nagelian Imp would react. But as it turns out, accepting a characterization of a morally authoritative subject does have normative consequences, so that the scope for reasonable and informed disagreement is at least reduced. (This is where the metaphysical sentimentalist moves from the ‘is’ of a certain kind of subject’s reactions to the ‘ought’ of fittingness happens.) For given the concerns and capacities of the Nagelian Imp, it is not an open question how he would want people to constrain and shape their other-involving plans.

The Imp, recall, is concerned, first of all, with his standing relative to others. So he wants people to rule out certain options with regard to him. Since he wants to make up his own mind, he wants others not to bypass his capacity to make choices, or manipulate it to someone else’s advantage, or to disable it by depriving him of basic necessities. When his plans rely on others doing their part, he wants those others to at least make an effort, or they will turn out to have acted in such a way that they might as well have manipulated him into doing his part. For brevity, let me adopt the popular, though potentially misleading phrase, and say that he wants others to rule out using him as mere means for their ends, and to make promoting his capacity to pursue form and pursue his ends as one of their own ends.[5] As Adam Smith puts it,

> What chiefly enrages us against the man who injures or insults us, is the little account which he seems to make of us, the unreasonable preference which he gives to himself above us, and that absurd self-love, by which he seems to imagine, that other people

[5] The echoes of Kant’s Formula of Humanity are obviously not a coincidence.

> may be sacrificed at any time, to his conveniency or his humour. (Smith 1759–1790/2002, 112–113)

What's more, the Imp wants others not to be in a position to use him for their ends, which is to say he wants there to be costs for others if they fail to rule out using him. Since the Imp occupies everyone's position serially and keeps track of the responses, adopting the strongest preference as his desire, he will be against anyone using another for her ends, and for protecting everyone from being used or neglected. By the standard proposed here, then, blame will be fitting response to anyone acting on a maxim of treating another as a mere means, or failing to treat them as ends. However, the Imp will not mind if an action benefits one person more than another, as long as this doesn't affect people's relative standing in this fundamental respect. Thus, if someone adopts a plan that aims exclusively at her own benefit or at the benefit of someone she cares about, the Imp does not necessarily desire that she change her plan even if she could do something else that would maximize benefit, considered impersonally. For the Imp is an *impartial*, not an *impersonal* spectator. And impartiality doesn't mean wanting the best outcome, but giving equal weight to everyone's grounds for resentment and other blaming attitudes. As Smith's formulation hints, not every preference given to oneself (or someone else) is "unreasonable", of the sort that would manifest a blameworthy quality of will. I'll return below to what these facts about the Imp's reactions means for the permissibility and impermissibility of actions.

3. Why the Blame-Wrongness Link Doesn't Hold

We now have a sketch of when negative reactive attitudes are fitting on neosentimentalist grounds. If it is indeed the case that an act is wrong if and only if it is fitting to blame the agent, in the absence of an excusing condition, it seems that we have the analysis that sentimentalists need to connect deontic properties with our attitudes. While wrongdoing doesn't entail blameworthiness because of the possibility of excuses and exemptions, blameworthiness does entail wrongdoing, if the Blame-Wrongness link holds. This link is thus crucial to the Gibbardian strategy of moving from fitting attitudes towards agents to the moral status of actions.

Alas, in recent years even nonconsequentialist ethicists have presented powerful arguments against the Blame-Wrongness Link, which consequentialists have always found problematic. Perhaps most notably, Thomas Scanlon (2008) argues that there is a fundamental difference between the two moral dimensions of *permissibility* and *meaning* of an action. On his view, blame and blameworthiness have to do with the meaning of an action, or what it says about a person's relations to others:

> [T]o claim that a person is blameworthy for an action is to claim that the action shows something about the agent's attitudes towards others that impairs the relations that others can have with him or her. (Scanlon 2008, 6)

This is compatible with what I said above about blameworthiness. However, Scanlon goes on the argue that questions of permissibility are strictly distinct from questions of blame, since they concern actions themselves and not the process whereby the agent arrived at the action. Not only is it possible to perform impermissible actions without blame, but it is also possible to be blameworthy for performing permissible actions. For a simple example, consider one of Frances Kamm's variants of the standard trolley case, in which it is possible to divert a trolley threatening to kill five people to a side track on which it will kill one person:

> Suppose that it is a bad person who sees the trolley headed toward the five. He has no interest in saving the five per se, but he knows that it is his enemy who will be the one person killed if he redirects the trolley. He does not want to be accused of acting impermissibly, however, and so while he redirects the trolley in order to kill the one, he does so only because he believes that (i.e., on condition that) a greater good will balance out the death. Hence, he would not turn the trolley unless he expected the five to be saved (Bad Man Case). His redirecting the trolley is still permissible, I believe, though he does it in order to kill his enemy. (2007, 132)

Here Bad Man is surely blameworthy, but his action is morally permissible (assuming it is such in the original trolley case). Nor does it seem necessary for permissibility that Bad Man expects the five to be saved – even if he didn't care about the five at all, turning would still be permissible. Surely one can do the right thing for the wrong reasons, and so merit blame. (For reasons of space, I'll leave aside here other kinds of purported counterexamples to the Blame-Wrongness Link, such as so-called suberogatory actions and doing the right thing accidentally or in the mistaken belief that it's wrong.[6])

Beyond intuitions about cases, critics of the Blame-Wrongness Link and mental state accounts of permissibility in general appeal to a difference between different contexts in which we can ask questions about the moral status of an action (Thomson 1999, Scanlon 2008). One context is criticizing others, in which case the agent's mental states such as intentions matter. But another is *deliberation and advice*. When deliberating, we ask whether an option is permissible or not. Here our own mental states don't seem to matter. In the context of deliberation, we're trying to decide what to do – intentions are the output of such deliberation, not inputs to it. So when we're deliberating, we need criteria for permissibility that don't depend on which intention we end up with afterwards. The same goes for advice to others – it's no use to say that something is permissible if you do it with one intention and impermissible if you do it with another intention. So it seems that permissibility of an action can't depend on intentions, or more broadly on the quality of will of the agent.

Here, in brief, is the challenge these considerations pose for sentimentalism:

1. Reactive attitudes are fitting on the basis of the agent's quality of will.
2. For sentimentalists, facts about moral permissibility and impermissibility are grounded in fitting reactive attitudes.

[6] For discussion of such cases, see Capes 2012.

3. If facts about moral permissibility and impermissibility are grounded in fitting reactive attitudes, they are determined by the quality of the will with which the agent acts.
4. Facts about moral permissibility and impermissibility are independent of the quality of the will with which the agent acts.
5. So, facts about moral permissibility and impermissibility are not grounded in fitting reactive attitudes.
6. So, sentimentalists can't account for moral permissibility and impermissibility of actions.

There are several ways that a sentimentalist might try to avoid the conclusion. The first premise is unassailable and accepted on all sides, and the third seems to follow from the first two. So the first salient possibility is rejecting the second premise. After all, sentimentalists don't think that all normative or evaluative facts are grounded in reactive attitudes in particular. Perhaps that is one way to understand what Hume is up to in his discussion of artificial virtues in *Treatise* 3.2 (Hume 1739–40/2006), when he switches focus from motives to consequences of rules for the general good. On this type of view, facts about permissibility are determined by rules it would be impersonally desirable for people to live by or internalize. Blame might then be fitting derivatively when one culpably violates such a rule. But it would be surprising and radically revisionary if the correctness of reactive attitudes depended on such impersonal quality of the will as following a useful rule. If you frivolously break a promise you've made to me, you've (pro tanto) wronged *me*. My response to you isn't some sort of general disapproval because you've violated a socially desirable rule, but resentment for what *you* did to *me*. What it is fitting for *other* people to feel is different. Were my resentment fitting because of your attitude toward beneficial rules and not because of your attitude towards me in particular, there would be no reason for any asymmetry between my attitudes and those of third parties. My resentment is a *second-personal* response, in Stephan Darwall's (2006) language. Indeed, reactive attitudes are arguably *essentially* second-personal, since they make no sense from the perspective of a detached observer who stands over and above human affairs. Rather, they are potentially fitting in the context of an interpersonal relationship, in which we make demands of each other and expect a response. This seems to rule out the alternative of the fittingness of blame depending on something other than the quality of the will the agent displays towards another person.[7]

To be sure, someone rejecting the second premise might argue that this is just more evidence that blame and permissibility must be judged by different criteria. Perhaps what I've said above about fitting blame is right, but has nothing to do with right and wrong. But this, too, amounts to biting a bullet, since it means that the blameworthiness of someone who does something wrong and the wrongness of her action have nothing to do with each other. To borrow a point Paul Hurley (2006) has made in a slightly different context, morality is supposed to make *demands* of us, not just provide standards for our actions. If what we're to blame for and what it's wrong for us to do match only *accidentally*, it's unclear in what sense morality demands us to avoid wrongdoing. Hurley puts the point in terms of our having *reason* to do the right thing, but on my view, this comes down to our being accountable for giving sufficient weight to doing the right thing in our practical thinking – in short, blameworthiness is conceptually prior to having reasons (see Kauppinen 2015). So

[7] As Remy Debes pointed out, sometimes following rules itself displays my attitudes towards other people. However, insofar as the rules in question are impersonally desirable ones, it is possible that there's a divergence between the attitudes the agent manifests towards people in general and the second-personal attitudes they manifest towards the person they're interacting with. The point I'm making is that it is the latter that makes *reactive* attitudes fitting. A sincere Stalinist who ruthlessly sacrifices a comrade in the name of progress, out of love for humanity, still does something the comrade may rightly resent.

other things being equal, I think it is better to pursue a different approach that doesn't leave it to mere chance that we're blameworthy for bad actions and praiseworthy for good actions.

So instead of denying the second premise, sentimentalists might join those who have resisted separating impermissibility and the agent's mental state, thus rejecting the fourth premise. To be sure, the case is far from closed, but I believe that the kind of considerations Scanlon, Thomson, Kamm, and others have put forward mean that defenders of mental state accounts of permissibility are on the back foot. I will therefore assume in the following that the fourth premise holds.

What is there left for a sentimentalist to do? I believe the best strategy is to reject the third premise. It doesn't, after all, strictly follow from the first two. This is because although reactive attitudes are fitting on the basis of the agent's quality of will, rightness and wrongness might be *indirectly* rather than directly grounded in them. That is, a sentimentalist need not appeal to a blameless quality of will in explaining the permissibility of an action. Instead, permissibility might be a matter of whether it is possible to perform the act in the circumstances without manifesting a blameworthy quality of the will. In such circumstances, even if one is to blame for one's actual quality of the will, one could have performed the same act without meriting blame, so that blame isn't *externally supported*, as I will put it in the next section. It is the external support for blame, or its absence, that will be crucial for explaining impermissibility and permissibility.

4. External Support for Blame

How could blame be somehow supported regardless of the agent's actual state of mind? My proposal is simple: sometimes we find ourselves in circumstances in which it is *not possible* to achieve a certain end by certain means without meriting blame, assuming that we are rational, informed, and lack any other kind of excuse. (I will give concrete examples soon.)

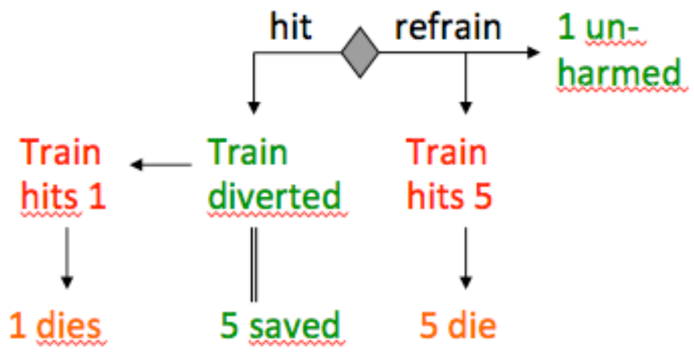
If that is the case, taking the means is morally impermissible. There is external support for blame for A-ing in C, if it is not possible to A in C without being to blame while fulfilling the conditions for accountability. At other times, we find ourselves in circumstances in which it is *possible* for us to achieve a certain end by certain means without meriting blame, even if we are rational, informed, and lack any other kind of excuse. If that is the case, taking the means is morally permissible.[8] There is no external support for blame for A-ing in C, if it is possible to A in C without being to blame, even if one fulfills the conditions of accountability. The key to the solution is that in the latter kind of scenario, even if we merit blame in virtue of our actual quality of the will, we *could have* performed the act we did blamelessly, without needing an excuse. Since external support is missing, the action is permissible, although the agent is blameworthy. Yet the permissibility or impermissibility of the action is still a function of fitness of being held accountable, albeit indirectly. What I call the Indirect Blame-Wrongness Link still holds.

This brief sketch raises various questions. Above all, what is it that determines whether it is possible (or not) to take some means to some end without meriting blame? Clearly, the answer to whether there is external support for blame must depend on what means to what ends are available in the agent's circumstances. To put it in more objective terms, the question is what kind of causal chains the agent can initiate and how desirable the outcomes of these causal chains are. So, whether there is external support hangs on the causal-evaluative structure of the agent's options, or what Frances Kamm, whose work I will draw on in what follows, calls the "objective correlative" of possible plans (Kamm 2007, 136).

Probably the best way to put flesh on these bones is to look at concrete cases, laying out their causal-evaluative structure and examining how it relates to potential

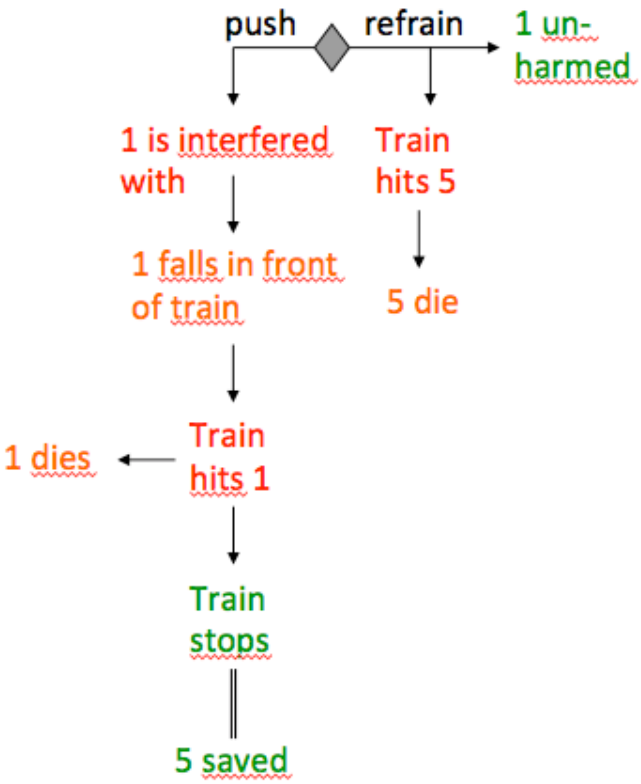
[8] This principle is directly inspired by Frances Kamm's (2007, ch. 5) non-state-of-mind theory of deontological constraints.

blameworthiness. I'm going to focus on the two most basic trolley cases, Switch and Footbridge. They are often criticized for being unrealistic, but they do have the virtue of throwing some morally relevant features of ordinary actions into sharp relief. Here is the first one, Switch:



In Switch, a bystander has the choice of either hitting a switch to turn a runaway train to a side track or refraining from action. If the bystander refrains, the train will hit and kill five people, while one person on the side track will be unharmed. (Note that harmful events, or events that involve someone in non-consensual harm, are marked in red, and good consequences are marked in green.) If the bystander hits the switch, the train is diverted to the sidetrack, which amounts to saving the five. Unfortunately, the train then hits the one person, who will die. Note here that it is the greater good (saving the five) that causes the lesser evil (the harm to and death of the one). In analyzing the case in terms of its causal-evaluative structure, Kamm maintains that what explains the permissibility of hitting the switch in this case that the harm being causally downstream from the benefit (Kamm 2007, 138ff). (I will return below to why and how this matters to permissibility.) She offers many other variants to illustrate the same idea. For example, it is permissible to save five people from being killed by the trolley, even if the result is that they tumble down a cliff and their combined weight kills one person.

Here is the causal-evaluative structure of the Footbridge case:



In the Footbridge case, too, the agent faces the choice between letting five die or killing one. But here the causal structure is different. While refraining from action has the same consequences, the only alternative is pushing a heavy person off a footbridge to stop the train before it hits the five, thus killing the one person. Here the event that involves the one in harm, or lesser evil* in Kamm's terms, is causally necessary to bring about the good event, stopping the train, which in context amounts to saving the five. Here doing bad things to one causes good things to others. Kamm maintains that this is what explains why pushing the heavy person is impermissible. She presents an impressive list of variants, such as Quinn's (1989) case of driving over an innocent person in an ambulance to save the lives of five patients (which involves harm as a side effect rather than means in mental state terms).

It is an obvious question at this point just *why* causal structure should have such moral significance. Again, Kamm does provide a persuasive deeper rationale appealing to the relative status of the people who are involved in the causal relationships. What happens in the Switch case, for example, the One person is *substituted* for the Five others as a victim when the bystander hits the switch. When harm to someone is causally downstream from the good to others, involving the person in harm doesn't serve the interests of the others. The good thing for the Five has already happened when the trolley is turned, and before it hits the One. In this important respect, the One and the Five are in the same position vis-à-vis each other. By contrast, in Footbridge, if the One is pushed down, she is made to serve the others in an important sense. As Kamm puts it, one person is *subordinated* to another, when specifying her causal role in a scenario "makes essential reference to his usefulness to achieving a good for that other person" (Kamm 2007, 165). So causal structure matters morally, because subordination matters morally.

But why does subordination matter morally? Why should it matter what kind of causal role one plays in producing an outcome? Is it just a brute normative fact? Kamm and other pure externalists about permissibility are plagued by such questions. Alastair Norcross, for example, is puzzled by Kamm's non-mental-state account:

> If the notions of respect or inviolability … really just amount to constraints on the permitted causal routes to and between harms and benefits, it is hard to see them as anything other than fetishizing certain causal processes. … How could this be relevant to justified moral concern for others? (Norcross 2008, 16)

This is where sentimentalism promises a further explanation in terms of accountability. When the only causal chain available to bring about an end involves subordinating one person to another, there is no way to pursue the end without gaining the blame of an impartial spectator of the Nagelian Imp type (unless, perhaps, the stakes are high enough). This isn't to say that a morally authoritative subject's reactions are *normatively* fundamental. To that extent, I agree with Kamm's claim that the "objective correlative" of plans (in other words, what I've called the causal-evaluative structure of the possible act) is primary. But the sentimentalist claim is that the fact that initiating certain causal chains is impermissible is *metaphysically* grounded in subjective evaluative reactions. It is no fetishism to care about whether possible actions involve people in ways that any impartial spectator would object to (or that you would yourself object to in the victim's shoes).

Consider the causal-evaluative structure of the trolley cases from this perspective. In section 3, I quoted Kamm's Bad Man case as a counterexample to the Blame-Wrongness Link. The Bad Man, recall, is in the Switch situation, but only turns the train in order to kill the One on the side track. Since we're assuming she's innocent, future life is good for her, and she hasn't consented, Bad Man is surely blameworthy because of his plan. But it would have been *possible* for someone in the Bad Man's place to act with the intention of saving the Five, and without wishing anything ill for the One or using the One as a means to save the Five.[9] In other words, it would have been possible to perform the action without protest from the Nagelian Imp.

In Footbridge, in contrast, it is *not possible* for the agent to involve the One in harm in a way that doesn't merit blame. More precisely, it is not possible for the agent to do so, insofar as she is rational, knows the relevant facts, and meets the general conditions of responsibility. No doubt she'll be more blameworthy if her end is to kill the One, and saving the Five is just a side effect. But even with the best of intentions, she'll have to intend to make the One her instrument, insofar as she's rational. She'll have to act with a quality of will that the Nagelian Imp would want her not to have. To use Kamm's (2007) terms, I think it's a credible conjecture that, other things being equal, if A-ing necessarily involves nonconsensual harm that is not downstream(ish) from a comparable good to more people, any plan involving A-ing, even with the best of intentions, requires making a person serve

[9] I am here assuming that Thomson's (2008) change of heart about the Switch case is a mistake, as FitzPatrick (2009) convincingly argues.

the good of others regardless of their will. Taking into account what I've said about the likely reactions of the Nagelian Imp, it seems that he would endorse blaming the agent for any such plan. Thus, it is very plausible that subordination (among other things), defined solely in terms of the causal structure involving harms and benefits to people, externally supports blame.

To sum up, I think these cases illustrate a two related general theses that we can formulate as follows:

> *External Support for Blame*
>
> Blame for A-ing in circumstances C is externally supported if adopting any plan involving A-ing in C would make blame fitting, assuming the agent is rational, informed, and meets the conditions of accountability.
>
> *Indirect Blame-Wrongness Link*
>
> A-ing in C is morally impermissible if and only if and because blame for A-ing in C is externally supported.

In the interest of keeping the things simple, I won't discuss other cases put forward as a challenge to the Blame-Wrongness link here. But what about the more principled objection that we should firmly distinguish between contexts of critical assessment of the agent, on the one hand, and advice and deliberation, on the other? A proponent of the Indirect Blame-Wrongness Link agrees that we need a notion of permissibility that is distinct from blameworthiness, and that when we deliberate or give advice, we don't consider what quality of the will we would be acting with. Rather, in deliberation and advice we consider the features that would or wouldn't lend external support for blame, for example whether taking certain means to our end would involve subordinating the good of some to the good of others. To take a standard case, if an air force commander asks the Prime Minister whether he should bomb a munitions factory to speed up the end of a just war, when the consequences include killing children in a nearby kindergarten, there's no suggestion that she should reply "it depends on whether killing the children is part of your plan for killing the war or whether it is just a regrettable side effect", although his blameworthiness does depend on his plans (cf. Thomson 1999). Instead, what is pertinent to moral advice and deliberation is whether the bombing is likely to achieve the goal of shortening the end of the war by way of killing the children (and the grief it causes) or by way of reducing available ammunition, and whether the alternative is that even more innocents die if the war is prolonged. So what the Prime Minister should say is that the air force may bomb only if doing so is necessary to avoid sufficiently awful consequences *and* destroying the factory is sufficient to promote this aim without involving innocents in harm (even if they will be harmed in actual fact). She might also say, truthfully but less informatively, that the air force may bomb only if it is possible to arrive at the decision by reasoning that doesn't make a morally authoritative subject resent them. Thus, with the Indirect Blame-Wrongness Link, the crucial distinction between contexts of deliberation and assessment is preserved, even though ultimately the criteria of permissibility derive from fitting reactive attitudes.

5. Conclusion: From Metaethics to Normative Ethics?

In this paper, I've been exploring the possibility of ambitious neo-sentimentalism, a form of sentimentalism that proposes to ground not only evaluative properties like being funny or being shameful but also deontic properties like being right and being wrong in fitting sentimental responses. My sketch is obviously not intended to be the final word on the matter. For one thing, I've only focused on the sentimentalist analysis of permissibility and impermissibility, and haven't said anything about how such fundamental notions as rights or reasons might be grounded in fitting attitudes (though for the latter, see Kauppinen 2015).

But even so, what I've said may seem overambitious to some eyes. I have, after all, breached one of the red lines of contemporary metaethics: I have argued that although we can defend a particular characterization of a morally authoritative subject in non-moral terms, a sentimentalist fitting attitudes analysis has substantive normative implications. This is to move from a metaethical 'is' to a normative ethical 'ought'.[10]

To be sure, there is precedent for such a move. I quoted earlier Hume's summary of his version of sentimentalist metaethics. This is how the passage continues:

> We then proceed to examine a plain matter of fact, to wit, what actions have this influence [of giving rise to approbation in a suitable spectator]: We consider all the circumstances, in which these actions agree: And thence endeavour to extract some general observations with regard to these sentiments. (Hume 1777/2006, 270)

The "general observations" are the normative claims that Hume makes, such as the claim that character traits that are beneficial or immediately pleasing either to the agent herself or others are virtues. This methodological comment fits with Hume's actual practice: he frequently makes observations about the "plain matter of fact" concerning people's reactions in suitable circumstances, and draws conclusions about what is actually right or wrong. The benefit of proceeding in this fashion in doing normative ethics is that there is no appeal to controversial moral intuitions but only to what follows from the putatively neutral foundational facts about morality. In this move, Hume is not alone: on a plausible reading, Kant and modern-day Kantians do just the same, only with a rationalist conception of metaethical foundations.

[10] I have left aside expressivist variants of sentimentalism. However, it is worth mentioning at this point that for a hybrid expressivist like Michael Ridge (2006), the question of who is a morally authoritative subject (an ideal advisor) is already a first-order question – if we disagree morally while agreeing on non-moral facts, our disagreement consists in the fact that we regard different kinds of subject as authoritative. Thus, what my variant of ambitious neo-sentimentalism treats as a metaethical question is already a normative question for Ridge, and the debate about whether the Utility Hare or Nagelian Imp is a better specification of the impartial spectator is fought at least partially on normative ground.

Although the view I have defended is not rationalist, there is a discernible Kantian flavor to the normative conclusions I claim will follow from the best kind of sentimentalist framework. That's no accident, since both my variety of the impartial spectator view and Kantianism focus on the possible quality of the agent's will, specifically the way in which the agent's reasoning and plans take the will of others into account. The key difference is that a sentimentalist is skeptical of deriving the requirement to never treat others as mere means, for example, from the very nature of practical rationality. Instead, it is founded on our contingent (though deeply rooted) tendency to resent those who treat us as having lesser worth, suitably idealized to serve a practical function for social animals like us.

There is a somewhat more modest conclusion to draw here, however. Instead of saying that the best form of neo-sentimentalism supports nonconsequentialist moral theory, we might say that for neo-sentimentalists, the metaethical and normative inquiries must proceed hand in hand. In this framework, someone whose first-order intuitions are consequentialist must dispute the characterization of the morally authoritative subject whose reactions ground moral properties, or argue that it is something like an impartial spectator's preferred rules rather than reactive attitudes that determine what is and isn't permissible. Although I've given some reasons here to go one way rather than another, I can't pretend that those considerations are in any way decisive. In any case, this kind of dispute is not just a matter of headbutting first-order intuitions or finding the best way to systematize them. It requires making the case that the Utility Hare, for example, is after all a more suitable model of a morally authoritative subject than the Nagelian Imp, and doing so requires reflection on such issues as human nature and the function of morality. I believe that shifting focus to such matters – pursuing a sentimentalist metaethics along with normative ethics – may offer a fruitful new approach to old debates concerning first-order questions about right and wrong.

References

Capes, Justin 2012. Blameworthiness Without Wrongdoing. *Pacific Philosophical Quarterly* 93 (3), 417–437.

D'Arms, Justin and Jacobson, Daniel forthcoming. Emotional Fittingness Without Cognitivism.

Darwall, Stephen 2006. *The Second-Personal Standpoint*. Cambridge, MA: Harvard University Press.

Ewing, A.C. 1947. *The Definition of the Good*. New York: Macmillan.

FitzPatrick, William 2009. Thomson's Turnabout on the Trolley. *Analysis* 69 (4), 636–643.

Gibbard, Allan 1990. *Wise Choices, Apt Feelings*. Cambridge, MA: Harvard University Press.

Hare, R.M. 1981. *Moral Thinking*. Oxford: Clarendon Press.

Hume, David 1739–1740/2006. *Treatise of Human Nature*. In Geoff Sayre-McCord (ed.), *Moral Philosophy*. Indianapolis: Hackett.

Hume, David 1777/2006. *Enquiry Concerning the Principles of Morals*. In Geoff Sayre-McCord (ed.), *Moral Philosophy*. Indianapolis: Hackett.

Hurley, Paul 2006. Does Consequentialism Make Too Many Demands, or None at All? *Ethics* 116 (4), 680–706.

Kamm, Frances 2007. *Intricate Ethics*. New York: Oxford University Press.

Kauppinen, Antti 2014. Fittingness and Idealization. *Ethics* 124 (3), 572–588.

Kauppinen, Antti 2015. Favoring. *Philosophical Studies* 172, 1953–1971.

Kauppinen, Antti forthcoming. Character and Blame in Hume and Beyond. In Iskra Fileva (ed.), *Perspectives on Character*. New York: Oxford University Press.

Nagel, Thomas 1979. Equality. In *Mortal Questions*. Cambridge: Cambridge University Press, 106–127.

Norcross, Alastair 2008. Off Her Trolley? Frances Kamm and the Metaphysics of Morality. *Utilitas* 20 (1), 65–80.

Nyholm, Sven 2012. *On the Universal Law and Humanity Formulas*. PhD Dissertation, University of Michigan.

Quinn, Warren 1989. Actions, Intentions, and Consequences: The Doctrine of Double Effect. *Philosophy and Public Affairs* 18 (4), 334–351.

Rawls, John 1971. *A Theory of Justice*. Cambridge, MA: Harvard University Press.

Ridge, Michael 2006. Ecumenical Expressivism: Finessing Frege. *Ethics* 116, 302–336.

Scanlon, Thomas 2008. *Moral Dimensions: Permissibility, Meaning, and Blame*. Cambridge, MA: Harvard University Press.

Smith, Adam 1759–90/2002. *The Theory of Moral Sentiments*. Ed. Knut Haakonssen. Cambridge: Cambridge University Press.

Strawson, Peter 1963/2003. Freedom and Resentment. Reprinted in Gary Watson (ed.), *Free Will*. Oxford: Oxford University Press, 72–93.

Tappolet, Christine 2011. Values and Emotions: Neo-Sentimentalism's Prospects. In Carla Bagnoli (ed.), *Morality and the Emotions*. Oxford: Oxford University Press.

Thomson, Judith 1999. Physician-Assisted Suicide: Two Moral Arguments. *Ethics* 109 (3), 497–518.

Thomson, Judith 2008. Turning the Trolley. *Philosophy and Public Affairs* 36 (4), 359–374.

Wallace, R. Jay 1996. *Responsibility and the Moral Sentiments*. Cambridge, MA: Harvard University Press.