Journal of the American Philosophical Association (2015) 617–634 © American Philosophical Association
DOI: 10.1017/apa.2015.6

Necessary Moral Principles

ABSTRACT: *Moral realism entails that there are metaphysically necessary moral principles of the form 'all actions of nonmoral kind Z are morally good'; being discoverable a priori, these must be (in a wide sense) logically necessary. This article seeks to justify this apparently puzzling consequence. A sentence expresses a logically necessary proposition iff its negation entails a contradiction. The method of reflective equilibrium assumes that the simplest account of the apparently correct use of sentences of some type in paradigm examples is probably logically necessary. An account is simple insofar as it uses few predicates designating properties easily recognizable in many different kinds of paradigm examples. I illustrate how reflective equilibrium uses these criteria to discover logically necessary nonmoral propositions such as 'if S remembers doing X, S believes that he did X' and 'if it is scarlet, it is red'. I then illustrate how exactly the same procedures can lead us from paradigm examples of apparently true sentences asserting that actions of some kind are good to discovering moral principles. I advance the contingent hypothesis that most contemporary humans, although they initially disagree about moral issues, have derived from different paradigm examples the same concept of moral goodness, which will ensure that the use of reflective equilibrium will lead to eventual agreement about moral issues. That strongly suggests that the moral principles eventually reached by reflective equilibrium are logically necessary ones.*

KEYWORDS: reflective equilibrium, moral realism, logical necessity, ethics, metaethics

1.

Moral realism as I shall understand it is the doctrine that some moral propositions, that is, propositions that attribute a moral property to some action, are true.[1] Moral supervenience, the highly plausible doctrine that I shall take for granted that the moral supervenes on the nonmoral, then becomes the more precise doctrine that

Thanks to anonymous reviewers for helpful comments on a previous version of this paper and to Roger Crisp for comments on a very early version.

1 I am concerned in this paper solely with propositions that ascribe 'thin' moral properties to possible or actual actions, that is, properties normally expressed by a predicate consisting of the adverb 'morally' and a general evaluative predicate 'good', 'bad', 'obligatory', or 'wrong' or ones definable thereby and with propositions that ascribe them on grounds independent of any agent's ability to do the actions being evaluated. For example, I am concerned with actions that are good, even if there are no agents able to do them, and with obligations agents still have, even if they are unable to perform them and so are not culpable for not doing so. Thus, I do not presuppose that 'ought implies can'.

moral properties strongly supervene on nonmoral properties in Kim's (1993: 80) sense of 'strongly supervene'.[2] In this sense a property of kind A supervenes on a property of kind B iff it is metaphysically necessary that for every object that has a property F of kind A, there is some property G of kind B, such that any object that has property G has property F and has property F *because* it has property G. So the supervenience of the moral on the nonmoral would consist in a metaphysical necessity that any particular action or kind of action is morally good or bad, right or wrong, because of some nonmoral property (normally, a conjunction of such properties) that it has. Thus, plausibly what Hitler did in ordering the German invasion of Poland in September 1939 was morally wrong because it was an act of attacking a peaceful independent country; what Florence Nightingale did in organizing a hospital in the Crimea in 1854 was morally good because it was an act of caring for the sick.

No action can be just morally good or bad; it is good or bad because it has certain nonmoral properties—namely, ones of the kinds I have illustrated. And any other action that had just those nonmoral properties would also have the same moral properties. I shall call any set of the most general metaphysically necessary moral propositions stating which moral property supervenes on which nonmoral properties entailing all other such necessary propositions a set of the true fundamental moral 'principles'. The conjunction of nonmoral properties that gives rise to the moral property might be a long one or a short one; and the examples I provided of the nonmoral properties that conferred on the actions of Hitler or Florence Nightingale the moral properties may well be too short a list. It may be that all acts of attacking a peaceful independent country are bad, or it may be that all acts of invading a peaceful independent country which is not harboring groups of foreign fighters about to attack your country or (a list of different circumstances in which it would not be bad to attack such a country) are bad. But it must be that if there is a metaphysically possible world W in which an action *a* individuated by a conjunction of all its nonmoral properties is bad, there could not be another world W∗ that is exactly the same as W in all nonmoral respects, but where *a* was not bad.[3] Therefore, for every true contingent moral proposition (e.g., 'action *a* was morally good', *a* being individuated by some unique description) there is a metaphysically necessary moral proposition (e.g., 'all actions of nonmoral kind Z are morally good') that together with a contingent nonmoral proposition ('action *a* was an action of kind Z') makes the contingent moral proposition true.

2 The doctrine of supervenience can be stated in a more general form, so as not to entail moral realism, as by Blackburn (1993).

3 I take no stance on whether the only true propositions of the form 'it is metaphysically necessary that any action of kind A is morally good [or whatever]' are ones where 'A' specifies all the properties possessed by any action of that kind (apart from those concerned with some agent's ability to do it)—a conjunction so large that it could never be put into words—or whether (to my mind, far more plausibly) there are many true propositions of this kind where 'A' specifies only a small subset of such properties. Both kinds of proposition seem to deserve the name of 'principle'. My thesis concerns 'supervenience', not 'resultance' (in the sense distinguished by Dancy [2004: 38–52]). The thesis does nevertheless, I believe, have the consequence that propositions asserting which moral properties 'result from' which non-moral properties are also logically necessary.

https://doi.org/10.1017/apa.2015.6 Published online by Cambridge University Press

In its strongest form in which I shall consider the doctrine that there are necessary moral propositions of this kind, I assume that a proposition is metaphysically necessary iff it is either 'logically' (in a narrow sense) or 'conceptually' necessary or as strongly necessary as either of these. I shall group together the 'logically' and 'conceptually' necessary and call any proposition whose modal status as metaphysically necessary, impossible, or possible can be determined a priori *logically* necessary, impossible, or possible. However, with this understanding of metaphysical necessity it seems initially to many thinkers that it is implausible to suppose that there are metaphysically necessary moral propositions, and that is a major reason why moral realism has not seemed very plausible to a substantial number of thinkers. (See Mackie [1977: 41] for the classic statement of this reason, and Horgan and Timmons [1992] for their endorsement of this as a crucial objection to moral realism.) The purpose of this paper is to argue that it is very plausible to suppose that there are metaphysically necessary moral propositions and, in particular (in my wide sense), logically necessary ones. For the purposes of this paper I shall understand 'moral' goodness, badness, and the like as overall goodness, badness, and the like, but the results of this paper can be applied easily to narrower conceptions of the moral, for example, to the kind of goodness involved in obligations to other sentient creatures.

Some recent writers have tried to make it plausible to suppose that there are a posteriori metaphysically necessary moral propositions. The examples by which Kripke (1980) and Putnam (1975) persuaded us that there are a posteriori necessary propositions such as 'Hesperus is Phosphorus' and 'water is H_2O' seem to exhibit a common pattern. An object (substance, property, or whatever) is picked out by a rigid designator, but our understanding of that particular rigid designator is that it designates an object that is the object it is in virtue of an underlying essence that it needs a posteriori work to discover. Having that essence, it could not fail to have that essence; therefore, any sentences stating that it has that essence or stating something entailed by it having that essence are necessary with as hard a necessity as logical necessity. Thus, when 'water' is understood—as Putnam supposed it was understood in the late eighteenth century—as a rigid designator of the transparent drinkable liquid in our rivers and seas, what makes the stuff that stuff is its chemical essence, namely, being H_2O. Having that essence, it could not not have that essence. But it needs a posteriori investigation to discover what that essence is. Thus, the suggestion is that moral principles might be like this.

Thus Russ Shafer-Landau (2003: 87–98) compares the relation of the moral to the natural (or physical) to the relation of 'chemical' to 'atomic' facts. David Brink (1989: 179) compares 'moral facts are constituted by natural facts' to 'tables are constituted by certain arrangements of microphysical particles'. Such examples are just like the 'water is H_2O' example. Both writers seek to bolster their case by comparing moral supervenience to the a posteriori metaphysical supervenience alleged by physicalism of the mental on the physical. But even if physicalism were true, that would show only that there is a large class of a posteriori necessary truths, not that moral principles belong to that class. That needs to be shown independently. Michael Smith (1994: 190–92) claims that just as we can know it is a necessary truth that the property A of a surface which actually (in our world)

https://doi.org/10.1017/apa.2015.6 Published online by Cambridge University Press

causes objects to look red does so, so there are necessary truths about the natural properties on which moral properties supervene. Strangely, he recognizes that the supervenience of redness on the reflecting properties of surfaces is a posteriori, whereas the supervenience of the moral on the non-moral is 'knowable a priori'; but he seems to think that this 'point of disanalogy' is relatively unimportant.

Yet, while clearly a posteriori investigation is needed to discover the truth of contingent moral truths, all arguments about the underlying moral principles rely on a priori considerations. When examples of particular situations (e.g., the trolley problem) are adduced in order to persuade us that some general moral principle is or is not true, it is quite irrelevant whether the examples are examples of an actual event or of an imagined event. What matters is what it would be right to conclude about which actions in that situation would be good or bad; whether or not the situation actually occurred is irrelevant. Accordingly, any metaphysically necessary moral principles must be logically necessary. Hence, either there are logically necessary moral principles, or moral realism is false.

Those who claim that fundamental moral principles are (in my wide sense) logically necessary sometimes claim that their necessity is a different kind of necessity from logical necessity (in a narrow sense) or conceptual necessity. Some writers have argued that the necessity of moral propositions is like that of fundamental epistemic propositions, which in effect (since in effect they claim that these propositions are true in all logically possible worlds) they claim to be a special kind of logical necessity (in my wide sense). Parfit (2011: 524–25) compares other normative propositions to 'normative truths about credibilities and epistemic reasons'. Wedgwood (2007: 153–73) argues that the possession of 'intentional' concepts and in particular the concept of belief entails a view about what are the rational procedures for acquiring and using them, and that rationality is a normative concept. Cuneo (2007) has written a whole book arguing that the supervenience of epistemic propositions on nonepistemic ones is just like the supervenience of moral propositions on nonmoral ones. By 'epistemic' propositions these authors have in mind propositions claiming that some belief is 'rational' or 'justified' or 'probably true', and by 'epistemic' principles they mean principles that determine when this is the case. But there seems to be a sharp distinction between the supervenience of moral propositions and the supervenience of epistemic propositions, in that the normative force of the supervenient epistemic proposition is hypothetical—for example, 'this belief is justified' entails 'if you want a true belief, you should hold this one'—whereas the normative force of the supervenient moral proposition is categorical—for example, 'this action is obligatory' entails 'whatever you want, you ought to do this'. For this reason it would be better not to argue for the logical necessity of moral principles on the basis of their similarity to epistemic principles.

Fine (2002) has taken the heroic stance of claiming that the necessity of moral principles, which he calls 'normative necessity', is sui generis, quite distinct from metaphysical necessity or natural necessity (the necessity possessed by laws of nature), but that leaves it very obscure how we can have access to such necessity, something much less obscure in the other two cases. I shall argue in this paper for the logical necessity of moral principles on the grounds of the similarity of ways

https://doi.org/10.1017/apa.2015.6 Published online by Cambridge University Press

in which we establish them to ways in which we establish obvious logical (in the narrow sense) and conceptual necessities. (Parfit [2011: 524–25] also compares normative truths to 'the necessary truths of logic and mathematics'). In order to do this, I need to devote the next section to putting forward a view of what it is for some proposition to be logically necessary and of how we can show that probably it is logically necessary. It is a view resulting from conceivability accounts of logical possibility (such as Yablo [1993] and Chalmers [2002]), which I suggest are not very different from 'understanding-based' accounts).

2.

Following Chalmers (2002: 147), I shall understand conceivability as 'a property of statements', that is, of token declarative sentences uttered (or whatever) on a particular occasion. Sentences express propositions, and the propositions are the propositions they are in virtue of the rules of the language governing the meanings of sentences. Sentences contain words; words express concepts that are contained in propositions. The rules of a language governing the meanings of words (and so of the sentences containing them) are of two kinds: syntactic and semantic. Syntactic rules relate a word to other words, describing what they entail and what is entailed by them, the kind of rules you find in a dictionary, such as 'a philatelist is someone who collects postage stamps or who knows about different kinds of postage stamps'. Semantic rules relate a word to objects or properties to which it paradigmatically applies, such as 'red is the color of ripe strawberries, British post boxes, and London buses'. By 'paradigm examples' of the application of a word or sentence I mean examples such that it would be inconceivable that that word did not apply to almost all of those examples (while allowing that it might not apply to a few of them). Some words have only syntactic rules for their use, for example, words introduced by arbitrary definition. But plausibly, unless the words to which a given word is related are themselves logically related to other words and they to other words and so on until we come to words that have semantic rules for their use, we wouldn't know what that given word meant. At any rate, the given word could in that case not designate a (metaphysically) contingent property possessed by some object or property. While different syntactic rules for a word entail that it expresses a different concept, different semantic rules need not have that consequence. For clearly two words can mean the same even if they are introduced by different paradigm examples of their correct application—for example, 'red' can be given a meaning by quite different sets of examples of red objects. But two words cannot mean the same if they don't apply to the same (paradigm or nonparadigm) objects, although it might need much investigation to show whether or not that is the case.

I shall assume for the next few paragraphs that speakers of a language agree about what would make a sentence true or false in any narrow set of circumstances to which it has been applied in the past or would be applied in future, and so in that sense they have a common understanding of the meanings of words and sentences. I shall call a sentence that expresses a logically necessary (possible/impossible) proposition, a logically necessary

https://doi.org/10.1017/apa.2015.6 Published online by Cambridge University Press

(possible/impossible) sentence. I shall equate being logically (im)possible with being 'ideally' '(in)conceivable', that is, 'conceivable on ideal rational reflection' (Chalmers 2002: 147); and I shall assume that a sentence being 'apparently' (in)conceivable ('prima facie' (in)conceivable) is, in the absence of counterreason, good reason to suppose that it is (in)conceivable and that a sentence being 'obviously' (in)conceivable is very strong reason to suppose that it is (in)conceivable. I shall however understand 'conceivable' as 'such that it makes sense to suppose that it is true', a sense that may be a wider sense than that espoused by Chalmers.[4]

Then we can come to see by a priori reflection on the public usage of the words and sentence forms involved that there are some declarative sentences that could not be true whatever the world was like, because it is obviously 'inconceivable' that there is any state of affairs that would constitute those sentences being true. The paradigm example of such a sentence is a self-contradictory sentence (e.g., one of the form 'both *p* and not-*p*'). But there are innumerable other examples of declarative sentences that are not explicitly self-contradictory but are also such that it is obviously inconceivable that there is any state of affairs that would constitute them being true. These include what used to be called 'category mistakes'—for example, 'the Prime Minister is a prime number' or 'I had two helpings of democracy for breakfast'. We can see too that there are some declarative sentences such that it is obviously inconceivable that there is any state of affairs that would constitute their being false, for example, the negations of self-contradictory sentences, and so these sentences must be true and hence logically necessary. We also come to see to which other sentences a speaker who utters a given sentence is explicitly committed in virtue of the rules of the language. If it is obviously inconceivable that (*p* and not-*q*), then in uttering *p*, a speaker is explicitly committed to *q*, and in that case (see Swinburne 2013:18) I shall say that *p* 'mini-entails' *q*. We can see also that various sentences other than logically necessary ones are obviously not logically impossible and so are logically possible.

A priori reflection can then lead us to discover the logical modality of other sentences that was not initially obvious. The direct way of discovering these things is by discovering entailments between sentences whose logical status is obvious. For example, *p* entails *q* iff (*p* and not-*q*) is logically impossible. One way of discovering that *p* entails *q* is to discover that *p* and *q* can be joined by a chain of mini-entailments, that is *p* mini-entails *s*, *s* mini-entails *t*, and so on until we reach a sentence that mini-entails *q*. We can discover that a sentence is logically impossible by discovering that it entails an obviously logically impossible sentence. We can discover that a sentence is logically necessary by discovering that its negation is logically impossible. And we can discover that a sentence is logically possible by discovering that it is entailed by some obviously logically possible sentence.

4 Thus Chalmers (2002: 153): 'A situation is [modally] coherently imagined where it is possible to fill in arbitrary details in the imagined situation such that no contradiction reveals itself. . . . I will say that S is positively conceivable when it is coherently modally imaginable'. Although sentences containing category mistakes and other sentences which we can determine to be false solely a priori, do, I believe, 'reveal', that is 'entail' a contradiction, my wider definition of conceivability does not assume this.

https://doi.org/10.1017/apa.2015.6 Published online by Cambridge University Press

While these methods will enable all investigators who make the same judgments about what is apparently conceivable and apparently inconceivable, to determine the logical status of many sentences (and so of the propositions they express), there will clearly be many sentences whose status they cannot determine straight off, for example, ones where they cannot immediately find an obviously conceivable sentence that entails the sentence in question or a route of proof by which they can deduce a self- contradiction from that sentence. In that case there is an indirect method that may help investigators to discover the logical modality (as defined above) of some propositions—the method that (in the context of discovering moral truths) Rawls (1999: 18) called 'reflective equilibrium'. Although Rawls may seem not to have understood his method in quite this way (but see the last paragraph of this paper), it is a method that enables us to discover some general principle governing the correct use of some type of sentence. The method assumes that the simplest account of the use of sentences of that type in various narrowly described sets of circumstances is most probably the account describing how such sentences would most probably be used in all circumstances; this account, then, most probably shows the logically necessary and/or sufficient conditions for the truth of a sentence of that type. Such an account is simple insofar as it uses few predicates designating properties easily recognizable in many different kinds of paradigm example, a condition I will illustrate with examples.

The application of this method to discover the logically necessary and/or sufficient conditions for the truth of a sentence of some type has been a prominent feature of 'analytic philosophy'. Consider the procedures whereby philosophers have tried to set out the logically necessary and sufficient conditions of 'S remembers having done X'. A first suggestion by Hume (T 1.1. 3 [1888]) was (in effect) that this was equivalent to 'S has a "lively" mental image of having done X which "preserves the original form" of S's awareness of having done X'. But then, it was objected, innumerable examples of usage showed that we allow that someone might have 'remembered' having done some action, even if he or she did not have a mental image of having done it. So it was suggested that almost all examples of usage showed that in order to 'remember' having done X, someone needs merely a true belief that he or she had done X. But then philosophers pointed out that we would not count a person S as having 'remembered' having done X merely because S had acquired the belief that she had done X from having read in a book that she had done X. So it was suggested that for S to 'remember' having done X, S's belief that she had done X must have been caused by S having done X. Then it was pointed out that even if S having done X caused her to write that she had done X in her diary or some letter that she had done X, and her reading this in the diary or letter caused S to believe that she had done X, we still 'would not say' that she 'remembered' having done X. Only if the route of causation went directly through S's body (in effect, her brain) and only if the belief was a basic belief not inferred from anything else would the resulting belief count as a memory. (This is a rough summary of the conclusion of Martin and Deutscher [1966]). As a result, an account of memory began to emerge along the lines of 'S remembers having done X iff S has a true basic belief that he did X, caused by a chain of causes in S's body, itself initiated by S having done X'. This account was supposed to be

https://doi.org/10.1017/apa.2015.6 Published online by Cambridge University Press

logically necessary and could be used to show sentences about personal memory to have various logical statuses—for example, that 'if S remembers having done X, then S has a true belief that he did X' is logically necessary. Similar processes led to accounts of the logically necessary and/or sufficient conditions of 'S knows that p', 'S perceived O', and the like.

In each case a simple extrapolation from a few examples was seen as giving a probably correct account of some logically necessary or sufficient condition for the truth of some type of sentence until investigators came across cases where a sentence of that type was clearly not being used in a way that satisfied that condition, and therefore they sought a different account. In seeking such an account, no one suggested that the only exceptions to a previous account were the exact narrowly described examples that did not satisfy the account. No one suggested that the only exception to the principle that for S to remember having done X his belief that he had done X must have been caused by having done X 'unless the causal route proceeded through S's diary or a letter written by S'. Why no one suggested that was because it would have been far too ad hoc and so too complicated a principle. Rather, philosophers tried to discover a new account using few predicates designating properties easily recognizable in many different kinds of paradigm example (such as 'directly through her body') that fitted both the old and the new examples.

However, an account of the logically necessary or sufficient conditions for the correct use of a word or sentence form, which captures almost all our usage but would need to become very much more complicated in order to capture one particular usage of that word or sentence form, would still be recognized as the correct account of the normal meaning of the word or sentence form, and the odd usage would then be considered one that uses the word or sentence form in a different sense from the normal sense. Thus S may sometimes be said to 'remember' or 'know' something that didn't happen just because S claims to remember or know it in the normal sense, and then S is said to be using these words in an 'inverted comma sense'.

The simplicity of a general principle underlying the paradigm examples of the application of a word may be a matter not of the fewness of the predicates involved (as in the previous examples) but of the 'easily recognizable' character of a property expressed by a single predicate. Suppose that the only red objects language users have seen are ripe strawberries, ripe plums, London buses, rubies, holly berries, and pink roses and suppose they are taught that these are paradigm examples of 'red' objects. Then, given that they can easily recognize the similarities of color in each of these different paradigm examples, they would see (as the simplest account of its meaning) the logically necessary general principle governing the use of 'red' as 'an object is red iff it is similar in color to actual ripe strawberries, ripe plums, London buses, rubies, holly berries, and pink roses). This in turn will lead these language users to conclude, when they come to observe what we would call different shades of red and name them 'scarlet' or 'crimson', that such sentences as 'if it is scarlet, it is red' are logically necessary and so express a logically necessary proposition. This example illustrates the point that the ability of language users to recognize necessary truths depends on the kinds of similarities between phenomena to which

https://doi.org/10.1017/apa.2015.6 Published online by Cambridge University Press

they are sensitive. If some language users could not recognize any similarities of color between the paradigm objects, they would conclude that there are six different unconnected senses of 'red'. From that it would follow that the sentence 'if it is scarlet, it is red' is logically impossible and so expresses a different proposition from the one expressed by the previous group of language users.

The third aspect of my definition of the simplicity of an account of the logically necessary and sufficient conditions for the correct use of a word or sentence form is that it uses predicates that designate properties recognizable in *many* different *kinds* of paradigm example. If language users were given only two kinds of paradigm example of red objects—for example, rubies and ripe plums—it would not be obvious whether only (what we call) dark red objects count as 'red' or whether all of (what we call) red objects count as 'red'. Neither extrapolation from the paradigm examples would be especially simple. Likewise, if we had only a few paradigm examples of 'S remembers having done X' or 'S knows that p', there would be many different but not especially simple ways of extrapolating from the examples to a general logically necessary principle about knowledge. The evident simplicity of one account becomes apparent only when the semantic rules provide many different kinds of paradigm example of correct application.

So far I have been assuming that all speakers of a language would regard the same examples of words and sentences as paradigm examples of their correct application, that is, as ones that it would be inconceivable to suppose that they did not apply to almost all of those examples. But of course most speakers will have been taught the meanings of words and sentences by different paradigm examples and sometimes by very different kinds of example. That may or may not lead them to acquire common concepts, expressed by intertranslatable words, depending on whether the concept some speakers derive from one set of paradigm examples is one they recognize as applying to the examples from which other speakers have derived their concept, that is, if reflective equilibrium leads to agreement about the different examples.

Suppose that there are two communities isolated from each other. One community lives in an environment where every object is of a different shade or mixture of blue or green: light blue, dark blue, blue-green, light green, dark green, etc. The community recognizes that there is something common to those different shades and they call it being 'colorato'. The other community lives in an environment where every object is of a different shade or mixture of red or green: light red, dark red, red-green, light green, dark green, etc. People in this community recognize that there is something in common to those different shades, and they call it 'coloré'. Suppose now that the two communities become acquainted with each other and their environments. Then it may happen that each community would come to recognize that what their objects of different shades have in common is the same kind of thing as what the objects of different shades of the other community have in common, which is the kind of thing we call being 'colored'. So the community members would realize that their understanding of their general term ('coloré' or 'colorato') is that it applies not merely to objects of shades in the environment of the other community already known to them, but also to ones not previously known to them. If that is what happens, the community would

https://doi.org/10.1017/apa.2015.6 Published online by Cambridge University Press

be showing that there was implicit in their previous understanding of 'coloré' or 'colorato' a natural fit to wider examples. If the first community came to recognize this, it would also come to recognize that the sentence 'if it's light red (or whatever they came to call that colour), it is colorato' is logically necessary and so expressed a logically necessary proposition. If the second community came to recognize this, it would come to recognize that 'if it is dark blue, it's coloré' expressed a logically necessary proposition. Then it would have turned out that 'colorato' and 'coloré' mean the same. Yet the processes by which members of the two communities originally learnt the concepts of 'colorato' and 'coloré' would not have included teaching them these logically necessary truths.

Alternatively, it might happen that one or both communities did not come to apply their words 'coloré' or 'colorato' to the new shades. This would show that implicit in their previous understanding of those words was a much narrower understanding of them than that of 'colored' (in our sense). In that case they would have to claim that sentences such as 'if it's light red, it is colorato' expressed not logically necessary but logically impossible propositions. Which outcome would result would depend on the kinds of sensitivity to color the teaching process has given them. Clearly, the greater the overlap between the paradigm examples of 'colorato' and 'coloré' objects, the more likely it is that each group would come to see these words as meaning the same.

In order to discover the meanings of color words and so the logical relations of color concepts to each other and, indeed, the meanings of any words that have semantic rules for their application and so for the concepts they express, communities will need in practice to have observed actual examples of differently colored objects (or whatever). But the conclusion they reach by reflecting on the concepts they acquire is logically independent of whether there are any objects having these colors (or whatever), and of whether the communities know this. Accordingly, like the conclusion about the nature of remembering, this conclusion is an a priori conclusion. (See Bonjour [1999: 35]: 'it is no objection to a claim of a priori justificatory status for a particular belief that experience is required for the acquisition of some of the constituent concepts'). The moral of this story is that consideration of new examples, actual or imaginary, can get us to recognize previously unrecognized modal truths about old words used in the same sense and so expressing the same concepts as before.

3.

It is well known that most attempts to show the truth of moral principles by the direct methods of deriving logically necessary truths, described in the previous section, are seldom successful for the kinds of reason that some attempts to derive logically necessary truths of other kinds also fail—for example, because disputants do not agree about what mini-entails what or when some sentence is obviously inconceivable. However, many philosophers recognize the method of 'reflective equilibrium' as a method of helping us to discover (probably) true moral principles. For example, Scanlon (2002:149) regards reflective equilibrium as 'the

https://doi.org/10.1017/apa.2015.6 Published online by Cambridge University Press

only defensible method of making up one's mind about moral matters and about many other subjects'. And he sees this as a method of getting us to recognize the correct principles and correct particular judgments: 'the fact that a given judgment does not fit with principles that account for most of our other judgments can lead us to change our mind about that judgment itself, and we may also be led to change our mind when we see that the only principles that *do* account for a given judgment are ones that are seen in other ways to be clearly mistaken' (148). I now illustrate this and show that the reason why this method works for moral principles is the same kind of reason as the reason why it works to show the necessary truth of nonmoral sentences. That makes it very plausible to suppose that the correct moral principles are themselves logically necessary truths.

Most children, although brought up in very different cultures, quickly acquire a concept of a kind of action it is important to do and of a kind of action it is important not to do, and they learn that some of the former are more important to do than others and that some of the latter are more important not to do than others. Children are praised and sometimes rewarded for doing actions of the former kind, and they are blamed and sometimes punished for doing actions of the latter kind. They thus acquire a concept of 'moral goodness' they recognize as applicable to all actions of the former kind and a concept of 'moral badness' they recognize as applicable to all actions of the latter kind These concepts are such that to have a belief that some action is morally 'good' (or 'bad') entails having some motivation to do (or not to do) the action. But children don't understand these concepts of moral goodness or badness as concepts of properties connected merely contingently with the concepts of the particular narrowly described kinds of action that exemplify them. Rather, they understand the former concepts as determinable concepts of which the determinates are the different kinds of morally good or bad action, just as 'colored' is a determinable of which red, green, and so on are determinates. So they would regard it as inconceivable to suppose that almost all of the kinds of action to which they originally applied the concept of moral goodness (or badness) are really not morally good (or bad). Someone who does allow that possibility for a concept has not got the concept that most children acquire and that I am discussing here. For there would be then no semantic criteria for the application to actions of 'morally good' (or whatever). And some of the syntactic criteria, illustrated above, that a 'morally good' action is one that is 'important' to do, and 'gets praised' can themselves only be understood if we understand 'important' as 'morally important', 'praised' as 'recognized as morally good', and so on. These moral terms can only be understood in terms of each other unless they are linked to paradigm examples illustrating the kind of importance that makes an action morally good, the kind of praise that recognizes a morally virtuous action, and so on. Otherwise 'importance' could be prudential importance, 'praise' could be praise for doing what the praiser likes to see done, and so on. While understanding moral terms *merely* in terms of action-guiding or similar notions is incompatible with moral realism, without the tie to action-guiding having a moral belief would just be *merely* having a belief that there are similarities of certain recognizable kinds between examples of morally good (bad) actions—and this constitutes an implausible moral naturalism.

https://doi.org/10.1017/apa.2015.6 Published online by Cambridge University Press

Further, the concepts of moral goodness, badness, and so on most children acquire are surely not ones they regard as applying merely to actions of narrow kinds, similar in almost all respects to those from which they originally learnt to apply them. At least as they get older, children are open to reflection on what they have been taught, to argument with members of other cultures, and to experiences of new situations, leading them to extend or retract the application of the concepts. Clearly, use of reflective equilibrium makes it possible for two groups who largely agree with each other about moral matters but have a few disagreements to reach agreement about disputed matters. They may both recognize a simple principle underlying some set of examples both groups recognize as paradigm cases of 'morally good' actions for which praise is appropriate. This has the (probably true) consequence that one kind of action that only one group recognizes as a paradigm case of a 'morally good' action is not like the others; rather, it is a kind of action for which praise is not appropriate, and so it is not 'morally good'.

Here is an example where the simplicity of a principle depends on the fewness of the predicates designating properties it lists, as with analyses of 'remember' and 'know'. Suppose that both groups have learnt that it is wrong for a person A to kill a person B unless B is about to kill A or some member of their tribe or unless A is acting on behalf of a court of law that has sentenced B to death for murdering a member of the tribe or unless B is an enemy combatant in war against their tribe. But one group, which I'll call the first group, unlike the second group, may also have been taught that it is not wrong for A to kill B in a duel with B if B has insulted A or a relative of A. Then the second group may point out to the first group that all the examples except the disputed one are examples of situations in which B has killed someone close to A or is about to do so. The second group then suggests that a simple principle underlying all the agreed examples is 'it is wrong to kill except as punishment for a killing or to prevent further killing (of a member of one's tribe)'. This simple principle treats life as so sacred that it must not be taken away except in compensation for life or to prevent further loss of life. The consequence of this principle is that it is wrong to kill in a duel merely to defend one's family's honor. The first group may then recognize this principle, and so move beyond what its members were taught when young, namely, that killing in a duel is not wrong. As with 'know' and 'remember' in the example above, investigators come to see a previous paradigm example as not an example of the same concept as the other paradigm examples. Of course, reflective equilibrium may eventually lead both groups to develop an even simpler principle: 'it is wrong to kill except to prevent further killing.'

Moral disagreement may turn not on whether a suggested principle uses few predicates, but—as with 'red'—on whether the predicates it uses designate properties easily recognizable in many paradigm examples. I take as an example of this a case where there is moral disagreement between two groups because although both groups agree that the moral goodness or badness of some action depends on certain properties it has, and they also agree about which of these properties make for the overall goodness of the action and which make for its badness, they disagree about which group of properties outweighs the other group. Each disputant group agrees that the considerations adduced by the opponent group

https://doi.org/10.1017/apa.2015.6 Published online by Cambridge University Press

have some force; that is, each group accepts that, for example, the other group's considerations would show the action to be morally good overall but for the other considerations it adduces against its moral goodness. Thus, a group opposing the euthanasia involved in helping a depressed person (who has no close friends or family) to commit suicide may argue that such an act is overall bad because of the sanctity of human life, the possibility of helping a depressed person to recover from his or her depression, the value of his or her overcoming that depression, and so on. A group advocating euthanasia may argue that helping a depressed person to commit suicide is helping that person to do what he or she clearly and firmly wants to do and what hurts no one else in any way. Both disputant groups appeal to considerations that the other group may admit to have some weight, but each group holds that the considerations it adduces outweigh those the other group presents. The consequence of this kind of disagreement is that for members of one group to persuade members of the other group to change their view will only require the members of the latter group on the basis of already agreed examples, to give a bit more weight to a feature they already admit to have some weight; one group does not need to persuade the other group to admit the relevance of some new good-making or bad-making feature.

Thus, the group opposing euthanasia might persuade the other group that the simplest principle underlying their many shared beliefs about its being good to help the sick, the unemployed, the unlovable, and so on is that it is always good not to 'give up' on helping people satisfy their basic *needs*. Hence, that first group might persuade the group supporting euthanasia that this principle is probably true and so has the consequence that one should not give up on helping the depressed to satisfy their basic need for happiness. Consequently, one should help them recover from their depression and not help anyone commit suicide. Alternatively, the group supporting euthanasia might persuade the group opposing it that the simplest principle underlying the many beliefs shared with the other group about it being always good to help those who want to get an education, to have children, or to travel abroad is that it is good to help people do what they firmly *want* to do when that hurts no one else. Accordingly, that group might persuade the other group that this applies to helping people commit suicide.

The issue may then turn on how easily recognizable the properties are that are designated in rival principles. It is often very obvious what a person 'wants' (that is, desires), but what a person 'needs' may often be very unclear; the latter depends on a view about what constitutes human flourishing, and even its defenders may be unclear about the application of this view. The issue will also depend on how many different kinds of paradigm example the rival principles fit. If there are a vast number of different paradigm examples where, disputants admit, it is good to help people to do what they want and relatively few paradigm examples where it is good to help people satisfy their basic needs—especially if there is a paradigm example of the former where 'wants' trump 'needs'—that shows that 'wants always trump needs' is the simpler moral principle. This parallels the 'colorato' case, where whether a red-green object is deemed to be 'colorato' will depend on how easy it is to recognize similarities between red-green objects and the previous paradigm examples of 'colorato' and on how large and varied that stock of previous examples

https://doi.org/10.1017/apa.2015.6 Published online by Cambridge University Press

is. The more varied the previous paradigm examples are, and the easier it is to recognize the similarities between them and the new examples, the more evident it will be that the latter are 'colorato'. In the moral case, as in the color case, a common kind of sensitivity to paradigm examples, whether color sensitivity or moral sensitivity, is required to recognize the relevant similarities and so to reach agreement, even if in the end the agreement is that the two principles are equally simple and that therefore when they clash, it is equally good to give priority to wants as to give priority to needs.

Clearly, just as the more two groups share common paradigm examples of 'colorato' and 'coloré', the more probable it is that use of the method of reflective equilibrium will lead to agreement on the logical truths about color. Likewise, the more shared examples of 'moral goodness' two groups have, the more probable it is that use of the method of reflective equilibrium will lead to agreement on moral principles. As with the color example, it is often in practice necessary for someone to have had personal experience of some situation in order to understand fully what is meant by someone doing some kind of action or having some kind of experience. One might need to experience acute depression oneself in order to understand that it is a necessary moral truth that it is good to cure acute depression by means of drugs that will not cause pain, delusion, or mental impairment. But again, if this is a moral truth, it is an a priori one because it is irrelevant to its status as a truth whether anyone ever does suffer from acute depression.

But surely, an objector may say, there is a big difference in the way people react to attempts to get them to change their views under the pressure of arguments appealing to reflective equilibrium when the arguments concern moral principles and the way they react when the arguments concern ordinary (that is, nonmoral) conceptual truths. The difference is that many people do not change their views about moral principles readily, whereas they do change their views about ordinary conceptual matters readily. And that suggests that there must be an important difference between the logical nature of conceptual truths and moral principles. There is, however, a ready explanation of the former difference that has nothing to do with the logical character of the propositions but has everything to do with the nature of humans, which makes it difficult for us to change our moral views in the light of good arguments. Our moral views are closely connected to our emotions, some of which keep us from recognizing moral views that we would otherwise naturally adopt. The natural emotion of anger at those believed to have molested children may lead us not to recognize their right to a fair trial, and so we may not adopt the principle that everyone is entitled to a fair trial. Our moral views are also closely connected with rival desires, some of which are desires to do actions that, if we accepted certain arguments, we would have to recognize as morally wrong and so would find ourselves under internal moral pressure not to do. If I conclude that there is a true moral principle that everyone ought to donate much of their income above a certain amount to providing food and shelter to those in distant countries, I may have to conclude that I ought to give a lot even when I desire not to do so. So again I refuse to acknowledge a strong argument in favor of there being such a moral principle. Humans are only partly rational creatures. It is for this reason that questions about whether some suggested moral principle is a true

https://doi.org/10.1017/apa.2015.6 Published online by Cambridge University Press

moral principle remain 'open questions' for a long time. Thus, there is no need to attribute the 'openness' of such questions to a difference between the logical nature of conceptual truths and moral principles.

But even if groups who have fairly similar moral views can resolve their differences by reflective equilibrium, what reason do we have for supposing that groups whose views are much further apart from each other can do so? My response is (tentatively) to put forward a contingent hypothesis that most twenty-first-century groups have enough moral views in common to make agreement between them possible, given enough time and willingness to understand each other. Almost all groups agree that it is good (though not always obligatory) to feed and otherwise care for their own (nondisabled) children and (mentally competent) parents and also some other members of their own society who need food or care, to keep their promises (not made under duress) except when very difficult to do so, not to steal what by community agreement belongs to another member of the society, and so on. Not merely are there some paradigm examples of moral goodness or badness on which almost all agree, but I suggest there is a chain between any groups of persons diametrically opposed on many moral issues of other groups who agree with one extreme group on most issues, different groups who agree with the semi-extreme group on most issues, and so on until we come to the group at the other extreme.

When reflective equilibrium has moved someone away from an extreme position to a slightly less extreme position, that person will regard many of the new examples of morally good and bad actions he or she comes to recognize as examples of the relevant concept just as satisfactory as the ones recognized previously and so as paradigm examples of such actions. And then that person will have much more in common by way of overlapping sets of paradigm examples with some others with whom he or she previously had less in common, and this will provide a basis for reaching agreement that did not exist previously. Accordingly, it is possible for some person to move by a rational process consisting of many such steps from one general moral outlook to a quite different one, say, from a narrow tribal understanding of which actions are overall good to a far wider one, while continuing to hold most of the views about the goodness of helping members of his or her own tribe, which provided that person's original paradigm examples of good actions. My ground for putting forward my contingent hypothesis is that over many centuries the human race has made much progress in coming to share moral views. No longer do most humans regard slavery or dueling, much less human sacrifice or suttee, as morally permissible. And that suggests that most of us do share a common sensitivity to what paradigm examples of 'morally good' (or whatever) actions have in common. It would follow that we share a common concept of morality, the logically necessary truths of which can be discovered by a priori reflection, although it may take a long time for all of us to agree on what these true fundamental moral principles are. One writer who emphasizes that although cultures start from different points, new experiences and the operation of reflective equilibrium lead to moral progress is Richard Boyd (1988). But Boyd seems to hold that in this process our concept of moral goodness itself changes, in the way that scientific concepts (such as the concept of a biological kind) change. The view I have been advocating is that most of us have a common concept of moral goodness and that progress

https://doi.org/10.1017/apa.2015.6 Published online by Cambridge University Press

normally consists in seeing that its extension is different from what we had believed hitherto.

Of course, it may sometimes be the case that those to whom new examples are presented and apparently simpler principles of morality are suggested show no tendency at all (even after prolonged exposure) to move toward agreement with most of us. That would suggest that these people do not have the same concept of 'morality', the same understanding of what it is for an action to be 'morally' good, bad, or whatever, and not merely that they do not have the same views as most of us about which actions are morally good. This is much like the conclusion we would reach about those exposed to new colors who did not recognize them as 'colorato'. The process of acquiring concepts of 'morally' good, bad, and so on would have produced in those not influenced in the ways illustrated by moral arguments a different concept of 'morality' from what it has produced in most of us. However, for the reasons given above, I do not think that that is often the case. And how little tendency to change 'moral' views after how much exposure to the 'moral' views of others suffices to show that someone has a different concept of morality? 'Having the same concept of 'morality' is, like most linguistic expressions, vague, and so there are borderline cases for its application. But there are plenty of definite cases of people who used to disagree about 'moral' matters, yet are shown by their subsequent agreement resulting from reflective equilibrium to have had the same concept of 'morality', a concept possessed, I have suggested, by most of the human race. I have sought to show that—given moral realism and so given that there are true moral principles—the similarity of procedures by which we establish what are the true moral principles to those by which we establish what are the true nonmoral conceptual principles makes it very plausible to suppose that both kinds of principle are conceptually and so (in my sense) logically necessary. This result alone, given the supervenience of the moral on the nonmoral, is needed to make moral realism plausible.

The unwillingness of many moral realists to recognize that there can be logically necessary moral principles seems to me to arise from failure to recognize the importance of semantic rules as well as syntactic rules for the use of words in determining the logical status of all sentences and thus also of the propositions they express. This is a failure to recognize that the application of semantic rules requires a certain kind of sensitivity to similarities between the paradigm examples. There is no difference between nonmoral conceptual necessities and moral principles in these respects. 'Anything red is colored' is surely logically necessary, but only because the meanings of both 'red' and 'colored' are fixed largely by paradigm examples, and language users recognize the same similarity between the paradigm examples of red things to which they are introduced, and the same similarity between the paradigm examples of colored things to which they are introduced, and recognize the former kind of similarity as a species of the latter. And since, in general, speakers of languages other than English recognize the same color distinctions as do English speakers, there are natural translations of 'red' and 'colored' into other people's language that have the consequence that a sentence of their language means the same as, and expresses the same logically necessary proposition as, 'anything red is colored'. So too, I have been suggesting, logically necessary moral sentences and the

https://doi.org/10.1017/apa.2015.6 Published online by Cambridge University Press

propositions they express are logically necessary because the meanings of 'morally good', 'morally obligatory', and so on are fixed partly by semantic rules, that is, by paradigm examples, to the similarities between which most of us have a similar sensitivity. Rawls may have recognized this when he wrote (1999: 44–45): 'It is obviously impossible to develop a substantive theory of justice founded solely on truths of logic and definition. The analysis of moral concepts and the a priori, however traditionally understood, is too slender a basis'. In other words, syntactic rules are not enough. However, he continued, 'If we can find an accurate account of our moral conceptions', which I take to mean that if we take account of the moral judgments we actually make, 'these questions of meaning and justification may prove easier to answer'. As I put it, semantic rules fill out what is meant by moral concepts and thereby (by the method of reflective equilibrium) enable us to determine their applicability to actions of different kinds.

RICHARD SWINBURNE
UNIVERSITY OF OXFORD
richard.swinburne@oriel.ox.ac.uk

References

Blackburn, S. (1993) 'Supervenience Revisited'. In *Essays in Quasi-Realism* (Oxford: Oxford University Press), 130–48.
Bonjour, L. (1999) 'Article on "A priori"'. In R. Audi (ed.), *The Cambridge Dictionary of Philosophy*, 2d ed. (Cambridge, UK: Cambridge University Press), 35.
Boyd, Richard. (1988) 'How to be a Moral Realist'. In G. Sayre-McCord (ed.), *Essays on Moral Realism* (Ithaca, NY: Cornell University Press), 181–228.
Brink, David. (1989) *Moral Realism and the Foundations of Ethics*. Cambridge, UK: Cambridge University Press.
Chalmers, David. (2002) 'Does Conceivability Entail Possibility?' In T. S. Gendler and J. Hawthorne (eds.), *Conceivability and Possibility* (Oxford: Oxford University Press), 145–200.
Cuneo, Terence. (2007) *The Normative Web*. Oxford: Oxford University Press.
Dancy, Jonathan. (2004) *Ethics Without Principles*. Oxford: Oxford University Press.
Fine, Kit. (2002) 'The Varieties of Necessity'. In T. S. Gendler and J. Hawthorne (eds.), *Conceivability and Possibility* (Oxford: Oxford University Press), 253–81.
Horgan, Terence, and Mark Timmons. (1992) 'Troubles on Moral Twin Earth: Moral Queerness Revised'. *Synthese*, 92, 221–60.
Hume, David. (1888) *A Treatise of Human Nature*. Edited by L. A. Selby-Bigge. Oxford: Oxford University Press.
Kim, Jaegwon. (1993) '"Strong" and "Global" Supervenience Revisited'. In *Supervenience and Mind* (Cambridge, UK: Cambridge University Press), 79–91.
Kripke, Saul. (1980) *Naming and Necessity*. Oxford: Blackwell.
Mackie, J. L. (1977) *Ethics: Inventing Right and Wrong*. Harmondsworth: Penguin.
Martin, C. B., and Max Deutscher. (1966) 'Remembering'. *Philosophical Review*, 75, 161–96.
Parfit, Derek. (2011) *On What Matters*. Oxford: Oxford University Press.
Putnam, Hilary. (1975) 'The Meaning of "Meaning"'. In *Mind, Language and Reality: Philosophical Papers*, vol. 2 (Cambridge, UK: Cambridge University Press), 215–71.
Rawls, John. (1999) *A Theory of Justice*. Rev. ed. Oxford: Oxford University Press.

https://doi.org/10.1017/apa.2015.6 Published online by Cambridge University Press

Scanlon, Thomas M. (2002) 'Rawls on Justification'. In S. Freeman (ed.), *The Cambridge Companion to Rawls* (Cambridge, UK: Cambridge University Press), 139–67.
Shafer-Landau, Russ. (2003) *Moral Realism: A Defence*. Oxford: Oxford University Press.
Smith, Michael. (1994) *The Moral Problem*. Oxford: Blackwell.
Swinburne, Richard. (2013) *Mind, Brain, and Free Will*. Oxford: Oxford University Press.
Wedgwood, Ralph. (2007) *The Nature of Normality*. Oxford: Oxford University Press.
Yablo, Stephen. (1993) 'Is Conceivability a Guide to Possibility?'. *Philosophy and Phenomenological Research*, 53, 1–14.

https://doi.org/10.1017/apa.2015.6 Published online by Cambridge University Press