Norm Acceptance and Fitting Attitudes

Howard Nye
University of Michigan

Abstract: I offer a way to distinguish between the kinds of reasons for attitudes that contribute to the instantiation of ethical concepts and the kinds that do not, thus solving what Rabinowicz and Ronnow-Rasmussen call the 'wrong kind of reasons [WKR]' problem for analyses of ethical concepts in terms of fitting attitudes. Intuitively, judgments about ethical-fact-making reasons for an attitude can, whereas judgments about other kinds of reasons cannot, directly cause one to have the attitude. I argue, however, that in order to clarify and defend this intuitive distinction, we should ultimately analyze judgments about fitting attitudes in terms of the acceptance of norms for attitudes. I contend that the best such analysis understands judgments about an agent's reasons as judgments about the prescriptions of the system of norms she deeply accepts. I call this view 'Norm Descriptivism', and argue that it best explains how judgments about reasons both guide attitudes and can be determined to be true or false via *a priori* reflective-equilibrium methods.

Norm Acceptance and Fitting Attitudes

Howard Nye

I. Introduction

Fitting attitude analyses of ethical concepts seek to analyze them in terms of the fittingness of attitudes like desires and emotions. For instance, A.C. Ewing (1939) may be read as arguing that we can understand judgments that a state of affairs is good as judgments that it is fitting to desire it.[1] Similarly, Allan Gibbard (1990) argues that we can analyze judgments that someone has done something morally blameworthy as judgments to the effect that it is fitting for him to feel guilt for what he has done, and fitting for others to feel angry at him for doing it.

I think that such fitting attitude analyses [hereafter 'FA-analyses'] are attractive because they offer us a straightforward way to explain the common normative features of ethical concepts - features they share with other normative concepts like REASON FOR BELIEF. In judging that a state of affairs is good, an act is blameworthy, or a belief is rational, we do not seem to be simply judging that the state of affairs, act, or belief have the kind of ordinary descriptive features that play a role in causal or geometric explanations. It is often said that these normative judgments are not merely descriptive but "prescriptive," in that to make them is to prescribe or think that one should have a certain response. The responses we think we should have in making each of the above judgments bear on what to do, but each kind of judgment is distinct in that it seems in the first instance to command or license a different kind of attitude. Just as judging a belief rational most immediately pertains to what to believe, judging a state good is distinctive in that it entails a judgment that one should desire it, and judging an act blameworthy is distinctive in that it entails a judgment that its author should feel guilt and others have reason to be angry with him.

These normative judgments, which entail that we have reason to have certain attitudes, seem to have important common properties. For one, it seems *coherent* (if in many cases obviously mistaken) to think that almost any kind of descriptively specified state, act, or belief is respectively good, blameworthy, or rational. It is coherent not only to think that states of happiness, achievement, or knowledge are intrinsically good, but indeed to think that racial purity, "women knowing their place," and the maintenance of tradition are intrinsic good-making features. We might usually think blameworthy (absent exculpation) only such things as harming

[1] Ewing (1939, 8-9) seemed to resist this characterization of his position on the grounds that 'desire' might be taken to mean an attitude one should have only towards what does not obtain or "a certain uneasy emotion." But in the text I intend 'desire' in what I think was the kind of sense for which Ewing was happy to concede "the definition in terms of desire merges into my definition."

others, failing to aid others, and violating the autonomy of other agents. But we will know only too well what someone means if she claims that such things as sexual practices, not doing what deities say, or uttering curse words are intrinsically blameworthy, or blameworthy quite apart from their effects on the welfare of any being or the autonomy of any agent.

Another common feature is that we determine which of the wide diversity of coherent normative judgments are true via an *a priori* method of reflective equilibrium. We use a similar reflective equilibrium method to analyze our concepts. We elicit intuitions about apparent platitudes that involve them and when things fall under them in concrete cases, and we seek the best unification and explanation of these intuitions in order to determine what our concepts actually are. But when we use this kind of method to determine which coherent normative judgments are true, we seek a best explanation of our substantive intuitions about normative principles and what to feel or believe in particular cases. The negations of these substantive intuitions seem perfectly coherent; they just seem false.

A final common feature of normative judgments is that they are intimately related to the guidance of our attitudes in a way that ordinary descriptive judgments do not seem to be. Judging that a state of affairs is good and the making the entailed judgment that one should desire it usually exerts a kind of direct causal pressure on one's actually desiring the state. Similar remarks go for judging an act blameworthy and coming to feel guilt (if one is the actor) or anger (if one is not). Of course, against the right profile of standing beliefs and desires, ordinary descriptive judgments will exert a kind of direct causal pressure on one's coming to have attitudes like desires and emotions. For instance, if one already desires a world in which welfare is distributed equally, one's coming to judge that policy X is most conducive to such equality will tend to cause one to desire that policy X be implemented. But judging that it is good for welfare to be equally distributed does not seem to require any background desire for whatever states are good for it to exert direct causal pressure on one's desiring such equality.

FA-analyses of ethical concepts like GOODNESS and BLAMEWORTHINESS enable us to subsume the above semantic, epistemic, and causal properties of ethical judgments under those of judgments about the fittingness, warrant, or justification of attitudes. These features are possessed not only by judgments about the fittingness of attitudes like emotions and desires, but also judgments of the warrant or justification of beliefs in judgments of epistemic rationality. It seems to be a desideratum on any theory of ethical judgments and any theory of judgments about reasons for attitudes it be at least compatible with both kinds of judgments manifesting these semantic, epistemic, and causal properties. For this reason, I think that the ability of FA-analyses to subsume these features of the former under these features of the latter is an attractive

theoretical virtue whatever one's views on what it is to judge that an attitude is fitting or warranted. Even if one thought that the notion of a fitting or warranted attitude resisted all further explanation or analysis, and that it is simply a brute fact that judgments that involve them have the above semantic, epistemic, and causal properties, it would still be good to keep the stock of brute facts as small as possible. This can be achieved by explaining the possession of these properties by ethical judgments in terms of their reducibility to judgments about the fittingness of attitudes like desires and emotions.

As attractive as they may thus be, FA-analyses face an important problem. Aptly dubbed 'the wrong kind of reasons [WKR] problem' by Rabinowicz and Ronnow-Rasmussen (2004), the problem is that intuitively some kinds of reasons to have attitudes like desires and emotions do not contribute to the instantiation of ethical concepts.[2] Suppose, for instance, that an evil demon were to threaten to harm your loved ones unless he detects that you desire that you have an even rather than odd number of hairs on your head. The fact that the demon will harm your loved ones if you do not desire a state in which you have an even number of hairs might seem to be a kind of reason to desire such a state, but it does not contribute to making the state good.[3]

The proponent of FA-analyses will of course want to distinguish the kinds of reasons to have an attitude that contributes to its fittingness, warrant, or appropriateness, of which her analyses speak, from these other reasons which do not. But there is a worry that what distinguishes those reasons to have an attitude that contribute to its fittingness from those that do not itself involves the concepts the FA-analyst is trying to analyze. Why do one's reasons to, say, desire states in which puppies are happy contribute to the fittingness of such desires, but one's reasons to desire states in which one has an even number of hairs in response to demonic threats do nothing of the kind? A natural explanation might *just be* that the former states are good and the latter states are not, and that what distinguishes fittingness from non-fittingness reasons for desires is that the former are the reasons one has to desire *good* states of affairs. But if this is what the FA-analyst must say to distinguish fittingness from non-fittingness reasons, she will run in a vicious circle when it comes time to explain which kinds of reasons for attitudes she

[2] Rabinowicz and Ronnow Rasmussen (2004) are explicitly concerned only with fitting attitude analyses of value or evaluative concepts like GOODNESS and ADMIRABILITY, though they note that the strategy can be applied to other ethical concepts like BLAMEWORTHINESS, and their problem readily generalizes to the case of all such fitting attitude analyses. If one is wondering what distinguishes value or evaluative from non-evaluative ethical concepts, I would venture that the former but not the latter entail the fittingness of emotions or desires that involve motivations to bring about various states of affairs (specified without certain *de se* representations), as opposed to motivations to do more particular things.

[3] A similar example, from which Rabinowicz and Ronnow-Rasmussen apparently draw the title of their paper, is due to Rodger Crisp (2000).

is talking about in her analyses. The problem for the FA-analyst of explaining what distinguishes fittingness from non-fittingness reasons to have attitudes, without running into the vicious circularity of invoking the ethical concepts she is trying to analyze, is the WKR problem.

In this paper I will argue that the FA-analyst's best hope of solving the WKR problem lies in observations about the kinds of causal influence that judgments of fittingness as opposed to non-fittingness reasons are capable of having on one's attitudes. Roughly, judging an attitude fitting is, where merely judging that one has non-fittingness reasons to have it is not, capable of causing one to have it directly, or without one's having to do anything in order to get oneself to have it. But I will contend that in order to make this thought precise and defend it from counterexamples, the FA-analyst should ultimately draw on a certain kind of analysis of what it is to judge that an attitude is fitting or warranted. This kind of analysis explains one's judging that it is fitting for one to have an attitude in terms of one's accepting or judging that one accepts norms that prescribe having it. These *norm acceptance analyses* of fittingness assessments offer us a way to explain their distinctive attitude guiding character. I will argue that this not only helps secure a solution to the WKR problem, but fits our intuitive picture of what makes normative judgments special and avoids a dilemma faced by rival accounts.

I will conclude by arguing for the superiority of a particular norm acceptance analysis, which I call 'Norm Descriptivism'. According to this analysis, judgments that it is fitting for an agent to have an attitude are judgments that the attitude is prescribed by the system of norms she deeply accepts. My contention will be that Norm Descriptivism offers us the best explanation of how fittingness judgments guide attitudes, what we are doing when we engage in basic inquiry into which attitudes are fitting, and how such inquiry can hook onto facts about fittingness. I will also argue that Norm Descriptivism best explains what is distinctive about attributing reasons to agents and which entities it makes sense to think subject to reasons.

Overall, my arguments seek to establish a relationship of mutual support between FA-analyses of ethical concepts and norm acceptance analyses of fittingness judgments. By solving the WKR problem for FA-analyses, norm acceptance analyses remove a major impediment to these otherwise attractive analyses of ethical concepts. At the same time, the fact that norm acceptance analyses can do important work for such theoretically attractive FA-analyses lends further support to norm acceptance analyses themselves. My arguments in favor of the superiority of Norm Descriptivism suggest that this relationship of mutual support is particularly strong between FA-analyses and this particular norm acceptance analysis of fittingness.

II. The WKR Problem: Failed Solutions and a New Approach

Rabinowicz and Ronnow-Rasmussen (2004, 2006) have convincingly argued that several attempts on behalf the FA-analyst to solve the WKR problem fail. To give the reader a sense of the difficulty of the WKR problem I will review the shortcomings of what I take to be three of the most natural attempts to solve it. I will then review a final, as yet problematic attempt to solve the WKR problem that I think is highly suggestive of the solution I will be proposing.

A first natural attempt to solve the WKR problem draws on Derek Parfit's (2001) distinction between what he calls "object given" and "state given" reasons for attitudes. Attitudes like desires and emotions have objects, or things they are about or directed towards. For instance, the object of a desire that a puppy is happy is a state of affairs in which he is happy, the object of one's guilt about lying is one's action of lying, and the object of one's anger at another person for lying is that person. Parfit calls a reason for an attitude "object given" just in case it is constituted by a fact about the attitude's object, while he calls a reason for an attitude "state given" just in case it is constituted by a fact, not about the attitude's object, but the state of one's having the attitude itself (Parfit 2001, 21-22). For instance, one's reason to desire a state of affairs constituted by the fact that it involves happy puppies is object given. On the other hand, one's reason to desire a state in which one has an even number of hairs constituted by the fact that a demon will harm one's loved one's if one fails to have this desire is state given. Since the former but not the latter kinds of reasons seem to be those that constitute the goodness of a state of affairs, it is natural to try to solve the WKR problem by identifying the reasons of which FA-analyses of ethical concepts speak as object rather than state given reasons.

Rabinowicz and Ronnow-Rasmussen (2004, 406-7) rightly point out that the instantiations of intuitively "Cambridge properties," like *being the object of a desire which is such that if one has it one's loved ones will be spared* seem to give us an object-given reason corresponding to each state-given reason and vice versa. To make this criterion work we would seem to need a way to restrict the relevant reasons to instantiations of non-Cambridge properties. But a deeper problem that Rabinowicz and Ronnow-Rasmussen raise for this attempt to solve the WKR problem is that we can have reasons to have attitudes, which do not seem to contribute to the instantiation of a corresponding ethical concept, but which are constituted by instantiations of intuitively *non*-Cambridge properties of the object of the attitude. Suppose that if a Greek deity detects that you are angry at him for engaging in homosexual intercourse, he will feel that he is worthless. As a result he will work very hard and effectively to end world poverty in order to vindicate himself. Intuitively, the kind of reason one has to be angry with the deity for engaging

in homosexual intercourse constituted by the fact that he has a disposition to end world poverty if you are does not contribute to the blameworthiness of his sexual act. But this reason certainly seems to be constituted by a fact about the deity who is the object of one's anger, or his instantiating a property that is at least as non-Cambridge as the property one's anger has of being such as to trigger his disposition to end world poverty should he detect it.

There is, however, still a way in which reasons like the above reason to feel angry at the deity relate to one's having the attitudes they are reasons to have, which might seem to distinguish them from fittingness reasons. While these reasons might be constituted by facts about the attitude's object, they still seem to "bring in" or be in part about the attitude they are reasons to have. A natural revision of the attempt to explain the difference between fittingness and non-fittingness reasons in terms of object as opposed to state given reasons is thus to attempt to explain it in terms of reasons that do not, as opposed to reasons that do, mention the attitude they are reasons to have. I think that this is essentially the solution proposed by Jonas Olson (2004). Rabinowicz and Ronnow-Rasmussen (2006) point out, however, that while all non-fittingness reasons may well mention the attitudes they are reasons to have, some fittingness reasons do so as well. They observe, for instance, that the fact that someone is indifferent to being admired may well be the kind of reason to admire her that contributes to her being admirable. Indeed, the fact that someone is indifferent to the very token of admiration that I am thinking of having might well be this kind of reason that contributes to her admirability.

One final attempt to solve the WKR problem along these lines is due to Rabinowicz and Ronnow-Rasmussen (2004) themselves. As should be clear from our above case of feeling angry at a deity for engaging in homosexual sex, our attitudes are not simply directed towards things like persons, but towards such things for or on account of certain features of them.[4] One thing that might seem to differentiate one's reason to feel angry at the deity when this will end world poverty from reasons that contribute to his blameworthiness is that the former seem to involve his having properties that are other than those it is a reason to feel angry at him for having. Rabinowicz and Ronnow-Rasmussen consider the general proposal that fittingness reasons to have an attitude towards an object for having a set of properties just are the object's instantiation of any of these properties, and reasons constituted by the object's instantiating other

[4] Perhaps our desires that states of affairs obtain are not simply directed at states of affairs, but rather at certain features of them, though this might seem more doubtful. Our desires for states of affairs certainly arise due to certain of their features, and we certainly take certain of their features to provide reasons for desiring them, but it would be another thing for certain features of the states, rather than simply the states, to be part of what they are desires for or about. For this reason it is doubtful whether this kind of solution could aspire to distinguish fittingness form non-fittingness reasons to desire that states of affairs obtain.

properties are non-fittingness reasons. Unfortunately, Rabinowicz and Ronnow-Rasmussen also show how this proposal is open to counterexample. Let us alter the case of the Greek deity and suppose that he will not be moved to end world poverty unless you feel angry at him, not for having engaged in homosexual acts, but for being such that he will respond to anger by ending world poverty. The fact that he will end world poverty if one is angry at him on this score would thus seem to count in favor of being angry at him *for being such that he will end world poverty under these conditions*. Although this consideration thus satisfies Rabinowicz and Ronnow-Rasmussen's proposed criterion for a fittingness reason, it does not seem to contribute to such anger's being fitting, or to the blameworthiness of the deity's disposition.

While these problems with such natural attempts to solve the WKR problem might seem disheartening for FA-analysts, some might think we were too hasty to allow that there is a genuine problem here, at least as Rabinowicz and Ronnow-Rasmussen understand it. Both Gibbard (1990, 36) and Parfit (2001, 27) insist that what I have been calling 'non-fittingness reasons' or 'reasons for attitudes that do not contribute to the instantiation of ethical concepts' are not in fact reasons for such attitudes at all. Gibbard and Parfit insist that these are merely reasons to want to have or try to get oneself to have the relevant attitudes. Thus, the fact that a demon will harm one's loved ones unless one desires that one has an even number of hairs is not a reason to desire that one has an even number of hairs, but simply a reason to want to or try to get oneself to have such a desire. Similarly, the fact that a deity will end world poverty if one is angry at him is not a reason to be angry with him, but rather a reason to want to or try to get oneself to have such anger.

Rabinowicz and Ronnow-Rasmussen (2004, 412) agree with Gibbard and Parfit that the considerations in question are reasons to want to and try to get oneself to have these desires and emotions, but they do not share the intuition that it is inappropriate to describe them as reasons for the desires and emotions as well. Much more importantly, Rabinowicz and Ronnow-Rasmussen note that whatever we decide to call the reasons of which FA-analyses do not intend to speak, the FA-analyst needs to explain what makes a consideration a member of this category in terms that do not reference the ethical concepts she is trying to analyze. Even if Gibbard and Parfit are correct, there is still a threat to the FA-analyst. This is that what makes a consideration a reason to, say, desire a state of affairs instead of a mere reason to try to get oneself to desire it might *just be* that it contributes to the state's goodness (Rabinowicz and Ronnow-Rasmussen 2004, 413-14). The Gibbardian or Parfittian FA-analyst is still in need of a way to distinguish

reasons for attitudes from reasons to get oneself to have attitudes that does not invoke the concepts she is trying to analyze.

Even if one does not initially share Gibbard and Parfit's linguistic intuitions, I think that it is highly instructive to ask why it is that they seem attracted to a description of non-fittingness reasons, not as reasons for the attitudes in question, but merely reasons to want to have or get oneself to have them. I think that an important part of the answer concerns the role that judging oneself to have reason for a given kind of response can (and typically does) play in causing one to have it. Judging that one has reason to desire a state of affairs because of such things as its involving happy puppies seems to be capable of causing one to desire it directly, or without one's having to do anything to get oneself to desire it.

But simply judging that a demon will harm one's loved ones if one doesn't want to have an even number of hairs does not seem capable of directly causing such a desire or feeling of anger. To respond to this second kind of reason one will have to do things like take pills, classically condition oneself, or rationalize oneself into thinking the attitudes fitting. What one's judgments about such reasons directly cause without behavioral intervention are only desires that one have the desire and behaviors undertaken in order to get oneself to have it. Thus, if one is attracted to the idea that judgments about reasons for a kind of response must be capable of causing one to have it directly or without one's having to do anything to bring it about, one might want to describe these other kinds of reasons as mere reasons to want or try to get oneself to have such desires and feelings of anger.

We can use these observations about the causal properties of judgments about various kinds of reasons to construct a solution to the WKR problem. We can in fact adopt this solution whether or not we agree with Gibbard and Parfit's linguistic intuitions and whether or not we insist that all kinds of reasons for a response are capable of directly causing one to have it. We can simply try saying that to judge that an attitude is fitting is capable of directly causing one to have it, while merely judging that one has non-fittingness reasons to (get oneself to) have it is not capable of causing one to have it without one's doing something to bring it about that one does. Whether or not we agree with Rabinowicz and Ronnow Rasmussen that both judgments are about reasons for the attitude in question, we have a criterion for distinguishing fittingness from non-fittingness reasons. This criterion certainly seems to categorize the above examples correctly, and it does not draw on the ethical concepts the FA-analyst is trying to explain. Indeed, the criterion involved here would seem to correctly distinguish fittingness from non-fittingness reasons for attitudes outside the ethical realm. It certainly seems, and has been noted

(see e.g. Kavka 1983, 36), that a feature that distinguishes evidential or epistemic reasons for belief from mere pragmatic reasons for belief is that judgments about the former can, while judgments about the latter cannot, directly cause one to have beliefs without behavior or activity undertaken in order to get oneself to have them.

III. Not Just Any Behavior Independent Process: Enter the Norm Acceptance Analyses

As attractive as it might seem to explain the difference between judging an attitude fitting and judging oneself to have non-fittingness reasons to (get oneself to) have it in terms of the former having and the latter lacking direct causal powers, the approach faces important problems. There are several ways in which judgments about *non*-fittingness reasons for attitudes can play a role in causing one to have them without one's having to do anything to bring it about that one does.

Consider, for instance, judgments about reasons to (get oneself to) have attitudes one thinks estimable to have or disestimable not to have. The judgment that it is cowardly not to feel angry at someone might cause one to feel such anger where mere considerations of having been wronged were insufficient. Intuitively these are not judgments about fittingness reasons for anger, or reasons that contribute to the blameworthiness of the person towards whom the anger is felt.[5] But such judgments about the disestimabilty of not feeling anger seem capable of causing one to feel it without one's having to do anything to bring it about that one does. More generally, judgments that it is fitting to esteem (or disesteem not) feeling F directly cause such (dis)esteem. But as Velleman (2002, 101) points out, this kind of esteem towards an ideal of feeling F involves emulating the ideal, or wishfully imagining that one feels F and acting as if one feels F too. This kind of acting out the role of feeling F can cause one to genuinely feel F without one's ever trying to make this the case.[6] The processes at work seem similar to those documented by Philip Zimbardo (2007) in his 1971 Stanford Prison Experiment, where ordinary male college students (not to mention Zimbardo himself) came to genuinely possess the attitudes characteristic of guards and prisoners just by playing at the roles.

There are other mechanisms by which judgments about non-fittingness reasons can cause one to have the relevant attitudes without one's having to do anything to make this the case. It

[5] The fact that failure to feel such anger at someone would be cowardly might be *evidence* that what he did was culpable, since it may be that only acts that are genuinely blameworthy can be such that it is cowardly to fail to feel anger at them. But this is distinct from the cowardliness of not feeling anger towards someone metaphysically *making it the case* that his action was blameworthy.

[6] Similar remarks go for disesteeming a negative ideal of not feeling F, acting out a role of being *un*like this negative ideal (and thus feeling F), and coming to feel F as a result.

certainly seems that with external apparatus or neurosurgery we could create cases in which agents' judgments about arbitrary kinds of reasons to (get themselves to) have attitudes would directly cause their have them, without these *ipso facto* becoming judgments about fittingness reasons. But we probably need not even move to such distant possibilities – actual psychic mechanisms such as those involved in classical conditioning and wishful thinking would probably suffice. The development of sufficient mental associations between one's having an attitude and the obtaining of preferred consequences may well be enough to cause one to have it.

To be sure there are differences between the way in which judging an attitude fitting causes one to have it and the way the other mechanisms we discussed cause one to have it. Judging an attitude estimable or disestimable not to have will usually take a period of time in which certain simulations take place in order for it to cause one to have the attitude. Similarly, forming sufficient emotional associations between, say, having a certain attitude and avoiding demonic threats for the former to be triggered by one's judgments about the presence of the latter would require a period of conditioning trials. But judging an attitude fitting seems to be capable of causing one to have it without any such delay of time, simulations, or conditioning trials.[7]

Another difference concerns what happens when the causal influence of the mechanism fails to be decisive. When we judge an attitude fitting but fail to have it, we experience a species of what psychologists refer to as "cognitive dissonance."[8] This cognitive dissonance is distinctive, however, in that it does not feel as though one is torn or of two minds about an issue. If anything, it feels rather like one is weak, deficient, or inadequate to the demands of reason. Since this kind of cognitive dissonance arises in response to having attitudes one thinks one should not have, and such attitudes are often called "recalcitrant" (see e.g. D'Arms and Jacobson, 2003), I call this kind of cognitive dissonance *recalcitrance dissonance*. Judging an attitude

[7] This is so even if the conditioning mechanism by which we came to associate attitudes with things like demonic threats were to work like so called "bait shyness," in which we become averse to a kind of food after only one subsequent bout of illness (see e.g. Zimbardo and Weber 1997, 210-211). Such a mechanism would still require one conditioning trial to operate, which marks a contrast between its causal influence and that of fittingness judgments.

[8] A complication here concerns whether we judge the attitude to be merely justified (which is the attitude we often take to anger as a response to blameworthy acts) as opposed to rationally required, or irrational not to have (which we might often think is true of preferences for better states of affairs and guilt in response to blameworthy acts one has performed oneself - at least when one has no more pressing matters to tend to, when little time has passed, etc.). When one merely judges an attitude fitting in the sense of justified, one need not experience this kind of cognitive dissonance should one fail to have it. The kind of cognitive dissonance in question is present when one fails to have attitudes one judges to be rationally required, including when one fails to have attitudes one would ordinarily take to be merely justified but such that in one's circumstances one has nothing better to engage with and that the reasons for having it (given that engagement is rational) are stronger than the reasons against. I am inclined to understand such assessments of rational requirements for attitudes as a joint function of answers to what Gibbard (1990, 126-127) calls questions of whether or not to engage one's attitudes in a certain way and the question of what attitudes to have given that it is rational to engage them.

fitting seems to be distinct, then, in that when its causal influence fails to be decisive one feels recalcitrance dissonance. One does not experience recalcitrance dissonance as a result of mechanisms like disesteem for not having an attitude or classical conditioning failing to be enough to cause one to have it.

I suspect that we cannot solve the WKR problem by simply saying that any judgment about reasons to have an attitude that exerts direct causal pressure on one's having it and has the above distinguishing features is a fittingness judgment. I suspect that with enough neurosurgery, external machinery, and divergence from our actual psychology we could construct cases in which judgments about non-fittingness reasons have these features that only judgments about fittingness reasons typically actually do. What I think we can do is use these features that distinguish fittingness judgments from other psychic processes to hone in on the particular mechanism involved in judging an attitude fitting. We can then use this mechanism to explain the difference between judging an attitude fitting and judging that one has non-fittingness reasons to (get oneself to) have it.

I think that fittingness judgments' mechanism of attitude causation is best understood in terms of a mind state we may call *norm acceptance*. We naturally talk about accepting norms for attitudes like beliefs and credences. We say, instance, that we accept modus ponens as a norm of deductive inference, that we accept norms that prescribe that *ceteris paribus* we believe the simplest explanation of a phenomenon, and that we accept norms that require that we conform our degrees of belief to the axioms of the probability calculus. When we speak in this way about accepting norms for belief, we seem to have in mind a state that is neither a belief about how our beliefs are nor a desire about how we would like them to be. Rather, just as beliefs and desires are states with the functional roles of combining to directly produce behavior, these states of norm acceptance have the functional role of directly revising attitudes like beliefs and desires.

There are at least two models of how accepted norms play this role. On the first, to accept a norm just is for one to have a distinctive kind of tendency to conform to what it *actually* prescribes, rather than what one thinks it prescribes. On this model, if one accepts a norm of the form 'Feel F in circumstances C!', then when one takes oneself to be in C there will be causal pressure in the direction of feeling F whether or not one *thinks* that any norm one accepts has these prescriptions. On this model, there is no psychological fact of the matter about what the norms one accepts prescribe beyond the patterns of attitudes they actually tend to cause one to have. I thus call this the *shallow model* of norm acceptance.

The shallow model still requires that we can identify states that play distinctive functional roles in a subject for it to be true that she accepts norms for attitudes. As we have seen, one

distinctive feature of states of norm acceptance is that they play the role of directly influencing attitudes like beliefs and desires much in the way the latter states play the role of directly influencing behavior. Another distinctive feature of such states is that the failure of their causal influence to be decisive gives rise to what I above called recalcitrance dissonance. A final and important feature of states of norm acceptance is their ability to combine with each other to constitute a subject's accepting a system of norms.[9] As our examples of epistemic norms should make clear, we do not accept only one norm for a kind of attitude; our norms for attitudes form a system that works together to prescribe a response. This system involves higher order norms that directly govern, not beliefs and desires, but the acceptance of other norms themselves. For a state to be one of norm acceptance requires that it be able to exert causal pressure on, or have causal pressure exerted on it by, other states of norm acceptance in a system.[10]

An alternative to the shallow model of norm acceptance is one according to which accepting a norm involves having a tendency to conform, not to its actual prescriptions, but to the prescriptions one represents it as having. On this model, one can accept a norm that genuinely prescribes feeling F in C, take oneself to be in C, but still not be influenced by the mechanism of norm acceptance to feel F if one does not represent this norm as prescribing feeling F in C. According to this model there are deep psychological facts about what the norms we accept prescribe above and beyond the patterns of attitudes we are caused to have by the mechanism of norm acceptance. According to it we have representations of what the norms we accept prescribe, these representations can be erroneous, and when they are erroneous it is the prescriptions we represent our norms as having, rather those they actually have, that determine the influence of norm acceptance on our attitudes. I call this the *deep model* of norm acceptance.

The deep model would allow us to say, for instance, that in cases of fallacious deductive reasoning we conform our attitudes to patterns like denying the antecedent without actually accepting norms that prescribe our doing so; we merely represent our norms as such. According to this model we have a way of representing the norms we accept without being able to tell what their prescriptions are for all specified cases. At the same time the representation's mode of

[9] And perhaps even more complex structures like her ruling out sets of systems of norms. On the notion of ruling out systems of norms, see (Gibbard 1990, chapter 5).

[10] Gibbard (1990, chapter 4) suggests an alternative way of understanding states of accepting norms for attitudes in terms of tendencies toward avowal and normative discussion. I fear, however, that the evolutionary story that would support such an account, according to which the adaptive benefits of normative thought and language concerned coordination via consensus formation, has little plausibility outside the moral case – for instance in the case of the adaptive function of our epistemic and prudential normative judgments. For this reason (as well as intuitions that concern the inessentiality of public language and communication to normative thought), I do not think that we can use such public language criteria to explain what it is to accept norms for attitudes.

presentation cannot be 'whatever the norms I accept prescribe'. Since one's having the relevant representations is part of what it is to accept a norm, this would be viciously circular, if not also highly implausible. Perhaps the states that represent what our norms prescribe must be such in virtue of their bearing a certain nomic relation to the states that encode the norms we accept. Candidates for this nomic relation might resemble the kind of information carrying under ideal conditions discussed by Dretske (1981) or the kind of asymmetric causal dependencies discussed by Fodor (1987, 1990, 1994). On the deep model states of norm acceptance are still distinctive in terms of their forming systems that directly influence attitudes and give rise to recalcitrance dissonance when their influence fails to be decisive. But according to this model the causal influence of accepted norms on attitudes like beliefs, desires, and other norms is exerted by representations of what attitudes are prescribed by the system of norms one accepts.

The shallow and deep models of norm acceptance each give rise to a natural way of explaining the direct influence that fittingness judgments have on our attitudes. These respectively identify judging that it would be fitting to have an attitude with (1) shallowly *accepting* norms that prescribe it, and (2) *judging* that the norms one deeply accepts prescribe it. I call (1) and (2) *norm acceptance analyses* of fittingness judgments. Various norm acceptance analyses can be given corresponding to whether one opts for (1) or (2) and what one says about what it is to make judgments about other agent's fittingness reasons. I shall focus on the following two norm acceptance analyses, which I think are the most plausible:

Norm Expressivism:
To judge that an agent, A, has fittingness reason to have attitude F is to shallowly accept a system of norms that prescribes having F in A's circumstances.[11]

Norm Descriptivism:
To judge that an agent, A, has fittingness reason to have attitude F is to judge that A deeply accepts a system of norms that prescribes having F.

[11] This version of Norm Expressivism is essentially Gibbard's (1990, 86-92) "second approximation" of Norm Expressivism. A version corresponding to Gibbard's final view would need to make reference to something like ruling out sets of ordered pairs of complete systems of norms and possible worlds (see Gibbard 1990, 94-99). As Gibbard points out, this final version of Norm Expressivism can better handle several problems the second cannot, including that of giving a semantics of normative language in embedded contexts that can solve the Frege-Geach problem. It is for the sake of simplicity and ready comparison with Norm Descriptivism that I explicitly discuss the second approximation. My second approximation talk could be replaced with third approximation formulations without any substantive effect on what I have to say.

Given the shallow and deep models of norm acceptance, Norm Expressivism and Norm Descriptivism each entail that an agent's judging that she has reason to have F in her current circumstances will exert the distinctive causal pressure of norm acceptance in the direction of her having F. [12]

By giving an independent account of what it is to judge an attitude fitting in terms of the attitude of norm acceptance, norm acceptance analyses like Norm Expressivism and Norm Descriptivism give us a way to solve the WKR problem. To judge it fitting for one to have an attitude is to accept or judge that one accepts norms that prescribe it. To judge the attitude to be supported by non-fittingness reasons, like 'the demon will harm my loved one's if I don't have the attitude' or 'I'll be a coward if I don't have the attitude' is to be in a different state of mind. Most plausibly this different state is one of accepting or judging that one accepts norms that prescribe being motivated to get oneself to have the attitude or disesteeming not having the attitude. This would explain why these other judgments exert the direct causal influence characteristic of norm acceptance on attitudes like motivation to get oneself to have an attitude and disesteem for not having the attitude, but not the attitude itself. If we like we can still follow Rabinowicz and Ronnow-Rasmussen and call these judgments that one has reasons for the attitude in question. But we will have found a way to distinguish them from fittingness judgments, in a way that explains which attitudes they are incapable of directly affecting, and in a way that does not invoke the ethical concepts the FA-analyst is trying to analyze.

IV. The Importance of Attitude Guiding Content: A Dilemma for Judgment Externalism

I think that Norm Expressivism and Norm Descriptivism enjoy significant theoretical support quite independently of their ability to solve the WKR problem. Above I noted that the effects judgments about reasons have on our attitudes seems to be an important part of what sets them apart as prescriptive rather than merely descriptive. It is thus an independently important advantage of Norm Expressivism and Norm Descriptivism that they can, unlike most other accounts, explain the distinctive way in which judgments about reasons influence our attitudes.

Most views according to which fittingness judgments are descriptive beliefs are what we might call 'judgment externalist'. According to these views there is no conceptual connection between an agent's judging it fitting that she feel F and there being causal pressure in the

[12] It is of course consistent with Norm Descriptivism that there could be entities that falsely believe that they deeply accept systems of norms, and thus do not make causally efficacious fittingness judgments. But as we shall see the Norm Descriptivist has a principled reason to deny that these entities are agents or the kinds of entities that have reasons and can reason their way to attitudes.

direction of her feeling F. I think that these kinds of accounts violate our intuitions about what makes normative judgments special and different from merely descriptive judgments. They also face an important dilemma. On the one hand, considerations of explanatory parsimony suggest that there are no descriptive but analytically irreducible facts about which attitudes are fitting, yet this seems to be a bad reason to embrace error theory about what to feel. On the other hand, attempts by judgment externalists to analytically reduce facts about which attitudes are fitting to other kinds of facts seem open to objections reminiscent of Moore's "open question argument."

Explanatory parsimony gives us reason to think that there exist only those descriptive facts that either figure into our best explanation of the total phenomena or get analytically entailed by it. The former presumably include the facts discussed by fundamental physics, while the latter include facts about averages and (arguably) things like color, heat, chemistry, and psychology.[13] There must, however, be some constraints on what will for the sake of the parsimony principle be allowed to count as the "total phenomena", or it would have no teeth. We might think that if we can explain such things as the appearance of presents on Christmas and Mommy and Daddy's telling us about Santa Claus without reference to facts about Santa, then parsimony dictates we should not believe there are any such facts. But what if the defender of Santa-facts objects that such things as Rudolph's being in the lead position on Santa's sleigh are phenomena that need to be explained, and that we *do* need facts about Santa to explain them?

One good response seems to be the following. Facts about the Santa, Rudolph, and flying sleighs are not only unnecessary for explaining things like the appearance of presents and why Mommy and Daddy tell us certain stories. Facts about such entities are also unnecessary for explaining whatever beliefs anyone has about them. The general lesson seems to be that considerations of parsimony dictate the following. If we do not need a certain kind of descriptive fact to explain anything else, and we can best explain all of our beliefs about such facts without invoking them (or an explanation that analytically entails them), then we should not believe that there are any such facts.[14]

[13] The example of averages is Harman's (1977). For an excellent treatment of the case that these other kinds of facts are analytically entailed by our best explanation of what there is, see Jackson (1998).

[14] See for instance Harman (1977) and Gibbard (1990, 2003). David Enoch (2007) has recently objected to this criterion, arguing that it is enough if belief in a kind of fact is indispensible to a "non-optional" project for it to be the case that we should believe that it exists. Enoch says that by 'non-optional' he is unsure whether he means projects from which "we *can*not disengage, or rather those we should not disengage, or perhaps some combination of the two." I am quite unclear what Enoch means by a project 'we cannot disengage'; whether read as a claim about psychological, metaphysical, or conceptual impossibility it does not seem that either of his two examples – deliberation and explanation – really qualify. I am also quite unclear as to why it would be at all plausible to claim that simply because we should engage in project P and project P requires belief in facts of kind F that we have *epistemic* reason to believe in facts of kind F. But what really baffles me is how, given that he is a descriptivist and

If we need normative facts to explain anything, it seems that it must be something about the attitudes, behavior, or normative judgments of agents. But to proximally explain agents' attitudes and behavior we need only their normative judgments, quite independent of their truth or falsity. To explain these judgments, we need facts about agents' acculturation, which may be largely explained by the normative views of those around them. But follow the chain of acculturation back, and it looks as though you need explain only how tendencies to make certain normative judgments got passed down via the mechanisms of biological or cultural evolution.[15] In the case of our capacities to form beliefs about such things as tables, chairs, and electrons, we need to posit the reliability of these mechanisms in tracking their subject matter to explain why they would enhance survival and reproduction and thus get passed down. But we do not need such further facts about fittingness judgments tracking a realm of irreducible normative facts *in addition to* their directly tracking such things as survival and reproduction in order to explain the mechanisms by which we make judgments about which attitudes are fitting (or which states of affairs are good, which actions are blameworthy, etc.).[16]

It looks, then, as though we will need to posit descriptive facts about which attitudes are fitting that were successfully "tracked" in evolution only if such facts are identical to something else we do need in our evolutionary story. This would be so if, for instance, facts about which states of affairs are fitting for an agent to desire are identical to facts about which states would have promoted her reproductive success in ancestral environments. The problem is not simply that it is preposterous to think an agent should desire all and only states of affairs that have this property. As with other kinds of descriptive facts, we should only believe in facts about the identity between fittingness facts and other kinds of facts if they enter into (or are analytically entailed by) the best explanation of something - at the very least our believing in such identities. We should, for instance, believe that facts about water are identical to facts about H_20 because this enters into (or is analytically entailed by[17]) our best explanation of what water is like and how we came to have the beliefs about it we do. But we should not, for instance, believe that

---

not an expressivist quasi-realist, Enoch can think that the deliberative project – that of figuring out what to do and why – is anything other than a sub-element of the explanatory project of figuring out what is the case and why.

[15] I actually do not think that anything depends upon the details of the true story of how we came to make the normative judgments we do. Irreducible normative facts would seem just as superfluous had we been set up to make normative judgments by deities, or had we been spontaneously generated a few moments ago by lighting hitting a swamp, or whatever. I stick to the actual evolutionary story for heuristic purposes.

[16] An argument like this is given by e.g. Harman (1977), Blackburn (1988), Gibbard (1990, 2003) and Street (2006).

[17] See Lewis (1970) and especially Jackson (1998) for a powerful case that this identity follows by analytic entailment from our best theory.

facts about the North Pole are identical to facts about Santa's actual home because no such facts enter into (or get analytically entailed by) our best explanation of how things are.

An identity between facts about which attitudes are fitting and facts about things like reproductive success would add nothing to our evolutionary story or any other part of our explanation of why people make the fittingness judgments they do. So unless the identity between fittingness facts and some other kind of facts follows analytically from our best explanation of how things are, considerations of parsimony entail that we should not believe there are any such identities. We have already seen that without such identities, considerations of parsimony entail that we should not believe there are any descriptive facts about which attitudes are fitting. Considerations of parsimony thus commit a judgment externalist who denies that there are analytic identities between fittingness facts and other facts to saying that fittingness judgments are beliefs about descriptive facts that do not exist. This means that these beliefs are just as mistaken and untrue as beliefs about Santa Claus. But surely the explanatory irrelevance of analytically irreducible facts about fittingness is a terrible reason to embrace this kind of error theory about what to feel and correspondingly what is good, blameworthy, estimable, and so on. So it cannot be that fittingness judgments are beliefs about such irreducible facts.

The only other option for the judgment externalist would be to hold that facts about fittingness are *analytically* reducible to some other kind of facts that we do need for explanatory purposes. Some of the most plausible reductions might analyze judgments that an attitude is fitting as beliefs that one would have the attitude if one were under some particular, non-normatively specified conditions.[18] Or the judgment externalist might hold that to judge that an attitude is fitting is to believe that it conforms to certain abstract rules (quite independently of whether one accepts them as norms), in the same way that this might be true of judging that a move conforms to the rules of a game like chess. The problem here is that tests reminiscent of Moore's "open question argument" suggest that these analyses are out of line with our intuitions about which fittingness judgments are coherent. It seems that for any descriptively specified conditions or set of abstract rules, one could coherently judge that one would have an attitude under the conditions, or judge that the attitude would conform to the rules, yet judge that the attitude is not fitting for one to have.

[18] For instance, a version reminiscent of Firth's (1952) account of moral judgments would hold that these conditions are those of full information and equal, vivid attention to all the (prosaically descriptive) facts, and a version reminiscent of Brandt's (1979) reforming analysis of judgments of rational desires would hold that these conditions are those of having undergone Brandtian "cognitive psychotherapy".

Gibbard (1990, 10, 22) in effect argues that the best diagnosis of the proneness of any such judgment externalist theory to the above problems is that judgments about which attitudes are fitting are essentially attitude guiding. Non-reductive forms of judgment externalism founder because fittingness judgments are more closely bound up with attitude guidance than with representing any class of explanatory facts. Analytically reductive forms of judgment externalism also founder because fittingness judgments are more intimately connected to attitude guidance than the descriptive beliefs that are offered as their *analysandi*.

Gibbard goes on to propose a Norm Expressivism as a natural alternative, arguing that the right lesson to draw from these problems with judgment externalism is that to judge an attitude fitting is not to make *any* descriptive judgment. Now, as we have seen, by employing the deep model of norm acceptance, Norm Descriptivism can, unlike the above judgment externalist views, both explain the attitude guiding character of fittingness judgments and maintain that these judgments are descriptive beliefs about what the system of norms one deeply accepts prescribes. The proponent of Norm Descriptivism can thus agree with Gibbard's diagnosis of the problems for judgment externalism without agreeing that the lesson to draw is that normative judgments are no species of descriptive judgments.

V. Reasons to Favor Norm Descriptivism

Norm Expressivism enjoys many theoretical advantages apart from its ability to explain how fittingness judgments guide attitudes which I have no room to discuss (see e.g. Gibbard 1990, 2003). If I do nothing else in this paper, I hope to have shown how Norm Expressivism can help solve the WKR problem for FA-analyses of ethical concepts, to the further credit of both Norm Expressivism and such FA-analyses. But as impressed as I am with Norm Expressivism, I do not think that it is true. In this section I sketch my reasons for preferring Norm Descriptivism.

Judgments about which attitudes are fitting not only guide attitudes. They are also states the truth and falsity of which we inquire into when we deliberate about how to feel – for instance when we deliberate about which states of affairs are good or which acts are blameworthy. Norm Expressivism might not have much trouble explaining what we are doing when we assume answers to basic questions about how to feel in fully stipulated circumstances and ask descriptive questions about which circumstances we occupy.[19] The problem is explaining what we are doing when we engage in basic inquiry about how to feel in a circumstance assuming that it is a certain

[19] Expressivist explanations of normative inquiry along these lines go back at least to Ayer (1936).

way. This kind of basic inquiry into what attitudes to have in a given circumstance is what normative philosophy specializes in.

It is far from straightforward what kind of account of basic first-personal normative inquiry can be given by Norm Expressivism, which identifies judging that one should have an attitude with simply shallowly accepting a system of norms that prescribes having it. Norm Expressivism seems largely motivated by an attempt to make sense of interpersonal normative discussion. It can explain changes of basic normative opinion as a result of a brute tendency of systems of shallowly accepted norms to alter on contact with one another.[20] But this explanation seems to leave little room for understanding what goes on when we try to figure out basic questions of what to feel for ourselves, or what happens when we change our minds about these matters as a result of first-personal inquiry.[21] Given Norm Expressivism's focus, it is not too surprising that the few things Gibbard (1990) says about basic first-personal normative inquiry look like attempts to explain it as a kind of rehearsal for interpersonal normative discussion:

> A person will engage in imaginative rehearsal for actual normative discussion; he practices by holding himself to consistency. The pressure for consistency need not be so strong as it is in good philosophical discussion, but it will be there, and it may be significant…In trying to decide what is rational, we are engaging our normative capacities to try to decide what norms to accept. We do this in normative discussion, actual and imaginative, as we take up positions, subject ourselves to demands for consistency, and undergo mutual influence (Gibbard 1990, 74-75, 81).

I fear, however, that it is unclear how the details of this account are supposed to go, and the more one tries to fill them the less it looks plausible. It would seem fantastic to suppose that deliberation is a process by which our minds fracture into parts that accept different systems of norms. The mere fact that one deliberates does not seem to entail that one is subject to conflicting pressures on one's attitudes.

Perhaps it is enough that we simply imagine talking to characters we think of as accepting a certain systems of norms. I think that this kind of imaginative rehearsal for normative discussion does take place, for instance in planning for political debates or presenting material to students. But this looks a lot more like deliberating about how to be most rhetorically or pedagogically effective in convincing someone of conclusions one already accepts than deliberating about whether these conclusions are correct. Of course, this kind of imaginative

[20] See Gibbard (1990) chapter 4.

[21] It is interesting to note that W.D. Falk (1963, 1986) claimed that in his day expressivist metanormative theories seemed primarily developed in order to account for the dynamics of normative discussion, which caused them to face problems in accounting for first-personal deliberation. In this respect little may have changed.

rehearsal for debate or presentation can sometimes *prompt* genuine deliberation about what to feel. But to the extent it does one seems to stop engaging in an imaginary normative discussion and to start doing something else that is not essentially tied to rehearsal for discussion.

A natural diagnosis of Norm Expressivism's difficulty explaining basic deliberation about which attitudes are fitting is the following. In general, when we inquire into whether or not P is true, we attempt to arrive at knowledge that P or that not P. But to know that P seems to require that P's truth figure into (or be analytically entailed by) the explanation of one's belief that P.[22] As such, deliberative inquiry into which attitudes are fitting seems to aim at a state where the truth of one's fittingness judgments explains (or is analytically entailed by what explains) them. This seems to require that fittingness judgments are descriptive beliefs with contents that can enter into or be analytically entailed by explanations of descriptive phenomena, rather than the non-descriptive states with which the Norm Expressivist identifies them.

But as we have seen, fittingness judgments cannot be any old kind of descriptive beliefs. Our intuitions suggest, and the dilemma for judgment externalism confirms, that unlike most kinds of descriptive beliefs, fittingness judgments are able to directly guide our attitudes. We have also seen how fittingness judgments seem analytically independent from most descriptive judgments, and we have noted how basic deliberation about which attitudes are fitting proceeds via *a priori* reflective equilibrium methods.

[22] Like Dretske's (1981) account, according to which instances of knowledge that P must be beliefs that P that are caused by the information that P, this requirement subsumes what might seem right with the causal theory of knowledge, but can do better by explaining the kind of causal-explanatory relationship that obtains between the truth of propositions about the future and knowledge of them. This formulation covers cases of knowledge that P where the truth of P is a cause of one's belief that P, but also covers cases where some other fact F jointly causes P and one's belief that P. In the latter kind of case, P's truth does not play a causal role in the explanation of one's belief, but is analytically entailed by it, and this is exactly the kind of causal-explanatory relationship that Dretske (1981) argues obtains between P's truth and one's belief that P in the case of knowledge about the future. Moreover, because P's causing B is presumably not the only kind of way in which P can enter into the ontic explanation of, or reason why it is the case that B, I think that my formulation also covers (where Dretske's does not) the kind of explanatory relationship that must obtain between the truth of necessary propositions and instances of knowledge of them. See also Gibbard (2003), chapter 13 on the notion of "deep vindication" and knowledge in the "more demanding sense" for a related formulation.

I think that this requirement can easily be extended to indeterministic cases, in which the ontic explanation of an event does not entail that it occurs. We just need to make provision for the indeterministic version of the relevant non-fluky metaphysical connection between belief that P and P's truth. I think the extension would be achieved by revising the necessary condition on knowledge that P so as to include not only the disjuncts (i) 'P's truth enters into the explanation of one's belief that P', and (ii) 'P's truth is analytically entailed by one's belief that P', but the additional disjunct (iii) 'a high objective probability of P's truth is analytically entailed by the explanation of one's belief that P'. Since nothing I say would be substantially altered by adopting this more accurate formulation that would cover the indeterministic case, I stick to the simpler formulation in the text, begging the careful reader to treat my use of it as shorthand for the more accurate one.

By embracing the deep model of norm acceptance, Norm Descriptivism can explain how fittingness judgments are descriptive beliefs the truth of which can explain our making them *and* how they have these distinctive features. According to the deep model of norm acceptance, accepting a system of norms involves its being the case that judgments that it prescribes an attitude exert causal pressure on one's having it. These judgments are descriptive beliefs, albeit ones that we can hold rather independently of most other descriptive beliefs.

I think that a plausible explanation of basic normative inquiry is the following. We have more immediate (though not infallible) access to what the norms we deeply accept prescribe in particular cases, and perhaps also to what kinds of differences between cases do or do not matter for what they prescribe. These more immediate representations of our norms' prescriptions are normative intuitions about what to feel in particular cases and principles about what differences between cases affect our reasons for attitudes. What we do in normative inquiry is elicit these intuitions and attempt to construct an account of what the system of norms we deeply accept prescribes that would best explain our having the intuitions we do. This will presumably include some intuitions having to be debunked as thrown up by causal processes other than accurate perception of what the system of norms we deeply accept prescribes. The epistemic access we have to the prescriptions of the norms we deeply accept through this reflective equilibrium process is in a sense an *a priori* form of access – the evidence set for what our norms prescribe in any fully stipulated situation does not consist of or rely on (external) sense experience.[23] By applying this method carefully, our beliefs about what our norms prescribe may very well be caused by facts about them.[24] Norm Descriptivism's identification of judgments about what to feel with beliefs about what attitudes the system of norms we deeply accept prescribes can thus explain how basic deliberation about what attitudes are fitting can hook onto facts about fittingness and yield *a priori* knowledge of them.

While Norm Descriptivism portrays fittingness judgments as analytically independent of most descriptive judgments, it does entail that one cannot coherently think both that the system

[23] That is, it does not rely on the experience of representations generated by those of our senses that transduce information about the external world. It might be, however, that normative intuitions are themselves quasi-perceptual representations, and it may (though it may not) make sense to speak of the processes which generate them as involving something similar to a sensory pathway that detects what is prescribed by norms we accept.

[24] Much in the way that applying the method of inference to the best explanation can yield knowledge of, say, physical phenomena, where intuitions about what our norms prescribe in clear cases play an analogous role to observations or external world perceptions in the physical case. This role includes these states conveying information, or in particular being caused by, facts about their subject matter; it is simply that where external sense perception involves content about and is caused by facts about the external world, intuitions about what one's norms prescribe have as content, and (often enough) causal origin facts about the norms one deeply accepts.

of norms one accepts prescribes feeling a certain way and that one should not feel that way.[25] I think, however, that the dual roles normative judgments play in guiding our attitudes and answering our deliberative questions is good evidence in favor of this analytic relationship. Suppose an agent were to ask herself "should I follow the prescriptions of the system of norms I accept?" The Norm Expressivist looks committed to saying that she is engaged in a kind of wondering that is not about whether any descriptive fact obtains. I find this obscure. Perhaps the idea is that she is signaling her receptiveness to the kind of brute influence the Expressivist claims normative discussion would have on her.[26] But it is very hard to see how asking for this kind of influence and receiving it could in any way answer the agent's deliberative questions about what to accept. I suspect that this is because such influence could not, on the Espressivist's picture, bring the agent closer to a state where facts about what she should accept enter into (or are analytically entailed by) the explanation of her views about what to accept.

I think that the only way for wondering about whether to accept or follow the prescriptions of the norms one accepts to have both the causal and the epistemic properties it has is for it to be an instance of wondering about what the system of norms one deeply accepts prescribes. Of course, it is far from *obvious* that this is what it is. So it is entirely understandable how someone could be led to ask whether she should have whatever attitudes the system of norms she deeply accepts prescribes, even if Norm Descriptivism is correct and the answer to the question is "yes, as a matter of conceptual necessity."

Norm Descriptivism also entails that one cannot coherently think of another agent both that the system of norms she deeply accepts prescribes feeling a certain way and that she should

[25] One might wonder in what sense it would be true if, as I suggested (in Section III), the representations Norm Descriptivism identifies as fittingness judgments represent what the norms we accept prescribe in virtue of their bearing certain causal-indicator kinds of relationships to the states that encode these norms. I suspect that matters here are analogous to how they are for analytic functionalists about qualia who think that our ordinary qualia concepts are phenomenal. In a sense the analytic functionalist will want to say that it's analytic that whatever states play the qualia-roles are qualia, even though we ordinarily represent these states with phenomenal concepts that do not make explicit reference to the qualia-roles. Presumably what is going on is that we have a concept of what a (phenomenal) qualia concept would have to be like, and that analysis of this concept tells us that the concept must stand in a kind of nomic relationship to whatever states play the qualia roles. Perhaps such analytic truths about a concept of a qualia concept that bridges phenomenal concepts and qualia-roles licenses talk of the relationship between qualia themselves and the qualia-roles as analytic. If this is right, and normative judgments are what they are in virtue of nomic relationships to accepted norms, then perhaps I am trying to do the same thing with our ordinary normative concepts and the concept of a normative concept (which speaks of such things as how normative judgments guide attitudes and we can access the truth and falsity of basic normative judgments via *a priori* reflective equilibrium methods). I am grateful to John Ku for helpful discussion about this matter.

[26] If there are no actual others with whom she is talking, perhaps the idea is that she makes this request or signals this receptiveness to influence to the imaginary interlocutors with whom Gibbard (1990) seems to suggest she rehearses for normative discussions. The suggestion that claims of normative uncertainty request or signal receptiveness to influence is also Gibbard's (personal correspondence).

not feel that way. It is crucial to bear in mind that another agent's having reason to respond in a certain way is distinct from its being good that she respond that way or one's having reason to try to get her to respond in that way. Unlike merely judging that the latter kinds of things are true, judging that someone has reason to have an attitude seems to involve thinking that she could correctly reason her way to having it. Correct reasoning is not just any old process whereby an entity comes to be as we have reason to want it to be. Were we to neurally alter a shark so that he no longer attacks innocents, he would not have thereby correctly reasoned himself to refraining from doing so. Correct reasoning seems rather to be a matter of going from what one accepts to what is genuinely prescribed by what one accepts. But if this is right, then how could we maintain that an agent has reason to have a response unless the system of norms she deeply accepts prescribes that she have it?

The dependence of reasons on reasoning seems important for explaining why it makes no sense to think that reasons apply to the responses of beings like infants and sharks (as we take them to be). To be sure these beings are capable of well being or welfare, of things going better or worse for them. But a response's being good or bad for a being is distinct from her having reason to have it. As Darwall (2002) has convincingly argued, to judge of a being's welfare is to judge of our reasons to want things for her out of sympathetic concern, which in no way requires her to be subject to reasons herself. To think beings like infants and sharks not only capable of welfare but also subject to reasons, we would seem to have to think them candidates for the kind of rational criticism involved in calling someone an idiot for making a foolish decision. Infants and sharks are of course capable of greater or lesser intelligence or learning ability, but it seems incoherent to hold them genuinely rationally criticizable on the assumption that their minds are as we take them to be. It is true that rationally criticizablility involves more than simply failing to respond to the reasons one has; an agent can fail do this but be rationally exculpated on account of diminished responsibility. But the way beings like infants and sharks lack rational responsibility for their behavior is importantly unlike that of, say, an otherwise psychologically typical adult human whose perennially recalcitrant emotions always sway her against her better judgment. Infants and sharks are incapable of this very kind of "better judgment" to which it seems one must be able to reason correctly to be subject to reasons at all.

Norm Expressivism entails that one can coherently judge of an entity both that she has reason to have a response and that there is no way she could correctly reason her way to having it from what she already (deeply or shallowly) accepts. This seems to put it at odds with our intuitions about the senselessness of holding the responses of entities like infants and sharks subject to reasons. The Norm Expressivist might try to prevent this by requiring, for instance,

that to (shallowly) accept a norm that prescribes having a certain response in a circumstance, that circumstance must be such that in it one can accept norms, or norms that prescribe having or not having that response.[27] But this looks *ad hoc*. Why can one shallowly accept norms that prescribe having a response in a circumstance only if one could accept norms (for or against the response) in that circumstance? This does not seem to fall out of the account of shallow norm acceptance. It rather looks tacked-on by the expressivist to accommodate intuitions about the incoherence of certain normative judgments that her view does not predict. Unlike these fixes to Norm Expressivism, the view that having reason for a response requires being able to correctly reason one's way to having it offers us a genuine explanation of why it is senseless to hold beings like infants and sharks (as we take them to be) subject to reasons. This view's consequent superiority, and its favoring Norm Descriptivism over Norm Expressivism, is important evidence of Norm Descriptivism's superiority.

I should conclude by considering the fact that Norm Descriptivism entails that it is conceptually possible for it to be fitting for different agents to have different attitudes in the same circumstances. I think that we do assume that differences in what each of us has reason to feel must be traceable to differences in our (norm-independent) circumstances. But I think that this is because we assume that we all deeply accept the same basic system of norms. We seem to confidently expect that, if we simply do things aright, our *a priori* reflective equilibrium methods will lead us to the same attitudes in the same circumstances. I believe that this is manifest in both professional philosophical contexts and the normative thought and talk of everyday life.

Our presupposition that we deeply accept the same system of norms could of course be mistaken, but I think that our losing confidence in it would mandate losing our confidence that differences in reasons trace to differences in norm independent circumstances. This would mean that we could not correctly reason our way to the same responses in the same circumstances. To insist on saying that one of us should respond in a way other than that to which she could correctly reason threatens to change the subject form reasons to something less philosophically interesting. It would at least yield the problems of either judgment externalism or expressivism.

Fortunately, I see no reason to think that our presupposition that we all deeply accept the same basic system of norms is mistaken. We seem to have no reason to doubt that differences in the attitudes parties think they should have will prove traceable to processes that some party would be able to regard as distortionary effects on the best explanation of her intuitive starting

[27] I am grateful to Allan Gibbard for these suggestions.

points. In fact, I think that there are good evolutionary reasons to think that we came to accept the same basic system of norms as a universal adaptation that allows our attitude systems to play their adaptive roles more flexibly.[28]

But suppose that at some point in the future we come across a race of space aliens who deeply accept different systems of norms for attitudes like desires, and that both we and they make ethical judgments, for instance about the goodness of certain states of affairs. I have been arguing that norm acceptance analyses of fittingness judgments can be used to solve the WKR problem for otherwise attractive FA-analyses of ethical judgments. In this section I have suggested that the Norm Descriptivism is in fact the best norm acceptance analysis for the job. But if we combine FA-analyses of ethical concepts with the Norm Descriptivist analysis of fittingness assessments, we need to ask: whose reasons for desire are we thinking about when we judge a state good?

This is actually just a special case of a question that *any* FA-analyst needs to answer. So long as it is coherent to think that differences in circumstances can make for differences in what one should desire (as surely it is), we must give an account of whose reasons for desire one thinks about when one judges a state good. Gibbard (1998) considers exactly this problem, and suggests that we can solve it by viewing goodness judgments as context sensitive:

> Fans of the Michigan Wolverines football team can exclaim together about what's good and what's bad as a game progresses, without bringing in the standpoint of anyone else. But the talk of good and bad is still neutral among participants in the conversation. The same talk would be tendentious in dealings with people from Ohio. It couldn't be regarded as seriously defensible, or if it were, the grounds would have to be different.
>
> Likewise I can think to myself about good and bad developments as, say, I compete for a job…though with luck I may be able to rope in a few friends to share my good news and bad. They can't, though, be equally friends of another competitor, or I'll have to qualify my language or make the conversation awkward for them. (Gibbard 1998, 254-255).

One way to spell out Gibbard's contextualist suggestion is to understand the concept of a GOOD STATE OF AFFAIRS to be equivalent to that of A STATE OF AFFAIRS THAT IT IS FITTING FOR EVERYONE TO DESIRE, but to understand the universal quantifier in the *analysans* to have a contextually determined quantifier domain restriction[29] that determines that (at least) the tokener and anyone he is talking to is in the domain.

The Norm Descriptivist can simply adopt this solution and apply it to contexts where she thinks differences in reasons are due to factors other than differences in (norm independent)

[28] Quartz and Sejnowski (2002) present a highly engaging discussion of how rapidly changing and increasingly social ancestral environments would have favored this kind of flexibility.
[29] Of the kind discussed by Stanley and Williamson (1995) and Stanley and Szabo (2000).

circumstances. One thing that Norm Descriptivism and this kind of contextualist FA-analysis would entail is that ethical talk could not be used truly in conversations with agents whose deeply accepted systems of norms prescribe attitudes different from those prescribed by one's own. But this, I think, is as it should be. As Gibbard's examples of the football fans and friends of competitors illustrate, evaluative talk breaks down where shared reasons for attitudes break down.[30] When such breakdown occurs, we have no choice but to fall back on talk about which attitudes are fitting for whom, without any presupposition that the same attitudes are fitting for everyone. But of course, this retreat from ethical talk to talk of individuals' fitting attitudes does not really give up anything central to our normative thought if, as the contexutalist FA-analyst has it, the former is simply the latter with a presupposition of shared reasons for attitudes.

[30] This is not simply the kind of breakdown that occurs because one's audience is in the wrong but cannot be brought (or expected) to see it. As Gibbard's examples illustrate, where one's audience's reasons for feeling genuinely diverge from one's own, there is no fact of the matter about whose reasons are "really" the right ones. Both sets of reasons seem relevant not only to the felicity of such claims but in fact to their truth. If one's evaluative talk is supported by only one such set of reasons, it seems awkward *because* it falsely entails that it is supported by all.

REFERENCES

Ayer, Alfred J. 1936. *Language, Truth, and Logic*. 2nd ed. London: Victor Gollancz.

Blackburn, Simon. 1988. How to Be an Ethical Antirealist. *Midwest Studies in Philosophy* 12: 361-375. Reprinted in n S. Darwall, A. Gibbard, and P. Railton, eds., *Moral Discourse and Practice*. New York: Oxford University Press.

Brandt, Richard B. 1979. *A Theory of the Good and the Right*. Amherst, New York: Prometheus Books.

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

D'Arms, Justin and Daniel Jacobson. 2003. The Significance of Recalcitrant Emotion (or, Anti-Quasijudgmentalism). *Philosophy: The Journal of the Royal Institute of Philosophy*, 52 (suppl.): 127-145.

Darwall, Stephen L. 2002. *Welfare and Rational Care*. Princeton: Princeton University Press.

Dretske, F.I. 1981. *Knowledge and the Flow of Information*. Cambridge: MIT Press.

Enoch, David. 2007. An Outline of an Argument for Robust Metanormative Realism. In Russ Shafer-Landau, ed., *Oxford Studies in Metaethics: Volume 2*. Oxford: Oxford University Press.

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

Falk, W.D. 1963. Action-Guiding Reasons. *The Journal of Philosophy* 60: 703-18. Reprinted in Falk, *Ought Reasons, and Morality: The Collected Papers of W.D. Falk*. Ithica: Cornell University Press.

Falk, W.D. 1986. On Learning about Reasons. In Falk, *Ought Reasons, and Morality: The Collected Papers of W.D. Falk*. Ithica: Cornell University Press.

Firth, Roderick. 1952. Ethical Absolutism and the Ideal Observer Theory. *Philosophy and Phenomenological Research*, 12: 317-345.

Fodor, Jerry A. 1987. *Psychosemantics*: the Problem of Meaning in the Philosophy of Mind. Cambridge, Massachusetts: MIT Press.

Fodor, Jerry A. 1990. *A Theory of Content and Other Essays*. Cambridge, Massachusetts: MIT Press.

Fodor, Jerry A. 1994. *The Elm and the Expert: Mentalese and Its Semantics*. Cambridge, Massachusetts: MIT Press.

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

Gibbard, Allan. 1998. Preference and Preferability. In Cristoph Fehige and Ulla Wessels, eds., *Preferences*. New York: Walter de Gruyter.

Gibbard, Allan. 2003. *Thinking How to Live*. Cambridge, Massachusetts: Harvard University Press.

Harman, Gilbert. 1977. *The Nature of Morality: An Introduction to Ethics*. New York: Oxford University Press.

Jackson, Frank. 1998. *From Metaphysics to Ethics: A Defense of Conceptual Analysis*. Oxford: Oxford University Press.

Kavka, Gregory S. 1983. The Toxin Puzzle. *Analysis*, 43: 33-36.

Lewis, David. 1970. How to Define Theoretical Terms. *Journal of Philosophy*, 67: 427-446.

Olson, Jonas. 2004. Buck-Passing and the Wrong Kind of Reasons. *The Philosophical Quarterly* 54: 295-300.

Parfit, Derek. 2001. Rationality and Reasons. In *Exploring Practical Philosophy: From Action to Values*, edited by Dan Egonsson, Bjorn Petersson, Jonas Josefsson, and Toni Ronnow-Rasmussen, 17-41. Aldershot: Ashgate.

Quartz, Steven R. and Terrence J. Sejnowski. 2002. *Liars, Lovers, and Heroes: What the New Brain Science Reveals About How We Become Who We Are*. New York: HarperCollins Publishers.

Rabinowicz, Wlodek and Toni Ronnow-Rasmussen. 2004. The Strike of the Demon: On Fitting Pro-attitudes and Value. *Ethics* 114: 391-423.

Rabinowicz, Wlodek and Toni Ronnow-Rasmussen. 2006. Buck-Passing and the Right Kind of Reasons. *The Philosophical Quarterly* 56: 14-120.

Stanley, Jason and Timothy Williamson. 1995. Quantifiers and Context Dependence. *Analysis* 55: 291-295.

Stanley, Jason and Zoltan G. Szabo. 2000. On Quantifier Domain Restriction. *Mind and Language* 15: 219-261.

Street, Sharon. 2006. A Darwinian Dilemma for Realist Theories of Value. *Philosophical Studies*, 127(1): 109-166.

Velleman, J. David. 2002. Motivation by Ideal. *Philosophical Explorations*, 5: 90-104.

Zimbardo, Philip G. 2007. *The Lucifer Effect: Understanding How Good People Turn Evil*. New York : Random House.

Zimbardo, Philip G. and Ann L. Weber. 1997. *Psychology*. 2nd Edition. New York: Longman.