Idealization and the Wrong Kind of Reasons*

John Brunero

I consider Antti Kauppinen's recent proposal for solving the wrong kind of reasons problem for fitting attitude analyses through an appeal to the verdicts of ideal subjects. I present two problems for Kauppinen's treatment of a foreseen objection, and construct a counterexample to his proposal as it applies to the wrong kind of reasons to admire someone. I then show how to construct similar counterexamples to his proposal as it applies to the wrong kind of reasons for other attitudes, including guilt and shame.

According to a simplistic fitting attitude analysis of *admirable*, what it is for someone to be admirable *just is* for admiration to be an appropriate attitude to have towards that person. But this analysis faces the "wrong kind of reasons" or "conflation" problem: it may sometimes be appropriate to admire someone without that person being admirable. For instance, if my admiring an evil dictator would somehow save 100 lives, it would be appropriate for me to admire him. But that doesn't make him admirable.

The fact that it would somehow save 100 lives is the "wrong kind of reason" for admiring someone. In contrast, Jimmy Carter's humanitarian work is the right kind of reason for admiring Carter; it points toward his being admirable. The wrong kind of reasons *problem* is to provide an account of the distinction between the "wrong" and "right" kind of reasons for attitudes in a way that allows for formulations of fitting attitude analyses that aren't vulnerable to such counterexamples.[1] (And one can't simply say what I said above – that the reasons of the right kind are those which point to someone's being admirable – since the whole point of a fitting attitude analysis of *admirable* is to explain that concept in other terms. Saying that the reasons of the right kind are those which point to someone's being admirable would be circular.) In short, we want to say that a person would be admirable only if there are *reasons of the right kind* to admire him. And the challenge is to say, without circularity, what makes a reason be of the right kind.

In this discussion, I'll consider a proposed solution from Antti Kauppinen's "Fittingness and Idealization."[2] Kuappinen aims to show how "old-fashioned sentimentalism" – specifically, the kind of idealizing sentimentalism found in the work of David Hume and Adam Smith, involving appeals to the verdicts of ideal critics, impartial spectators, and the like – can solve three problems: the alleged lack of a tenable motivation for idealizing, a second problem Kauppinen calls "the many kinds of fittingness problem," and the wrong kind of reasons problem. I'm going to set aside the first two and consider whether Kauppinen solves the third. I'll (§1) outline his proposal, (§2) present two problems with a response Kauppinen provides to a foreseen objection, and (§3) show how his proposal has trouble with cases in which it's useful to both *have* and *successfully manifest* a sentimental response that's nonetheless unfitting. Kauppinen's proposal does well in dealing with the examples he considers, but cannot sufficiently generalize to solve the wrong kind of reasons problem.

§1.

On Kauppinen's solution, we must consider whether an ideal moral subject – the impartial spectator[3] on the sentimentalist account – *endorses* the agent's having some attitude or instead merely *approves of the agent having* the attitude. Kauppinen tells us that "in fully endorsing an attitude, the ideal spectator endorses manifesting [the attitude] with success."[4] What this involves depends on the specific attitude. Manifesting admiration with success involves emulating the person admired. Manifesting guilt with success involves making amends. Manifesting fear with success involves fleeing. Kauppinen notes that the ideal spectator "may *approve of* someone's having an attitude without *endorsing* it, if merely having the attitude has beneficial consequences."[5] Return to the case

in which my admiring the dictator would somehow save 100 lives. In Kauppinen's view, the impartial spectator would (merely) approve of my admiring the dictator – since having that attitude has beneficial consequences – but would not endorse my admiring the dictator, since manifesting admiration with success would involve emulating the dictator, and "[it] would be awful if I did the sort of things he does."[6]

Kauppinen thinks that this insight provides the key for discovering the right kind of reasons *in general*:[7]

> I believe the best way to capture the relationship between ideal subjects' attitudes and our (right kind of) reasons is to say that for *r* to be a pro tanto (right kind of) W reason for S to have Y is for any W-ly ideal subject to take *r* to favor S's manifesting Y with success. … By 'manifesting' I mean doing or feeling whatever the attitude essentially motivates or disposes one to do or feel – for example, fear essentially motivates one to flee, shame to hide, guilt to make amends, admiration to emulate, intention to act, and desire to attend to and try to bring about its object. Motivating or disposing one to perform such actions is essential to the attitudes because it is part of what makes them the attitudes they are.[8]

Consider how this applies to the *right* kind of (moral) reasons to admire. Kauppinen provides an example: Mandela's integrity is a right kind of reason for Joan to admire Mandela. And his account predicts this, since "any sympathetic impartial spectator [would] take Mandela's integrity to count in favor of Joan's manifesting admiration by emulating Mandela."[9] And, as we've already seen, it yields the right predictions for some *wrong* kind of reasons to admire: a sympathetic impartial spectator wouldn't take the fact that my admiring the dictator would somehow save 100 lives to favor my manifesting admiration by emulating him.

On Kauppinen's view of the right kind of (moral) reasons to admire, for *r* to be a right kind of moral reason for S to admire someone is for a morally ideal subject to take *r* to favor S's manifesting admiration of that person by emulating her. We aren't given a complete account of exactly when the morally ideal subject would, and would not, take *r* to favor S's manifesting admiration of a person by emulating her. But this much is clear: if the morally ideal subject wouldn't take *r* to favor S's emulating the person, then the morally ideal subject wouldn't take *r* to favor S's manifesting admiration by emulating this person, and so *r* would not be a right kind of reason to admire this person. And that's enough to predict that my reason to admire the dictator – namely, that doing so would somehow save 100 lives – isn't a reason of the right kind, since this fact isn't a reason for me to emulate the dictator.

§2.

Kauppinen anticipates an objection: what if the demon makes it such that to save 100 lives, one must not only admire the dictator, but also emulate him? In this case, it's not clear why the morally ideal subject would not take the fact that 100 lives would be saved by my admiring and emulating the dictator to favor my manifesting admiration with success. And so it's not clear that we have grounds for saying that the 100 lives saved isn't a reason of the right kind to admire the dictator. (And, of course, it *isn't* a reason of the right kind; it doesn't point toward the dictator's being admirable.)

In reply, Kauppinen argues that "without further qualification, an impartial spectator will still not approve of emulating the dictator, since that would mean going around ordering underlings to kill and terrorize people, among other things."[10] Kauppinen then anticipates an objection to this reply: what if I'm not able to harm

anyone, and so emulating the dictator wouldn't cause any harm? In that case, Kauppinen argues, the impartial spectator would only be approving of my *trying* to emulate the dictator, not my *emulating* the dictator. And remember that, on Kauppinen's proposal, for a moral reason *r* to admire someone to be a reason of the right kind, the morally ideal subject must take *r* to favor S's manifesting admiration by emulating – not merely trying to emulate – the person.

Let's return to Kauppinen's reply to the initial objection – specifically, his claim that the ideal spectator "will still not approve of emulating the dictator, since that would mean ordering underlings to kill and terrorize people, among other things." First, note that on Kauppinen's official formulation, quoted in §1 above, we shouldn't be interested in whether the morally ideal subject would approve of emulating the dictator. Rather, we should be interested in whether the morally ideal subject would *take r to favor* S's manifesting admiration with success. And it's not clear that the impartial spectator would *not* take the fact that my manifesting admiration with success would save 100 lives to favor my manifesting admiration with success. Perhaps the impartial spectator would take one fact (that manifesting admiration with success would save 100 lives) to favor manifesting admiration with success, while taking another fact (that manifesting admiration with success would involve ordering underlings to terrorize and kill people) to *count against* manifesting admiration with success. Of course, the ideal spectator would take the latter fact to favor not manifesting admiration with success *more strongly* than the former fact favors doing so. But that doesn't mean the ideal spectator wouldn't take the former fact to favor doing so.[11] And if we think the ideal spectator would take this fact to favor manifesting admiration with success, we lose our license to say that this isn't a reason of the right kind.

A second, more serious, problem concerns a slight variation on the example. Suppose we increase the number of lives saved by my manifesting admiration with success. Perhaps some demon now promises to save a million lives. And suppose the dictator is such that by emulating him, you would cause some minor harms or annoyances, insignificant in comparison to the million lives saved. In Kauppinen's example of Kim Jong-un, emulation involves "ordering underlings to kill and terrorize people." Let's replace Kim Jong-un with someone who also isn't admirable, but is such that emulation doesn't involve anything quite that extreme. Note that in Kauppinen's example of Kim Jong-un, we could say that the morally ideal subject wouldn't take the lives saved to favor manifesting admiration with success, since emulation would involve relatively significant harms. But now that we've altered the example to remove the relatively significant harms, there isn't the same ground for thinking that a morally ideal subject wouldn't take the lives saved to favor manifesting admiration with success. So, we don't have the same ground for declaring that the million lives saved is not a right kind of reason for admiring the dictator.

Perhaps there are some other grounds for declaring this reason isn't a reason of the right kind on Kauppinen's formula – that is, some other grounds for thinking that a morally ideal subject wouldn't take the fact that my manifesting admiration with success would save a million lives to favor my manifesting admiration with success. But Kauppinen's account of the morally ideal subject provides us with no indication of what those grounds might be.[12] And, of course, we can't say that the morally ideal subject would not take this fact to favor my manifesting admiration with success because the fact doesn't point to a way in which the dictator is admirable. That would introduce the kind of circularity that, as we noted earlier, must be avoided.

In summary, Kauppinen's analysis does well for the examples he considers. He is able to distinguish Joan's right kind of reason to admire Mandela from my wrong kind of reason to admire Kim Jong-un. In the former case, a morally ideal subject would take Mandela's integrity to favor Joan's manifesting admiration with success (by emulating Mandela), while in the latter case, a morally ideal subject wouldn't take the 100 lives saved to favor my admiring Kim Jong-un with success (by emulating Kim Jong-un). But if we alter the example so that my manifesting admiration with success would save a million lives and involve the commission of minor harms, then there aren't the same grounds available for saying that a morally ideal subject wouldn't take the fact that a million lives would be saved by my manifesting admiration with success to favor my manifesting admiration with success. And so there aren't the same grounds available for saying that this fact isn't a reason of the right kind to admire the dictator. It appears that Kauppinen's proposal doesn't have the resources to deliver the result that this fact isn't a reason of the right kind. And any successful solution to the wrong kind of reasons problem should be able to deliver this result.

§3

Related to this last concern, we might wonder how well Kauppinen's strategy generalizes to other fitting attitude analyses – perhaps *shameful* in terms of appropriate shame, *blameworthy* in terms of appropriate blame, *guilt-worthy* in terms of appropriate *guilt,* and so forth. Recall that Kauppinen's strategy is to claim that, for reasons of the wrong kind, an ideal subject would not take *r* to favor S's manifesting Y with success. Let's consider reasons to feel guilty about some action you performed in the past. On Kauppinen's view, guilt essentially motivates or disposes one to make amends.[13] If the ideal subject

takes some fact to favor one's feeling guilty, but that fact isn't also taken to favor one's making amends, then it would be a reason of the wrong kind.

But does Kauppinen's strategy account for all of the wrong kind of reasons to feel guilty? One wrong kind of reason to feel guilty is that doing so will help smooth things over. Suppose I presented a criticism of a friend's work in a polite and constructive way, but my friend took it badly, and is upset with me. I've done nothing for which guilt would be appropriate or merited. So, there are no reasons of the right kind for me to feel guilty. But there may nonetheless be a reason – perhaps an insufficient one – for me to feel guilty: in feeling guilty, I will likely be led to express those emotions and perform those actions that will help smooth things over. (Let's suppose I'm so transparent that if I didn't feel any guilt, it's unlikely that I could say, "I'm sorry about having upset you" without coming across as insincere or, worse, sarcastic.) Such instrumental reasons to feel guilt are familiar, but yet they are clearly reasons of the wrong kind, since they don't point toward my guilt as merited or fitting. But Kauppinen's strategy doesn't allow us to dismiss them as reasons of the wrong kind, since we can't say, "But that's not taken by the ideal subject to favor making amends, and so it's a reason of the wrong kind." The fact that it would smooth things over *would* be taken by the ideal subject to favor *both* my feeling guilty and my making amends. So, it's not clear how Kauppinen's proposal provides the resources to exclude such reasons to feel guilty as reasons of the wrong kind.[14]

The problem here is similar to the problem we saw in the previous section. In the modified dictator case (where the demon's incentive requires that I both admire *and emulate* the mostly harmless but unadmirable dictator in order to save a million lives), we cannot say what we said in Kauppinen's original Kim Jong-un case: "But the demon's

incentive isn't taken by the ideal subject to favor emulation of the dictator, and so isn't a reason of the right kind to admire the dictator." The incentive *would* be taken to favor both admiration and emulation of the dictator. And in this case, we cannot say, "But the fact that it will smooth things over wouldn't be taken by the ideal subject to favor making amends, and so isn't a reason of the right kind to feel guilty." This fact *would* be taken to favor both feeling guilty and making amends.[15] So, Kauppinen needs to provide some other basis for predicting that such reasons are not of the right kind. And it's not clear what that could be.

Once we see how these examples work, we could construct similarly structured challenges for other attitudes. For instance, Kauppinen notes that shame essentially motivates or disposes one to hide. If I live in a town full of bigots, one reason for me to be ashamed of my despised sexual preferences is that being ashamed, rather than proud, will motivate or dispose me to behave in ways that will save my family a lot of grief from the bigots. This, of course, is the *wrong kind* of reason to be ashamed, since the fact that it'll spare my family a lot of grief doesn't point to my sexual preferences being *shameful*. But the fact that it'll spare my family a lot of grief would be taken to favor *both* my being ashamed and my doing what shame essentially motivates one to do, namely hide. Indeed, it would be taken to favor my being ashamed precisely *because* being ashamed would motivate or dispose me toward behaviors, like hiding, that would spare them from grief. (It's of course compatible with this that there are weightier reasons of the right kind to be unashamed.) So, the fact that it would spare them grief would be taken to favor my manifesting shame with success. But we still need to deliver the result that this is a reason of the *wrong* kind to be ashamed.

§4.

In conclusion, Kauppinen's proposal does well for the examples he considers (Mandela; Kim Jong-un) but cannot generalize sufficiently to solve the wrong kind of reasons problem. His response to the scenario in which the reward of 100 lives saved is attached to my both admiring and emulating the dictator takes advantage of specific features of that example that could easily be altered. And his proposal cannot deal with similarly structured examples regarding the successful manifestation of other attitudes, including his own examples of guilt and shame. So, we should reject Kauppinen's solution to the wrong kind of reasons problem.

* I am grateful to anonymous referees and the editor of *Ethics* for helpful comments and advice.

[1] This problem has received a good deal of attention, and there are now several proposals (and objections to those proposals) for solving the problem in the literature. Some important discussions of the problem can be found in Justin D'Arms and Daniel Jacobson, "Sentiment and Value," *Ethics* 110 (2000): 722-748, Wlodek Rabinowicz and Toni Rønnow-Rasmussen, "The Strike of the Demon: on Fitting Pro-Attitudes and Value" *Ethics* 114 (2004): 391-424, Philip Stratton-Lake, "How to Deal with Evil Demons: Comment on Rabinowicz and Rønnow-Rasmussen," *Ethics* 155 (2005): 788-798, Jonas Olson, "Buck-Passing and the Wrong Kind of Reasons," *Philosophical Quarterly* 54 (2004): 295-300, Wlodek Rabinowicz and Toni Rønnow-Rasmussen, "Buck-Passing and the Right Kind of Reasons," *Philosophical Quarterly* 56 (2006): 114-120, Pamela Hieronymi, "The Wrong Kind of Reason," *The Journal of Philosophy* 102 (2005): 437-457, Sven Danielsson and Jonas Olson, "Brentano and the Buck-Passers," *Mind* 116 (2007): 511-522, Gerald Lang, "The Right Kind of Solution to the Wrong Kind of Reason Problem," *Utilitas* 20 (2008): 472-489, Mark Schroeder, "Value and the Right Kind of Reason," *Oxford Studies in Metaethics, Vol. 5.*, ed. Russ Shafer-Landau (Oxford: Oxford University Press, 2010), 25-55, and Mark Schroeder, "The Ubiquity of State-Given Reasons," *Ethics* 122 (2012): 457-88.

[2] Antti Kauppinen, "Fittingness and Idealization," *Ethics* 124 (2014): 572-588.

[3] According to Kauppinen, "the optimal moral point of view is that of any informed, impartial, and sympathetic spectator with otherwise normal emotional tendencies" (581).

He acknowledges that this is a thin description. But he thinks that the distinction between the *endorsement* of, and the *mere approval* of, an attitude by the ideal spectator will provide us with enough resources to go on to solve the wrong kind of reasons problem.

[4] Kauppinen 2014, 584. Although the word "fully" appears here, it is never explained. No distinction is drawn between full and partial endorsement. Additionally, the concept of endorsement isn't here explained in other terms, given the way "endorses" appears twice in this sentence. However, a more precise formula, which I quote below, appeals to which facts the ideal subject *takes to favor* the successful manifestation of an attitude. I'll focus on that formula in this paper.

[5] Kauppinen 2014, 584.

[6] Ibid., 584.

[7] This is necessary, since the wrong kind of reasons problem isn't just a problem for fitting attitude analyses of *admirable*, but for fitting attitude analyses in general.

[8] Kauppinen 2014, 584.

[9] Ibid., 584.

[10] Ibid., 586.

[11] Perhaps this problem could be solved by amending the account so as to require, for the existence of a right kind of reason, that the ideal spectator take the fact that 100 lives would be saved by my manifesting admiration with success to *sufficiently* favor my manifesting admiration with success. So, instead of "for *r* to be a pro tanto (right kind of) W reason for S to have Y is for any W-ly ideal subject to take *r* to favor S's manifesting Y with success" we would have "for *r* to be a pro tanto (right kind of) W reason for S to have Y is for any W-ly ideal subject to take *r* to *sufficiently* favor S's manifesting Y with success."

But this threatens to rule out the possibility of merely *pro tanto right* kind of reasons – that is, reasons of the right kind which count in favor of S's having Y but are outweighed by reasons of the right kind which count against S's having Y – since in such cases the morally ideal subject wouldn't take *r* to *sufficiently* favor S's manifesting Y with success.

It has been suggested to me that this worry might be averted if we instead say "for *r* to be a pro tanto (right kind of) W reason for S to have Y is for any W-ly ideal subject to take *r* to, *ceteris paribus*, sufficiently favor S's manifesting Y with success." The idea here is that it's still true of the merely *pro tanto* (right kind of) reasons that an ideal spectator would take them to, *ceteris paribus,* sufficiently favor S's manifesting Y with success. There are two difficulties here. First, it's hard to understand how the *ceteris paribus* clause functions in the formulation, and what *ceteris paribus sufficiency* amounts to. Secondly, and more importantly, this proposal re-introduces the original problem, since, presumably, the fact that 100 lives would be saved would be taken by the ideal spectator to, *ceteris paribus,* sufficiently favor my manifesting admiration with success, and so this fact would have to count as a right kind of reason to admire the dictator, which is implausible.

[12] As I noted in footnote 3 above, Kauppinen provides a thin description of the morally ideal subject, as impartial, sympathetic, and well-informed. But nothing about being impartial, having a sympathetic nature, or being well-informed, seems to provide the resources to deliver the verdict that we need – namely, that the morally ideal subject wouldn't take the fact that a million lives would be saved if I manifested admiration with success to favor my manifesting admiration with success.

[13] See again the passage quoted in §1 above.

[14] Kauppinen endorses, at 567-577, Mark Schroeder's observation that the wrong kind of reasons are *idiosyncratic*. (See Schroeder, "Value and the Right Kind of Reason," esp. §3). Regarding our example, one might note that it's not the case that *any possible agent* who presents criticism of a friend's work would have a reason to manifest guilt with success. I have this reason only because my friend is particularly sensitive. But Schroeder's observation isn't incorporated into Kauppinen's view. On Kauppinen's view (see the passages quoted in §1 above), what matters is whether an ideal subject would take some fact to favor *my* manifesting guilt with success. It doesn't matter whether the ideal subject would take some fact to favor *other* actual or possible agents manifesting guilt with success.

[15] The two cases aren't analogous in every respect. For instance, in our dictator case, the incentive is tied to one's both admiring and emulating the dictator, such that it's impossible to get the benefits by emulation alone. In the guilt case, however, it's possible for me to smooth things over merely by making amends, without feeling any guilt; it's just that, given my transparency, it's highly unlikely that I'd do so without also having the relevant guilty feelings. But this difference isn't important. On Kauppinen's formula, what matters is whether the ideal subject would take *r* to favor S's manifesting Y with success. And, in both of these examples, that is the case. The ideal subject would take the lives saved to favor manifesting admiration by emulation, and the ideal subject would take the smoothing over of things to favor manifesting guilt by making amends.