CHAPTER 12

MORAL FUNCTIONALISM

FRANK JACKSON AND PHILIP PETTIT

1. Moral Functionalism: The Basic Idea and a Reason to Believe It

Moral functionalism is a thesis about the meanings of moral terms, originally inspired by analytical functionalism about mental state terms. According to analytical functionalism, terms for mental states get their meanings from their roles in a theory, the theory known as folk psychology. According to moral functionalism, ethical terms get their meanings from their roles in a theory, a theory we might call folk morality (see, e.g., Jackson 1992 and Jackson and Pettit 1995). "Is morally right," for example, is true of X just if X has the property that plays the "is morally right" role in folk morality. We spell out what this comes to shortly.

Folk morality has a tripartite structure. It has *input clauses*, clauses that go from matters described in nonmoral terms to matters described in moral terms; they concern what kinds of actions, motivations, policies, etc., are morally right and wrong, and what kinds of results are morally good and bad. Killing is typically wrong. Inflicting suffering is bad. Telling the truth is very often what ought to be done. Folk morality also has *internal role clauses*, clauses concerning the interconnections between matters described in moral terms. Illustrations are inevitably controversial, but many find plausible a clause like: what is morally best out of the options available to an agent settles what the agent ought to do, the morally right action for the agent. Others object that agents are not always obliged to do what is best—the objectors often have in mind cases where what is best is very demanding. However, although they are objecting to the example just given, they aren't objecting to the very existence of internal role clauses. For example, our objectors likely allow that what is best is always morally permissible. Finally, there are the *output clauses* of folk morality. They concern the connections between what agents believe about what they ought or ought not to do and what's good and bad, on the one hand, and what they in fact do, on the other. A simple example is: agents who

believe that they ought to cease eating meat tend to become vegetarians. Some insist that there is some kind of conceptual connection between believing that something is what one ought to do and being at least inclined to do it. Others maintain that all that's true is that often (and maybe not often enough) agents tend to do what they believe they ought to do.[1]

We take it as obvious that there is such a thing as folk morality. It isn't controversial that people have beliefs about what kinds of actions are right and which are wrong, about how to reason using ethical terms and concepts, and about the influence of a person's moral opinions on their behavior. What's contentious are the details. The illustrations of input, internal role and output clauses given above are sketches, subject to one or another qualification by one or another theorist, but the picture in the broad is not. But why think, as moral functionalism maintains, that folk morality—or some suitable descendent of it (more on what this means below)—delivers an account of the meanings of moral terms?

One reason comes from reflecting on the nature of debates over the merits of one or another ethical theory. Consider, for example, what happens when philosophers (and nonphilosophers, if it comes to that) attack classical utilitarianism. Sometimes they point to the counterintuitive verdicts utilitarianism gives concerning what we ought to do in various cases described in nonmoral terms. They observe, for example, that utilitarianism tells us there is nothing wrong *as such* with punishing those we know to be innocent, and that the special place we give to the interests of those closest to us cannot be justified. Here the critics are targeting what utilitarianism says about input clauses. Sometimes critics of utilitarianism point out that it requires an agent to do that which is best out of the options available to them, no matter how demanding this might be, but surely, say the critics, sometimes it is morally permissible to do less than the best when doing the best is unduly onerous. Here the critics are targeting the implications of utilitarianism for an internal role clause. Sometimes critics of utilitarianism argue that no one could possibly live up to the demands of a theory that requires them to do what they believe is best, impartially considered (i.e., in a way that gives no special status to their own concerns). Here the critics are targeting the implications of utilitarianism for an output clause.

We all know how utilitarians respond. They grant what's common ground, namely that there is, prima facie, a clash between utilitarianism and what is often tagged "commonsense morality" (which is in effect another name for folk morality) but, they argue, careful reflection tells a different story. After reflection, it becomes clear that the implications of utilitarianism are ones we should embrace.

What matters for our purposes here is not who wins the familiar debate over utilitarianism that we have just reminded you of. What matters is that it is the debate that needs to be had.[2] Experiments in the sense meant in the social and physical sciences are not

[1] See, e.g., Brink (1989, ch. 3) for a detailed account of this debate.

[2] And is had in, e.g., Smart and Williams (1973), with Williams as the opponent and Smart the supporter of utilitarianism; for attempts to defang one or another objection from commonsense morality to consequentialist theories, see, e.g., Kagan (1982), Jackson (1991), and Pettit (2015, ch. 7).

going to decide whether or not utilitarianism is true, though they may tell us, for example, how popular it is. What is required—and what is in fact done—is to marshal the implications of utilitarianism for input clauses, internal role clauses and output clauses, and then to ask whether or not the implications are, after careful reflection, intuitively acceptable.

We grant that some utilitarian writings read as if what's on offer is in part a revisionary stipulation. But when it comes to defending the stipulation, what happens is the process we have just described. The stipulation is defended by supporting its implications for input clauses, internal roles clauses, and output clauses, often combined with suggestions about how to explain away the prima facie clashes with folk intuitions in one way or another, including in evolutionary terms. And, as those who offer the stipulation are well aware, its reception is precisely a function of how plausible their audiences find the defence being offered.

Much the same goes for debates over Kantian theories, virtue theories, reasons-first theories, modifications of utilitarianism that seek to ameliorate the putative clash with commonsense morality, and so on. The parties to the debates survey the verdicts of each theory for certain sorts of cases (i.e., for input clauses), the implications of each for which transitions between matters described in ethical terms are valid (i.e., for internal role clauses), and what each theory implies about the motivational force of the judgments we express in moral terms (i.e., for output clauses). They then perform a "compare and contrast" exercise, urging that it is their favorite theory that wins out in the end. The protagonists may not describe what they are doing in quite the way we have but, we submit, this is in fact what is going on. The simplest argument for moral functionalism is that it explains why this is what happens. The debate is all about finding the theory that delivers the best fit with our considered judgments—understood so as to include the 'explainings away' we mention above—concerning input clauses, internal role clauses, and output clauses.

One way to highlight our message is to imagine that utilitarians succeed in defusing all the famous objections. They convince us, for instance, that when we think matters through carefully, it is clear that sometimes, though not usually, it is right to punish the innocent, and that we always ought to do what's best, and that the belief that an action maximizes expected happiness has the exactly the right kind of connection to motivation. If that happened, it would be, as they say, game over. What more could one ask for by way of vindicating utilitarianism? Moral functionalism explains why this would be the case. What would have been established is that utilitarianism makes true the clauses that give moral terms their meaning. No wonder that it would be game over.

We can make essentially the same point from the other side of the fence, so to speak. Surely many who are certain that utilitarianism is false are certain precisely because they are convinced that utilitarians cannot explain away the prima facie clashes with commonsense or folk morality. Moral functionalism explains the relevance of this conviction: if correct, it means that the meanings of the moral terms imply that utilitarianism is false.

In what follows, we survey the implications of moral functionalism for a number of live issues in ethics (in one case, the issue is perhaps better described as one that should be more alive than it is) including, of course, for moral realism. It is an interested survey. We will be suggesting that the implications are plausible ones. In what follows, we use "moral" and "ethical" interchangeably.

2. The Supervenience of the Ethical on the Nonethical

The supervenience of the ethical on the nonethical differs from the supervenience of the mental on the physical in two important ways. First, it is far less contentious. Although many affirm that the mental supervenes on the physical, it is far from common ground among philosophers of mind. It is, however, rare to come across someone who denies that the moral supervenes on the nonmoral. Nearly everyone grants that if two actions, states of affairs, policies, etc., are exactly alike in nonethical ways—ways we can specify without using ethical terms—they are exactly alike ethically. Or, to say it the other way around, a difference in ethical nature demands a difference in nonethical nature: for example, if one action ought to be punished and another ought not, the actions must differ in some respect we can specify in nonethical terms. Second, it is necessary and a priori that the ethical supervenes on the nonethical, whereas the supervenience of the mental on the physical is, at best, an a posteriori truth, and (many would add) at best a contingent truth.[3]

Moral functionalism can explain the necessary a priori supervenience of the ethical on the nonethical because it makes possible a reductive analysis of the ethical in terms of the nonethical. It allows us to exploit the Ramsey–Carnap–Lewis way of defining theoretical terms (see Lewis 1970) in a way which makes it transparent how the nonmoral a priori entails the moral, so explaining the necessary a priori supervenience of the moral on the nonmoral. Here is how the account runs, in outline. We can think of folk morality as a longish conjunction of the input clauses, internal role clauses and output clauses that give the meanings of moral terms according to moral functionalism. Let $\mathbf{T}(M_1, \ldots, M_n)$ be the sentence that gives the theory, where the M_is are all the moral terms. Folk morality will be satisfied just if there are properties standing in the required relations, if, that is, $(\exists x_1) \ldots (\exists x_n)\, \mathbf{T}(x_1, \ldots, x_n)$, where each x_j is in M_j's place in **T**. (This is the "Ramsey sentence" for folk morality.) But if, as moral functionalism holds, each M_i is defined by its place in **T**, then we can specify what it is to be M_i, in terms of that very place. Here is how it looks for the case of being right

[3] See, e.g., Lewis (1994, p. 52).

(A) y is right if and only if $(\mathrm{E}x_1) \ldots (\mathrm{E}x_n)$ [y has x_i & **T** $(x_1, \ldots, x_n)$]

where x_i replaced "rightness" in **T**. To say it without symbols: to be right is to have the property that fills the "rightness" place in folk morality. *Mutatis mutandis* for being what ought not to be done, being morally permissible, being good, and so on.

The important point, in the current context, is that (A) specifies what it is to be right in nonmoral terms; there are no "M_i"s to the right of the "if and only if" in (A). It thus explains how the nonmoral can a priori entail the moral for the case of being right. *Mutatis mutandis* for the other moral properties. This is how moral functionalism explains the necessary a priori supervenience of the moral on the nonmoral.

Have we given a naturalistic account of the moral properties? The nonmoral terms that appear on the right-hand side of (A) presumably count as naturalistic. They contain no ethical expressions. (A) can, therefore, be regarded as a naturalistic reduction of being right. However, (A) says nothing about the property that fills the "rightness" place—that is, that plays the rightness role, and the same goes for the other properties that make the Ramsey sentence true. This suggests that a "Moorean" could insist that *these* properties, the ones that play the roles, cannot be captured without using ethical terms. In this context, our Moorean might draw attention to the familiar point that analytical functionalism is not, as it stands, a version of materialism or physicalism about the mind. It becomes a version of materialism when combined with a thesis about the kind of properties that stand in the relationships definitive of being in one or another mental state.

This would, however, be a mistake. (A)—and the corresponding accounts for the other ethical terms—imply that there are *no* properties that can only be captured using ethical terms; this is because they are reductive analyses of ethical language. What might be true, for all that (A) and its partners say, is that the properties that make the Ramsey sentence true are not natural properties in another sense; they may not be the kinds of properties that figure in one or another account of what our world is like to be found in the natural sciences. How likely this is, and further issues raised by Mooreanism, are discussed in the next section.

3. How We Learn the Meanings of Moral Terms

An account of what a word in a public language means should explain how it can come to have that meaning and how that meaning can come to be a shared meaning, at least in many cases. That's required if we are to use words to express agreement and disagreement. People don't agree or disagree merely by uttering the same or different words. Well not quite. Perhaps I hear the sentence "John Doe lost on a TKO." I may have only a rough idea what a TKO is. I may, nevertheless, accept what I hear and agree with it in

the sense that I am confident that John Doe lost on what those more knowledgeable about boxing than I am mean by a TKO. That's the message of Hilary Putnam's division of linguistic labor (Putnam 1975), or so it seems to us. This does not, however, mean that I give "TKO" the same meaning as the knowledgeable. If it did, there'd be no point in proceeding to ask what a TKO is, which is of course what one typically does, after which one says that *now* one knows what the phrase means.

It is important that the key ethical terms have shared meanings. Agreements and disagreements over the morality of abortion, mercy killing, and our moral obligations to future generations are among the most important agreements and disagreements we have. What is more and obviously, we want more than agreement and disagreement about words; we want our agreements and disagreements to be substantive (agreement in the sense illustrated by the TKO example is not what we want).

It follows that an important desideratum for any account of the meaning of ethical terms is that it can give an account of how we came to acquire them that explains how it is possible for them to have shared meanings, and this needs to be an account that is consistent with the fact that the meanings we give our words is a contingent a posteriori matter. The importance of this desideratum can be overlooked. The reason may be the assumption that intending to mean what others mean is in itself enough to create a strong presumption that one does in fact mean what others mean—some seem to read Putnam as saying this—but intending to do so and so is one thing and in fact doing it is quite another. Intending to mean what others mean is not special among intentions in being presumptively successful.[4] Moreover, we want it to be the case, at least sometimes, that English speakers and, say, Croatian speakers can agree and disagree over moral questions, and for this to be revealed by the words that come from their respective mouths, pens, and keyboards, despite the fact that English and Croatian speakers use different words to express their opinions. This requires that the words in their different languages have the same meanings. But a typical speaker of English and a typical speaker of Croatian do not intend to mean the same by, for example, "morally good" and "moralno dobro." That's an intention they will *acquire* if and when they become competent in both languages.

According to moral functionalism, learning the meanings of moral terms is acquiring a mastery of the input, internal role and output clauses we talked about earlier. We learn which actions, policies, etc., receive which ethical labels (the input side of the story). We learn how to reason using the ethical labels (the internal role side of the story), and we learn what kinds of reactions are appropriate and are to be expected from those who

[4] Incidentally, intending to mean what others mean by a term isn't necessary for meaning what they mean. R is, we may suppose, a mathematical genius living in a poor village. As a result of limited access to texts, he wrongly believes that mathematicians use "prime number" for any positive integer divisible by itself and one alone. He resolves, however, that in his work he will use "prime number" for any positive integer divisible by itself and one alone with the exception of one, realizing that this is the theoretically better notion. Unwittingly, he is using "prime number" with the meaning mathematicians give it despite intending not to.

apply various ethical labels to actions available to them (the output side of the story). We trust that these remarks will resonate with parents who seek to introduce their children to ethical ways of thinking and talking. Don't we parents point to examples of what's morally good, wrong, etc., highlighting the features that make them so? Don't we indicate the kind of behavior called for by judgments of what's right and wrong? Don't we seek to give our children a sense of how to reason about ethical issues using ethical terms?

How can this story be a story about shared meanings, given how much disagreement there is over ethical issues; how, that is, can we find enough agreement over input, internal role and output clauses to allow this story to be one about shared meanings? There are two ways to respond to this fair question. The optimistic one is implicit in our earlier discussion of the debate over utilitarianism. There we pointed out what it would take for utilitarians to win; they would need to defuse the famous problem cases for utilitarianism. They would need to show that what looks bad on the face of it—"how can you possibly hold that the interests of my children have no special claim when I ask myself what I ought to do!"—isn't when one thinks the issues through. We made the same observation about, e.g., Kantian views. Suppose it is the Kantians who find themselves in the happy position of being able reconcile what their view says with folk morality. By the time the big books have been written and all the *t*'s crossed and *i*'s dotted, it becomes clear that a Kantian style ethical theory delivers the intuitively satisfying answers for the input, internal role and output clauses. In that case, the Kantians would be the winners. As we argued earlier, what is going on when theorists defend their favorite position in ethics is, in effect, an exercise in reconciling the answers their position delivers on the key input, internal role and output clauses with our preanalytic but considered intuitions (beliefs). So, the first response to the fair question is that, at the end of the day—and in this context we might talk of reflective equilibrium—we will find very substantial agreement over the key input, output, and internal role clauses. Our problem, as philosophers seeking the one true theory in ethics, is to find the crucial insights and ways of framing matters that tell us which overarching theory—utilitarianism, Kantianism, virtue theory, contract ethics, idealized desire ethics, . . .—delivers the goods.

This view about what it takes to find the one true theory might sensibly be combined with an awareness that, in the process of searching for a way to make best sense of the famously abundant and far from uniform intuitions on display in books and papers on ethics, we will find it important to distinguish folk morality from what we might call *mature folk morality* (as in Jackson 1998, p. 133), where mature folk morality is the result of facing up to the fact that parts of extant folk morality may embody attitudes we rightly realize need to be jettisoned and may contain internal inconsistencies (think, e.g., of the way harvesting subjects' first up responses to variations on the trolley problem often deliver responses that are at war with each other). Our task then becomes that of finding the best theory that makes overarching sense of mature folk morality, and the moral functionalist account of the meanings of moral terms will need to be framed in terms of mature folk morality—the theory we will end up converging on—whatever that turns out to be, precisely. In discussion, some have expressed scepticism about the possibility

of convergence. That's understandable perhaps, but we note that those who work in ethics must include a large number of optimists. For surely very many of those who give talks and write books and papers on ethics are doing so because they think they have some chance of convincing people—their audiences and their readers—that their claims about what's right and wrong are correct, or that what they say about moral reasoning is correct, or that what they say about the behavior of those with one or another moral opinion is correct. They must, that is, believe that there is some chance of agreement on input clauses, internal role clauses and output clauses (and of course on the relevant empirical considerations in the background). Here they may well have recourse to the 'explainings away' we mentioned earlier. For isn't this belief a major rationale for publishing those books and papers, and for giving those talks?

The more pessimistic response is to grasp the nettle and allow that we don't all give the key ethical terms the very same meanings, and this isn't going to change as time passes; the talk above of converging over time on an agreed theory, call it mature folk morality if you like, may be a hope that explains the publishing and advocacy behavior of many who work in ethics but it is a vain hope all the same. We insist that, as we say early on, there is such a thing as folk morality and there will be such a thing as mature folk morality—surely we are improving—but grant that, at the end of the day, what may become clear is that there is no *single* theory to call "mature folk morality." What we have, it may turn out, are a number of different, internally consistent theories that agree about a lot in the sense of having input clauses, internal role clauses and output clauses that are similar in many ways, but there may not be enough uniformity between the theories to imply that the ethical terms as defined by one of the theories have the very same meanings as those defined by another of the theories. There will be enough similarity to allow for substantive, and not merely nominal, agreement and disagreement on many issues as expressed using the terms as defined by the different theories, but this will not be true for all issues. Given what we say early on in this section about the importance of shared meanings for ethical terms, grasping the nettle will have a significant cost—sometimes what appears to be substantial agreement or disagreement will be merely nominal—but perhaps it will turn out that this is a cost we have to pay. Philosophers sometimes seem gripped by the idea that there is one true ethical theory, perhaps the one they have devoted their life to first finding and later defending. Maybe there is, but maybe there isn't, or maybe there is a degree of indeterminacy.

We close this section by addressing two further worries that regularly come up in discussion, which, as we will see, connect with the issues lately canvassed. The first worry might be expressed in the following words:

> It cannot be the case that the meanings of ethical terms are given by the kind of network account you have offered. The meanings of central moral terms like "morally right," "morally impermissible," "morally evil," etc. remain *fixed* as we debate the input, internal role and output clauses you regard as meaning-giving. For example, when we argue over a candidate input condition like that intentional killing not in self-defense is always morally impermissible, the meaning of "morally

> impermissible" remains a constant. It isn't a function of where we end up on the question. The same goes for any other input condition that might be nominated, and the same goes for the internal role and output clauses that might be nominated as part of the network account of the meaning of moral terms. The meanings of moral terms are one thing; the verdicts we frame in moral terms are quite another.

We grant the initial appeal of this line of thought, as is evinced by the fact that it often comes up, in one way or another, but insist that it does not withstand scrutiny. To start with, it is a general point about meaning that the meanings of expressions and the verdicts we deliver using those expressions are intimately linked, as, e.g., Locke (1689/1975, Book III), says. We use words to express how we *take* things to be. For example, physicists use "Electrons exist" to make a claim about how they believe things to be (the verdict they have come to), and the claim they make is a function of the meaning they give the word "electron." Or think of what happens when we go to a philosophy talk. We recover what the speaker believes—their verdicts about this or that issue—to the extent that we *understand* the words that come from their mouth or appear on their slides. It isn't true that the verdicts we express using words float free of the meanings of those words.[5] A verdict on whether a certain action is right or wrong, for example, will be a function of the meaning of those moral terms and of the ascertained empirical facts about the action.

In any case, if the meanings of ethical terms float free of input, internal role, and output clauses, we need to ask what, in that case, does fix the meanings of the ethical terms? We saw earlier the problems attendant on relying on intentions to use the words as one's fellow language users use them to fix meaning—having an intention is one thing, fulfilling it is quite another. Moreover, intentions to mean what others mean, even when successful, radically underdetermine meaning: samenesses and differences underdetermine what the samenesses and differences hold between.

We suspect that what lies behind the line of thought in question is a view in metaphysics, the one we called "Moorean" earlier; the view that there are irreducibly non-natural properties, where to be irreducibly non-natural is to be a property that can only be picked out using ethical terms. We can then see, runs the thought, how the meanings of ethical terms can be independent of input, internal role and output clauses. Their meanings are fixed by being connected to these non-natural properties. But there are well known problems for any account of this sort. One is that it rests on an implausible view about the nature of our world. There is no reason to believe that the posited properties are instantiated. A second problem is that, even if they are instantiated and we somehow come to know this, how could we come to have justified beliefs about their distribution, and how could we explain how ethical terms come to pick them out in a

[5] "Can't competent speakers of English differ in their verdicts concerning which substances are poisonous without thereby giving 'poisonous' different meanings, consistently, that is, with their being in real disagreement?" Well, no. For they need to agree on what it takes to be poisonous; their disagreement needs to be limited to which substances have what it takes.

way that ensures that one person's use of "is morally required" picks out the same property as another's use of "is morally required," or picks out the same property as some phrase in Croatian. A third problem comes from the earlier noted supervenience of the ethical on the nonethical. The ethical *depends* (in the sense of Broad 1968) on the nonethical, in addition to supervening on it. It is impossible to have an ethical nature without having a relevant nonethical nature (a nature we can describe without using ethical terms). Good acts must also be acts that increase happiness, honor a promise or whatever. The upshot is that the ethical and the nonethical are locked together in a way that makes it very hard to believe that they are distinct in the way Mooreans hold that they are.[6]

The second of our two further worries relates to our observation that moral functionalism does not offer a guarantee that we all mean the same by moral terms; in consequence, we may have to allow that some (not all) agreements and differences over ethical questions are merely verbal. The worry is that this is a special, and undesirable, feature of moral functionalism. Our reply is that it is a feature of any and every account of moral language, be it expressivist, Moorean, causal, etc. What our words mean is a contingent, a posteriori matter, as we say above; that should be granted independently of whether one is a moral functionalist, an expressivist, a Moorean, a causal theorist, etc. All theorists in ethics must allow that it is an open question whether or not what one person means by, say, "morally right" is the same as what another means by "morally right," where an open question means one that has to be investigated empirically and one that may receive an affirmative or a negative answer. The same is true for whether or not what one English speaker means by "morally good" is the same as what some given speaker of another language, Croatian, say, means by "moralno dobro." In this case, the key point is especially obvious: compiling English–Foreign Language dictionaries is an exercise in linguistic fieldwork.

How one thinks the investigation should go will depend on one's views about moral language. Perhaps an English-speaking expressivist will approach the issue somewhat as follows: "Reflection on my use of 'is morally good' tells me that I use it to express a certain pro-attitude. Having this attitude manifests itself in various ways and I note that these manifestations are to be found in those around me, and also in those who speak one or another language other than English when they utter . . . ." An English-speaking Moorean will be concerned instead with how they can be confident that the properties they hold are picked out by their use of "is morally good," "is wrong," etc., are the same properties as are picked out by those terms by their fellow English speakers, and will want assurance that their English words pick out the same properties as do various terms in other languages. Arguably, as we suggest two paragraphs back, the special nature of the properties a Moorean holds are picked out by their own usage makes finding good reasons to hold that others pick out the same properties especially challenging.

[6] For more on this and related points, see Jackson (1998, ch. 5) and Streumer (2017, ch. 2).

But we are speculating. It is really for supporters of those views to explain how they can be confident that the meanings of moral terms (as they take them to be) are the same—or enough alike—across a language community and between different language communities to ensure that agreements and disputes framed in moral terms are genuine and not merely verbal. We have given the answer we like, and are discussing the issue as it arises more generally to make the point that there is no special problem here for moral functionalism.

4. The Utility of Moral Language

Why is the historical-causal theory of reference for proper names so plausible? The answer is that it explains why proper names are so useful, and, thereby, why language evolved so as to contain them. The platforms at Penn Station get assigned different numbers to avoid confusing one platform with another, and the numbers are then used to pass on information about where and when a train is arriving. That's common knowledge. In the broad, the same goes for proper names. People, cities, streets, buildings, etc., get assigned names to assist in distinguishing one from another. These distinguishing marks help us avoid confusing different people, cites, streets, etc., and are available to assist in passing on information about people, cities, streets, etc. This is as much common knowledge as our remarks about Penn Station's platform numbers. Philosophers of language debate whether this common knowledge should be viewed as a version of causal descriptivism, or as a version of a causal theory of reference for proper names, viewed as importantly different from a description theory. That debate is by the way here. What is important for our concerns is that it is clear what purpose is served by having names in our languages, and that this allows us to explain why they evolved so as to contain them.[7]

We should expect the same of an account of ethical terms. It should make it clear why they are useful—what purpose or purposes they serve—and thus of why our languages (English, Croatian, Chinese, etc.) evolved so as to contain them. Arguably, this desideratum has of late not received the attention it deserves (but see Sterelny and Fraser 2017). It has, however, a history in writings that go under the banner of evolutionary ethics. Here is a simplified version of a style of evolutionary ethics to be found in Alexander (1891). The feeling of pain is useful because it has two properties: it draws attention to bodily damage, and it tends to cause behavior that minimizes the damage. In similar vein, runs the theory, the feeling of obligation is useful because it has two properties: it is caused by the availability of possible actions of a kind such that performing them would have great utility in certain circumstances (typically, ones where cooperation and coordination between agents with different interests is important now or in

[7] The theory is due to Kripke (1980); we are not suggesting he would agree with the way we present it. Our presentation draws on Kroon (1987) and Jackson (2010).

the future—Alexander talks of the importance of an "equilibrium" in society between competing interests), and it tends to cause the performance of these actions. The fact that it has these two properties together explains why we evolved to experience it. And, of course, once we evolved to have the feeling of obligation, it was useful to have words for the kinds of actions that provoke it, and to grasp the category or categories to which the actions that provoke it belong. We have, accordingly, an evolutionary explanation of how we came to have words like "morally required" in our language, and to have the concept of the morally required, for we have an explanation of why ethical terms (and concepts) are useful.

This is the merest sketch of the theory Alexander outlines, but even so the problem with it is obvious. There is no feeling of obligation akin to the feeling of pain. What we do have are *beliefs* that certain actions available to us are ones we ought to perform. If we fail to act on them, we sometimes have "pangs of conscience" but often we do not, and, in any case, the pangs do not motivate in the way that pain does. What is more, beliefs that certain actions are ones we ought to perform are not typically caused in the way that feelings are. How hot a chilli tastes is a function of its chemistry, whereas our moral judgments are responses to how we take something to be—that it is an infliction of needless pain, that it is the keeping of a promise, etc.

Nevertheless, Alexander gives us an item on the agenda for any account of the meanings of moral terms: explain why, on the account in question, it would be useful to have moral terms and why we might have evolved to have them. Now there are a number of extant, interesting accounts of how moral concepts and terms evolved, directed precisely to the question of what makes them—the concepts and the words—useful. What is important for what follows are not the details of one or another account but what is in common between them: ethical terms, concepts and ways of thinking evolved through the need to adjudicate between competing interests, something that became especially pressing when we formed communities in ways that enhanced our chances of survival by encouraging cooperation.[8] Hermits can do what they like. They do not have to balance what they want against what others want. The downsides are that if bad things happen, they have no one to turn to for help, and that projects that require many hands to the wheel are beyond them. But, once we started to live in communities, a pressing issue became how to balance what one person or group wants against what another person or group wants, and how to regulate our behavior as members of a community in ways that promote the interests of the community as a whole and, thereby, at least sometimes, our own individual interests. We needed to find acceptable ways of adjudicating between competing interests in ways that enhance cooperation and lead to positive outcomes. When this started to happen was when we started to think ethically.

Although this lightning sketch prescinds from the details of how the evolution of thinking in ethical terms helps resolve disputes and aids co-operation, we can say this

[8] Thus Sterelny and Fraser (2017, p. 1003) talk of "principles of action and interaction that support forms of cooperation."

much in the broad. Three things must be true if an account anything like this is to explain the evolution of morality. The first is that there need to be properties with the following feature: acting so as to promote their instantiation has good effects for those who belong to communities. (What about properties whose *suppression* has good effects? We can think of these in terms of promoting the instantiation of their nonoccurrence.) The second concerns thinking in terms of these properties. It had better be possible to do so. Human agents are thinking agents. The third concerns motivation. It had better be the case that the belief that some course of action has one or another of these properties has, as a rule, some tendency to cause behavior that leads to their instantiation. One question is which properties are such that their instantiation *would* have good effects; a second and distinct question is the mechanism by which these properties might come to be instantiated. For intentional agents like us, the mechanism will often involve the tendency of beliefs about these properties to cause actions that realize them.

Suppose that all three requirements obtain. How could this be an account of the evolution of ethical ways of thinking? There is no mention of ethics as such in any of the three conditions. But if moral functionalism is true, there is mention of ethics—implicitly. The first condition was in effect that we need input clauses—they specify the actions, policies, etc., that have the properties whose instantiation would have utility; the second condition tells us in effect that we need internal role clauses—they tell us how to reason in terms of these properties; the third condition tells us in effect that we need output clauses—they tell us how the needed properties might come to be instantiated through the actions of intentional agents. The upshot is that the existence of properties that satisfy the input, internal role and output clauses of moral functionalism is exactly what would be required for some kind of evolutionary account of the emergence of moral ways of thinking to make sense. And the utility of ethical terms will then lie in the utility of having terms for the properties in question.

We started this section by noting that any account of ethical language should explain its utility. Talk of utility naturally invites evolutionary reflections, and we have seen how moral functionalism makes good sense from an evolutionary perspective. But, for those unimpressed by evolutionary ways of thinking about ethics, we note that the key point about how moral functionalism explains the utility of moral language can be made independently. It is a matter of record that describing matters in ethical terms makes things go better, especially when dealing with problems that arise from the fact that we live in communities and need to adjudicate between competing claims for a share of limited resources. It is very hard to believe that this is an accident. But if it isn't an accident, there must be a story to tell about the properties our ethical terms are picking out which explains this happy result. The situation is akin to that with names. It is a matter of record that assigned names—to cities, streets, people, etc.—are very useful. This fact calls for explanation and, as we say above, the explanation will advert to some version or other of a historical-causal theory. What's the right story in the case of ethical terms? First, they need to pick out properties whose promotion would make things go better, and properties that do the opposite. Second, they need to pick out properties we can reason about. (A feature of the way using ethical terms makes things go better is their

role in facilitating deliberations about what ought to be done.) Finally, they need to pick out properties we tend to promote, and properties whose instantiation we have some tendency to suppress. The story will, therefore, have input clauses—clauses that tell us where the properties in question are to be found; internal role clauses—clauses that tell us how to reason about the properties; and output clauses—clauses that tell us about their motivational properties, both pro and con. All three are essential if moral terms are to be of use to us in negotiating our interactions with others in our communities in the ways distinctive of debates framed in moral terms. This is how moral functionalism can explain the utility of ethical language, independently of the question of how it evolved.

We close this section by noting that a focus on making sense of how ethical terms and ways of thinking did evolve—or would have evolved the way things might well have been—allows us to reshape the way we presented moral functionalism earlier. We presented it as a reductive analysis of moral terms, one that allows us to find a place for moral properties—the properties the moral terms pick out—within a naturalistic picture, and we noted why this would be a good thing to do. Our focus, however, wasn't on strict fidelity to our current moral concepts—that was the point of the distinction between folk morality and mature folk morality, and our suggestion that it would be best to analyze moral terms via their place in mature folk morality. There is, though, another way of thinking of moral functionalism. We can view it as what we would end up with if we asked after the *genealogy* of ethics. The key idea can be explained using the example of money.

No one, we take it, thinks that the concept of money is *sui generis*. There will be a reductive analysis of the following form

(B) X is money if and only if X is . . .

where the words after "if and only if" do not contain "money" or its equivalent. (When we say that there will be such an analysis, we are not suggesting that writing it down will be an easy task.) But, of course, there is a story to be told about how money came into existence, a story that will advert to its utility in assisting with the exchange of goods, etc. An (idealized) account of this kind is to be found in Menger (1892). This means that a good question to ask is, What would X need to be like in order for it have evolved in the way that money did, and for it to play the kind of role that money plays in society? And we should expect an answer to this question to deliver something like what appears after "if and only if" in (B), and a plausible thought is that a fruitful way of thinking about (B) is as an answer to the genealogy question, in the sense that it tells us what is required of money in order for it to have evolved in the way that it did, or for it to be such that it would have evolved in any society much like ours.

Likewise, as we emphasize above, there is a story to be told about how ethical ways of thinking and talking evolved. And, as we said, we can say this much in the broad. It will be a story about (i) there being properties whose instantiation promotes survival in the kinds of situations in which ethical ways of thinking in fact evolved, and properties that do the opposite; (ii) our being able to reason about these properties; and (iii) our coming

to be such that beliefs about these properties have at least some motivational force, pro or con. Clauses (i), (ii) and (iii) correspond, respectively, to the input, internal role and output clauses of moral functionalism. The upshot is that asking after the genealogy of ethical terms and ways of thinking—how they came into existence the way things were or would have come into existence the way things might well have been—will deliver biconditionals like (A), and the corresponding ones for the other ethical terms.[9]

5. What It Takes to Be a Moral Realist

We take the *cognitivist* part of moral realism to be the thesis that there are moral properties, understood as the claim that predicates like "is morally wrong" and "is morally good" ascribe properties. The *realist* part of moral realism adds that the properties in question are, on occasion, actually possessed; being morally wrong is in fact a property of, say, needless killing, and being morally good is in fact often a property of, say, donating to charity.

How substantive is the cognitivist part of moral realism? We have the predicate "is morally wrong" in English, and can form the expressions "being morally wrong" and "the property of being morally wrong" from that predicate. Given that, it might be asked how could anyone doubt the existence of moral properties? Our answer to this good question is that the debate isn't about words and isn't about the constructions our language allows us to make from words. It is about whether or not certain words and phrases are good for telling us about how things are. Take, for example, the word "house." Imagine we have divided all the objects there are into those in the extension of "house" and those not in its extension. Is there a difference between the items in the two sets over and above the difference with regard to whether or not they are in the extension of "house"? Of course there is. There is a way something has to be for "house" to apply to it, and this way is over and above belonging to the extension of "house" and is the information about how things are that the word is good for delivering. If that were not true, the information the use of the word makes available would be limited to a fact about word usage, and it isn't. Cognitivism in ethics affirms, as we understand it here, that the same goes for ethical terms. There is, for example, a difference between actions in the extension of "is morally wrong" and those not in the extension of "is morally wrong," *over and above* the difference in whether or not they belong to the extension of "is morally wrong." The debate over the nature of the property of being morally wrong is, we urge, to be understood as the debate over this difference, and in particular the nature of the acts inside the extension of "is morally wrong."

We mentioned earlier Moorean views about the metaphysics of morals. If they are correct, the only words we have for the moral properties are expressions containing

[9] For more on the genealogical way of thinking in the case of ethics, see Pettit (2018).

moral terms. Any attempt to pick out moral properties in nonmoral terms is bound to fail. This is the sense in which Mooreanism is inconsistent with naturalism in ethics, on one understanding of naturalism. All the same, actions in the extension of one or another moral term will have something in common that outruns their falling under the extension of the term in question, namely, the very properties Mooreans hold we can only talk about by using ethical terms. In this sense, on Moorean views moral terms mark real and not merely nominal divisions among, for example, actions, and Moorean views are a species of cognitivism.

Moral functionalism likewise is a species of cognitivism. To fall under such and such a moral predicate is to have the property that plays the relevant role in the input, internal role and output clauses. Earlier, in §2, we spelt out what this comes to for "is right." Is moral functionalism a species of realism? Not as such. Realism about some given moral property will be the claim that there is a property, in the sense of an *instantiated* property, that fills the role definitive of that property. Is there, for example, a property that some actions in fact possess that fills the bill for being right that we gave in §2? As we note in that section, this question comes down to the question as to whether or not the Ramsey sentence for mature folk morality is true.

This means that one might embrace moral functionalism and proceed to use it as a platform for the denial of realism concerning, for example, the property of being right. One might, for example, insist that the only plausible version of an output clause for being right is that an action is right just if it has a property which is such that believing an action has that property *entails* desiring that the action be done, and proceed to argue that, as there is no such property instantiated in our world (and maybe, for Humean reasons, there could not be), nothing has the property of being right.

We favor a less hard-line approach. We have already commented on the manifest utility of moral language and how understanding its utility goes hand in hand with understanding how we came to acquire moral concepts. This makes it hard to believe that moral terms mark out empty categories. The situation is akin to that which obtains with analytical functionalism about the mind. The manifest usefulness of its folk psychological categories makes it hard to believe that they are empty. Advances in neuroscience may suggest important refinements and extensions, but elimination is very unlikely. We talked earlier of the nature of debates in ethics. They are, we suggested, best seen as attempts to find occupants for the roles we moral functionalists talk about. There is plenty of give and take, perhaps some explaining away, and sometimes a certain amount of bullet biting, but it is a matter of record that there is plenty of constructive engagement. We think the message is that, somewhere or other, somehow or other, there are instantiated properties to be found that near enough fill the roles that moral functionalism says need to be filled (while granting the possibility, noted earlier, that there may be irresolvable differences). Even Mackie, who argued that the roles were not filled—his error theory (1977) is in part based on insisting on demanding specifications of the roles of folk morality and mature folk morality, with the addition of the claim that nothing fills these demandingly specified roles—felt free to write a book with two apparently inconsistent parts. The first lays out his argument for an error theory. The second part is

an essay on various topics in ethics. How could he have thought that this was a sensible enterprise? (We are not alone in asking this question.) He took it for granted that there were near enough occupants of the roles to allow the second half of the book to be worth writing. At any rate that is, we urge, the way to make good sense of what Mackie is doing in the second part of the book given the thesis he argues for in the first part.

In a number of places earlier in this chapter, we highlight the fact that what our words mean is a contingent a posteriori fact. Here we are suggesting that we might, if it is needed, make adjustments—sensible ones—on what ethical terms mean according to moral functionalism to ensure that realism comes out true. This is really no different from what happened when it was discovered that atoms could be split. The sensible semantic decision was to relax the clause that insisted that atoms had to be more than just very hard to split, so allowing us to avoid going eliminativist about atoms.[10]

References

Alexander, S. 1891–2. "Is the Distinction between 'I' and 'Ought' Ultimate and Irreducible?" *Proc. Aristotelian Society* 2(1): 100–107.

Brink, David O. 1989. *Moral Realism and the Foundations of Ethics*. Cambridge: Cambridge University Press.

Broad, C. D. 1968. "Certain Features in Moore's Ethical Doctrines." In *The Philosophy of G. E. Moore*, edited by P. A. Schilpp, 43–67. La Salle, IL: Open Court.

Jackson, Frank. 1991. "Decision–Theoretic Consequentialism and the Nearest and Dearest Objection." *Ethics* 101(3): 461–482.

Jackson, Frank. 1992. "Critical Notice of Susan Hurley, *Natural Reasons*, Oxford University Press, 1989." *Australasian Journal of Philosophy* 70(4): 475–487.

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

Jackson, Frank. 2010. *Language, Names, and Information*, Oxford: Wiley-Blackwell.

Jackson, Frank, and Philip Pettit. 1995. "Moral Functionalism and Moral Motivation." *Philosophical Quarterly* 46: 82–86.

Kagan, Shelly. 1982. *The Limits of Morality*. Oxford: Oxford University Press.

Kripke, Saul. 1980. *Naming and Necessity*. Cambridge, MA: Harvard University Press.

Kroon, Fred. 1987. "Causal Descriptivism." *Australasian Journal of Philosophy* 65(1): 1–17.

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

Lewis, David K. 1994. "Reduction of Mind." In *Companion to the Philosophy of Mind*, edited by Samuel Guttenplan, 412–431. Oxford: Blackwell.

Locke, John. [1689] 1975. *An Essay Concerning Human Understanding*, edited by Peter H. Nidditch. Oxford: Clarendon Press.

Mackie, J. L. 1977. *Ethics: Inventing Right and Wrong*. Harmondsworth: Penguin.

Menger, C. 1892. "On the Origin of Money." *Economic Journal* 2: 239–255.

[10] We have discussed moral functionalism with colleagues and friends over many years. Some have been sympathetic, some less so; thanks are due to them all, especially Michael Smith and David Lewis. We are also indebted to the editors for convincing us of the need to add to §3.

Pettit, Philip. 2015. *The Robust Demands of the Good: Ethics with Attachment, Virtue and Respect*. Oxford: Oxford University Press.

Pettit, Philip. 2018. *The Birth of Ethics: Reconstructing the Role and Nature of Morality*. New York: Oxford University Press.

Putnam, Hilary. 1975. "The Meaning of 'Meaning.'" In *Mind, Language and Reality*, 215–271. Cambridge, UK: Cambridge University Press.

Smart, J. J. C., and Bernard Williams. 1973. *Utilitarianism: For and Against*. Cambridge: Cambridge University Press.

Sterelny, Kim, and Ben Fraser. 2017. "Evolution and Moral Realism." *British Journal for the Philosophy of Science* 68: 981–1006.

Streumer, Bart 2017. *Unbelievable Errors: An Error Theory about All Normative Judgements*. Oxford: Oxford University Press.