The accidental error theorist

Richard Joyce

Penultimate draft of paper appearing in *Oxford Studies in Metaethics* vol. 6, R. Shafer-Landau (ed.), 2011.

Introduction

The moral error theorist holds that morality is flawed in the same way as the atheist holds that religion is flawed: The discourse aims at the truth but systematically fails to secure it. Of the various possible arguments that might lead one to this form of moral skepticism, John Mackie's are best known and most influential (Mackie 1946; 1977; 1980). Mackie's skeptical case targets both moral prescriptions and moral values, and in both cases it is their putative association with a kind of *objectivity* that bothers him. He is not an error theorist about prescriptions and values *per se*; he is always careful to identify the target of his doubt as "*objective* prescriptions" and "*objective* values." It is possible that one might be a moral error theorist for entirely different reasons (see Joyce 2007: 52; Joyce forthcoming), but here we will stick with Mackie's grounds.

The moral error theorist's opponents can be divided broadly into three types—one of which will be the focus of this paper. First, some opponents are noncognitivists, who deny that moral discourse aims at the truth. For all those remaining opponents we can use the label "success theorists"—those who both embrace cognitivism and hold that moral discourse often succeeds in hitting the truth.[1] There are many ways of taxonomizing success theorists, but for present purposes—clarifying types of opposition to Mackie's error theory—they can be divided into two. Some success theorists respond to Mackie's argument *head-on*: They argue that the troublesome concepts *objective prescription* and *objective value* are, when properly understood, perfectly defensible. Other success theorists pursue a *concessive* strategy: They concede that Mackie's target concepts are indeed philosophically indefensible but they nevertheless resist the error theory by maintaining that substantial tracts of moral discourse remain unscathed by the skeptical critique.

Were this paper concerned with the head-on response, we would proceed to discuss what Mackie means by "objective prescription" and "objective value," investigate the merits of his famous Argument from Queerness against such phenomena, and so forth. But given that this paper is concerned solely with the concessive strategy, many of the details of that discussion can be put aside. The concessive success theorist and the error theorist *agree* that objective prescriptions and objective values are too queer to be admitted into our ontology, and since the present paper takes this agreement as its point of departure, we need not pause to wonder about its precise content or its reasonableness. (This focus should not be taken to imply that I take a dismissive attitude towards noncognitivism or the head-on strategy against the error theory—though I do confess to thinking that the concessive route is the most promising.)

The concessive opponent will allow that *if* morality consisted only of objective

[1] The term "success theory" comes from Sayre-McCord (1986).

prescriptions and/or objective values**,** then morality should indeed go the way of phlogiston and astrology. But, he insists, the antecedent doesn't hold. First, it may be argued that a notion like *objectivity* is just too recherché to be attributed to vernacular moral concepts; everyday moral discourse is committed to nothing of the sort. Second, it may be maintained that even if references to objective prescriptions and objective values are distinctly present in ordinary moral thinking, these are just two elements among others; the problematic elements could be discarded while leaving us with a perfectly serviceable and unflawed moral system. Third, it may be pointed out that even if our actual morality consisted entirely of these elements, it would not follow that they are an *essential* feature of morality; again, after excising the defective elements we would be left with a robust normative system still deserving of the name "morality." Hence, from the concessive strategist the error theorist faces three kinds of accusation of narrow-mindedness: She is criticized for projecting her own philosophical extravagances onto the ordinary folk; she is charged with seeing only a part of morality and assuming it to be the whole of morality; and she is accused of seeing only the contemporary Western moral tradition and assuming that morality *must* be this way.

These accusations of narrow-mindedness cannot stand unsupported, however. It is incumbent on the concessive opponent of the error theory to identify those elements of normativity that (A) survive the skeptical critique and (B) are sufficient to constitute a morality. In other words, the viability of the concessive strategy depends on the viability of the positive theory on offer. It is accepted by both relevant parties that the offering must not imply the existence of objective prescriptions or objective values (at least of the kind targeted by Mackie); the principal question is whether it can be recognizably a *moral* system.

In this paper I will identify and critically assess several metaethical theories that can be interpreted as offering a concessive response to Mackie's error theory. The first perspective is the dispositional view of moral properties, the discussion of several versions of which will comprise the bulk of this paper. The dispositional theorist can be interpreted as agreeing with Mackie that objective prescriptions and objective values are too weird to be philosophically tolerated, and as responding that we should therefore identify moral properties with a certain class of *non-objective* prescriptions and values. The second perspective is virtue ethics, which will be discussed more briefly towards the end of this paper. The virtue ethicist can be interpreted (though with some strain, I admit) as agreeing with Mackie that objective prescriptions and objective values are too weird to be philosophically tolerated, and as responding that we should therefore begin our ethical inquiry not trying to make sense of prescriptions and values, but rather begin with normative questions about *character*.

This paper is not so ambitious as to try to show that these programs cannot succeed—I will not get even close to that conclusion—but I do want to show that neither side-steps the error theory as easily as is sometimes thought. In particular, I shall argue that proponents of these views are often playing directly into the hands of the moral skeptic by unwittingly championing error theoretic views.

How is it possible to be an *accidental* error theorist? If Mackie is correct, then everyone who participates sincerely in moral discourse (which is, presumably, the vast majority of speakers) makes commitments that render the moral error theory true; and since speakers do not intend for this to be the case, there is some sense of "accidentality" involved. Yet to have

the error theory be true of you—or (speaking more carefully) true of a discourse in which you engage—is not to *be* an error theorist. To be an error theorist is to take a metaethical stance; it is to offer a concrete opinion on the nature of moral discourse. But how could one perform such a purposeful and reflective action *accidentally*? What I have in mind is that some moral philosophers offer metaethical views that are intended to be versions of success theory but which, unwittingly, commit them to an error theory. To give a toy example, suppose that a metaethicist defends a divine command theory (DCT): identifying the moral property of *moral obligation*, say, with *whatever is willed by God*. Such a person presumably will be a theist. But suppose that in fact atheism is correct and there are no gods. Thus, since the predicate "…is willed by God" has an empty extension, then so too will the predicate "…is morally obligated." And if we assume that this is so of all other moral predicates too, then all sentences of the form "X is M" (where M is any moral predicate and X names something actual) will be false. The divine command theorist will have purposefully and reflectively asserted a metaethical view, but one which inadvertently commits him to an error theory.

Notice that it is not just any old problem with the DCT that would lead to this result. Most traditional objections to the theory can be categorized as casting doubt on the acceptability of identifying moral properties with properties pertaining to God's will. If one or more such objection were to prove entirely victorious, then although the divine command theorist would be shown to be *in error*, he would not thereby be shown to be *an error theorist*. By contrast, the result upon which I am focusing depends crucially on that identity claim (or weaker biconditional, as the case may be) being accepted and embraced (if only for the sake of argument), and then the failure of the right-hand side implies an inadvertent error theoretic commitment.

Notice also that I am ignoring a potentially interesting distinction between those divine command theorists who, were they to come to believe that the right-hand side fails, would nevertheless maintain the DCT identity claim (thus acquiescing to the error theory), and those who would rather retreat from or revise this claim (thus avoiding the error theory). Those with the former disposition are in some sense more strongly "committed" to the error theory than those with the latter disposition. However, I propose to disregard this distinction; my interest here is not in how the proponent of the DCT would respond upon coming to believe in the failure of the right-hand side. Perhaps the existence of God's will is so strongly an article of faith for a given person that nothing could bring her to believe that the right-hand side suffers from failure of reference. Such a person may be committed to the error theory not in the sense that anything could get her to acknowledge the error theory, but in the sense that she advocates a claim from which it follows (it is at least reasonable to suspect, even if *she* cannot be brought to suspect it) that no moral predicate is instantiated.

The dialectic can be simplified as follows. Suppose Ernie accepts proposition P but rejects proposition Q. Bert, by contrast, supports Q, and also believes proposition R, which, when combined with P, implies Q. Ernie's natural reaction is to deny R, and perhaps nothing will budge him from this denial. If Ernie could be gotten to accept R, he would face a choice: He could grudgingly admit Q or he could retreat from P. (We'll assume that denying "If P&R then Q" is unacceptable.) It might be protested that if Ernie's disposition is the latter—to reject Q by rejecting P—then he is not *really* committed to Q at all. My attention, however, is

not directed at Ernie's possible reactions, but rather at Bert when he advocates R and observes its implications. Bert could, of course, simply argue for R; perhaps this is the real nub of their disagreement. But supposing that Bert is aware that Ernie's whole motivation for embracing P was to avoid Q—indeed, Bert has suffered Ernie's objections to Q which take the form "Not Q, because P instead!"—then we could hardly fault Bert for taking glee in defending himself by pointing out that by embracing P Ernie has "inadvertently" committed himself to Q. (That there may be an even stronger kind of commitment Ernie might have to Q—that he is disposed, upon accepting "If P&R then Q," to grudgingly consent to Q—is another matter.)

Similarly, the divine command theorist will presumably believe that the predicate "...is willed by God" has a non-empty extension, and if we believe otherwise then this could be the topic of debate. But a more playfully provocative way of voicing our objection would be to accuse the divine command theorist—a would-be success theorist—of being an inadvertent advocate of moral skepticism. (That this is provocative is revealed by one's temptation to add an exclamation mark; that it is playful is revealed by the fact that one is disposed to deliver the objection with a grin.) My contention is that this happens in metaethics more often than is generally acknowledged. The number of unwitting moral error theorists is probably larger than the number of witting moral error theorists.

Response dependent morality

The advocate of a response dependency (RD) account of morality is a conspicuous example of the kind of opponent of the error theorist that has been under discussion. The RD theorist likely agrees with Mackie that *objective* normativity would be unacceptably odd—if, at least, by "objective" we mean something like (when applied to properties): *possessed irrespective of anybody's attitudes or psychological responses*. But the RD theorist is unimpressed with Mackie's attempts to convince us that any such robust objectivism "has been incorporated in the basic, conventional, meanings of moral terms" (Mackie 1977: 35). The RD theorist argues that there is nothing unacceptable in the idea of a *non-objective* morality—a morality that is, in some manner (which Mackie would deny), constituted by our psychological responses to the world; a morality that we *make*.

When it comes to the RD theorist's positive proposals, there is a great deal of variation, and my intention is not to criticize the program *tout court* but rather to point out some pitfalls along well-traveled paths. My target will be restricted to versions that make central the idea of a *disposition*. There are numerous versions of metaethical dispositionalism, many of which do not fall foul of the problems I will raise. The ones that do fall foul I will divide into two: negligent dispositionalism and optimistic dispositionalism. These terms will be explained shortly.

According to the dispositionalist RD theorist, moral properties are to be identified with dispositional properties, where the dispositions in question concern the generation of some kind of psychological response. The standard equation is this:

Moral goodness = the disposition to produce R in S in C

where "R" denotes a psychological response, "S" a type of subject, and "C" a set of circumstances (and where each of the three variables can be specified independently of the others). (See Johnston 1989, 1993; Lewis [1989] 2000; Casati & Tappolet 1998.) The substantive variation among different versions of dispositionalism arises from the different ways in which these three variables might be filled in (and the logical relation between the two sides).

Although I am classifying such theories as conceding a retreat from moral *objectivity*, there nevertheless remains a sense in which such properties retain a kind of objectivity. The disposition to produce R in S in C might be instantiated in an object even if there are no minds in existence (no S's having R), and thus goodness, on this model, would be existentially mind-independent. (See Pettit 1991.) However, the model renders goodness non-objective in at least some other sense: The concept of the disposition to produce R in S in C cannot be articulated without making reference to a mental event (R), and thus goodness remains conceptually mind-dependent. To the extent that "response independent" is legitimately used as a synonym for one kind of "objectivity" (a kind that might be associated with primary as opposed to secondary qualities), it can hardly be denied that RD theorists embrace some kind of non-objective morality. We can make the simplifying assumption that whatever kind of moral objectivity RD theorists reject is precisely the kind that Mackie thinks is an essential but problematic aspect of moral discourse. (Recall that Mackie likens moral phenomenology to the perception of primary qualities: 1980: 34.) These theorists concur with Mackie that *that* kind of moral objectivity is unavailable.[2]

How could a proponent of dispositionalism be an accidental error theorist? Simply this: If the descriptive phrase on the right-hand side of the equation fails to denote a property, or denotes a property that is uninstantiated in the actual world.[3] We must, of course, be careful to distinguish an uninstantiated dispositional property from a non-manifest dispositional property. Consider the disposition to squeal if kicked unexpectedly. This dispositional property might be instantiated by something—a small dog, say—even if the creature is never kicked and never squeals; even if, that is, the disposition never becomes manifest. I am not looking for non-manifest dispositional properties; I am looking for descriptive phrases that purport to denote actually instantiated dispositional properties (whether manifest or not) but fail to do so.

---

[2] The possibility of different kinds of objectivity reveals a shortcoming in my earlier classification of cognitivist opponents of the moral error theory into *head-on* versus *concessive* strategists. Suppose that Mackie successfully refutes a certain kind of objectivity for morality—we'll call it "type-A objectivity." Suppose that there is at least one other kind of objectivity possible: type-B. Someone might start out pursuing a concessive strategy: agreeing with Mackie that there are no such things as objective [type-A] prescriptions, while nevertheless denying that type-A objectivity is an essential feature of morality. But now suppose that this same person goes on defend the existence of objective [type-B] prescriptions. We could interpret this now as an instance of a head-on strategy: maintaining that there *are* such things as objective prescriptions, while insisting that Mackie has misconstrued their nature.

[3] Sometimes Mackie is interpreted as claiming that there is something *incoherent* about moral predicates, such that the error theory holds *necessarily*. I do not think this is a correct reading of his position (see Joyce & Kirchin 2010: xvi), but in any case, irrespective of Mackie's views on the matter, the most natural characterization of the moral error theory will allow that holding moral properties merely to be *actually* uninstantiated suffices to satisfy the criteria.

One way that this might occur is when the descriptive phrase is incomplete. Consider the phrase "the disposition to squeal." Nothing has this disposition—nothing *can* have this disposition—for no disposition is picked out; the description is only partial. One no more succeeds in picking out a property with the phrase "the disposition to squeal" than one would succeed in denoting an object using the partial definite description "the book that is between the."

For completeness, a dispositional description needs to specify a stimulus event (e.g., being kicked unexpectedly), a manifestation event (e.g., squealing), and conditions of stimulus. Often the last item can be specified tacitly. We could point at a particular small dog and ask "Would this dog squeal if I kicked it?" No circumstances are mentioned, but they are nevertheless implied: The question might assume that we are referring to the circumstances that the dog is actually in as we point at him, or assume, albeit vaguely, that we are referring to the "typical" circumstances in which one might encounter this small dog (thus excluding circumstances where there is no oxygen present, where the dog is exhausted from already being kicked, where the dog is wearing a little suit of armor, etc.). The latter might trump the former. We might point to a particular dog that is wearing a little suit of armor, acknowledge that if we were to kick it here and now it would not squeal, but nevertheless maintain that it has the disposition to squeal if kicked, in as much as it would squeal if kicked in *ordinary* (sans armor) circumstances.[4]

Now let us turn to moral dispositionalists who neglect to specify some elements of the disposition to such an extent that their dispositional phrase in fact fails to denote any property. Suppose moral goodness were identified with the disposition to produce approval in observers. I will assume that "approval" is an attitude that can be adequately specified, so will assume that the variable R is unproblematic. When the type of subject (S) is simply "observers," then the immediate question prompted is "*Which* observers?" Literally anything could produce approval in *some* observers (after all, we haven't yet restricted ourselves even to *human* observers), and I take it that any analysis of moral goodness that renders literally everything morally good can be rejected. One solution is to restrict the type of observer that is relevant; another solution is to modify the account in a relativistic direction (where "*o*" ranges over observers):

($\forall o$) Moral goodness (for *o*) = the disposition to produce approval in *o* in C.

The latter solution might have the consequence that everything is morally good *relative to someone*, which is not quite so unsightly as the result that everything is morally good *period*.

Now turn attention to variable C. Its importance (noted earlier) is brought out by imagining how things would stand were it absent. Suppose we restrict our attention to a

[4] It should also be noted that there can be a certain arbitrariness as to whether aspects of the disposition are specified as elements of R or S or C. Consider, e.g., the trait of full information. We might speak most naturally of the disposition to produce R in *fully informed Ss* in C. But we could instead pack the feature in question into the circumstances: speaking of the disposition to produce R in S in *circumstances that provide full information*. We even might speak of the disposition to produce *fully-informed R responses* in Ss in C. On many occasions such differences in how the disposition is described are of no ontological significance.

particular observer, Mary. Does anything at all have the disposition of producing approval in Mary *period*? I should say not. Certainly some things have produced approval in Mary in the past, and certain things can be reliably expected to produce approval in her again. If we consider Mary encountering or reflecting upon certain things (say, acts of generosity) then we might be justified in supposing that she will feel approval. However, in making such observations we will inevitably be including an understanding of Mary's *circumstances*—if only a tacit and vague presupposition of "ordinary circumstances." The fact that we can easily imagine circumstances in which Mary might encounter an act of generosity without feeling approval—because she's being chased by a tiger, for example—demonstrates that in order to specify the dispositional property some restriction must be placed on the circumstances of stimulus, on pain of no dispositional property being picked out.

Prinz's relativistic sensibility theory

In his recent defense of relativistic sensibility theory, Jesse Prinz identifies moral properties as "powers to cause emotions in us" (2007: 89). The "us," it turns out, is left open: After toying with "normal observers" and observers who have "knowledge of relevant facts, and are not under emotional or cognitive influences that are not relevant to the case at hand" (ibid: 91), Prinz opts to drop all restrictions. He is unfazed by the fact that X may prompt one emotion in one observer and another emotion in another observer; Prinz simply embraces the relativistic view that X may be morally good relative to some observers and bad relative to others.

What of *circumstances* in which a moral property causes an emotion in an observer? Given the lack of restriction on "observer," we cannot make use of the aforementioned solution of considering observers in "ordinary" circumstances, for what are the "ordinary circumstances" of *observers*? One might respond that "ordinary circumstances" can themselves to be relativized to types of observer: When the observer in question is a Martian we mean those circumstances that are ordinary to Martians; when the observer is a Cro-Magnon we mean what is ordinary for them; and when the observer is Mary we mean yet other set of ordinary circumstances. Not least among the glaring problems with this response is that for any given observer there is no fact about the level of generality at which these categories should be drawn. When the observer in question is Mary being chased by a tiger, say, do we look to the "ordinary circumstances" of *a human observer*, or those of *a human observer being chased by a tiger*, or those of *a running human*, or those of *Mary when frightened*, or what?

Seemingly the only restriction Prinz places on circumstances is that the observer must be "in good epistemic conditions" (ibid: 102). But if our worry is that in failing to specify circumstances Prinz has provided a description of the disposition that is incomplete to such an extent that it fails to denote any property at all, then this slight narrowing of the space of possibilities is unassuaging. X may cause Fred in good epistemic conditions *on Monday* to feel approval, while X may cause Fred in equally good epistemic conditions *on Tuesday* to feel disapproval. Thus the question "Does X cause Fred, when in good epistemic conditions, to feel approval?" has no answer; one must appeal to the inquirer for a more precise question.

Prinz is content to leave circumstances unspecified because (it turns out as his theory

develops) the relevant psychological response is not an *emotion*, but a *sentiment*—where "sentiment" is a term of art denoting a dispositional state: the disposition to have an emotion. On a given occasion Mary might lack the emotion of anger while still having the sentiment of anger. If the relevant sentiment is properly defined, including circumstances of stimulus, then Prinz might not need to specify circumstances in the broader description of the disposition. In other words, the seemingly incomplete description "the disposition to produce response R in subjects of type S" might pass muster if it turns outs that "R" surreptitiously specifies circumstances—e.g., that "R" is defined as something of the format "the disposition to have emotion E *in circumstances C*."

Let me address some potential puzzlement about this before making my principal criticism. The puzzlement arises because we now seemingly have two dispositions in play: The moral property is a disposition (a "power") and the relevant observer's response is a disposition (a "sentiment"). Whenever we have a response dependency theory, we always have options about whether to discuss dispositions in the world or in the individual. A dispositional view of color, for example, might claim that redness is the disposition to produce red-sensations in ordinary viewers under optimal viewing conditions. Alternatively, one might say that, for any *x*, *x* is red if and only if ordinary viewers have the disposition to experience red-sensations when observing *x* in optimal viewing conditions. Nobody need deny that both dispositions simultaneously exist. John Heil usefully observes that "manifestation of a disposition is a manifestation of reciprocal dispositional partner. ... A salt crystal manifests its disposition to dissolve in water by dissolving in water. But this manifestation is a manifestation of both the salt crystal's disposition to dissolve in water *and* the water's reciprocal disposition to dissolve salt" (Heil 2005: 350). In the redness case, all parties can agree to there being dispositions both in the world and in the subject; the pertinent dispute is over whether to identify the former disposition as the property of redness. In the moral case, despite Prinz's tendency to focus on the internal disposition (the sentiment), it is clear that he also wants to identify moral properties with dispositions (though he seems to like the old word "powers") (see Prinz 2007: 89, 92, 107).

Let us return to the matter of specifying circumstances of stimulus, the need for which Prinz thinks he can bypass by making the psychological response itself a disposition (ibid: 91-2). But this strategy successfully avoids the charge of incompleteness only if the description of the sentimental disposition is itself complete, and unfortunately the problem just reiterates here, for in his discussion of sentiments Prinz says hardly anything about the circumstances relevant to sentiment dispositions. He mentions that fear of flying is something that manifests itself only when on a plane (ibid: 85), but when he comes to the moral sentiments the need to specify circumstances (if only roughly) seems to have been overlooked. He characterizes resentment, for example, simply as the disposition to feel "bitterness, anger, or contempt" (ibid: 86). But does *anyone* have the disposition to feel the occurrent emotion of bitterness *period*? Do you? The natural question is "At what?" But even if "At what?" could be answered—suppose it's specified that we're asking whether you have the disposition to feel bitterness towards ex-lovers—the next question is "In what circumstances?" You might feel occurrent bitterness towards ex-lovers in certain circumstances but not in other circumstances. (It would be a sad fate indeed if you felt bitterness towards ex-lovers under *any*

circumstances.)[5]

For reasons that should now be familiar, Prinz's description of the sentiment of resentment, as it stands, fails to denote any property at all in any possible world, and thus, if we take the definition at face value, he has offered an error theory of resentment. And since he has tied moral properties to these sentiments, then if we take *that* definition at face value, he has also offered a moral error theory. My objection here is not simply that Prinz has left his description of the disposition somewhat vague and open-ended. If that were the problem, then virtually every reference to a disposition ever made would be at fault. The problem is, rather, that the description is incomplete in a striking manner that leaves me (at least) with *no idea* how it should be finished, and thus I do not feel inclined to grant the benefit of the doubt that Prinz's description picks out (even vaguely and open-endedly) any property whatsoever.

I confess that I don't really expect this accusation to stick; charging Prinz with unwittingly offering an error theory is really just a cheeky way of pointing out some significant gaps in his metaethical account. Nevertheless, it should lead us to wonder whether accidental error theorists might appear elsewhere on the metaethical landscape.

Firth's ideal observer theory

When a theory of moral dispositionalism offers a description of the relevant disposition that is incomplete to such an extent that we can know without further investigation that there is no such property, I will call this *negligent* dispositionalism. By contrast, *optimistic* dispositionalism is when the description of the disposition leaves it open whether it denotes any actually instantiated property; the advocate of the theory assumes or hopes that it does, but there are serious grounds for doubt. I classify Roderick Firth's *ideal observer theory* as an example of optimistic dispositionalism.

Firth identifies moral goodness with the disposition to prompt approval in the ideal observer, who in turn is defined as omniscient, omnipercipient, disinterested, dispassionate, consistent, and in other respects normal (Firth 1952, 1955).[6] The question with which we are concerned is whether there is *anything* that has this disposition.

The term "the ideal observer" is intended neither to refer to an *actual* individual nor a possible *individual*. In so far as the characteristics provided are sufficient to locate anyone in modal space, they will presumably locate a number of individuals. Thus the phrase "the ideal observer" is less like "the president of the USA" and more like "the blue whale." When we

---

[5] Prinz also talks of a person's sentiment remaining steady despite the tendency for it to manifest in an occurrent emotion diminishing (2007: 97-98). Frequent exposure to homeless people, for example, may reduce the frequency or intensity of our sympathetic emotions, while our disposition to feel such emotions towards the homeless remains intact. But this in itself reveals a problematic disregard for the role that conditions of stimulus play in defining the disposition in question. If at time *t* passing a homeless person produces strong sympathy in Mary, whereas at *t+1* the same stimulus condition does not produce that emotion (but rather a somewhat more extreme exposure is needed in order to prompt Mary's sympathetic emotions), then the dispositional property Mary instantiates at *t* is not the same dispositional property as she instantiates at *t+1*, and thus the sentiment has not remained steady.

[6] Firth doesn't actually set out to define *moral goodness* in particular, but rather refers generally to "any moral predicate." He also postpones specifying the relevant kind of reaction, preferring to speak of the ideal observer's "ethically-relevant reaction." I use the term "approval" for brevity.

say “The blue whale lives in the Southern Ocean” we are not referring to an individual whale, but to a kind. This introduces at least a touch of oddity to Firth’s theory, for we are supposed to take a token action and consider the response of a *kind* of individual to that action. By analogy, suppose we pointed to a particular school of krill and wondered whether *the blue whale* would have reaction R to that token. What would that mean? Any blue whale? Some blue whales? Most blue whales? A typical blue whale?

I don’t think that there is a settled answer to these kinds of question; it varies with conversational context. When we say “The blue whale is the largest animal ever to live” we don’t mean *any* blue whale. The existence of a stunted blue whale considerably smaller than, say, an average sperm whale would not prompt us to retract the statement. Nor must we mean *some* blue whales. If our stunted blue whale individual never surpassed 40 feet long, we would not on that account claim “The blue whale does not surpass 40 feet in length.” Nor does it seem correct that we must always mean the *typical* blue whale. If marine biologists were to observe an exceptional whale stay submerged for over an hour, then, even if they were aware that they had witnessed a unique record-breaking event—something that no other blue whale could accomplish—they would not hesitate to claim subsequently “The blue whale can stay submerged for over an hour.”

In the case of the ideal observer, the problem posed by analogous questions (i.e., some? all? most?) would recede considerably if there is a convergence in the ideal observers’ relevant responses. But this is exactly the point at which I would like to place pressure on Firth’s theory.

We may have some justified beliefs about the effect upon our attitudes of having less information versus having more information, of being calm versus being emotionally aroused, of being selfish versus being generous, and so forth, but we really have *no idea* what a creature would be like with the ideal observer’s extreme characteristics.[7] For all we know, complete disinterestedness might lead to the coldest kind of consequentialist calculations, whereby appalling sacrifices will be countenanced for the greater good. Maybe a spot of genocide really would work out for the best *eventually*, and perhaps it is precisely the observer’s “idealized” psychology that liberates him or her from those emotions that usually cause us to turn away from that possibility appalled. Or perhaps the ideal observer would be indifferent to the “greater good”; perhaps he or she would be confused by the very idea.

We should also be wary of a lurking fallacy of assuming that because all instances of moral disagreement that we have ever encountered have been due to a deficiency of X (e.g., true information) among interlocutors, providing X “to an extreme degree” (Firth 1952: 321) will lead to convergence. This is like saying that because a death was caused by a lack of oxygen in the room, death would have been avoided had the room been filled with 100% oxygen.

The point in which I am interested is not so much that all ideal observers might turn out to be monsters by our standards, but that the characteristics provided by Firth are insufficient to determine any particular pattern of attitudinal responses. Just as his list is too general to pick out an individual in modal space but rather picks out a kind, so too it may be too general to

---

[7] Firth’s addendum that the ideal observer is “otherwise normal” seems of little use here. It brings to mind someone describing a divine being as all-seeing, all-knowing, all-powerful, infinitely loving, eternally existing, the creator of the universe ... but *otherwise just like you and me*.

pick out a kind with a determinate pattern of (dis)approval but rather picks out a kind with attitudinal variation in this respect. There may be no fact of the matter about what an ideal observer would approve of, any more than there is a fact of the matter about whether the ideal observer prefers vanilla to chocolate ice cream. Consider an actual token action ϕ. If we examine the closest possible worlds at which there are ideal observers, then perhaps some of them disapprove of ϕ, some are neutral, and some even approve (even when in the same circumstances). If this is so, then does ϕ instantiate the dispositional property of prompting approval in the ideal observer? If we mean *all* ideal observers, then the answer is "no."[8] It may be that *some* actions are like ϕ in this respect or it may be that *all* actions are. Despite its strength, I do not find the last possibility absurd—it does not seem implausible that there exists *nothing* about which equally ideal Firthian observers will agree—in which case *nothing* will be morally good or bad.

Firth in fact admits to one of the premises of this argument. When pressed upon the question of convergence by Richard Brandt (Brandt 1955: 408-9), Firth admits that if there could be two ideal observers with different or opposed reactions to an act, "it would follow … that the act in question would be neither right nor wrong" (Firth 1955: 415). Firth rejects the antecedent, however, by claiming that divergent attitudinal responses imply differences in the traits used to identify the ideal observers.

But Firth is far from convincing on this point. The traits that he provides for the ideal observers clearly aren't sufficient to ensure a convergence on favorite ice cream flavor; perhaps they aren't sufficient to ensure a convergence on attitudes of approval and disapproval either. The question of what shared psychological traits are sufficient to ensure a convergence in (dis)approval is to a large extent it is an empirical matter, many of the details of which remain unknown. There is a growing body of literature revealing that the things that can influence an individual's morally relevant attitudes can be quite surprising. We might not have supposed, for example, that a person's tendency to act dishonestly can be enhanced by her wearing sunglasses or being placed in a dimly lit room (Zhong et al. 2010). Nor might we have guessed the effect of hand-washing on a person's moral evaluations (Schnall, Benton, et al. 2008). We might not have appreciated how easy it is to manipulate someone's moral opinions by placing him in a messy environment—e.g., in the presence of a dirty tissue (Schnall, Haidt, et al. 2008).

Firth thinks that the characteristics he uses to pick out the ideal observer in modal space are sufficient to (A) ensure convergence and (B) get intuitively correct results (e.g., the ideal observers do not turn out to all be Nazi sympathizers); he argues that all instances of moral disagreement which he has observed or which he can imagine are the result of differences in belief, or selfish interests, or self-referential emotions. Let us consider the plausibility of this in light of one of the empirical cases just cited: Suppose that there are two people making

---

[8] For further discussion of this point, see Carson 1984, 1989, and objections by Taliaferro 1988. If, alternatively, we interpret the question as asking whether the action would prompt approval in *some* ideal observers, then the answer is presumably "yes." The problem with this, however, is that it is not unreasonable to suspect that for just about anything there is *some* ideal observer that approves of it. This would not make the proponent of ideal observer theory an unwitting error theorist, but would nevertheless be a kind of reductio: of implying that just about everything is morally good. Relativizing one's ideal observer theory (see Carson 1984, 1989) is one obvious response, though such a reaction brings its own set of problems.

moral judgments about ϕ, and one of them is in the presence of a disgust-prompting dirty tissue which influences him to judge ϕ negatively, while the other is not. Certainly the two have different beliefs (one thinks "There's a dirty tissue" while the other does not think this), but this fact doesn't mean that they cannot both be equally ideal by Firth's lights. As far as beliefs go, Firth says that the ideal observer must be "omniscient with respect to the non-ethical facts"; but this is not to say that all ideal observers must have *the same* beliefs. We should presume that both ideal observes will be well-informed about experimental psychology: They will know about all of the aforementioned studies, including the "Schnall, Haidt, et al. 2008" paper which demonstrates the influence a dirty tissue may have on a subject's moral attitudes. Moreover, we can assume that the disgusted person realizes that he is being manipulated in just this way. But does this knowledge make his disgust (and connected moral assessment) dissipate? Perhaps; perhaps not. Attitudes prompted manipulatively in the setting of a psychology lab often survive the debriefing session (Ross et al. 1975; Nisbett & Ross 1985). Even medical placebos sometimes work in conditions of full information (Lee & Covi 1965; Aulas & Rosner 2003). It is not my intention to persuade anyone that the disgust *would* remain in situations of full-information, only that it is an empirical matter against which Firth would be imprudent to bet the farm.

If neither of our two imaginary persons must be suffering from doxastic failing, Firth may instead claim that the subject whose moral assessment is influenced by the dirty tissue has become less dispassionate or less disinterested. This can seem plausible if one equates dispassionateness with lack of emotion, for it seems highly likely that the dirty tissue influences moral assessment only via arousing the emotion of disgust. However, such an equation would be a mistake. Firth's notion of dispassionateness pertains to the absence of "*particular* emotions," which are defined as emotions that are "directed toward an object only because the object is thought to have one or more essentially particular properties" (1952: 340)—where "particular properties" are those "which cannot be defined without the use of proper names" (ibid: 338).[9] I see no ground for assuming that the disgust one might feel at the presence of a dirty tissue must take the form of a particular emotion; one's emotion might be directed at dirty tissues—or, more likely, the associated bodily fluids—in general.

The more general point to which I should like to draw attention is the fact that at present nobody knows too much about the psychological mechanisms through which these kinds of subtle influences on morally relevant attitudes work, and it would be a hasty empirical bet to assume that differences in attitude must always entail a difference in the traits of the Firthian ideal observer. So on the question of whether he is committing himself to an error theory, Firth becomes a hostage to empirical fortune.[10]

---

[9] For convenience, Firth includes pronouns such as "I," "here," and "this" as proper names.

[10] Alternatively, Firth could try to rule out these kinds of influence on ideal observers' attitudes by specifying the hypothetical *circumstances* of the ideal observers' judgment in a way that excludes such possibilities. In other words, he could say that X is morally right iff the ideal observer would feel approval towards X in circumstances where there are no dirty tissues nearby, where he is in well-lit conditions, where he has not been prevented from washing his hands, and so on. It is, however, difficult to see how that "…and so on" is going to be cashed out. Moreover, it is worth noting that one of the curious things about Firth's dispositionalist account is that the *conditions* of stimulus are never mentioned. It is as if he thinks that they just don't matter at all. But of course

Scanlon's hypothetical contractualism

Contemporary versions of hypothetical contractualism face analogous challenges. Here, it is not the attitude of a (hypothetical) kind of individual that counts, but rather the collective response of a (hypothetical) group of persons. Many objections to such views have been voiced in the literature, but it is seldom appreciated that the dispositional description proposed by contractualist may simply fail to denote any property at all.

Thomas Scanlon, for example, writes that "an act is wrong if its performance under the circumstances would be disallowed by any set of principles for the general regulation of behaviour that no one could reasonably reject as a basis for informed, unforced general agreement" (Scanlon 1998: 153). Unlike Firth, Scanlon at least wears his quantifiers on his sleeve: Instead of speaking somewhat mysteriously of "*the* set of principles that would be adopted by *the* group," he makes clear that he means "*any* set of principles that could not be reasonably rejected by *any* group." But even so, is there *any* action that would be disallowed by any such set of principles? The question, again, is one of convergence.

Suppose, for the sake of introducing the argument, we were just talking about sets of principles that would be accepted or rejected by possible *groups of humans*, with no further qualification imposed on what kind of humans; we are, in other words, including Vlad the Impaler and his henchmen, kamikaze pilots, drunken vikings, suicidal nihilists, the woefully stupid, the willfully annoying,[11] and so forth. Presumably, the "sets of principles for the general regulation of behavior" that these human groups might endorse will vary wildly and may not bear much resemblance to those sets that will tempt civilized folk. Is there *any* action that would be disallowed by *any* of these sets of principles? I see no grounds that should incline one to answer in the positive. But if this is correct, then, if by "…is morally wrong" we mean "…is such that its performance under the circumstances would be disallowed by any set of principles for the general regulation of behaviour that no group of humans would reject," there exists nothing that is morally wrong.

Clearly, then, all that stands between Scanlon and the error theory is the all-important restriction he places upon the type of sets of principles admitted: those that no group "could reasonably reject as a basis for informed, unforced general agreement." And our attention should immediately be drawn to the word "reasonably." The natural supposition is that it is this qualification that acts as the principal bulwark against a slide to moral skepticism, but along with this supposition comes the suspicion that Scanlon cannot simply help himself to the notion in advance of having refuted that very skepticism. If what is "reasonable" in this context implies a substantive moral framework, then Scanlon clearly begs the question.[12] Perhaps to our taxonomy of negligent dispositionalists and optimistic dispositionalists, we

---

they do; and one thing I've already argued is that any dispositionalist who neglects to specify conditions of stimulus (at least a ballpark estimate) is on the fast track to an error theory.

[11] Among the ranks of the willfully annoying I include those imaginary humans who choose principles of behavior on the sole basis of refuting popular metaethical theories.

[12] "It would clearly render [Scanlon's] position uselessly circular if the fact that a putative principle permitted agents to act wrongly were to be adduced as a reasonable ground for rejecting it; for the procedure is supposed to help us identify what courses of action are wrong" (Baldwin 2002: 99).

should add question-begging dispositionalists. But it isn't begging the question of which I wish principally to accuse Scanlon, but rather an ungrounded optimism. Or, perhaps speaking more carefully, I suspect him of flitting between begging the question and ungrounded optimism without finding a stable point between.

Scanlon considers the following imaginary case:

> Suppose that Jones has suffered an accident in the transmitter room of a television station. Electrical equipment has fallen on his arm, and we cannot rescue him without turning off the transmitter for 15 minutes. A World Cup match is in progress, watched by many people, and it will not be over for an hour. Jones's injury will not get any worse if we wait, but his hand has been mashed and he is receiving extremely painful electrical shocks. Should we rescue him now or wait until the match is over? Does the right thing to do depend on how many people are watching – whether it is one million or five million or a hundred million? (Scanlon 1998: 235)

Scanlon goes on: "It seems to me that we should not wait, no matter how many viewers there are, and I believe that contractualism can account for this judgment" (ibid.). He backs up this opinion by appeal to the *unfairness* of imposing such a sacrifice upon Jones in order to avoid the proportionally lesser inconvenience to each individual viewer. However, in order to establish that leaving Jones to suffer is morally wrong (by Scanlon's own lights), it is insufficient to show that endorsing a set of principles that disallows the imposition of this suffering is a reasonable choice. (I myself feel confident that it is reasonable, as does Scanlon, as, presumably, do most readers.) Rather, it must be shown that any group of persons (aiming at informed, unforced general agreement) that rejected any such set of principles *would be unreasonable*. This latter conclusion evidently doesn't follow from the former; from the fact that someone could reasonably accept X it hardly follows that anyone who rejects X is unreasonable. But Scanlon has nothing else to say to convince us of the crucial proposition, and it does not seem difficult to generate doubt about it.

Empirical evidence reveals a great deal of cross-cultural variation in norms of *fairness* (see Henrich et al. 2004), and it is entirely possible that we could locate actual normative frameworks that will reject principles prohibiting us from waiting till the match is over before rescuing Jones. I will, however, stick with an imaginary case in order to make the point crisply. Suppose a group of persons (let's call them "Stoic sages") believe in some ubiquitous providential divine plan—in such a way that they consider everyday "accidents" to serve some great (though mysterious) purpose, interference with which is to be avoided. Holding this belief, centrally and sincerely, does not obviously exclude the possibilities that these Stoics may seek practical principles and may value informed general agreement. (After all, the real Stoics did maintain distinct political views.) When faced with the unfortunate trapped Jones, the Stoics feel disinclined to step in to upset the unfolding of the divine plan; in fact, they steadfastly reject any set of principles that demands or even permits interference. Is the Stoics' rejection of any such set of principles *unreasonable*?

In a sense, of course it is. The problem, however, is that this sense is one that is ineliminably informed by our own substantive moral beliefs and principles, and thus is not one to which Scanlon can appeal, on pain of endorsing a question-begging dispositionalism. But

the other horn of the dilemma is no less damaging: If Scanlon employs a thinner and less morally-loaded notion of *reasonableness*, then there seems no grounds to exclude the Stoics' choice as unreasonable. My firm suspicion is that we can pull this move over and over again, for *any* action one cares to mention: We can always locate a hypothetical group of persons with sufficiently wacky beliefs about how the universe works, or harboring sufficiently aberrant desires, or committed to sufficiently bizarre values, or inhabiting sufficiently atypical circumstances, that they will be willing to reject any set of principles we care to imagine, *without our being able to make the charge of unreasonableness stick*. If this is so, then there exists *no* set of principles for the general regulation of behavior that *no one* could reasonably reject as a basis for informed, unforced general agreement.

It is possible that Scanlon, and hypothetical contractualists in general, might yet find some kind of plausible rejoinder to this line of objection. My point is that they are yet to do so, and so at this stage we must consider the conviction that the requisite notion of *reasonableness* is forthcoming to be another instance of *optimism* being the only thing standing between endorsement of the theory and the moral error theory.[13]

Virtue ethics

The virtue ethicist is another prominent example of the kind of concessive opponent of the moral error theorist that has been under discussion. The virtue ethicist may share many or all of Mackie's misgivings about objective values and objective prescriptions. "Certainly," she may concede, "if morality were like *that* then we should all be error theorists; but morality is not like that, or, at least, need not be like that." And the virtue ethicist will then point to the Ancient Greeks—and Aristotle, in particular—as providing an exemplar of an ethical system that got along very nicely without all those problematic objective values and prescriptions ruining everything. (The *locus classicus* of this view is Anscombe 1958.) Unlike deontological ethical theories that begin with the action-oriented question "How ought one to act?", or teleological ethical theories that begin with the value-oriented question "What is of intrinsic value?", virtue ethics begins with agent-oriented questions like "What kind of person should one be?" or "What is the good life for a human being?"[14] It is thought that this fundamental difference in starting point promises to immunize virtue ethics from the kinds of error theoretic worries that dog deontological and teleological theories. But is this true? Does

---

[13] I am also inclined to accuse Michael Smith (1994) of supporting an optimistic dispositionalism. Smith argues that S has a normative reason to ϕ iff a fully rational counterpart of S would advise S to ϕ. He then argues that ϕing is morally right only if *everyone's* fully rational counterpart would come down on the side of ϕing. The latter is a conceptual claim, leaving open the substantive question of whether everyone's normative reasons do in fact converge in the necessary manner. Smith is here consigned to a footnote because I have criticized him on this point before (Joyce 2001: 88-95; see also Sobel 1999), and in any case what I have said against Firth and Scanlon gives a pretty good hint of what I'll say again against Smith. What is distinctive about Smith in the present context is that, if faced with the failure of the convergence premise, he seems willing to embrace the error theory. (See Smith 1994: 187-189; 2002; 2006). Doubts about convergence also lead me to regard Frank Jackson (1998) as a potential accidental error theorist. See Robinson (2009) for criticism of Jackson's presuppositions about convergence.

[14] Deontological theories put duty first, and define value in relation to duty. Teleological theories (e.g., utilitarianism) put value first, and define duty in relation to value. See Broad 1930: 277 ff.

virtue ethics really represent a smooth escape route from the threat of a moral error theory?

It suffices for an answer in the negative if serious doubt arises as to whether there even *are* any of the entities that the virtue ethicist refers to as "virtues." If there simply aren't any such things, then all the virtue ethicist's distinctive assertions—such as "The virtue of honesty is an important part of human flourishing" or "Albert Schweitzer was more virtuous than Albert Speer"—and all deontological and teleological talk that virtue ethicists allow as derivative upon virtue talk—will fail to be true. In other words, if the virtue ethicist bases his or her theory on the claim

($\forall x$) *x* is a virtue iff *x* is a P,

but it turns out that nothing satisfies the predicate "...is a P," then the virtue ethicist is in fact proposing an error theory.

One possible way that this might happen is if the virtue ethicist's conception of a virtue ineliminably presupposes a badly flawed image of human psychology. Gilbert Harman and John Doris have both argued that the existence of the kind of entrenched personality traits upon which virtue ethics depends is cast into doubt by empirical evidence in support of "situationism" in social psychology (Harman 1999, 2000; Doris 2002). This is a controversial claim (see Merritt 2002; Sreenivasan 2002), and it is not a strategy that will be further explored here.

Rather, I have doubts about the virtue ethicist's starting presuppositions: the eudaimonia-oriented questions—*What kind of person should one be? What is the good life for a human being?*—which are supposed to provide such a trouble-free point of departure when compared to the rival deontologists' and teleologists' guiding inquiries. My understanding is that when we direct these questions at traits of character (asking, e.g., *What kind of character traits must one cultivate in order to be the kind of person one should be?*) then the virtues have the theoretical role of *answers* to these questions. But my contention is that it is entirely possible that these questions have no answers and thus there are no virtues (as conceived by the virtue ethicist).

This concern can be brought into focus by parodying the virtue ethicist's questions: *What kind of ice cream flavor must one prefer in order to be the kind of person one should be? What ice cream preference contributes to the good life for a human being?* Let us assume, not unreasonably, that it is acceptable to choose ice cream flavors on the basis of gustatory whim. It is *possible* that certain flavors are better for one's health, or better for the environment (if they use sustainable ingredients, say), or better for the wider community (if their production eschews exploitative practices, say)—but let's assume that all such potential complications come to naught and that one can select on the basis of taste alone. Then I would know what flavor *I* should prefer,[15] but there would be no flavor that "one" should prefer, and no flavor that one must prefer in order to be the kind of person that one should be. To equate the predicate "…is P" with "…is the ice cream flavor that one must prefer in order to be the kind of person that one should be" would be to endorse (perhaps unwittingly) an error theory about

[15] Ben & Jerry's Rainforest Crunch—now, alas, discontinued. (See Ben & Jerry's "Flavor Graveyard.")

P-discourse.

In order for the virtue ethicist's questions to fare better, the kind of life that "one should live" cannot be similarly a matter of whimsical choice and cannot change from individual to individual; it must be grounded in something shared by all humans. This is a problem, since there are many images of the good life: the life of the Buddhist monk, of the hedonistic consumer, of the intellectual, of the Stoic sage, of the noble savage, and so on. At this point, the Aristotelian virtue ethicist will often appeal to *human nature* in order to privilege one kind of "good life" that is shared by all. The virtue ethicist Rosalind Hursthouse writes: "A virtue is a character trait that human beings, given their physical and psychological nature, need to flourish (or to do and fare well)" (1995: 68). Clarity demands that we distinguish the kind of flourishing that (supposedly) derives from human nature from any alternative and competing visions of human flourishing that an individual or group might (or might not) embrace, and since the pertinent difference here appears to be that some visions of flourishing might be *chosen* while the one derived from human nature is bestowed upon us whether we like it or not, I shall refer (somewhat clunkily) to the nature-given account of the good life as "non-chosen human flourishing." This kind of human flourishing should be no more troubling (the virtue ethicist avows) to our naturalistically-inclined philosophical temperaments than the notions of *antelope flourishing* or *petunia flourishing*, which can be derived from accounts of antelope nature and petunia nature, respectively. "A correct conception of the virtues must be at least partly shaped by a correct conception of healthy growth and development which in part constitute our flourishing" (Swanton 2003: 60). The virtue ethicist will, moreover, stress the *social* nature of our species, in the expectation that the more prosocial virtues, like generosity and friendship, will be contributors to non-chosen human flourishing. "We are naturally sociable creatures who like to have friends and want to be loved by friends and family" (Hursthouse 1987: 226).

The core of my skepticism about this is that there remains abundant room for reasonable doubt that the facts of "human nature" are going to play out in the determinate way that the virtue ethicist assumes. One can allow (if only for the sake of argument) that it is legitimate to speak of "human nature" and hence "human flourishing," but nevertheless humans are the most psychologically plastic organisms we have ever encountered, and thus the "end" of human flourishing may provide only a minimal constraint on lifestyle decisions, and no constraint at all on character traits. Humans are without doubt obligatorily gregarious organisms, and so one might reasonably claim that living in some sort of community of fellows is an "end" that has been conferred upon humans by nature. But what degree of specificity of character traits is determined by this "end"? Hitler had loyal and sincere admirers; Genghis Khan was surrounded by good mates; perhaps even Jack the Ripper was a solid family man. The idea that the sociality inherent in human nature cannot be satisfied in a restricted domain while coupled with cold disregard and astounding cruelty towards anyone lying outside the favored sphere strikes me as a romantic misapprehension. To put the point provocatively: It is not foolish to declare that Hitler's character traits were just as true to his nature as a social organism as Mother Teresa's.[16] (Let us not forget that the Nazi war machine

[16] Though it *would* be foolish to think that my saying this indicates any glimmer of tolerance towards Hitler.

required an enormous amount of interpersonal cooperation, much of which was motivated by strong prosocial feelings.) If this is so, then there may be no specific set of character traits that is underwritten by our social nature.

One might be tempted to respond that there is surely *something* in common between Hitler and Mother Teresa with respect to their social skills: some very general and minimal interpersonal faculties operative in any human who manages to have any kind of successful relationship with his or her fellows. Maybe, then, these very minimal traits might count among the virtues? But this is hardly a line that the virtue ethicist will find attractive, for if even Hitler and Jack the Ripper turn out to have the social virtues then we've surely seriously lost track of the point of endorsing virtue ethics. Second, it is far from obvious that the kinds of minimal social skills manifest by anyone capable of maintaining any sort of meaningful interpersonal relationship are going to count as *character traits* in the requisite manner. The Aristotelian virtue ethicist will usually embrace a view of character traits according to which they are "relatively long-term stable dispositions to act in distinctive ways ... involving [inter alia] habits of desiring" (Harman 1999: 317). Even if humans are by nature social organisms, and thus need certain traits in order to flourish as social organisms, it doesn't follow that there is a set of *character traits* (in the sense just described) that humans need in order to flourish as social organisms—any more than there is an ice cream preference we need in order to flourish as social organisms. Thus, if by *virtue* the virtue ethicist specifically means "a character trait needed for non-chosen human flourishing," then we are once more looking at a potential error theory.

None of this is to deny that we are free to create and embrace more substantial visions of flourishing and the good life, for many of which the cultivation of specific character traits will certainly be necessary (either causally or constitutively). For example, one might maintain that the good life consists of living like a Buddhist monk—a kind of life for which (a) the claim that it is "natural" for humans is highly implausible, and (b) certain character traits, like irenic acceptance, are necessary. The problem, though, is that one individual's or community's robust vision of the good life will differ from another's, and the character traits needed to succeed at one life may diverge from those needed to succeed at another. It may well be true that for any person, *x*, there exists a good life, *y*, and exists a certain set of character traits, *z*, such that *z* is necessary for *y*. But we must be careful not to commit a quantifier-shift fallacy of flipping this round and thinking that there exists a set of character traits, *z*, and exists a good life, *y*, such that for any person, *x*, *y* is *x*'s good life and *z* is necessary for *y*.

We have seen that so long as *a virtue* is defined as "a character trait necessary for non-chosen human flourishing" there may be no such thing. I haven't tried to establish that there is no such thing, but merely to expose the presence of reasonable doubt. The virtue ethicist may attempt to alleviate this doubt by weakening the definition to "a character trait that *tends to* contribute to non-chosen human flourishing." ("[T]he claim is not that being virtuous guarantees that one will flourish. … Virtue is only a reliable bet; it will probably bring flourishing" (Hursthouse 1987: 230).) But this does not necessarily help. After all (sticking with a parody that no doubt grows repetitive but remains instructive): there is no ice cream flavor preference that even *tends* to contribute to non-chosen human flourishing. So why assume that there is any such set of character traits?

Rather than answer that question directly, let me try another tack. Suppose that there *are* some character traits that tend to contribute to non-chosen human flourishing. I noted earlier the multiplicity of alternative visions of the good life, and we can suppose that there are also sets of character traits that tend to contribute to these alternative images. One version of the good life encourages friendliness, say, while another version (seemingly as legitimate as the first) urges remaining aloof; one encourages turning the other cheek while another urges vengeance, and so forth. Will the virtue ethicist be satisfied merely to persuade us of the terminological stipulation that one among these competing sets—the one that is "nature-given"—can be given the label "the moral virtues," and leave it at that? I should think not. The moral virtues are expected to carry some practical weight, some normative force, some extra authority. Suppose this is considered to be not merely a contingent feature of virtue, but is taken to be an additional essential quality; a virtue is defined not merely as a character trait that probabilifies non-chosen human flourishing, but a trait whose cultivation carries more normative weight than any other character traits which probabilify any alternative chosen visions of flourishing.

However, I again find myself hesitant to believe that there exist *any* character traits that enjoy this attribute. Indeed, if any kind of substantive normativity is made a defining feature of the virtues then this may lead straight to an error theory, for it is hard to see how non-chosen human flourishing will supply those traits that conduce to its satisfaction with any normative relevance whatsoever.

Let us think of this in terms of proper functions. The proper function of a wire coat hanger is to hang up clothes, which gives license to a variety of normative-sounding language, such as "A coat hanger ought to support clothes," and "A good coat hanger supports clothes well." This may provide one with various reasons *if one wants to hang up clothes*, but if one has no such interest, but rather has an interest in, say, retrieving one's dropped keys from the drain, then using the coat hanger as a fishing hook (destroying it in the process) is entirely legitimate. This would not be an instance of there being two competing reason-conferring functions—one of which outweighs the other. Rather, given one's interests, the proper function of the coat hanger—what the hanger is "supposed to do"—carries no weight in one's deliberations whatsoever. There is not a slightest drop of true normativity (independent of the agent's antecedent interests) that can be squeezed from the proper function of the coat hanger.

Suppose that non-chosen human flourishing really does require the cultivation of certain prosocial character traits. But there will also be a range of alternative ends that a person might genuinely prefer (say, the life of selfish hedonism, or the life of an ascetic hermit, or the life of a reclusive intellectual) which require the cultivation of different sets of character traits. The question is not merely why the former end must trump any of the latter, but why the former end, in and of itself, constitutes any kind of practical consideration at all. If a hermit withdraws from society—letting his prosocial character traits wither, abandoning the end of nature-bestowed human flourishing in favor of an alternative vision of flourishing—must this constitute *a mistake*, any more than the person who uses a wire coat hanger to retrieve her keys makes a mistake?

Lest it be thought that the example of a hermit is too extreme and unusual to be bothersome, we should remind ourselves that the same point could be made of any lifestyle

one cares to think of. We can imagine someone who prefers the end of living by-and-large like an upstanding citizen but with occasional self-serving exceptions (even at serious cost to others) when the chances of incurring punishment are low. Just picture any behavior at all that seems intuitively morally wrong—whether mild or dramatic: We can imagine someone constructing and preferring a lifestyle that occasionally allows this behavior, and cultivating the set of character traits conducive to that end. The question to which I am drawing attention is: Why, in such circumstances, does the "natural" end of human flourishing, from which the preferred end deviates either a little or a lot, furnish any practical authority at all?

One might be inclined to answer that organisms are more likely to achieve well-being and fulfillment if they satisfy the ends laid down by nature (where this is not a trivial claim derived from using "well-being" and "fulfillment" as synonyms of "human flourishing"). But, first, this is an empirically dubious claim. Natural selection may well have forged us as creatures that *strive* to achieve a sense of well-being; but there would be little evolutionary mileage in creating creatures that actually *achieve* a sense of longstanding well-being upon attaining fitness-enhancing goals. The plausibility of the hypothesis that the true road to well-being and fulfillment is to live like a Buddhist monk is in no way undermined by the observation that such a life probably represents a dramatic departure from the kind of "human flourishing" laid down by nature. Second, even if it were true that non-chosen flourishing reliably leads to well-being and fulfillment, if this is the sole ground for recommending the character traits conducive to that end then it is not the flourishing *per se* that matters but the states of well-being, etc., that reliably accompany it. What, then, would there be to exclude the discovery of an alternative (possibly vicious) means of achieving that same well-being, perhaps more efficiently and abundantly?

I conclude, therefore, that if a virtue is defined as a character trait that is necessary for, or probabilifies, non-chosen human flourishing *and thus has normative weight*, then there are additional grounds for doubting that there exist any such things at all. The may be versions of virtue ethics that do not include such a claim—that are not even based on any claim like "The virtues are those character traits that tend to contribute to human flourishing"—and I freely admit that the objections I have raised here do not apply to any such versions.

Conclusion

My ambitions have been more modest than they might appear. I have discussed several well-known metaethical theories in a critical voice, underlining the places where proponents, if they are not careful, will commit themselves to a moral error theory. It bears repeating that my calling these philosophers "accidental error theorists" is not to be taken too literally; it is really just a slightly mischievous way of drawing attention to a pattern of defect in their theories. This is less a damning critique of these types of theories so much as a plea for greater specificity. I do not mean to suggest that adequate specificity may not be supplied which will allow these theories to avoid the pitfalls I have set before them. It should, however, be underlined that even if moral dispositionalists and virtue ethicists can avoid the charge that they are unwittingly error theorists, it doesn't follow that they thereby defeat the error theorist. Even if, for example, we are satisfied that in the equation "Moral goodness = property P" the

right-hand side succeeds in denoting an instantiated property, the error theorist can still object to the adequacy of the equation as a whole. Indeed, this latter kind of argument may well be the strongest strategy for the error theorist to pursue. My intention has been merely to show that some metaethical theories threaten to disintegrate before the need for debating the plausibility of the identity claim even arises.

REFERENCES:

Anscombe, G.E.M. 1958. “Modern moral philosophy.” *Philosophy* 33: 1-19.

Aulas, J., & Rosner, I. 2003. “Efficacy of a non blind placebo prescription.” *Encephale* 29: 68-71.

Baldwin, T. 2002. “The three phases of intuitionism.” In P. Stratton-Lake (ed.), *Ethical intuitionism: Re-evaluations*. Oxford: Oxford University Press: 92-112.

Brandt, R. 1955. “The definition of an ‘ideal observer’ theory in ethics.” *Philosophy and Phenomenological Research* 15: 407-413.

Broad, C.D. 1930. *Five types of ethical theory*. New York: Harcourt, Brace and Co.

Carson, T. 1984. *The status of morality*. Boston: D. Reidel.

Carson, T. 1989. “Could ideal observers disagree?: A reply to Taliaferro.” *Philosophy and Phenomenological Research* 50: 115-124.

Casati R. & Tappolet, C. (eds.). 1998. *European Review of Philosophy* 3: *Response-Dependence*.

Doris, J. 2002. *Lack of character: Personality and moral behavior*. Cambridge: Cambridge University Press.

Firth, R. 1952. “Ethical absolutism and the ideal observer.” *Philosophy and Phenomenological Research* 12: 317-345.

Firth, R. 1955. “Reply to Professor Brandt.” *Philosophy and Phenomenological Research* 15: 414-421.

Harman, G. 1999. “Moral philosophy meets social psychology: Virtue ethics and the fundamental attribution error.” *Proceedings of the Aristotelian Society* 99: 315-331.

Harman, G. 2000. “The nonexistence of character traits.” *Proceedings of the Aristotelian Society* 100: 223-226.

Heil, J. 2005. “Dispositions.” *Synthese* 144: 343–356

Henrich, J., Boyd, R., Bowles, S., Camerer, C., Fehr, E., & Gintis, H. 2004. *Foundations of human sociality*. New York: Oxford University Press.

Hursthouse, R. 1987. *Beginning lives*. Oxford: Wiley-Blackwell.

Hursthouse, R. 1995. “Applying virtue ethics.” In R. Hursthouse, G. Lawrence, & W. Quinn (eds.), *Virtues and reasons: Philippa Foot and moral theory*. Oxford: Clarendon Press: 57–75.

Jackson, F. 1998. *From Metaphysics to Ethics*. Oxford: Oxford University Press.

Johnston, M. 1989. “Dispositional theories of value.” *Proceedings of the Aristotelian Society*, suppl. vol. 62: 139-174.

Johnston, M. 1993. “Objectivity refigured: Pragmatism without verificationism.” In J.

Haldane & C. Wright (eds.), *Reality, representation, and projection*. Oxford: Oxford University Press: 85-130.

Joyce, R. 2001. *The myth of morality*. Cambridge: Cambridge University Press.

Joyce, R. 2007. "Morality, schmorality." In P. Bloomfield (ed.), *Morality and self-interest.* Oxford: Oxford University Press: 51-75.

Joyce, R. Forthcoming. "The error in 'The error in the error theory.'"

Joyce, R. & Kirchin, S. (eds.). 2010. *A world without values: Essays on John Mackie's moral error theory.* Dordrecht: Springer Press.

Lewis, D.K. [1989] 2000. "Dispositional theories of value." In his *Papers in ethics and social philosophy*. Cambridge: Cambridge University Press: 68-94.

Mackie, J.L. 1946. "A refutation of morals." *Australasian Journal of Psychology and Philosophy* 24: 77-90.

Mackie, J.L. 1977. *Ethics: Inventing right and wrong*. London: Penguin.

Mackie, J.L. 1980. *Hume's moral theory*. London: Routledge and Kegan Paul.

Merritt, M. 2002. "Virtue ethics and situationist personality psychology." *Ethical Theory and Moral Practice* 3: 365-383.

Nisbett, R. & Ross, L. 1980. *Human inference: Strategies and shortcomings of social judgement*. Englewood Cliffs, NJ: Prentice-Hall.

Park, L. & Covi, U. 1965. "Nonblind placebo trial." *Archives of General Psychiatry* 12: 336-345.

Pettit, P. 1991. "Realism and response-dependence." *Mind* 100: 587-626.

Prinz, J. 2007. *The emotional construction of morals*. Oxford University Press.

Robinson, D. 2009. "Moral functionalism, ethical quasi-relativism, and the Canberra Plan." In D. Braddon-Mitchell & R. Nola (eds.), *Conceptual analysis and philosophical naturalism*. Cambridge, MA.: MIT Press: 315-348.

Ross, L., Lepper, M., & Hubbard, M. 1975. "Perseverance in self-perception and social perception: Biased attributional processes in the debriefing paradigm." *Journal of Personality and Social Psychology* 32: 880–892.

Sayre-McCord, G. 1986. "The many moral realisms." *Southern Journal of Philosophy*, suppl. vol. 24: 1-22.

Scanlon, T. 1998. *What we owe to each other*. Cambridge, MA.: Harvard University Press.

Schnall, S., Benton, J., & Harvey, S. 2008. "With a clean conscience: Cleanliness reduces the severity of moral judgment." *Psychological Science* 19: 1219-1222.

Schnall, S., Haidt, J., Clore, G., & Jordan, A. 2008. "Disgust as embodied moral judgment." *Personality and Social Psychology Bulletin* 34: 1096-1109.

Smith, M. 1994. *The moral problem*. Oxford: Blackwell.

Smith, M. 2002. "Exploring the implications of the dispositional theory of value." *Philosophical Issues* 12: 329-347.

Smith, M. 2006. "Is that all there is?" *Journal of Ethics* 10: 75-106.

Sobel, D. 1999. "Do the desires of rational agents converge?" *Analysis* 59: 137-147.

Sreenivasan, G. 2002. "Errors about errors: Virtue theory and trait attribution." *Mind* 111: 47-68.

Swanton, C. 2003. *Virtue ethics: A pluralistic view*. New York: Oxford University Press.

Taliaferro, C. 1988. “Relativising the ideal observer theory.” *Philosophy and Phenomenological Research* 49: 123-138.

Zhong, C., Bohns, V., & Gino, F. 2010. “Good lamps are the best police: Darkness increases dishonesty and self-interested behavior.” *Psychological Science* 21: 311-314.