Chapter 22

The Wrong Kind of Reasons

Howard Nye

Forthcoming in *The Routledge Handbook of Metaethics*,
(eds.) Tristram McPherson and David Plunkett

1. Introduction

An important class of metaethical views seek to explain what it is for something to fall under an ethical or evaluative category in terms of its being something towards which we ought, or have reasons, to have certain kinds of attitudes. These *Fitting Attitude* [FA-] *Analyses* might hold, for instance, that an outcome's goodness consists in its being an outcome we should desire or have "pro-attitudes" towards (Ewing 1939); or that an action's moral blameworthiness consists in the fact that the agent who performed it ought to feel guilty, and that others are justified in feeling resentment or indignation at her, for doing it (Gibbard 1990, ch 3 and 7).[1]

FA-analyses are attractive as a way of helping to provide a simple and unified picture of how values and reasons fit together, by explaining the distinctive normative or prescriptive features of ethical and evaluative categories as stemming from the deontic categories of oughts or reasons in terms of which they are explicated (cf. Rabinowicz and Ronnow-Rasmussen: 391-2; Olson 2004: 295). The judgment that an outcome is good (or that an action is blameworthy) seems to "speak in favor" of desiring (or feeling guilt and anger towards) it, in much the same way that the judgment that a belief is warranted by one's evidence speaks in favor of the belief, or that a judgment that one should plan or intend to do something speaks in favor of intending to do it. Ethical and evaluative judgments seem to have the distinctive properties of this broader

class of judgments about what attitudes we ought, or have reasons, to have – such as their being "essentially [coherently] contestable" (cf. D'Arms and Jacobson 2000b), in the sense that it seems conceptually coherent to affirm *or* deny most of them; our having to determine which of these coherent judgments are correct by means of *a priori* philosophical arguments; and their exerting a significant (if defeasible) influence on the attitudes we come to have (Gibbard 1990; Persson 2007; Raz 2009). FA-analyses can compactly explain this by holding that ethical and evaluative judgments simply *are* specific instances of judgments about what attitudes we ought, or have reasons, to have. The alternatives to FA-analyses hold either that ethical and evaluative categories are distinct from but substantively related to those of reasons for attitudes like emotions and desires (Dancy 2000; Zimmerman 2007), or that what it is to have reasons for such attitudes just is for their objects to have the relevant ethical and evaluative properties (Orisi 2013). It is much less clear on these views how ethical and evaluative categories are related to other normative categories like those of reasons for belief and intention, and how and why they share the same central normative features.

These attractions of FA-analyses are compatible with whatever one takes to be the best fundamental metaethical (or 'metanormative') account of judgments or facts about what attitudes we ought, or have reasons, to have – such as robust realism, expressivism, or reductive realism (all three of which have been favored by different proponents of FA-analyses, such as Ewing 1939; Gibbard 1990; and Brandt 1946; respectively). FA-analyses are thus a kind of "non-fundamental" metanormative account, which seek to explain the metaphysics and / or semantics of one normative category in terms of another normative category, without committing to any particular view about the underlying metaphysics or semantics of the explaining category.

One of the main problems for FA-analyses is that – at least at the outset of inquiry – there seem to be reasons for or against having the relevant attitudes towards things that do not bear on whether those things fall under the relevant ethical or evaluative categories. For instance, suppose that an evil demon credibly threatened to harm your loved ones unless you desired to have an odd number of hairs on your head (cf. Crisp 2000). The fact that the demon will spare your loved ones just in case you desire that you have an odd number of hairs seems to be a reason – or to make it the case that you ought – to desire that you have an odd number of hairs. But it does not make your having an odd number of hairs good. Similarly, the fact that you were causally involved in a someone's death, even though you took all reasonable precautions to avoid it, seems to be a reason to feel guilt for what you did, or a factor that makes guilt on your part appropriate (Greenspan 1992). But it does not seem to make it the case that your conduct was blameworthy.

Proponents of FA-analyses [*FA-analysts*] will want to insist that, although considerations like the demon's threat or your causal involvement in the death in some sense count in favor of your having (or at least wanting or getting yourself to have) the relevant attitudes, they are not the kinds of reasons intended by FA-analyses. To mark the intended distinction, we can say that FA-analyses hold that something falls under an ethical or evaluative category just in case it is *fitting* to have the relevant attitudes towards it, and that the reasons or oughts constituted by things like the demon's threat or your causal involvement in the death do not bear on the fittingness of the relevant attitudes (hence the name 'fitting attitude analyses' – cf. D'Arms and Jacobson 2000b: 746; Rabinowicz and Ronnow-Rasmussen 2004: 422-3). But herein lurks a problem for the FA-analyst – namely that, unless some further argument or account is offered, it might seem that the only thing that distinguishes considerations that bear upon an attitude's

fittingness from those that do not is that the former, but not the latter, make for the instantiation of the ethical or evaluative category that the FA-analyst is trying to analyze. Why, for instance, does the fact that an outcome involves a puppy's being happy contribute to the fittingness of a desire for it, while the fact that a demon will harm one's loved ones if one does not desire some outcome *not* contribute to the fittingness of a desire for *it*? A natural explanation is that an outcome's involving the puppy's happiness *contributes to its goodness*, while its being such that desiring it will spare one's loved ones does not. But if this is what the FA-analyst must say to distinguish fittingness from non-fittingness reasons, her account of the kind of reasons for attitudes she is talking about in her analyses looks viciously circular.

FA-analysts thus face a problem of explaining what distinguishes fittingness reasons to have attitudes from non-fittingness reasons with respect to those attitudes, without running into the vicious circularity of invoking the ethical concepts they are trying to analyze in the first place. This is the *Wrong Kind of Reasons* [WKR] *problem* with which I will be concerned in this chapter. I believe that this is a maximally general way of understanding the distinctive problem, described by this phrase in the literature, which is faced by all FA-analyses – regardless of whether they speak explicitly of a distinct category of fittingness (or simply different kinds of 'reasons'), allow that non-fittingness reasons with respect to an attitude are genuine reasons for the attitude (or insist that they are only reasons to "want to" or "make ourselves have" the attitude), and whether they offer a reductive account of the distinction between fittingness and non-fittingness reasons (or take the category of fittingness as more or less primitive). On the other hand, because solutions to this problem are constrained by the FA-analyst's need to distinguish the 'right' from the 'wrong' kind of reasons without invoking corresponding ethical and evaluative categories, it is (at least potentially) a more specific problem than that sometimes

referred to as a 'WKR problem' of distinguishing different kinds of reasons for attitudes, such as epistemic from non-epistemic reasons for beliefs, and reasons to intend to do things that do as opposed to do not constitute reasons to act out of them.

I think that the main responses to the WKR problem can be divided into three general kinds of approaches. The first and to date most popular *material approach* seeks to explain the distinction between fittingness and non-fittingness reasons in terms of what *features of considerations* make them eligible as opposed to ineligible to serve as fittingness reasons. Perhaps the best known version of the material approach, inspired by Parfit (2001), holds that fittingness reasons to have an attitude must be "object-given," or cite features of the attitude's object, while non-fittingness reasons with respect to the attitude are "state-given," or cite features of the state of having the attitude. A second, *constitutivist* approach seeks to distinguish fittingness from non-fittingness reasons by holding that, while fittingness reasons for an attitude pertain to the concerns or commitments constitutive of having the attitude at all, non-fittingness reasons with respect to the attitude stem from concerns external to the attitude's essence. A third, *formal approach* seeks to explain the distinction between fittingness and non-fittingness reasons in terms of *features of judgments or inferences* about these different kinds of reasons. Perhaps the most natural version of this view is that our attitudes can respond *directly* to what we take to be fittingness reasons for them, while they can only respond indirectly to what we take to be non-fittingness reasons.

In the next three sections I will explore each of these approaches to solving the WKR problem. As I will conclude in the fifth and final section, I believe that each approach answers to distinctive motivations and faces distinctive problems. In many of the cases considered in the literature, there seems to be a common structure that separates fittingness from non-fittingness reasons, which appropriately motivates material approaches. But proponents of material

approaches seem to overlook other kinds of cases – like that of feeling guilt in response to one's causal involvement in harm from an unavoidable accident – where the reasons need not conform to this structure. Constitutivist approaches seem applicable to both the cases that animate the material approach and many of the cases it ignores. But there are concerns about whether FA-analysts can explain the idea of the relevant attitudes' constitutive aims without incurring vicious circularity. Formal approaches promise maximal generality without vicious circularity. But the features of judgments about fittingness reasons adduced by existing formal approaches do not seem sufficient to distinguish them from all judgments about non-fittingness reasons.

2. Material Approaches

Material approaches, like most existing solutions to the WKR problem, have been focused on cases that resemble (in a sense I will clarify) that of Roger Crisp's demon, who (on my telling) threatens to harm your loved ones unless you desire to have an odd number of hairs. I believe that material approaches have made genuine progress on the problem of giving a compelling, non-circular explanation of what distinguishes fittingness form non-fittingness reasons in such cases. Initially, Wlodek Rabinowicz and Toni Ronnow-Rasmussen (2004) surveyed and rejected several material approaches, including the highly influential, Parfit-inspired:

> **Object-Given vs. State-Given Proposal**: Consideration *R* is a fittingness reason to have attitude *A* towards object *O* [i.e. *R* is the kind of reason to have *A* towards *O*, the existence of which FA-analyses should say makes it the case that *O* falls under some ethical or evaluative category] only if *R* cites properties of *O*. If *R* cites properties of the state of having *A* towards *O*, then *R* is not a fittingness reason to have *A* towards *O*.

Since the fact that *desiring to have an odd number of hairs will spare one's loved ones harm* cites a property of the state of desiring an odd number of hairs, this proposal seems to correctly avoid counting it as a fittingness reason to have this desire. It also seems to correctly count, for instance, the fact that *a puppy's having companions will make her happy* as a fittingness reason to desire her having companions, since the fact cites a property of the outcome of the puppy's having companions, which is the object of the desire for which it is a reason.

However, as Rabinowicz and Ronnow-Rasmussen (2004: 407-8) observed, there are cases in which non-fittingness reasons to have certain attitudes seem to cite properties of the objects of those attitudes, such as

> *The Admiration-Demanding Demon*. This demon credibly threatens to harm your loved ones unless you admire him.

Here, the fact that *the demon will harm your loved ones unless you admire him* is surely a *non*-fittingness reason to (make yourself) admire him. But because it cites a property of the demon, the object-given vs. state-given proposal seems to incorrectly count it as a fittingness reason to admire him.

In response, Jonas Olson (2004: 299) proposed, in effect, that one must strengthen this proposal's account of fittingness reasons to

> **Olson's Proposal**: *R* is a fittingness reason to have *A* only if *R* makes no reference to *A*.

Since the fact that *the demon will harm your loved ones unless you admire him* makes reference to admiration, Olson's proposal correctly avoids counting it as a fittingness reason. But, as Rabinowicz and Ronnow-Rasmussen (2006: 118) observed, Olson's proposal seems to go too

far, since there are cases where *fittingness* reasons to have attitudes make reference to those attitudes, such as

> *Modesty*. Allan is indifferent to whether we admire him.

The fact that *Allan is indifferent to our admiring him* seems to be a fittingness reason to admire him – such modestly genuinely contributes to his admirably. But because this fact makes reference to admiration, it appears that Olson's proposal cannot count it as such.

Generalizing a proposal by Gerald Lang (2008), in response to important criticisms of Olson (2009), Lars Samuelsson (2013) argues that we can solve the WKR problem through the appropriate interpretation of

> **Samuelsson's Proposal**: *R* is a fittingness reason to have *A* only if *R* does not cite or depend upon the consequences of having *A*.

Samuelsson does not here wish to restrict 'consequences' to causal consequences (398-92), since, like Rabinowicz and Ronnow-Rasmussen (2004: 403-4), he observes that there can be cases like those of

> *Mental State Axiologies*. These views hold that it is intrinsically good that we have certain mental states, even if their objects lack intrinsic value (e.g. one might think deep immersion in a project is intrinsically good, even if the project lacks intrinsic value).

Such views might coherently hold, for instance, that the fact that *one's attitude towards grass counting would be one of deep immersion* is a *non*-fittingness reason to (make oneself) have that attitude – an intrinsic reason to (make ourselves) have it that does not contribute to grass-

counting's intrinsic value – even though the fact makes no reference to the attitude's causal consequences. But Samuelsson's proposal can allow this if we understand "the consequences of having an attitude" to include non-causal consequences, such as the bringing about of the state of affairs of one's having it. So interpreted, Samuelsson's proposal seems right that (a) the axiologies in question take us to have reasons to (make ourselves) have the relevant attitudes *in virtue of* our having reasons to value the state of affairs of our having them, and (b) this distinguishes these reasons to (make ourselves) have attitudes from fittingness reasons (e.g. the fact that *companions would make a puppy happy* as a reason to desire that she has them), which do not seem to depend upon the value of the state of affairs of our having them. With fittingness reasons, the direction of explanation may often go the other way around: we have reasons to value the state of affairs of our having certain attitudes *in virtue of* our having fittingness reasons to have them (Nye, Plunkett, and Ku 2015: 16).

I believe that the non-fittingness reasons to (make ourselves) have attitudes upon which material approaches have focused all resemble those provided by Crisp's demon, in that they depend upon our reasons to care about the consequences of those attitudes (broadly construed). There seem, however, to be non-fittingness reasons that do not fit this pattern, which material approaches appear to have ignored. These may include the example from Section 1 of the fact that *your conduct was causally involved in someone's death despite all reasonable precautions* as a genuine reason to (make yourself) feel guilt for your conduct, but not a fittingness reason to feel guilt that constitutes your blameworthiness. A defender of Samuelsson's proposal might insist that you have this reason because feeling such guilt partially constitutes having a good moral character, and you should desire the state of affairs of having a good moral character (cf. D'Arms and Jacobson 2014: 30). But it seems coherent – and indeed quite plausible – to think

instead that, much like a fittingness reason to feel guilt (e.g. that *you deliberately harmed someone for fun*), the fact that *you were causally involved in someone's death* is a reason to (make yourself) feel guilt quite independently of the consequences of doing so, *including* the intrinsic value of the state of affairs of your feeling guilt or being morally good. One might insist that, although mere causal involvement in harm does not make guilt fitting (in the sense that entails blameworthiness), it makes guilt "morally appropriate", and that such appropriateness-making reasons cannot be explained by – but may instead help explain – which states of feeling guilt we should value.

D'Arms and Jacobson have explored many other examples of non-fittingness, "appropriateness-making" reasons to (make ourselves) have or omit having attitudes, some of which at least arguably do not depend upon the value of the states of our having them. These include factors that would make it petty to envy someone's accomplishments, even if they are genuinely good and this reflects poorly on oneself (2000a: 73-4; 2014: 24-5), factors that make genuinely funny jokes morally inappropriate (2000a: 80-1; 2014: 10-19), factors that make it admirable or noble to be unafraid of genuinely fearful odds (2000a: 85), and factors that make it morally virtuous to pity morally deserved suffering (2014: 38-40). Simply at the level of what features a reason does or does not concern (in relation to the object of an attitude or the upshot of having it), there seems to be no common feature that this wide variety of "appropriateness-making" reasons share with the broadly "pragmatic" reasons upon which material approaches have focused. This suggests that material approaches will not be successful in explaining what distinguishes all non-fittingness reasons from fittingness reasons.

It is important to understand that proponents of material approaches cannot solve this problem by arguing that, as a substantive ethical matter, there are no appropriateness-making

reasons to (make ourselves) have attitudes that do not depend upon our reasons to care about the broad consequences of having those attitudes. FA-analyses are metaethcial theories, which hold that *what is at issue* between rival substantive ethical and evaluative views is what attitudes we have fittingness reasons to have. As such, they must provide interpretations of all coherent, or at least all reasonably plausible, ethical views as views about fittingness reasons. This requires solving the WKR problem of distinguishing fittingness from non-fittingness reasons in a way that works for all coherent, or at least all reasonably plausible, ethical views – many of which are substantively false.

3. Constitutivist Approaches

A natural way of trying to articulate what separates the fittingness reasons for attitudes that constitute something's falling under an ethical or evaluative category from both pragmatic and appropriateness-making reasons is the following. It might seem that certain evaluative concerns or questions are somehow internal to or constitutive of having the attitudes at all. For instance, envy might seem essentially to portray its object as having something valuable, your relative lack of which reflects badly on you (D'Arms and Jacobson 2000a: 64). Other evaluative concerns, including both whether envy would have good consequences and whether it would be morally inappropriate, petty, or ignoble, seem to be "external" to envy, in the sense that they go beyond how envy portrays its object (D'Arms and Jacobson 2000a: 73-4), or what we commit ourselves to in envying it (cf. Hieronymi 2005). If we can draw such a distinction between the concerns constitutive of as opposed to external to an attitude, we can try solving the WKR problem by appealing to some version of

> **The Constitutivist Proposal**: Fittingness reasons for attitude *A* are considerations that show *A* to "fit" its object, by showing *A*'s object to have the features that *A* is constitutively concerned with. Non-fittingness reasons to (make oneself) have *A* do not speak to *A*'s constitutive concerns (cf. D'Arms and Jacboson 2000a; 2006; Hieronymi 2005; Schroeder 2010).

In addition to offering a convincing explanation of what separates fittingness reasons from both pragmatic and appropriateness-making reasons, it seems that this approach can make sense of what is at issue between many different coherent ethical views. Suppose, for instance, that a member of an honor-loving, acetic warrior caste lives a life of immense enjoyment by killing a record number of opponents in wars of conquest. Many of us might think that the fact that

> *E*: He enjoyed his life so much

makes his life enviable in a respect, but the fact that

> *K*: He killed so many opponents in wars of conquest

is a reason not to (allow ourselves to) envy him, which *does not* make his life less enviable. But his fellow warriors, who think that killing opponents in wars of conquest is the highest good and despise the love of pleasure as lowly, might think instead that the fact that *K* makes his life enviable, while the fact that *E* is a reason not to (allow themselves to) envy him that does not make his life less enviable. The constitutivist proposal can give the following compelling explanation of the difference between our thoughts. We take *E* to bear upon envy's constitutive concerns by showing the warrior to have something valuable that we lack, while we take *K* to

bear upon considerations external to envy (*viz*. whether the envy is morally appropriate). His fellow warriors, on the other hand, take *K* to bear on envy's constitutive concerns by showing him to have something valuable that they lack, while they take *E* to bear upon considerations external to envy (*viz*. whether the envy would involve a cowardly desire for a pleasant life).

The main problems for constitutivist approaches concern whether they really can draw a distinction between evaluations that are constitutive of as opposed to external to attitudes, in a way that solves the WKR problem for FA-analyses. First, there are problems about what it is for an evaluation to be constitutive of an attitude. The most natural idea here is that part of what it is to have the attitude is to make the evaluation, either the form of a judgment (Foot 1963), an entertained thought (Greenspan 1988), or a percept-like "construal" (Roberts 1988). Because D'Arms and Jacobson (2003) argue (quite convincingly) against all of these ideas, their talk of attitudes "presenting" things as falling under constitutive evaluations seems metaphorical. One way they suggest of cashing out the metaphor is that evaluations constitutive of attitudes must play "deep" and "wide" roles in the psychology of almost all humans, in that they fit at least many of our deeply entrenched tendencies to have them, and speak to a wide variety of our other concerns (2006: 116-18). But as D'Arms and Jacobson themselves recognize, this entails a highly controversial view, on which contingent psychological facts limit the range of eligible judgments about the fittingness of different attitudes.

In defense of their attempt to tie standards of an attitude's fittingness to our actual propensities to have it, D'Arms and Jacobson suggest that because the purpose of such standards is to regulate the attitude, we should accept standards that have "significant traction" with our propensities to have it (118). But there are at least two problems with this defense. The first is that judgments about fittingness can substantially regulate our responses even when they fail to

prevent us from having attitudes that we judge to be unfitting, namely by causing us to *avoid acting out of* such recalcitrant attitudes (cf. Nye 2009: 149-53). The second is that D'Arms and Jacobson's defense of their view seems to rely upon a potentially objectionable form of "rule pragmatism" about fittingness standards by portraying them as something we can or should construct to achieve certain ends. Pragmatic considerations about what will achieve various ends we care about can look just as irrelevant to the question of what fittingness standards are genuinely correct as they do to whether a token attitude is fitting.

A second set of problems concerns whether the evaluations held to be constitutive of attitudes can themselves be described in ways that will avoid making the FA-analyses that cite the attitudes viciously circular (cf. Rabinowicz and Ronnow-Rasmussen 2004: 420-2; Persson 2007: 6n4; Louise 2009: 360-2). Evaluations of undesirable differences in possession do not seem to make reference to the category of the enviable. But problems quickly arise when we consider what evaluations could be said to be constitutive of the attitudes referenced by FA-analyses of core ethical and evaluative categories, like that of *good outcomes*. What evaluation could be said to be constitutive of attitudes like desire for an outcome? There seems to be no candidate, other than "that the outcome is good." But it looks viciously circular for an FA-analysis to say that "What it is for an outcome *O* to be good is for desires that *O* to be fitting, what it is for a desire that *O* to be fitting is for *O* to meet the constitutive standards of desires that *O*, and what it is for *O* to meet the constitutive standards of desires that *O* is for *O* to be good."

More subtle vicious circularities may arise for FA-analysts' attempts to specify the constitutive concerns of other attitudes. For instance, D'Arms and Jacobson (2006: 109) suggest that shame constitutively evaluates its object as a "social disability." They recognize, however, that this is "vague and potentially misleading" and that it would be "more accurate, but rather

less edifying" to say that shame constitutively evaluates its object as shameful. It seems that one can with perfect coherence think that intuitively non-social and non-disabling traits are shameful, and our only grip on D'Arms and Jacobson's intended sense of 'social disability' may be that of something that is, well, shameful. But if this is the case, their constitutivist account of shame's fittingness seems to make the FA-analysis of shameful things as things it is fitting to be ashamed of viciously circular, in virtue of having to appeal to shamefulness in explaining what it is for shame to be fitting.

There may be certain attitudes, like envy, for which plausible constitutive concerns can be articulated without even tacit reference to their corresponding ethical categories. But these still seem to make at least tacit reference to other ethical or evaluative categories – e.g. *someone's having something **valuable**, your lack of which reflects **badly** on you*, makes two such references. If an FA-analyst's ambitions extend to giving FA-analyses of all ethical and evaluative categories, including those needed to explicate the allegedly constitutive concerns of attitudes like envy, then it seems that she cannot globally employ a constitutivist account of fittingness without incurring the vicious circularities described above. But perhaps constitutivism could be still be used more locally. For instance, D'Arms and Jacobson (2000b, 2006) suggest sympathetically that one might defend FA-analyses of what they call "sentimental values," which correspond to the fittingness of natural emotion kinds (constituted by robust psychological syndromes of attention, physiological response, and motivation), while eschewing FA-analyses of core ethical and evaluative categories like *good outcomes*. Such "local" deployments of constitutivist accounts of fittingness should, however, have something to say about why they are appropriate for distinguishing fittingness from non-fittingness reasons in the cases of interest, even though alterative explanations of similar differences between kinds of reasons must be

given in other cases. This may include an explanation of what is attractive about being a "local" FA-analyst – distinct from the "global" motivation presented in Section 1.

4. Formal Approaches

Pamela Hieronymi suggests a way of understanding the concerns "constitutive of" an attitude, which has the potential to avoid at least some of the difficulties described in Section 3. According to Hieronymi (2005: 447-50), an answer to a question is "constitutive of" an attitude just in case "settling [the] question amounts to forming the attitude." Thus, we might say that in envying a celebrity's fame, one has settled in the affirmative the question as to whether her fame is something valuable, one's relative lack of which reflects badly on oneself. As Hieronymi herself notes, this claim might sound too strong, since one can, for instance, recalcitrantly envy the celebrity's fame despite one's judgments that it is *not* valuable, or that one's relative obscurity does *not* reflect badly on oneself. But Hieronymi (2005: 454n34; 2009) wants to insist that there is a sense in which one has "settled" the relevant question simply by having the attitude, since having it makes one "vulnerable to the questions and criticisms which would be satisfied by considerations that bear positively on the question." For instance, if Hieronymi agreed that envy is constituted by commitments about relative differences in advantage that reflect badly on oneself, the idea would presumably be that, in envying the celebrity – even recalcitrantly – one can be asked "why did you envy her?," and one's answer should indicate a consideration (such as her fame and one's relative obscurity) that one "took to" make it the case that she has something valuable, one's relative lack of which reflects badly on oneself.

One might object to Hieronymi's apparent assumption that there is a literal, explanatorily powerful sense in which even recalcitrant attitudes must involve "taking" certain ethical claims

to hold, despite one's judging such claims to be false (on the strength, for instance, of D'Arms and Jacobson's 2003 objections to similar "quasi-judgmentalist" views). But even if one does, Hieronymi's emphasis on the relationship between taking a consideration to be a fittingness reason for an attitude and actually having the attitude suggests an importantly different kind of approach to the WKR problem. Even if our attitudes do not always respond to what we take to be the fittingness reasons for or against having them, there seems to be a crucial attitude-guiding difference between taking a consideration to be a fittingness reason as opposed to a non-fittingness reason. This is that we *can* – and, barring special circumstances, do – have attitudes in response to what we take to be fittingness reasons for them *directly*, without our first having to do something to bring it about that we have them. Simply coming to judge that object *O* has a feature, which you already take to make attitude *A* fitting, can cause you to have *A* towards *O*. For instance, if you already accept that an outcome's having the property

PS: preventing enormous suffering at the cost of convenience and some gustatory thrills

makes it fitting to desire it, then coming to learn that everyone's becoming vegan will have property *PS* can directly cause you to desire everyone's becoming vegan.

Similarly, simply coming to judge that a feature, which you already take *O* to have, actually makes *A* fitting, can also cause you to have *A* towards *O*. For instance, you might initially realize that everyone's becoming vegan has the property

PAS: preventing enormous non-human animal suffering at the cost of convenience and some gustatory thrills for humans,

but start off *not* thinking that an outcome's having *PAS* is a fittingness reason to desire it. But, upon further consideration of whether factors like a being's intellectual ability (with reference to profoundly intellectually disabled humans) and bare biological species membership (consisting as it does of mere history of phylogenetic decent, phenotype-independent genotype, or psychology-independent morphology) should make such a difference, you might come to think that an outcome's having *PAS* actually *is* a fittingness reason to desire it. This too can directly cause you desire everyone's becoming vegan.

On the other hand, it seems that we can respond to what we take to be *non*-fittingness reasons to (make ourselves) have attitudes only *indirectly*, by doing things to bring it about that we have them. Thus, those of us who take the fact that

TH: A demon will harm our loved ones if we do not admire him

to be a reason to (make ourselves) admire him, which does not contribute to his admirability, cannot admire him simply by coming to believe that *TH* or that *TH* is a reason of this kind. To respond to *TH* by admiring the demon, it seems that we must do something like psychologically condition ourselves to love such demons, take mind-altering drugs, or selectively attend to considerations that make him seem genuinely admirable.

This ability to respond directly to what we take to be fittingness as opposed to non-fittingness reasons corresponds to similar distinctions between judgments about reasons with respect to attitudes like intentions and beliefs (Persson 2007: 4-5, 12-13; Raz 2009: 39-40, 50-2; Skorupski 2010, ch 10; Parfit 2011, app. A). If we take a consideration to favor intending to φ and *actually φ-ing*, we can respond to it directly by intending to φ (e.g. we can intend to be vegan directly in response to taking the consideration that *being vegan avoids complicity in enormous*

suffering at relatively trivial personal cost to favor intending to be – and actually being – vegan). But as Kavka (1983) observed, it does not seem that we can intend to φ simply in response to taking our intending to φ to have good consequences if they do not translate into reasons to actually φ (e.g. we cannot now intend to take a toxin tomorrow that will make us sick for a day simply because a mind reader will financially reward us for having this intention today). Similarly, we can believe that *p* directly in response to taking there to be epistemic or "truth-related" reasons to believe that *p* (e.g. we can believe there are no gods directly in response to taking the positing of them to add complexity without any additional predictive or explanatory power). But as Pascal (1670) observed, it does not seem that we can believe that *P* simply in response to taking the belief that *P* to have good consequences (e.g. you cannot believe in gods simply by taking your having this belief to have good expected consequences – as Pascal suggested, you must do things that will "naturally make you believe," like acting as if you had religious beliefs and attending religious rituals).

Several authors take these observations, together with the plausible idea that we must be able to have attitude *A* in response to what we take to be reasons *for A*, to be principled grounds for concluding that fittingness reasons for conative attitudes, along with epistemic reasons for beliefs and action-relevant reasons for intentions, are the only genuine reasons for these attitudes, while pragmatic reasons with respect to an attitude are merely reasons to want or get ourselves to have it (Persson 2007; Raz 2009; Parfit 2011; Rowland 2014). Whatever the merits of this (I think largely terminological) conclusion, one can use the idea of direct response to apparent reasons to try to solve the WKR problem in a highly general way, which also distinguishes judgments about epistemic reasons for beliefs and action-relevant reasons for intentions from pragmatic reasons for these attitudes.

One might try saying, for instance, that *R* is a fittingness reason to have *A* only if one can have *A* directly in response to *R* (Persson 2007; Raz 2009). Unfortunately, Jennie Louise (2009) and Andrew Reisner (2009) argue convincingly that, depending upon how we interpret 'can have', this either makes what genuine reasons we have too dependent upon the contingencies of our psychology, or fails to distinguish fittingness from non-fittingness reasons. What the forgoing observations about responsiveness to apparent reasons seem to support, however, is a way of distinguishing *judgments about* fittingness reasons from *judgments about* non-fittingness reasons. Moreover, since FA-analyses are ultimately trying to explain what is at issue between different ethical and evaluative views in terms of their *holding different views* about fittingness reasons, FA-analyses can – and it seems should – solve the WKR problem by explaining the difference between *judgments about* fittingness and non-fittingness reasons. This, it seems, might be achieved by appealing to

> **The Response to Apparent Reasons Proposal**: One judges that *R* is a fittingness reason to have *A* only if one can have *A* directly in response to one's judgment about *R*. If one can have *A* only indirectly in response to one's judgment about *R*, then one's judgment is that *R* is a non-fittingness reason to make ourselves have *A*.

D'Arms and Jacobson (2014: 27-30) have, however, argued that judgments about appropriateness-making reasons pose a dilemma for this proposal. On the one hand, if we understand "having *A* directly in response to one's judgment about *R*" to amount to no more than the judgment's causing one to have *A* without one's having to deliberately instill A in oneself, then certain judgments about appropriateness-making reasons seem to falsify the proposal. For instance, it seems that one's judgment that it would be cold or uncaring not to feel guilt for one's

causal involvement in a harm can cause one to feel guilt for it, without one's having to deliberately try to instill the guilt – even if one takes oneself to be blameless. Similarly, it seems that one's judgment that a joke is morally offensive can directly cause one not to be amused by it even if one takes such moral considerations to be irrelevant to whether the joke is genuinely funny. On the other hand, one might insist that these purported examples of attitudes *directly responding* to apparent non-fittingness reasons are merely examples of attitudes *being directly caused* by apparent non-fittingness reasons. But it would be viciously circular to say that one's attitudes cannot count as direct *responses* simply because they are being caused by judgments about non-fittingness reasons, and it is unclear how else one can distinguish between an attitude's "directly responding to" as opposed to "being directly caused by" a judgment.

Hope that the relevant kind of "response to" judgments can be non-circularly distinguished from merely "being caused by them" may be found in Persson's (2007: 5-7) suggestion that we should look to the inferences or "valid patterns of reasoning" in which judgments about fittingness as opposed to non-fittingness reasons can participate. Along these lines, Way (2012) suggests that fittingness reasons do *not*, while *non*-fittingness reasons *do* "transmit across facilitating attitudes," or:

> **Ways' Proposal**: *R* is a fittingness reason for attitude *A* only if one *cannot* infer from *R* that the fact that *attitude B facilitates attitude A* is the same kind of reason for *B*. If *R* is a reason with respect to *A*, and one can infer from *R* that *B*'s facilitating *A* is a reason of the same kind for *B*, then *R* is a non-fittingness reason to bring about *A*.

For instance, the fact that *outcome O will make Bugsy happy* is a fittingness reason to desire *O*. If one irrationally hated Bugsy, and the only way one could get oneself to desire *O* was to first

admire the people Bugsy admires, this would *not* contribute to the fittingness of admiring them. On the other hand, if one had the non-fittingness reason to (make oneself) desire an odd number of hairs constituted by the fact that a demon will harm one's loved ones unless one desires this, and the only way to instill this desire was to first desire that nothing be symmetrical, this *would* seem to give one a non-fittingness reason to (make oneself) desire that nothing be symmetrical.

Unfortunately, while Way's proposal may distinguish fittingness reasons from *pragmatic* reasons to make oneself have them, it does not seem to distinguish fittingness reasons from *appropriateness-making* reasons. It is at least coherent, and indeed quite plausible, to think that we *cannot* infer from

(i) *R* contributes to the appropriateness of having *A*, and

(ii) *B* facilitates having *A*

that (ii) contributes to the appropriateness of having *B*. For instance, suppose that

(i') joke *M*'s offensive portrayal of minorities makes it morally inappropriate to be amused by *M*, but

(ii') the only way to suppress one's amusement at *M* is to be amused by joke *S*, which is both unfunny and portrays women in an offensive way.

It is at least coherent, and indeed quite plausible, to think that one cannot infer from (i') and (ii') that (ii') contributes in the least to the moral appropriateness of being amused by *S*. Similarly, suppose that

(i*) your friendship with Abigail, and envy's focus on relative positions, make it inappropriately petty to envy Abigail's accomplishments, but

(ii*) the only way to avoid envying Abigail's accomplishments is to envy your equally good friend Brittany's accomplishments.

It is at the very least coherent, and I think extremely plausible, to think that one cannot infer from (i*) and (ii*) that (ii*) makes it any less inappropriately petty to envy Brittany's accomplishments.

5. Conclusion

I have thus surveyed the primary motivations and problems for what I take to be the main approaches to solving the WKR problem. Material approaches seem to have made progress articulating features of pragmatic reasons to make ourselves have attitudes that distinguish them from fittingness reasons to have them. But existing material approaches seem to have ignored appropriateness-making reasons, and it is unclear how they can be extended to accommodate them. Constitutivist approaches hold the promise of distinguishing fittingness from appropriateness-making reasons, but face problems explaining the alleged "constitutive evaluations" of attitudes in ways that avoid vicious circularity. Constitutitivism might avoid certain (but not all) of these problems by limiting its scope to solving the WKR problem for FA-analyses of a special set of ethical or evaluative categories, but this would be inconsistent with the ambitions and motivations of most FA-analysts, and require an alterative rationale. Formal approaches that look to the attitude guiding or inference-licensing features of judgments about fittingness as opposed to non-fittingness reasons have the potential to explain the distinction

between such judgments in a highly general way, which distinguishes similar judgments about reasons for attitudes like beliefs and intentions from judgments about pragmatic reasons for them. But existing formal approaches seem to have problems distinguishing judgments about fittingness reasons from judgments about appropriateness-making reasons.

Some might be tempted to conclude that these difficulties show the WKR problem to be insoluble, which would appear to have significant theoretical implications. Explaining the core normative features of ethical and evaluative categories by understanding them as categories of reasons for attitudes might make for a simple and unified picture of our normative categories, and justify focusing our fundamental metanormative efforts on understanding the foundational category of reasons. But if FA-analyses cannot avoid a viciously circular reference to ethical and evaluative categories in delineating the sorts of reasons that are supposed to do the explaining, this initially attractive theoretical perspective and strategy of metanormative research looks untenable.

I think, however, that FA-analysts can learn from the problems I have surveyed for existing approaches to solving the WKR problem, and that prospects for a solution remain bright. FA-analysts need to pay more attention to the theoretical motivations for FA-analyses, and the variety of coherent judgments – especially about appropriateness-making reasons – that they need to adequately interpret. Along these lines, I believe that there is in particular room for further fruitful developments of formal approaches that pay more attention to the functional and inferential roles of judgments about fittingness reasons beyond direct attitude causation and the non-transmission of reasons across facilitating attitudes (cf. Gibbard 1990, ch 4; Nye 2009, ch 6).

Acknowledgements

I am grateful to Justin D'Arms, Daniel Jacobson, John Ku, Tristram McPherson, David Plunkett, and Lea Schroeder for extremely helpful comments on an earlier draft of this chapter.

References

Anderson, Elizabeth. 1993. *Value in Ethics and Economics*. Cambridge: Harvard University Press.

Brandt, Richard. 1946. Moral Valuation. *Ethics* 56: 106-121.

Crisp, Rodger. 2000. Review of Value... and What Follows by Joel Kupperman. *Philosophy* 75: 458-462.

Dancy, Jonathan. 2000. Should We Pass the Buck? *Royal Institute of Philosophy Supplement* 47: 159-173

D'Arms, Justin and Daniel Jacobson. 2000a. The Moralistic Fallacy. *Philosophy and Phenomenological Research* 61: 65-90.

———. 2000b. Sentiment and Value. *Ethics* 110: 722-748.

———. 2003. The Significance of Recalcitrant Emotion. *Philosophy: The Journal of the Royal Institute of Philosophy*, 52 (suppl.): 127-145.

———. 2006. Anthropocentric Constraints on Human Value. In R. Shafer-Landau (ed.) *Oxford Studies in Metaethics*, Vol 1, New York: Oxford University Press, 99-126.

———. 2014. Wrong Kinds of Reasons and the Opacity of Normative Force. Forthcoming in R. Shafer-Landau (ed.) *Oxford Studies in Metaethics*, Vol 9, New York: Oxford University Press.

Ewing, A.C.. 1939. A Suggested Non-Naturalsitic Analysis of Good. *Mind* 48:1-22.

Foot, Philippa. 1963. Hume on Moral Judgment. In D. Pears (ed.), *David Hume*, London: McMillan.

Gibbard, Allan. 1990. *Wise Choices, Apt Feelings: A Theory of Normative Judgment*. Cambridge, Mass.: Harvard University Press.

Greenspan, P.S. 1988. *Emotions and Reason*. London: Routledge & Kegan Paul.

———. 1992. Subjective Guilt and Responsibility. *Mind* 101: 287-303.

Hieronymi, Pamela. 2005. The Wrong Kind of Reason. *The Journal of Philosophy* 102: 437-457.

———. 2009. The Will as Reason. *Philosophical Perspectives* 23: 201-220.

Jacobson, Daniel. 2011. Fitting Attitude Theories of Value. In E. Zalta (ed.), *The Stanford Encyclopedia of Philosophy*, Spring 2011 Edition, URL = <http://plato.stanford.edu/archives/spr2011/entries/fitting-attitude-theories/>.

Kavka, Gregory. 1983. The Toxin Puzzle. *Analysis* 34:33-36.

Lang, Gerald. 2008. The Right Kind of Solution to the Wrong Kind of Reason Problem. *Utilitas* 20: 472-489.

Louise, Jennie. 2009. Correct Responses and the Priority of the Normative. *Ethical Theory and Moral Practice* 12: 345-364.

Nye, Howard. 2009. *Ethics, Fitting Attitudes, and Practical Reasons*. Ph.D. Dissertation, University of Michigan.

Nye, Howard, David Plunkett, and John Ku. 2015. Non-Consequentialism Demystified. *Philosophers' Imprint* 15: 1-28.
Olson, Jonas. 2004. Buck-Passing and the Wrong Kind of Reasons. *Philosophical Quarterly* 54: 295-300.
———. 2009. The Wrong Kind of Solution to the Wrong Kind of Reason Problem. *Utilitas* 21: 225-232.
Orisi, Francesco. 2013. What's Wrong with Moorean Buck-Passing? *Philosophical Studies* 164: 727–746
Parfit, Derek. 2001. Rationality and Reasons. In D. Egonson, J. Josefson, B. Petterson, and T. Ronnow-Rasmussen (eds.), *Exploring Practical Philosophy*, Aldersot: Ashgate, 17-39.
———. 2011. *On What Matters: Volume One*. Oxford: Oxford University Press.
Pascal, Blaise. 1670. *Pensées*. W. F. Trotter (Trans.), London: Dent, 1910.
Persson, Ingmar. 2007. Primary and Secondary Reasons. In T. Rønnow-Rasmussen, B. Petersson, J. Josefsson and D. Egonsson (eds.) *Homage á Wlodek*, URL = <http://www.fil.lu.se/hommageawlodek>.
Rabinowicz, Wlodek and Toni Ronnow-Rasmussen. 2004. The Strike of the Demon. *Ethics* 114: 391-423.
———. 2006. Buck-Passing and the Right Kind of Reasons. *Philosophical Quarterly* 56: 114-120.
Raz, Joseph. 2009. Reasons: Practical and Adaptive. In D. Sobel and S. Wall (eds.) *Reasons for Actions*, Cambridge: Cambridge University Press, 37-57.
Reisner, Andrew. 2009. The Possibility of Pragmatic Reasons for Belief and the Wrong Kind of Reasons Problem. *Philosophical Studies* 145: 257-272.
Roberts, Robert. 1988. What an Emotion Is: A Sketch. *The Philosophical Review* 97: 183-209.
Samuelsson, Lars. 2013. The Right Version of 'the Right Kind of Solution to the Wrong Kind of Reason Problem'. *Utilitas* 25: 383-404.
Scanlon, T.M. 1998. *What We Owe To Each Other*. Cambridge: Harvard University Press.
Schroeder, Mark. 2010. Value and the Right Kind of Reason. In R. Shafer-Landau (ed.), *Oxford Studies in Metaethics*, Vol 5, New York: Oxford University Press, 25-55.
Skorupski, John. 2010. *The Domain of Reasons*. Oxford: Oxford University Press.
Way, Jonathan. 2012. Transmission and the Wrong Kind of Reason. *Ethics* 122: 489-515.
Zimmerman, Michael. 2013. The Good and the Right. *Utilitas* 19: 326-353.

Note

[1] Those who have encountered something like the idea of these analyses through Scanlon (1998, ch 3) often refer to them as 'buck-passing accounts'. But as Jacobson (2011, §2.1) has observed, what Scanlon actually describes as a 'buck-passing account' incorporates elements that are quite distinct from the view that we can explain something's falling under an ethical or evaluative category in terms of there being reasons to have certain kinds of attitudes towards it. These include the view that something's value does not provide reasons to value it over and above its value-makers (which opponents of explaining value in terms of reasons for attitudes can accept), and the possible view that we can explain something's value in terms of reasons to *act* towards it in certain ways. This latter idea may be problematic from the standpoint of explaining value in terms of reasons in at least two ways. First, a feature of something (e.g. someone's bravely helping others) may make for two very different kinds of value (e.g. it may make her both morally admirable and aesthetically inspiring), in a way that can be captured only by citing its status as a reason for two different kinds of attitudes (e.g. moral admiration and aesthetic appreciation), as there may be no

bifurcation in the acts it is a reason to perform. Second, one may not want to build into the very idea of something's value the idea that we should act towards it in certain ways. One may want to explain the connection between values and reasons to act in terms of other principles, such as a general connection between fitting motives and reasons to act (cf. Anderson 1993; Skorupski 2010; Nye, Plunket, and Ku 2015).

Fortunately, most of the literature on the Wrong Kind of Reasons problem that addresses itself to 'buck-passing accounts' is concerned with them as FA-analyses in the sense I explain. But because talk of 'buck-passing' in this context can still threaten to obscure the core issues, I shall avoid all further use of it.