New Blackfriars (2024), **105**, 116–130
doi:10.1017/nbf.2023.22

CAMBRIDGE UNIVERSITY PRESS

ARTICLE

How to Define 'Moral Realism'

Richard Swinburne

Faculties of Theology and Philosophy, University of Oxford Oriel College, Oxford, UK
Email: richard.swinburne@oriel.ox.ac.uk

(Received 5 September 2023; accepted 5 September 2023)

Abstract

The paper considers three possible definitions of what it is for an action to be 'morally' good: (1) that it is overall important to do; (2) that it is overall important to do in virtue of a universalisable principle; and (3) that it is overall important to do in virtue of a universalisable principle, belonging to a system of such principles, which includes almost all of certain moral fixed points. I defend (3) and show how we can reach such a system, starting from the basic beliefs with which we find ourselves, through the process of reflective equilibrium. Moral realism is then the doctrine that there is such a system of true moral beliefs. My optimistic view is that all human communities could eventually reach the same such system. But, if they cannot, then there will be two (or more) different such systems, and so two (or more) different senses of 'moral realism'.

Keywords: moral goodness; moral realism; universalisability; Russ Shafer-Landau; Terence Cuneo; reflective equilibrium; moral sense

1. First definition

I understand by a 'moral proposition' a proposition asserting or implying that some action or kind of action is 'morally' good (or, more particularly, 'morally' obligatory) or 'morally' bad (or, more particularly, 'morally' wrong).[1] Moral realism is the doctrine that some moral propositions are true. In this paper, I shall assume moral cognitivism, that is, that 'moral beliefs' are beliefs that some particular moral proposition is true; but I shall not assume moral realism. The English word 'moral' (and, I suspect, almost all words of other languages normally so translated into English) may be used in many different senses. This paper distinguishes several different senses of 'moral' and ends by defending one sense which I claim is the sense which is implicitly assumed by

[1]This article was published previously in the Iranian journal, *Journal of philosopjhical theological Research*, and is republished here with the appproval of the editior of that journal. Thanks to Terence Cuneo for clarifying the account by him and his collaborators of their understanding of 'conceptual truths'. I have taken the examples of moral disagreements which I discuss in the concluding pages of this paper, from an earlier paper of mine 'Necessary Moral Principles', *Journal of the American Philosophical Association*, 1 (2015), 617–34.

most of those who dispute about 'moral' issues and is a sense in which many moral propositions and so moral realism are knowably true.

There are various narrow senses of 'moral' in which only actions of certain kinds can be assessed as 'morally good' or 'morally bad'. In one sense, for example, 'morality' concerns one's actions only insofar as those actions are part of one's private or social life – how one interacts with one's spouse, parents, and children and keeps promises to one's business associates and customers. In any such sense, an action of that particular kind is morally good only if it is important that one should do it, morally obligatory only if it is *very* important that one should do it, morally bad only if it is important that one should not do it, and morally wrong only if it is morally obligatory not to do it – only in so far as it is an action of the relevant kind. (The concept of 'moral obligation' needs to be made a little more precise than this, but I shall ignore that point in this paper for reasons of space.) In the sense in which 'morality' concerns only one's private or social life, plausibly adultery is morally wrong and caring for one's parents in their old age is morally obligatory. But such a limited scope for morality leaves it open that an action might be morally bad in a particular limited sense, but overall more important to do than some incompatible morally good action, and conversely. For example there is an issue of whether or not it is overall more important that a spy should be faithful to his own wife than that he should seduce the wife of some enemy general in order to discover the enemy's plans for conquering the spy's country and subjecting it to a tyrannous oppression.

Philosophers are rightly more interested in wider senses of 'moral' concerned with the overall importance of different actions, rather than with their importance considered as actions of a certain kind.[2] Hence I shall be concerned in future with morality in senses which include the following requirements: an action is morally good only if it is important overall to do it; an action is morally bad only if it is important overall not to do it; an action is morally obligatory only if it is very important overall to do it; and an action is morally wrong only if it is morally obligatory (in that sense) not to do it. (Again, the concept of moral obligation needs to be made somewhat more precise than this, but I ignore this point for reasons of space.) (Henceforward, I ask the reader to assume that what I say about some sense of 'morally good', 'morally bad', 'morally obligatory', or 'morally wrong' applies with obvious changes to each of these expressions, unless I specify otherwise. I shall not always mention each of them in connection with what I say about some sense of 'moral'.) We sometimes talk of actions being good for one person (or persons) and not good for another person (or persons). If I give a lot of money to help refugees as a result of which I will have to eat very cheap food in the future, it may seem that this action was good for the refugees but bad for me. To suppose that there are morally good actions involves assuming that the goodness or badness of an action for me can be weighed together with its goodness or badness for others to reach an overall weight, either that it is morally good overall or morally bad overall or neither.

Any view that holds that an action is morally good to the extent to which it is important overall to do it, entails 'moral internalism', the view that moral beliefs are necessarily motivating. Believing that some action is morally good to do (and

[2]See Cooper (1970) for the narrower and wider senses of 'morality', distinguished in this way.

https://doi.org/10.1017/nbf.2023.22 Published online by Cambridge University Press

especially if we believe it to be obligatory) entails that we will desire (= it will incline us, influence us) to some degree to do it, even if in the end we yield to counter-influences which lead us not to do it. And believing that some action is morally bad (and especially if we believe it to be wrong) entails that we will desire to some degree not to do it, even if in the end we yield to influences which lead us to do it. It makes no sense to suppose that a person could believe some action is important to do, and yet have not the slightest desire to do it.

But surely, the moral externalist since she believes that moral beliefs do not entail desires) may respond, there could be and perhaps are totally amoral persons having no inclination whatsoever to do actions which are morally good who still have the same beliefs about which actions are morally good (or whatever) as morally good persons. The internalist, however, denies that the amoral person can have the same beliefs as the morally good person. She claims that the amoral person may have learnt to call the same actions 'morally good' as does the morally good person, but the amoral person has learnt to do this simply by noticing which are the non-moral actions which the morally good person calls 'morally good' (or 'bad', 'obligatory', or 'wrong') and calling those actions 'morally good' (or whatever). Thus, the amoral person may have noticed that the morally good person calls actions of feeding starving people 'morally good', and so the amoral person also calls those actions 'morally good'. But the difference between the amoral person and the morally good person is that the morally good person calls those actions 'morally good' because he believes them important to do, while the amoral person calls those actions 'morally good' because the good person calls them 'morally good'. The amoral person is using the expression 'morally good' in what has been called an 'inverted comma' sense. Internalists (e.g., Smith 1994: 68–71) have compared the situation of the amoral person to the situation of the person born blind from birth who has a reliable method of detecting which objects sighted people call 'red', 'green', or whatever; for example, she may have a machine which she can read by touch, which detects the colours of objects by the wavelength of the light reflected from them. Then the blind person can correctly predict which objects a sighted person will call 'red', 'green', or whatever, but she does not mean the same by these words as does the sighted person. I conclude that, for similar reasons, the amoral person does not mean by 'morally good' what the morally good person means and has no conception of moral goodness and so no moral beliefs.

The discussion so far allows for a very wide sense of moral goodness in which the criterion of overall importance is the only criterion of moral status. In this widest sense, which I shall call my first definition, an action is 'morally good' iff (= if and only if) it is 'overall important to do', 'morally bad' iff it is 'important not to do', and so on. It will seem obvious to the vast majority of people that in this sense there are some moral propositions which are evidently true. For example, it will seem obvious to the vast majority of people that 'Hitler did something very wrong in attempting to exterminate the Jews' or 'what Mother Teresa did in founding a community to care for the poor of Calcutta was morally good'. But it will also seem obvious to the vast majority of people that in this sense of 'morally good', there is no agreed way of discovering which of disputed moral propositions is true. And that leaves moral realism open to a verificationist objection that we could not understand some expression unless we could point to some object to which it applies or define it in terms of other expressions that can be eventually understood in virtue of their application to some object to which we could point. If there are no agreed cases of 'morally good' (or whatever) actions,

https://doi.org/10.1017/nbf.2023.22 Published online by Cambridge University Press

and 'morally good' cannot be defined in terms of other expressions for which there are agreed cases, we would not understand what 'morally good' means.

Quite independently of this verificationist objection, it is, however, natural to suppose that the very wide definition of a 'moral' proposition as one which affirms the overall importance of doing (or not doing) some action, with which I have so far operated, does not fully capture what is normally supposed to be involved in moral beliefs and so the moral propositions which they affirm. Surely one would not call anyone's belief that it was very important that he should stand on his head three times a day a 'moral' belief unless he held that there was some reason why he ought to stand on his head three times a day, in the form of there being something about him which made it obligatory for him to stand on his head three times a day, and then it would follow that anyone who was like him in that respect would have a similar obligation.

2. Second definition

The example just considered suggests that an action can be morally good only if it is justifiable by a reason of a 'universalisable' kind. It was Kant who drew the attention of the modern world to the importance of this criterion by his claim that moral obligation involves acting 'as if the maxim of your action were to become through your will a universal law of nature' (Kant 1785, 89). Put in my terminology, this is the requirement that an obligation to do some action is a moral obligation only if it is justifiable by a general principle to the effect that that action is an action of a certain non-moral kind, such that all actions of that kind are 'morally obligatory'. Although Kant was concerned almost entirely with moral obligation, this definition can clearly be extended to 'moral goodness' and the other species of moral propositions.

In these senses, which I shall call my second definition, an action is morally good (or bad, or whatever) iff the action is overall important to do (or not to do, or whatever) for a reason that the action is of a certain non-moral kind and all actions of that kind are overall important to do (or not to do, or whatever). The non-moral kind must be describable by (conjunctions or disjunctions of) universals, that is, properties, the essence of which does not involve any particular individual (person, time, or place). Thus, plausibly what Hitler did in ordering the German invasion of Poland in September 1939 was morally wrong because it was overall very important not to do for the reason that it was an act of invading a peaceful independent country; what Florence Nightingale did in organising a hospital in the Crimea in 1856 was morally good for the reason that it was overall important to do because it was an act of caring for the sick. The conjunction (of disjunctions, or disjunction of conjunctions).

of non-moral properties on which the moral property depends might be a long one or a short one, and the examples which I provided of the non-moral properties which confer on the actions of Hitler or Florence Nightingale the moral properties are almost certainly too short a list. It may be that all acts of invading a peaceful independent country are morally wrong, or it may be that all acts of invading a peaceful independent country that is not harbouring groups of foreign fighters about to attack the invading country or [a list of different circumstances in which it would not be wrong to attack such a country] are morally wrong, and so generally.

Since a particular action is morally good or bad for the reason that it has certain non-moral properties, it follows that any other action which had just those non-moral properties would have the same moral properties. If what Hitler did in ordering the

https://doi.org/10.1017/nbf.2023.22 Published online by Cambridge University Press

German invasion of Poland in September 1939 was morally wrong because it was an act of invading a peaceful independent country which is not harbouring groups of foreign fighters about to attack the invading country, it must be the case that any other action of invading a peaceful independent country which is not harbouring groups of foreign fighters about to attack the invading country is also morally wrong.

So every particular moral proposition (e.g., 'action *a* was morally good', *a* being picked out as the action it is in virtue of some of its non-moral properties) can be true only if it is entailed by a conjunction of a true general moral principle (e.g., 'all actions of non-moral kind Z are morally good') together with a true non-moral proposition ('action *a* was an action of kind Z'). Thus, if the particular moral proposition 'John ought not to smoke cigarettes' is true, it might be true in virtue of being entailed by the conjunction of a true general moral principle 'all humans ought not to smoke cigarettes' and a true non-moral proposition 'John is a human'. But if one or other of these conjuncts is false, the particular moral proposition would have to be entailed by some other such conjunction if it is to be true. General moral principles may themselves hold in all conceivable circumstances, that is, be true in all possible worlds (describable by non-moral propositions) or be true only in some possible worlds including the actual world. I shall call ones which are true in all possible worlds, fundamental moral principles. If 'genocide is always morally wrong' does not apply in a world where the only way to save most of humanity is to kill all members of a certain race who have an infectious disease that will kill most other humans unless all of that race are eliminated, then 'genocide is always morally wrong' is not a fundamental moral principle. But, in that case, it still might be a fundamental principle that 'genocide is always morally wrong unless it is the only way to save the lives of most other humans'. A fundamental moral principle has all the exceptions to it built into it, and so it applies in all possible worlds. I shall call propositions true in all possible worlds 'metaphysically necessary' propositions.

If what makes some particular moral proposition true is the conjunction of a particular non-moral proposition and a general principle P, which is true only in some possible worlds including the actual world, then there will still be a metaphysically necessary and so fundamental principle of the form '{in such-and-such conditions (defined by non-moral propositions). including the present conditions}, P is true'. Thus, if what makes the particular moral proposition 'John ought not to smoke cigarettes' true is that it is entailed by the conjunction of the true non-moral proposition 'John is a human' with the true general moral principle 'all humans ought not to smoke cigarettes' which is true only in certain conditions, including the present conditions, then there will be a principle true in all possible worlds of the form that 'all humans ought not to smoke cigarettes [in those conditions]'. That fundamental principle might be that 'in possible worlds where humans smoking cigarettes damages their health, all humans ought not to smoke'. If that principle is a fundamental principle, it is a very narrow fundamental principle, and it is natural to suppose that such narrow fundamental moral principles are true in virtue of being entailed by wider fundamental moral principles such as 'all humans ought not to pursue activities that damage their health'. But whether or not narrow fundamental principles are always entailed by wide fundamental principles, it remains the case that, on this second definition of 'morally good' ('bad, obligatory or wrong'), if there are any true moral propositions, there are metaphysically necessary moral principles. Any set of the most general fundamental

https://doi.org/10.1017/nbf.2023.22 Published online by Cambridge University Press

moral principles that entail all other fundamental moral principles constitutes the true moral system.

Since fundamental principles apply in all conceivable circumstances, to discover what they are, we must clearly conceive of different circumstances, and if, by so doing, we can discover what is conceivable, that discovery will be the discovery of an a priori truth. I will call metaphysically necessary propositions, which are discoverable a priori, 'logically necessary' propositions. So any fundamental moral principles are logically necessary principles. The way to work out whether some suggested moral principle such as 'all humans ought not to pursue activities that damage their health' is a fundamental moral principle is to reflect, as far as we can, on the possible worlds that there might be and then to reflect again whether it really seems that that principle still holds in all such worlds. If we think that we cannot have any idea about which moral principles would hold in worlds very different from our own, then our knowledge of fundamental moral principles will be confined to those which contain a clause 'in all worlds similar to ours', and it may prove very difficult to fill out in detail in what respects a world has to be similar to ours in order for some general principle to hold in it. So, to revert to my earlier example, we may plausibly claim that 'in all worlds similar to ours, genocide is always morally wrong unless it is the only way to save the lives of most humans'. It remains irrelevant whether there will ever be a race, all of whose members suffer from a disease, such that the only way to save the lives of most humans is to kill all members of that race. But since, I suggest, it seems plausible to most people that there are moral propositions of the kind that actions are good to do for reasons of a universalisable kind, and so that there are logically necessary fundamental moral principles, it ought to seem plausible to them that on this second definition moral realism is true. But is there now a way to discover what the fundamental moral principles are, and so meet the verificationist objection?

It is normally agreed that the class of 'logically necessary' propositions in my wide sense includes both 'logically necessary' propositions in the narrow sense of axioms or theorems of certain undisputed systems of formal logic and 'conceptually necessary' propositions. Conceptually necessary propositions are naturally understood as ones true in virtue of the meanings of words in the sentences that express them – for example, 'all squares have four sides' or 'if it's red, it's coloured' are true in virtue of the meanings of such words as 'square', 'side', 'red', and 'coloured'. But fundamental moral principles are not axioms or theorems of any undisputed system of formal logic. So either moral realism is false, or fundamental moral principles are conceptual truths, or they are logically necessary truths of some totally new kind. Some philosophers, such as Fine (2002), have taken the latter route; but that still leaves it open to the verificationist objection that there is no agreed way to discover what the necessary truths are. But on the view that fundamental moral principles are conceptually necessary truths, there is an agreed way to discover what those principles are – by analysing the meanings of 'morally good', 'morally bad', and so on. So let us explore the possibility that fundamental moral principles are conceptually necessary truths.

3. Third definition

Terence Cuneo and Russ Shafer-Landau (2014) and Bengson, Shafer-Landau, and Cuneo (2023) have taken a first step in exploring this possibility by postulating that there

https://doi.org/10.1017/nbf.2023.22 Published online by Cambridge University Press

are many 'moral fixed points'. (Frank Jackson and Philip Petit [1995] have postulated that there are some 'commonplaces' that have a somewhat similar role. I shall follow the terminology of Cuneo and Shafer-Landau.) The moral fixed points are principles about narrow kinds of actions such as (my examples) 'it is wrong to torture others just because they have inconvenienced you', 'it is wrong to rape a child solely to indulge one's lust', 'it is obligatory to protect one's children from lethal danger', and 'it is obligatory not to break a promise on which another person is relying simply for the sake of convenience'. These propositions are 'ones which we hold must find a place in any system of moral norms that applies to beings like us, in worlds similar to our own' (Cuneo and Shafer-Landau 2014). And they go on to list many similar propositions (2014, 405) which almost all humans (who believe that there are 'moral' truths) would agree to be moral truths. What makes some moral proposition a conceptual truth, they hold, is that it belongs to the essence of a moral concept, such as 'wrong', that, necessarily, it applies to certain non-moral concepts. For that reason, they hold, it belongs to the essence of 'wrong' that, necessarily, it applies to 'raping a child solely to indulge one's lust'. Their grounds for claiming that such propositions are true are 'the implausibility of holding otherwise' (Bengson, Shafer-Landau, and Cuneo, 2023); they thus appear to assume that things are as they seem to be in the absence of counter-evidence. That we should believe that things are as they seem to be in the absence of counter-evidence is what I have called 'the principle of credulity'; and, as I argue, without this principle, we could have no justified beliefs.[3] However, Cuneo and Shafer-Landau provide no full account of what it is to belong to the essence of a concept and so of how we could know that that concept does apply to certain non-moral concepts. So I suggest that they should endorse the definition which I suggested is the natural one, that conceptually necessary propositions are ones which are true in virtue of the meanings of the words in sentences which express them. This is the one which I shall assume in considering the suggestion that fundamental moral principles are conceptually necessary. A sentence being true in virtue of the meanings of the words contained in it has the consequence that the negation of that sentence, and so the proposition which the negation expresses, entaila contradiction.A proposition is 'coherent' (= 'logically possible' (in the wide sense which I am assuming) or conceivable') iff it does not entail a contradiction. Similarly, a sentence being false in virtue of the meanings of the words contained in it has the consequence that that sentence itself entails a contradiction.

So, on this view, a fundamental moral principle is a conceptual truth iff its negation entails a contradiction. But it is implausible to suppose that the negation of any general principle of the fixed points kind (even if they all include a clause like 'in worlds sufficiently similar to ours') entails a contradiction. It seems even more implausible to generalise this into a claim that the negation of any moral principle, whether of the fixed points, or not, is true iff its negation entails a contradiction, and false if it itself entails a contradiction. It is overall important that every male human should soon after reaching puberty rape some female human' may seem to us manifestly false, but it seems implausible to suppose that it entails a contradiction. So moral principles

[3]for a brief justification of this claim that we need the principle of credulity in order to have any rational beliefs, see Swinburne (2013: 42–44). This principle is similar to the principles called by other philosophers, 'phenomenal conservatism', 'epistemic conservatism', and (in a favourable sense) 'dogmatism'.

https://doi.org/10.1017/nbf.2023.22 Published online by Cambridge University Press

cannot be conceptual truths if they are understood merely in the way developed in my second definition, as propositions asserting that actions are overall important to do (or not to do) for reasons of a universalisable kind.

The alternative way to understand the view that fundamental moral principles are conceptually necessary is to understand the concept of 'moral' goodness, as picking out a kind of goodness which is the kind it is, not merely in virtue of satisfying the formal conditions of being important for universalisable reasons, but also in virtue of the kinds of action described in terms of their non-moral properties, which can be coherently said to be 'morally good'. Yet clearly we do not mean by a 'fundamental moral principle' one which states that certain specified kinds of action are 'morally good' ('bad', obligatory', or 'wrong'), because that would have the consequence that anyone who held a belief that a kind of action was 'morally good' that was not in the list of specified actions was simply using 'morally good' in a different sense from other people. And that would rule out the possibility that anyone could disagree with anyone else about fundamental moral principles or make any progress in coming to know what are the fundamental moral principles. Yet we do assume that such disagreement and progress can occur. The restriction on the kinds of actions which can coherently be said to be 'morally good' must have a more elastic form. It will have the form that an action can count as 'morally' good (bad, obligatory, or wrong) only if it is an action of a kind which is connected in a certain sort of way with actions which many other humans also recognise as good (or whatever) because they are overall important (to do or not to do) for universalisable reasons.[4] The task now is to define the kind of connection. So an initial way to regard Cuneo and Shafer-Landau's claim is as a proposal to understand a suggested fundamental general principle asserting that a kind of action is 'morally good' as asserting that an action is important for reasons of a universalisable kind and also consistent with the moral fixed points, so chosen because almost all other humans believe them important (to do or not do) for reasons of a universalisable kind. Then to deny a moral fixed point would entail a contradiction.

So far, so good. But any list of fixed points drawn up by anyone would be disputed by someone else, and people hold fixed points with different degrees of conviction. And surely not any principle consistent with the moral fixed points, asserting that an action is overall important for reasons of a universalisable kind, such as the principle that everyone ought to stand on their head three times a day, should count as a possible fundamental moral principle. I suggest that if we are to pursue the route of defining a fundamental moral principle partly in terms of its content, we need an account which recognises that beliefs about fundamental moral principles belong to systems of beliefs satisfying the formal criteria, not all of which have exactly the same fixed points, but which are connected to each other by considerable overlaps of content between many of them, and also because they are open to a common method of resolving moral disagreements. I will now explain what I mean by 'the formal criteria', a 'system of beliefs', 'overlap of content' and 'common method of resolving disagreements'.

[4]Rawls may have recognised this limitation 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'. He continued, 'we can find an accurate account of our moral conceptions', which I take to mean that if we take account of the moral beliefs we actually have, 'these questions of meaning and justification may prove easier to answer'.

https://doi.org/10.1017/nbf.2023.22 Published online by Cambridge University Press

I mean by a suggested fundamental moral principle satisfying the 'formal criteria', that it is a principle about the overall importance of a kind of action for universalisable reasons in all possible worlds. These will include principles that a kind of action has this overall importance in any world like ours in certain specified respects (since propositions necessary in one world are necessary in all worlds). I mean by 'a system of beliefs' all the beliefs satisfying those formal criteria held by some contemporary group of humans of some significant size. Two different systems can 'overlap' in two different ways. First, they may contain most of the same beliefs and only a few different ones; humans who disagree about a few contested issues often agree about many other issues. Second, they may agree that the overall goodness or badness (obligatoriness or wrongness) of many kinds of action (in any world like ours in some specific respect) arises from their possession of largely the same good and bad non-moral properties, while disagreeing only as to which group of properties outweighs the other group. Thus, the opponent of the euthanasia involved in helping a depressed person to commit suicide may argue that such an act is overall bad because of the sanctity and value of human life, the possibility of helping a depressed person to recover from their depression, the value of their overcoming that depression, and so on. The advocate of euthanasia of this kind may argue that helping a depressed person to commit suicide is helping him or her to do what they clearly and firmly want to do and what hurts no one else in any way; and that because the action has this character it is an overall good action. Each disputant may readily admit that the considerations which their opponent adduces have some weight, while holding that the considerations which they adduce outweigh those which their opponent adduces. I suggest that most principles believed by groups of contemporary humans of significant size to be fundamental moral principles, including, of course, principles widely agreed to be moral fixed points, belong to systems which overlap with several other systems in one or other of these two ways.

Those who have beliefs of this kind do not understand their belief that some kind of action is 'morally good' merely as meaning that it belongs to the finite set of their actual beliefs which satisfy the formal criteria. They understand their belief that that kind of action is 'morally good' as a belief that that kind of action has some further property which they can recognise, and which is not definable merely as the property possessed by all the actions which they currently believe to be 'morally good'. Hence, they must allow that they could come to discover that some other kind of action (described in non-moral terms) has that property, and also that some action which they previously believed to have that property did not have it. So, I suggest, one could not have a belief that some kind of action is 'morally good' (or whatever) unless one believed that one's beliefs about which actions are morally good (or whatever) are open to at least minor revision. This is G.E. Moore's (1903: 20–21) point that whatever purported 'definition' of 'good' anyone offers, it would always be an 'open question' whether whatever satisfies the definition really is good. But, I now suggest, it does not follow that one cannot have a good reason, even if it is seldom a conclusive reason, for believing some moral proposition. And clearly people change their views about which actions are 'morally good' (or whatever) in the light of experience and reason.

As a result of experience, people may change their view that some kind of action is 'morally good' because they change their view about the truth of some non-moral proposition which, when conjoined with some moral principle which they believe, entails the view that that kind of action is morally good. If they had previously believed

https://doi.org/10.1017/nbf.2023.22 Published online by Cambridge University Press

that 'it is morally wrong to smoke cigarettes' because they also believed that 'all humans ought not to pursue activities that will damage their health' and that 'smoking cigarettes damages health', were to learn that smoking cigarettes does not damage their health, that would lead to a change in their moral view that it is wrong to smoke. But a change in the belief that some principle is a fundamental moral principle and so holds in all conceivable circumstances can only be reached by doing more conceiving, that is, by a priori reflection. I now claim that there is a procedure implicitly applied by those who seek to make progress in understanding which suggested fundamental principles satisfying the formal criteria for moral beliefs are probably true, explicitly formulated by John Rawls (1999: 20–21) as 'reflective equilibrium' and recognised as an important such procedure by many moral philosophers. I now describe this procedure.

We begin with our different basic beliefs (resulting from what we have been taught in childhood and our subsequent life experiences) about the 'moral' status (good, obligatory, bad, or wrong) of actions of some fairly narrow type. We hold these beliefs with different initial degrees of conviction, believing almost all the fixed points with almost total conviction. We then look for some more general principle applicable in many different conceivable circumstances, from which it would follow that at least most of those beliefs are indeed true. In so far as that principle is a simple principle which entails that most of our basic beliefs are true and that hardly any of them are false, we come to believe that principle. That principle will then increase our conviction in the basic beliefs entailed by it and decrease our conviction in any basic beliefs inconsistent with it. We then go on to consider the moral status of other possible kinds of actions whose moral status is entailed by our principle and see whether we find ourselves naturally inclined to believe that they have that moral status. To the extent to which we are so inclined, that adds to our conviction in the truth of the principle and inclines us to abandon beliefs inconsistent with it. A system of morality, like a system of science, metaphysics, or epistemology, is probably true in so far as it has few simple general principles about the morality of kinds of actions picked out by a few readily recognisable properties and also entails many recognisably true basic beliefs and no recognisably false basic beliefs about the morality of actions of narrower kinds with which we find ourselves. As in science, metaphysics, and epistemology, so also in ethics, simplicity has to be balanced against the evidence – in the case of ethics, the basic beliefs about the morality of actions of narrower kinds, with which we find ourselves. If a basic belief is one that some narrow kind of action is only *fairly probably* good (or whatever), then the believer will feel the force of a suggested necessary contrary simple moral principle, which entails most of their other basic beliefs, and abandon the basic belief inconsistent with it. But if the believer's basic belief about the original narrow kind of action is a very strong one, then he or she will be right to retain that belief and not recognise the more general principle. We aim by this process, if at all possible, to reach not merely fundamental principles including a clause 'in all worlds similar to ours' but ones without any such restrictive clause. Frequently, moral progress is facilitated by moral argument – others present to us objections to our views.

Here are two examples of that process at work – one where the issue turns on the simplicity of a suggested principle, and one where the issue turns on the number of beliefs about different narrow kinds of action which are entailed by two rival principles, one of which may involve more 'readily recognisable' properties than the other.

https://doi.org/10.1017/nbf.2023.22 Published online by Cambridge University Press

For my example of the first kind, I suppose that John initially believes several narrow principles about when it is legitimate to kill: that it is wrong for a person A to kill a person B unless B is about to kill A or a relative of A, or has killed a relative of A, or A is acting on behalf of a court of law of A's country that has sentenced B to death for murdering a citizen, or B is an enemy combatant in a war with A's country, or A is fighting a duel with B, because B has insulted A or a relative of A. Suppose that Mary also initially believes all these narrow principles except the last one. Then Mary may ask John to reflect that a far simpler principle which would entail most of the narrower principles is that 'it is wrong for A to kill B except in a court-authorised retribution (that is, capital punishment) for a killing of a fellow-citizen, or to prevent further killing of a fellow citizen'. 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, and so not in a duel. It has the consequence, which both John and Mary may recognise, that it was arbitrary to confine A's right to kill in order to save A's life or that of a relative, except in war where one is fighting to save the lives of fellow citizens. And it also has the consequence, which both John and Mary may recognise, that it was arbitrary to allow anyone to kill in retribution for killing a relative without any permission, but only to kill other citizens in retribution for a killing with the permission of a court. Hence, while having most of the same moral consequences as the earlier narrow principles, and so, in virtue of its simplicity, providing a reason why those narrow principles are true, it provides significant reason for correcting them in ways that are likely to appeal to many moral inquirers.

Further reflection on the sacredness of life as what underlies the new general principle might lead John and/or Mary to develop an even simpler principle, 'it is wrong to kill any person except to prevent the killing of another person', whether the latter person is a fellow citizen or not, which would have the consequence that any retribution for a killing, whether court-authorised or not, is wrong. But if either John or Mary believe very strongly that killing in a duel or capital punishment for murder is justified, they will rationally resist the move to the simplest principle.

Here is an example where moral disagreement turns on whether a suggested more general principle involves 'readily recognisable properties' and explains significantly more initial basic beliefs than its rival. I described earlier a dispute about the morality of helping a depressed person to commit suicide where its supporters and opponents agree that the moral goodness or badness of doing so depends on the same non-moral properties which it has, some of which make for the overall goodness of the action and some of which make for its overall badness, but disagree about which group of properties outweighs the other group. John may claim, against Mary, that the simplest principle underlying their many shared beliefs about it being good to help the sick, the unemployed, the unlovable, etc. is that it is always good not to 'give up' on helping people satisfy their basic *needs*. He might then claim that this has the consequence that one should not give up on helping the depressed to satisfy their basic need for happiness, and so on helping them to recover from their depression; and so that one should not help anyone to commit suicide. By contrast, Mary might urge that the simplest principle underlying their many shared beliefs about it being always good to help those who want to get an education, have children, or travel abroad is that it is good to help people to do what they firmly *want* to do, when that hurts no one else, and so claim that it follows that we should all help people to commit suicide if that is what

https://doi.org/10.1017/nbf.2023.22 Published online by Cambridge University Press

they firmly want to do. Also, Mary may claim that it is often very obvious what a person 'wants' (that is, desires), but what a person 'needs' may often not be (in my phrase) a 'readily recognisable property'; and that many of our good actions are actions of helping people to do what they want, whereas far fewer of our good actions are actions of helping people to do what they need (even if we agree on what their needs are). Then John may recognise the easy applicability of Mary's principle and its ability to explain the goodness of a large number of actions which he recognises as good, and so he may change his view about the morality of euthanasia. Or Mary may recognise the improbability of some of the consequences of her principle.

But why should we believe that the principles reached by this procedure are probable conceptual truths? My answer is that it is just the same procedure as the procedure for establishing general principles about the logically necessary and sufficient conditions for the application of some central epistemological or metaphysical concept used by 'analytic' philosophers, especially in the mid-20th-century period. That is what the procedure is – a procedure for establishing conceptual truths.

I illustrate this claim by the example of how philosophers proceeded to discover the logically necessary and sufficient conditions for 'S remembers having done X'. A first suggestion by Hume (1739: Book1, Part1, section 3) 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, and innumerable examples of usage showed, that we allow that someone might have 'remembered' having done some action, even if they did not have a 'lively' mental image of having done it. Yet clearly all examples of usage showed that in order to 'remember' having done X, someone needs more than merely a true belief that they had done X. Philosophers pointed out that we would not count someone as having 'remembered' having done X merely because they had acquired the belief that they had done X from having read in a book that they had done X. So then it was suggested that for S to 'remember' having done X, S's belief that he had done X must have been caused by him having done X. Then it was pointed out that even if S having done X caused him to write in his diary that he had done X, and his reading this later in his diary caused him to believe that he had done X, we still 'would not say' that he 'remembered' having done X. Only if the route of causation went directly through S's body (in effect, his brain), and only if the belief was a basic belief not inferred from anything else, would the resulting belief count as a memory. And so 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.[5] The application of processes of this kind led to accounts of the logically necessary and sufficient conditions of 'S knows that p', 'S perceived O', and such like.

In each case, a principle which fitted the range of examples so far considered was seen as giving a probably correct account of the logically necessary and sufficient conditions for the truth of any sentence of the relevant form, such as 'S remembers having done X', until investigators came across cases where a sentence of that type was clearly not being used in a way which satisfied those conditions, and so they sought a different account. In seeking a new account, they sought one which was not a mere conjunction

[5]This is a rough summary of the conclusion of Martin and Deutscher (1966).

https://doi.org/10.1017/nbf.2023.22 Published online by Cambridge University Press

of kinds of examples in which a sentence of the form 'S remembers having done X' was true. Rather, they looked for one which was simple in using few predicates designating properties easily recognisable in almost all different kinds of examples. In summary, an account of the logically necessary and sufficient conditions for the use of such a sentence was judged a probably correct one in so far as it was a simple one which fitted very many examples of that use well, and very few examples badly. It follows that the conjunction of such a probably correct principle with a proposition inconsistent with it would be a contradiction, and so the principle was probably a necessary conceptual truth.

So I suggest that the evident truth-conduciveness of the 'analytic' method of determining the logically necessary and sufficient conditions for the application of some concept, such as 'remembers doing X', is strong reason for supposing that the same method will reach probable truth in ethics. The 'analytic' method assumed that all speakers were using the relevant expressions with the same logically necessary and sufficient conditions for their application, that is that the expressions were not ambiguous. The evidence for this was that almost all speakers agreed about many paradigm examples of when someone could correctly said to have 'remembered' doing some action; what philosophers were trying to discover was the general principles which explained the application of the crucial expression to those and other examples. Likewise, the method of reflective equilibrium, as used to analyse moral expressions, can be applied only to groups who have many of the same paradigm examples of 'morally' good, bad (or whatever) actions, in the form of 'fixed points'. That is good reason to suppose that the groups mean the same by 'morally good' (or whatever).

In that case, if humans gradually become aware of the moral beliefs of others, both of those whose beliefs differ only slightly from theirs and of others whose beliefs are more distant from theirs, the continual practice of reflective equilibrium should lead to a far greater agreement between them. When reflective equilibrium has moved someone away from one moral system to an overlapping moral system, that person will regard many of the new examples of morally good and bad actions which he comes to recognise as examples of the relevant concept just as satisfactory as the ones recognised previously and still recognises. And then that person will have much more in common by way of overlap with the views of some person with which they originally had much less in common, which will provide a basis for reaching agreement which did not exist previously. And so it is possible for someone to move by a rational process consisting of many such steps from one general moral outlook to a fairly different one, say from a narrow tribal understanding of which actions are overall good to a far wider one – although they will continue to hold most of the 'fixed points' which they already believed, and they will have reached that result by a shared process which they recognise as truth-conducive. What will impede progress will be the fact that members of different groups start with very different initial basic beliefs about the morality of particular kinds of actions, which they have learnt from their parents and peers. But if the use of reflective equilibrium leads some members of these groups to change their views, then there will be fewer parents and peers who hold the original basic beliefs, and so fewer people who will acquire these beliefs from their parents and peers. I shall mean by 'ideal reflection' the continued use of the method of reflective equilibrium over many centuries, interacting with the initial basic 'moral' beliefs of all other such systems. My optimistic view is that, in view of their agreement on the 'moral fixed

https://doi.org/10.1017/nbf.2023.22 Published online by Cambridge University Press

points' and the fact that systems of 'moral' beliefs have considerable overlap with each other, if human communities become more and more aware of each other's views and follow the rules of reflective equilibrium which I have analysed, ideal reflection will lead to near-universal agreement about what are the true 'moral' principles. Hence my third and final suggested definition which defines the 'moral goodness' (or whatever) of an action partly by the way in which it could be shown to be very probable that the action is 'morally good:' an action is morally good (or bad, or whatever) iff the action is overall important to do (or not to do, or whatever) because (when conjoined with some true non-moral proposition) this follows from a fundamental principle of a universalisable kind, belonging to a system of such principles which includes almost all the moral fixed points; when a suggested fundamental principle is one which would be shown to be very probably true by ideal reflection.

Alas, human communities may not become more and more aware of each other's views, and humans are subject to many temptations to irrationality which inhibit the unbiased use of the method of reflective equilibrium. My optimistic view, however, is that by pursuit of ideal reflection of the kind which I have illustrated, all humans would reach the same views about what are the true 'moral' principles. Given that optimistic view, all humans have a common moral 'sense', which enables us as a result of our initial 'moral' instruction to recognise special sorts of properties – 'moral goodness', 'moral badness', 'moral obligation', and 'moral wrongness', possessed by many different actions, many of which actions are not ones which we were originally taught had those properties. On this third definition, a moral proposition is a proposition which asserts that an action is morally good (or bad, or whatever) on this definition. A moral belief is a belief that some action is morally good, bad, obligatory or wrong on this definition. Moral realism is the doctrine that some moral propositions (on this definition) are true – and since there are very many moral propositions, such as those affirming moral fixed points, which are very probably true, it is immensely probable not merely that moral realism is true, but that it is knowably true.

But if my optimistic view is mistaken, and ideal reflection would never lead to one very probable universal system of morality, but its use by different groups beginning with different initial moral beliefs would lead to two or more different moral systems, then there would be as many different senses of 'morally good' (or whatever) as ideal reflection would produce, and so two or more different senses of 'moral realism'. For each such system, there will be a definition corresponding to my third definition with the consequences which follow from it, and different groups of humans will have different 'moral senses' sensitive to different properties of actions. Yet, given that agreement on 'moral fixed points' is involved in my definition of a moral principle and given the very considerable overlap of moral systems with each other, I suggest that it is implausible to suppose that ideal reflection would have other than one clear result.

I should add that this account of the fundamental moral principles is perfectly compatible with a religious view that many of our obligations to do particular actions are due to the will of God. This is because it is surely a fundamental moral principle that if someone gives you a gift on condition that you use it in a certain way and you use the gift, then you have an obligation to use the gift in that way. If there is a God, then our life is a gift from him, and so if he has made it a condition that we should use it in a certain way, then we have the obligation to use it in that way as long as we live.

https://doi.org/10.1017/nbf.2023.22 Published online by Cambridge University Press

References

Charles B. Martin and Max Deutscher, 'Remembering'. *Philosophical Review*, 75 (1966), 161–96.
David Hume, *A Treatise of Human Nature* (London: John Noon, 1739).
Frank Jackson and Philip Petit, 'Moral Functionalism and Moral Motivation', *Philosophical Quarterly*, 45 (1995), 20–40.
G.E. Moore, *Principia Ethica* (Cambridge: Cambridge University Press, 1903).
Immanuel Kant, *The Groundwork of the Metaphysic of Morals*, trans. by H.J. Paton (London: Hutchinson and Co, 1785).
John Bengson, Terence Cuneo, and Russ Shafer-Landau, 'Conceptual Moral Truths', in *Oxford Handbook of Moral Realism*, ed. by P. Bloomfield, and D. Copp (Oxford: Oxford University Press , 2023), pp. 317–345.
John Rawls, *A Theory of Justice* (rev. ed. Oxford: Oxford University Press, 1999).
Kit Fine 'The Varieties of Necessity', in *Conceivability and Possibility*, ed. by T.S. Gendler and J. Hawthorne (Oxford: Oxford University Press, 2002), pp. 253–81.
Michael Smith, *The Moral Problem* (Oxford: Blackwell, 1994).
Neil Cooper, 'Morality and Importance', in *The Definition of Morality*, ed. by G. Wallace and A.D.M. Walker (London: Methuen and Co, 1970), pp. 91–97.
Richard Swinburne, *Mind, Brain, and Free Will* (Oxford: Oxford University Press, 2013).
Terence Cuneo and Russ Shafer-Landau, 'The Moral Fixed Points', *Philosophical Studies*, 171 (2014), 399–443.

Cite this article: Richard Swinburne, 'How to Define "Moral Realism"', *New Blackfriars*, 105 (2024), 116–130.
https://doi.org/10.1017/nbf.2023.22