META-ETHICS AND NORMATIVE COMMITMENT

James Dreier
Brown University

Introduction

The inspiration for this paper is an article of Ronald Dworkin's: "Objectivity and Truth: You'd Better Believe It". I will not dwell on the details of Dworkin's paper, but let me set out the issue broached there that I want to explore.

Dworkin's paper argues against what he calls an 'Archimedean' position about ethics, one which seeks a point outside ethics proper to pass external judgments on the status of ethical propositions, saying that they are not objective, or not really true, or the like, while avoiding taking a normative position of its own. Attempts to establish Archimedean positions are quite well known. For instance, Emotivism seeks such a position, and John Mackie's Error Theory, and various forms of Subjectivism. The point is that these are not theories of what to do at all. They are supposed to be theories of the nature of moral advice, what sort of a thing it is. And they all claim that it lacks a certain preferred status: it is not the sort of thing that can be objectively correct.

One of Dworkin's main strategies is to try to draw the Archimedean into making recognizably ethical judgments. For instance, if the Archimedean says that the judgment, that slavery is unjust, is not true, then Dworkin takes this to be a (very implausible) moral judgment, namely, the judgment that there is nothing unjust about slavery. Fair enough (I think). But Dworkin makes a sweeping claim. He says that there is no Archimedean position at all. There is no way to have a purely meta-ethical theory, one without any substantive normative implications.

Earlier in this century just about all of the better known meta-ethical theories were supposed to be 'pure' meta-ethical theories, carrying no normative moral implications. Besides the Archimedean theories (or, if Dworkin is right, the would-be Archimedean theories) I mentioned, there was also Moore's theory. Moore did, of course, include a lot of substantive normative ethics in *Principia Ethica*, but the meta-ethical theory there seems to be independent of any particular moral view. One could agree with Moore that goodness is a *sui generis*,
simple, non-natural property, without agreeing with his broadly eudaimonistic consequentialism. Moore's theory was not Archimedean, in Dworkin's sense, since it in no way attempted to cast any doubt on the objectivity or reality of moral properties, but it was still 'pure'.

But nowadays one hears more and more doubts that there is such a bright line between meta-ethics and normative ethics, doubts that there are or could be any pure meta-ethical theories. In this paper I will investigate these doubts. I think they are ill-founded. I think there are meta-ethical theories that carry no normative moral commitments whatsoever.

Dworkin's paper considers explicitly a couple of the theories I think really are Archimedean. One of these theories (really a family of theories) is Expressivism. Dworkin argues that Expressivist theories, including Simon Blackburn's Quasi-realism and Allan Gibbard's Norm Expressivism, are failures as Archimedean theories. I will not pursue that particular matter. The other family of theories sometimes goes by the name of Secondary Quality theory, and the particular version that seems to me most clearly Archimedean is one that says that a thing is morally wrong if and only if it tends to cause in us as we actually are a feeling of moral disapproval. Dworkin insists that this theory is not Archimedean at all. I will explain briefly what is in dispute. The bulk of my paper is devoted to developing enough theoretic understanding of the idea of a theory's carrying moral commitment to be able to address the dispute in a rigorous way, and I will return to this special form of Secondary Quality theory at the end of the paper.

Dworkin considers a Secondary Quality theory according to which moral wrongness is a disposition to 'outrage' normal observers. He writes,

> But someone who holds that moral properties are secondary properties does take sides in actual or potential substantive disputes. Suppose we discovered that, contrary to our expectations, contemplating genocide does not in fact outrage even most normal people. Genocide would not then be morally wrong on that dispositional account, though, of course, many people would think it was. [101]

It is not entirely clear which of two points Dworkin is making here. He might be suggesting that for all we know, most people are not outraged by the idea of genocide, so that the Secondary Quality theory entails that genocide is not in fact morally wrong. It is not easy to see what significance this suggestion has, since it is extremely implausible. Or he might be asking us to consider a mere possibility, a possible world in which most people would not be revolted by genocide. Then the point would be that the Secondary Quality theory has very implausible consequences, consequences that are themselves counterfactuals. The counterfactual consequence in point would be this one:

(1) If most normal people were not outraged by genocide, then genocide would not be morally wrong.
This counterfactual is not true. Since one version of the Secondary Quality theory entails it, that version cannot be true. I think this is a good argument against the version of the Secondary Quality theory in question.

However, there is another version that does better. It is the theory that identifies moral wrongness with the disposition, not to outrage whichever people happen to exist, but to outrage *us actual moralizers*, to outrage us as we actually are.\(^3\) This version, which ties moral wrongness to our *actual* dispositions, is more plausible than the simpler version. When we consider a merely possible situation, asking whether genocide would be morally wrong in that situation, this 'actualized' Secondary Quality theory tells us to consider whether we ourselves, as we actually are, feel outrage at the genocidal acts, the merely possible ones. And of course we do, we find genocide morally horrible.

But Dworkin does consider this ‘fix’, and he rejects it. Here is what he says:

> The dispositional account might, it is true, take a different form. It might hold, for example, that what makes genocide wrong is the reaction, not of whichever kind of people happen to exist from time to time, but of us, that is, of people with the physiological structure, basic interests, and general mental dispositions that people actually have now. In that case, it would no longer follow that genocide would cease being wicked if human beings developed very different general interests or different neural wiring. But some plainly substantive and controversial claims would still follow: for instance, that genocide would not have been wicked if economic or other circumstances had been different as human reactions evolved, so that creatures with our general interests and attitudes had not been revolted by genocide. \[102\]

I find it difficult to interpret this objection. After all, ‘creatures with our general interests and attitudes’ would indeed be revolted by genocide, since revulsion at genocide is one of our general interests and attitudes. Creatures who shared many of our interests and attitudes but were not revolted by genocide would simply not be relevant, according to the 'actualized' Secondary Quality theory, to the determination of moral wrongness. So there is a dispute here, but it is hard to know what further can be said. At the most general level, it is a dispute over whether meta-ethical theories, and in particular Secondary Quality theories, carry normative moral implications. I would like to be able to resolve the dispute, but there is an obstacle. It is not easy to say what it is for a theory to carry normative moral implications. One difficulty is in saying which implications are normative moral implications, but this difficulty is not grave for our purposes, I think. More problematic is to explain the relevant notion of a commitment. One simple explication says that something carries normative commitments just in case its logical implications include some normative statements. But this turns out to be a useless explication. In the first section I will explain why it is useless. In the second and third sections I will suggest an alternative. Then I will examine one kind of theory that I will claim to be Archimedean, and finally I will return to Dworkin’s claims.
I. Why the Normative Commitments of a Theory are Not Just Its Normative Implications

A. Hume's Law

To explain what a normative commitment is, what it means for a (meta-ethical) theory to carry normative commitments, we have to explain both what it is for something to be normative and what it is for a theory to carry something as a commitment. We could kill both birds with one stone if only we could rely on Hume's Law:4

(H) There is no logically valid argument with only non-moral premises and a normative moral conclusion.

If we could rely on Hume's Law, then we could suppose that a commitment of a theory was anything it logically entailed. And since we can identify with confidence at least certain normative statements, we could test others for normativity by checking to see whether they entailed any of the paradigmatically normative statements. If a meta-ethical theory turned out to logically entail a normative statement, then the theory itself would be a normative moral statement, by Hume's Law. Do we need to worry about what exactly a 'normative moral statement' is? Here we are interested in normative moral commitments, and there are plenty of paradigms. Rather than providing a list, I'll assume that basic, paradigmatically moral statements are predications of moral predicates, especially 'wrong', 'right', 'morally permissible', 'evil', 'good'. As long as we can rely on Hume's Law, we can check other statements for normative moral status by seeing whether they can be used as premises in arguments (all of whose other premises are clearly descriptive) with paradigmatically moral conclusions.

If we are going to answer the question about meta-ethical theories in a fully rigorous way, we need to characterize meta-ethical theories rigorously. I doubt that the common philosophical understanding of meta-ethical theory is capable of sharp definition, though, so I will focus on a certain central core of meta-ethical theories, which I will call 'Standard'. A Standard meta-ethical theory is an instance of the schema

(Standard)□ (x) (Mx ↔ Dx)

where 'Mx' stands in for a "moral formula" and 'Dx' for a "descriptive formula". By a "moral formula" I mean one whose instantiations with ordinary names (or other singular terms) are moral sentences. A descriptive formula is one that is not moral; typically a Standard theory will have a clearly descriptive, clearly non-moral component for 'Dx'.

4 Hume's Law is a statement in the form of a modal principle: if no moral premise leads to a moral conclusion, then no non-moral premise can lead to a moral conclusion. This is widely considered to be a foundational argument against the possibility of deriving moral conclusions from non-moral premises.

[Note: The author's reference to Hume's Law appears to be in error. Hume's Law is actually a principle related to causality, not normativity. The correct principle for discussing normative commitments is more likely to be Moore's observation that 'good' cannot be defined in terms of descriptive statements.]
For example, here is familiar sort of contractualism expressed in Standard form:

\[(\text{Contract})\Box (x) \; (x \text{ is wrong } \leftrightarrow x \text{ would be disallowed by any set of rules which could be agreed to by all members of society})\]

Any instance of (Standard) logically implies some very good candidates for normative moral statements. For each instance implies an instance of

\[D \to M\]

I'll call such instances, \(D \to M\) conditionals. These are typically very good candidates for normative moral statements just on intuitive grounds. For instance, consider

(2) If Peter eats meat, then he is morally reprehensible.

Hume's Law gives us another excellent reason to think of \(D \to M\) conditionals as invariably normative moral statements, since together with a descriptive premise \((D)\) they always entail a moral conclusion \((M)\). Indeed, one might even say that this is the whole point and function of a \(D \to M\) conditional: it is an inference ticket granting its bearer passage from a descriptive premise to a normative moral conclusion.

So if we could rely on Hume's Law, we would have a quick and satisfying argument to the conclusion that all Standard Theories carry normative commitments. Unfortunately, we cannot rely on Hume's Law.

B. Hume's Law Refuted

The following alarmingly simple refutation of Hume's Law is due to A. N. Prior (Prior 1960).

Let us suppose, as seems safe, that the negation of a non-moral proposition is itself non-moral. At the very least, it seems reasonable to suppose that if a sentence contains no moral vocabulary at all, then it is non-moral. In the demonstration below, we may assume that \(D\), a non-moral proposition, is chosen so as to contain no moral vocabulary; so neither does its negation; so its negation is also non-moral.

We show that Hume's Law is false by Cases.

Take some moral proposition, \(M\), and some non-moral one, \(D\). Consider their disjunction, \(D \lor M\). Is it moral, or non-moral?

(i) Suppose it is moral. But it is logically entailed by \(D\). So Hume's Law is false.
(ii) Suppose it is non-moral. But together with \( \neg D \), it entails \( M \). So Hume’s Law is false.

We cannot rely on Hume’s Law. So we cannot use it to decide whether a given complex statement is a normative moral statement, and in particular we cannot use it to decide whether a \( D \rightarrow M \) conditional is a normative moral statement. We need another approach.

C. Hume’s Kernel Salvaged?

We could try a conservative revision, fixing up Hume’s Law to save the kernel of truth while avoiding Prior’s refutation. One such revision is due to Tommas Karmo (Karmo 1988). Karmo suggests that we classify a sentence as moral (‘ethical’, he says) at a possible world, \( w \), if and only if it is true at \( w \) according to one moral standard, and false at \( w \) according to another. For the notion of a ‘moral standard’, we consider the class of uncontroversially moral sentences. And for our purposes, this could just be the class of atomic sentences that are predications of moral predicates. (The characterization of moral predicates we leave open. While there are controversial examples, we will be concerned only with the uncontroversial ones, like ‘is wrong’, ‘is morally good’. ) A moral standard, then, is (or more intuitively, is determined by and determines) a consistent assignment of truth-values to uncontroversially moral sentences.

With Karmo, we classify sentences as moral only at a world. An example explains why.

Consider

(3) Benito is evil or New Zealand is a Communist Republic.

In our world, this sentence counts as moral, because it is true according to any moral standard that assigns the value True to “Benito is evil” and false according to any other moral standard. But in a possible world in which New Zealand is a Communist Republic, (3) is not moral, because it is true according to any moral standard. Intuitively, Karmo’s scheme says that a proposition is non-moral (at a world) if and only if you can tell whether it is true (at that world) without any moral investigation.

A counterintuitive feature of Karmo’s taxonomy is that it fails to close the class of moral sentences under converse entailment (and fails to close the class of non-moral statements under entailment). That is, it may happen that although a sentence is moral, it is entailed by a sentence which is not moral. How this can happen should be obvious; (3), counted as moral at our world, is entailed by “New Zealand is a Communist Republic”, which is not moral. In our context this counterintuitive feature cannot be counted as an objection, though, since it amounts to the point that Karmo’s taxonomy does not respect Hume’s Law.
So, as Karmo organizes things, Hume's Law fails. But a relative succeeds:

(K) There is no sound argument with only non-moral premises and a moral conclusion.

Karmo's Law, together with his classification of statements, manages to avoid Prior's refutation by making sure that at least one true premise in any "bridge-crossing" valid argument will be false. For example, the valid argument from "New Zealand is a Communist Republic" to (3) has a false premise in this world. Bring the same argument over to a possible world in which New Zealand is a Communist Republic and the argument becomes sound, but the conclusion gets classified as non-moral relative to that world. If we consider instead the disjunctive syllogism from (3) and "New Zealand is not a Communist Republic" to the conclusion, "Benito is evil", we find that at our world it may be sound (it is sound just in case Benito is evil), but that (3), one of its premises, is counted as a moral statement. Bring the argument to a world in which New Zealand is a Communist Republic and the premises will all be non-moral, but the argument will no longer be sound.

This particular salvage does not seem very helpful for our purposes. For one thing, Karmo's classification counts Newtonian mechanics as having moral consequences, merely because Newtonian mechanics is false. (All false statements will have Karmo-moral consequences.) To see this, consider the following material conditional, which is a logical consequence of Newtonian mechanics:

(Y) If Newtonian mechanics is false, then eating yams is morally wrong.

(Y) is plainly a Karmo-moral implication of Newtonian mechanics.

We were hoping for a sense of carrying moral commitment in which metaethical theories do and Newtonian mechanics does not carry any. We won't get such a sense by letting moral commitments be Karmo-moral implications.

There is another mismatch between the intuitive idea of something's carrying moral commitment and its Karmo classification. The idea of an assertion 'committing' one to something is not quite the same as the idea of a proposition's having the something as an implication. The examples I'm thinking of are relatives of Moore's Paradox. They have nothing in particular to do with moral commitments, but they illustrate the more general phenomenon. Moore noticed that

(M) It is raining, but I don't believe that it is raining.

has the flavor of a contradiction, even though it is perfectly possible that it be true (on any occasion of utterance). I suggest that this flavor of contradiction has to do with the commitments carried by an assertion of (M). A contradiction implies every proposition, it is (maximally) too strong a claim to be true. (M)
is not like that. But (M) overcommits one, it commits one to too much, it is too strong to be asserted. The relative I want to consider is

(N) Either it is raining, or I don’t believe that it is raining.

(N) stands to (M) roughly as a tautology stands to a contradiction. Where (M) is maximally overcommitting, (N) is so noncommittal as to be trivial. Anyone could assert (N), no matter what his beliefs. Suppose (following (Stalnaker 1978)) we think of a conversation as having the primary function of informing the various interlocutors about the beliefs of the others. Then (N) fails to inform. We start out, let’s suppose, with no idea of what other conversants believe. My representation of George’s beliefs is the big set of all the possible worlds—not that I take him to have such weak beliefs, but that I am leaving open all the belief sets he might have. I do not, from the beginning, eliminate any possibilities. When George makes an assertion I use it (assuming I take the assertion to be sincere) to whittle down the set of possible worlds that might be George’s belief set. When he asserts that it is raining, I rule out all of those worlds in which it is not raining. But when he asserts (N), I can’t rule anything out. Can’t I? Can I rule out this possibility: that George believes it is not raining, but believes that he believes that it is raining? Such a failure of transparency in belief is hard to imagine, but I don’t mean to take a stand on the issue here. As will emerge in the remainder of the paper, what matters to the conception of a ‘moral commitment’ (of an assertion) is what a person’s views must be in order for the person to be willing to make the assertion. So the point of the analogy is that George should be willing to assert (N) whether he believes that it is raining (since in that case he can infer (N) from its first disjunct) or not (since in that case he ‘safely’ asserts something true). George can recognize that (N) cannot commit him to anything false that he does not already believe. So (N) might be considered especially noncommittal, even though it does imply something substantial (namely, that George does not falsely believe that it is raining).

Moore’s Paradox and its relative have to do with assertions about belief. The main groundwork that I want to lay, though, has to do with a different conception of the commitments of an assertion. The type of examples I’ll focus on are drawn from a different philosophical literature, the literature on ‘analytic contingencies’.

II. Assertions and Their Commitments

Let’s start with this example.

(4) Jamie’s first grade teacher is identical to Jamie’s actual first grade teacher.
(4) certainly does not express a necessary truth. What it says is not true at every possible world. For there is a world at which Jamie’s first grade teacher is Arnold Schwarzenegger, and Arnold Schwarzenegger is not identical to Jamie’s actual first grade teacher (who was, for the record, Mrs. Proctor). So the proposition expressed by (4) is false at such a world. Of course, the proposition that would be expressed by (4) if it were uttered at such a world is true. Still, (4) actually expresses some proposition. Let’s use sets of possible worlds as (to represent) propositions. Then (4) expresses a set containing all those worlds which are like the actual one in respect of who teaches me in first grade (that is, all those worlds in which Mrs. Proctor is my first grade teacher).

Now this proposition implies, in the sense of strict implication, that my first grade teacher is a woman (taking a person’s sex as essential to her, controversially but harmlessly for illustration). But there is a fairly ordinary sense of an assertion’s committing one to something in which asserting (4) does not at all commit anyone to the proposition that my first grade teacher is a woman. In fact, in this same fairly ordinary sense, asserting (4) is perfectly noncommittal. One might assert it, you could say, without fear of contradiction. How exactly do commitment and implication come apart?

We may appeal to some apparatus developed by theorists at the intersection of semantics and pragmatics. For many purposes we can represent propositions by sets of possible worlds: intuitively, a proposition is represented by the set of all worlds at which that proposition is true. Assuming (pretending, really) that we have a standard enumeration of the possible worlds, we can think of a proposition as a row of cells each containing a ‘T’ or an ‘F’, depending on whether the proposition in question is true or false at the world corresponding to that cell. Clearly, a row-style-proposition is recoverable from a set-style-proposition, and vice versa.

Some sentences express different propositions in different contexts. The sentence

(5) Yesterday was Wednesday.

expresses different propositions when asserted on different days. Some of these propositions are (timelessly) true, and others are false. If I utter the sentence on Tuesday, what I say is false, and what I said continues to be false in the days ahead. Of course, the very same sentence in some sense turns true two days later, but the proposition I expressed doesn’t.

We can’t represent the full semantic value of a sentence by a set of possible worlds. For compare:

(6) Jamie’s first grade teacher is a woman.
(7) Jamie’s actual first grade teacher is a woman.

The first of these sentences could be represented by the set of all worlds in which I have a woman for a first grade teacher. The second cannot, though. I
suppose the proposition expressed by the second is the set of all worlds at which Mrs. Proctor exists. (As long as Mrs. Proctor had existed, she would have been a woman.) But this set of worlds cannot represent the semantic value of the sentence, for if that very same sentence had been uttered by me at a world in which Arnold Schwarzenegger were my first grade teacher, it would have expressed a proposition containing no worlds at all.

We may represent sentences containing indexicals, like ‘actually’, by two dimensional matrices. Each row in the matrix will be a proposition, so each column will be labeled by a world. A given row gets filled in with ‘T’s and ‘F’s depending on whether its proposition is true or false at the world labeling the column. The rows are labeled by contexts of utterance. A sentence expresses a proposition at one context, and maybe another at a different context. What are contexts? For our immediate purposes, we may include just worlds, again, as important features of context.

Then for simplicity let’s restrict our attention to just three worlds. At world $i$, Jamie’s first grade teacher is Mrs. Proctor. At world $j$, Jamie’s first grade teacher is Arnold Schwarzenegger. At world $k$, Jamie’s first grade teacher is David Kaplan. Here is the matrix representing (4).

<table>
<thead>
<tr>
<th></th>
<th>$i$</th>
<th>$j$</th>
<th>$k$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$i$</td>
<td>T</td>
<td>F</td>
<td>F</td>
</tr>
<tr>
<td>$j$</td>
<td>F</td>
<td>T</td>
<td>F</td>
</tr>
<tr>
<td>$k$</td>
<td>F</td>
<td>F</td>
<td>T</td>
</tr>
</tbody>
</table>

What is salient about this matrix is that its diagonal is full of T’s. Following Stalnaker let’s speak of the strip of T’s and F’s along the diagonal of a matrix as the diagonal proposition of the given sentence. The diagonal proposition of (4) is the necessary proposition. (Again, we are ignoring worlds at which I have no first grade teacher, for simplicity.) Since we’d decided that (4) is an especially noncommittal sentence to assert, we might speculate that its vacuity derives from the necessity of its diagonal. And I think this is correct.

Some theorists have said that we think of sentences as (expressing?) things knowable a priori exactly when those sentences have necessary diagonal propositions. Without going into subtleties, the reason for putting it this way is fairly simple. Without knowing anything about which world I am in (without knowing anything about the world that I don’t already know about all possible worlds), I can deduce that such a sentence is true in my world. This is a bit odd, since I cannot deduce a priori the proposition that the sentence expresses. The suggestion is that it is really sentences that are knowable a priori, or perhaps a-proposition-relative-to-a-sentence (see (Wong 1996)). Whatever plausibility accrues to the proposal, I propose to commandeer it for a sense of
an assertion's committing one to something. Think of it this way. Suppose I assert a sentence with this matrix:

\[
\begin{array}{ccc}
  i & j & k \\
  i & F & F & F \\
  j & F & T & F \\
  k & T & F & T \\
\end{array}
\]

And suppose for a moment that we are in fact at world \(k\); \(k\) is the actual world. One might think that my commitment is: we are at world \(i\) or world \(k\). For those are the worlds at which the proposition I asserted is true. But this seems wrong. Am I really committed to our not being at world \(j\)? Is my commitment really fulfilled if we are at world \(i\)? If you could convince me that we are at world \(i\), then I would be forced to withdraw my assertion. For if we were at world \(i\), I would have just asserted something false. So in that sense, I am committed to our not being at world \(i\). Suppose instead you were to convince me that we were at world \(j\). I should not want to withdraw my assertion, then, for what I asserted would be true. And similarly if I become convinced that we are at world \(k\), I will not withdraw my assertion. A reasonably natural way to put it is this: I would withdraw my assertion if and only if I came to believe that in my world, what I asserted was false. I would stand by my assertion if and only if I thought that in my world, what I asserted was true. (This way of putting it is ambiguous. The more formal way of putting it is univocal.) Formally, I would withdraw my assertion just in case I came to believe that I was at a world with an \(F\) at the diagonal proposition of the sentence I asserted, and stand by if I came to believe that I was at a world with a \(T\) at the diagonal. So it does seem right to say that my commitment is precisely to the set of worlds true along the diagonal of the sentence I assert.

III. Moral Commitment

A. Informally

Investigating the idea of ‘moral commitments’, we concluded that the notion of a statement’s having moral implications was not really what we had in mind. We concluded, for example, that Newtonian mechanics has moral implications (merely by being false).

Though it is not entirely useless for us, Karmo’s classification fails to capture a crucial sense of a statement’s being moral, or having moral implications. Suppose that you hear me assert confidently:

(8) Either tea drinking is common in England, or all New Zealanders ought to be shot.
Knowing me as you do, you won’t jump to any hasty conclusions. You ask me my grounds for the outrageous assertion, and I explain that I know very well that tea drinking is common in England. Then you won’t draw any conclusions about my moral views. You will take my assertion to carry no moral commitment; my ground for it is entirely non-moral. Of course, Karmo’s scheme does count this statement as non-moral relative to our actual world. But now suppose you hear me assert:

(9) Either the senator is a Cherokee or all New Zealanders ought to be shot.

Again you inquire, and again I explain my grounds: I am extremely confident the senator is a Cherokee. Then again you take my assertion to carry no moral commitment. But this time Karmo’s scheme classifies it as moral (since the senator is not, in fact, a Cherokee). The difference is that your assessment depends on what world you think I take us to be in, whereas Karmo’s depends only on what world we are in. This is not to disparage Karmo’s classification. Indeed, it has pointed us to a fairly satisfactory solution. Asserting a proposition, we might say, is committing oneself morally whenever that proposition counts as Karmo-moral relative to the world the speaker believes he is in.

There are a number of complications involved in filling out this account. For one thing, a speaker does not generally believe that he is in a particular world. To believe that I am in some particular world is to have unimaginably detailed beliefs. Rather, speakers’ beliefs are better representable by a set of worlds. Then we might consider whether a sentence that is Karmo-moral relative to some of my belief worlds but not others is morally committing or not. But I will not consider that question. Second, we should admit that we are idealizing people’s epistemic states by representing their total system of beliefs even by sets of possible worlds. Thinking of beliefs that way is taking them to be more systematic than real people’s beliefs are. A person might believe two things which are incompossible; then we would represent his ‘belief system’ by the empty set. But this looks too crude. It just seems wrong to say of such a person that he believes everything. Again, I will ignore this problem, and continue to use the idealization.

Most important for our purposes, the proposition asserted does not seem to be quite right as the measure of a speaker’s commitments, moral or otherwise. We have just seen why not. If someone asserts that Jamie’s first grade teacher is identical to Jamie’s actual first grade teacher, we should not take her to be committed to the proposition expressed, but to the diagonal proposition of the matrix of the sentence spoken. For, as we noted, the assertion does not seem to be at all committing. Under no circumstances would the speaker be required to withdraw it.

To measure moral commitment, we need something like the diagonal of our matrices. But the matrices are not exactly the ones we use to represent the
semantic values of ordinary, non-moral sentences (those containing no moral vocabulary at all).

B. Formally: Moral Systems in Place of Possible Worlds

To begin with we dispense (for the time being) with possible worlds, focusing on a different aspect of context and a different index of truth evaluation. Remember that on Karmo’s scheme, a statement is counted moral (relative to a world) iff it is true according to some moral standards and false according to others (at that world). This is the test that we’ll use for a statement’s being morally committing, too. The rough idea is that if I assert a statement that is true relative to some moral standards but not others, you will be able to draw some conclusions about my moral standards. I will have committed myself to standing by one or another of those moral standards that count the statement true. So as you listen to me make assertions, you will be able to narrow down the class of moral standards that could make my assertions true.

Since (to start with) we will ignore the plurality of possible worlds, we are really interested in another feature of contexts of assertion: the speaker. Here, of course, the idea is that the same sentence might express now one proposition, now another, depending on who utters it. We combine these two variable features (moral standards, delivering truth values from moral propositions, and contexts, delivering propositions from indexical sentences) in a matrix to represent the semantic value of a moral sentence. For example, here is the matrix for the sentence,

\[
\begin{array}{ccc}
  m_1 & m_2 & m_3 \\
  a_1 & F & T & F \\
  a_2 & T & F & F \\
  a_3 & T & T & F \\
\end{array}
\]

Here the \( a_i \) are speakers, and the \( m_i \) are moral standards. And in particular, we’ll let \( a_1 \) be G. E. Moore, \( a_2 \) be Ray Charles, and \( a_3 \) be the Pope; and \( m_1 \) counts as wrong making fun of the blind and the Catholic, \( m_2 \) counts making fun of philosophers and Catholics wrong; \( m_3 \) permits making fun of anyone. We get a moral claim, something that is true or false according to various moral standards, given a context, and in particular the speaker in the context. The context supplies a referent for ‘me’. The standards provide a truth valuation.
C. Morally Noncommittal Assertions Are True on the Diagonal

In the matrix above there was no particular connection between the indexing of the moral standards, the \( m_i \), and the indexing of the persons. The subscript of the one was not particularly related to the subscript of another. We could have written the indices in any old order. But now I want to impose another restriction.

Let us arrange the matrices so that for each \( i, m_i \) is the moral standard held by \( a_i \). What exactly does this mean? What determines what your moral standard is or mine? I don’t want to give any developed theory. The moral standard of the speaker is supposed to be analogous to her ‘belief world’. In practice, a speaker hasn’t any particular world in mind as the actual one, but a set of worlds; and in practice, we don’t have full moral standards but only partial ones. But we are interested in the commitments of a speaker. We are interested in explaining in what way her assertions commit her to some subset of all the possible moral standards. We might think of a speaker’s particular moral standard as being that standard which she would come to accept in reflective equilibrium.

Now consider the sentence:

\[
(11) \text{Abortion is wrong if and only if it is counted wrong by my moral standard.}
\]

The matrix for this sentence is:

\[
\begin{array}{ccc}
& m_1 & m_2 & m_3 \\
\hline
a_1 & T & T & F \\
\hline
a_2 & T & T & F \\
\hline
a_3 & F & F & T \\
\end{array}
\]

We’ll specify only that \( a_1 \)’s moral standard (which is standard \( m_1 \)) permits abortion, as does \( a_2 \)’s, and \( a_3 \)’s does not. So the first row is the moral proposition that abortion is wrong if and only if it is counted wrong by \( a_1 \)’s moral standard; this is counted true (because the right side simply is false and the left side is counted false) by both \( m_1 \) and \( m_2 \), and counted false (because the right side remains simply false, while the left side is counted true) by \( m_3 \). And similarly for the other two rows.

The diagonal of the matrix is all T’s. So, I say (for the moment), the sentence itself (the assertion of it) is morally noncommittal. Why is this the proper criterion? A person may confidently assert a sentence whose diagonal moral proposition is true, no matter what her moral standard. Even if she is not quite
settled about what her moral standard is, she can be sure that whatever it turns out to be, what she has said will be true according to that moral standard. She has incurred no unwanted moral commitment.

We could think of it this way. Suppose my moral standard is standard M, one which strictly forbids abortion. Then by uttering the sentence, I have said that abortion is wrong if and only if it is wrong according to standard M. (This is not quite right, but it is good enough for the moment, since we are ignoring the plurality of possible worlds.) Well, abortion is wrong according to standard M. So by committing myself to the biconditional, I am committing myself to abortion's being wrong. But since it is wrong, according to my moral standard, that is a commitment I am happy with. From the perspective of the audience, things are a bit complicated. Supposing that you don’t know what my moral standard is, you don’t know exactly what moral standards make the moral proposition I have uttered true and which ones make it false. Since in fact my moral standard permits abortion, all and only those moral standards which permit abortion make that moral proposition true. But you can’t tell that just by hearing me. Now, if you learned more about my moral standard, you would discover that what I said is true according to just those moral standards according to which abortion is permissible. Then you could specify my moral commitment. You could narrow down the class of moral standards to which I am committed by pruning away all those standards that prohibit abortion. But of course, if you were able to learn that much about my moral standard, that is exactly what you would have learned! You would have been able to prune away all the abortion-prohibiting standards from my commitment set. So my assertion adds nothing. That is why it is perfectly noncommittal. Or in any case, it is perfectly noncommittal as long as we do not consider other possible worlds. The picture changes when we start to consider them. The next section explains how the picture changes.

D. Boggling Dimensions: Worlds and Moral Standards Together

We have (for a short while) been ignoring the plurality of worlds. Our matrix, the way we represent the semantic value of a moral proposition, includes no dimension for possible worlds. But we will want to represent modal moral sentences, things like

(12) Although capital punishment is wrong, it could have been permissible.

So now we have to complicate our model. We could add two dimensions, a dimension of possible worlds each for context and evaluation, and we would have four dimensional matrices. But we can just add one, with a simplifying trick that keeps the dimensionality of the matrices small enough to visualize.

We’ll think of a moral proposition as delivering a proposition (a set of possible worlds, or an assignment of truth values at each world) given a moral
standard. So it is a propositional function, taking moral standards as arguments. Then a moral proposition is itself a matrix, with worlds as one dimension and moral standards as the other. And for contexts we’ll take not speakers, or worlds, but speakers-at-worlds (or if you prefer, ordered pairs of worlds and speakers at those worlds). Then a sentence containing moral terms gets represented by a three dimensional matrix. A cell in the matrix is specified by a context, a world, and a standard, and is filled in with a truth value. Though three dimensional matrices are not so easy to visualize, and the graphic value of using them as heuristics is thus smaller than that of two dimensional matrices, we can think of their dimensions as Height, Depth, and Length. Each horizontal row is a proposition, so the worlds (of evaluation) are arranged along the Length. Each row is a section of both a horizontal and a vertical slab; a slab is a two dimensional matrix. Vertical slabs are moral propositions. A moral proposition delivers a proposition given a moral standard; so the moral standards are arranged along the Depth. And contexts are arranged along the Height; each context picks out a slab, a moral proposition.

Next, ordering. Contexts are speakers-at-worlds. For each \( i \), let \( w_i \) be the world in \( c_i \), the \( i \)th context, and \( m_i \) be the moral standard of the person-at-world \( i \). This means that the sequence of worlds along the Length contains repetitions of worlds. Worlds will occur at more than one position. For there will be many contexts containing the same world (one for each speaker, if we are to be complete). Formally there is no real objection to this feature, though it does prevent us from thinking naturally of the rows of cells as propositions in the most straightforward way. But the rows are still easily convertible to (and from) sets of possible worlds, which is all that matters technically. Moral standards recur, too, in their sequence, since two speakers-at-worlds might have the same moral standard.

Now, what is the formal characteristic of a sentence’s matrix that marks it as morally noncommittal (to assert)? If any old speaker at any old world could assert the sentence sincerely, no matter what her moral system, then assertion of the sentence reveals nothing. So if, for each world \( w^* \) and speaker \( a^* \), the moral proposition (a slab having Height and Depth but no Length) expressed by the sentence at the context \( a^* \) at \( w^* \) is true at \( w^* \), according to the moral system held by \( a^* \) at \( w^* \). Hard to parse. But not so hard to visualize; morally noncommittal sentences have T’s all along their major diagonal. (Which diagonal is the major one? We haven’t said which corner we start with. One corner is the origin, labeled by the first context, the first world, the first moral standard. The major diagonal runs from that corner and cuts all three dimensions.)

By way of illustration, here is a morally noncommittal sentence.

\[
\text{(13) Necessarily, abortion is wrong if and only if my actual moral standards count it wrong.}
\]

In a context (that is, relative to a speaker at a world), this sentence expresses a moral proposition (a slab, or a function from moral standards to plain old prop-
ositions). Suppose it is spoken by Karl at world \( w_k \). Then it expresses the moral proposition that takes moral standards to the proposition that necessarily, abortion is wrong-according-to-the-given-standard iff Karl's moral standards at \( w_k \) count abortion wrong. Then this moral proposition takes Karl's own moral standards, \( m_k \), to the proposition that necessarily, abortion is wrong according to \( m_k \) iff abortion is wrong according to \( m_k \). Since the embedded biconditional is indeed true at every world, the proposition is true. We conclude that it is true independent of the choice of \( k \) (and independent of any facts about Karl); then the matrix for the sentence is true all along the major diagonal. So it is morally noncommittal. And intuitively this seems right, too. For no matter who uttered it, no matter what his moral standards, no matter which world he turns out to be in, under no circumstances would he be compelled to withdraw his assertion.

Although (13) is an example of a morally noncommittal sentence, it is not a theory. With technical apparatus in hand, we now return to meta-ethical theories, including some that are, by our criterion, morally noncommittal.

IV. Secondary Quality Theories

Here is a simple version of what we earlier called Secondary Quality theory.

\[(SQ) \Box (x) (x \text{ is wrong } \leftrightarrow x \text{ tends to cause in us a feeling of moral disapproval})\]

I leave vague the terms ‘us’ and the expression ‘feeling of moral disapproval’. Particular versions of the Secondary Quality theory will arise by specification. I believe that the assertion of a theory in the same family as (SQ) is not morally revealing. Let me start with a clearer case.

\[(SQ^*) \Box (x) (x \text{ is wrong } \leftrightarrow x \text{ is wrong according to my moral standards})\]

Here, one might think, is a standard meta-ethical theory whose assertion lacks moral commitment. For instantiations will yield biconditionals which in turn imply basic moral judgments only when the biconditionals are combined with propositions about my moral standards. All we could infer about a person's moral commitments, from the fact that she asserted (SQ*), would be that her moral standards say that a certain thing is wrong just in case it is wrong according to her moral standards. And this is a logical truth. Not to say that the instantiations are themselves logical truths! But the implications of her sincerely asserting (SQ*), or its instantiations, are vacuous.

One might think so. But only if one has ignored the necessity operator of (SQ*). In the previous section we went to some trouble to combine the moral with the modal, and found a way of characterizing the morally noncommittal modal assertions. We now apply the formalism.
Without harm, though not strictly correctly, let's call such sentences as the following 'instantiations of (SQ*)':

(SQ*i) □(eating meat is wrong ↔ eating meat is wrong according to my moral standards)

Suppose that according to my actual moral standards, eating meat is not wrong. Now consider a world, w, in which I have standards according to which eating meat is wrong. I assert (SQ*) in the actual world. I am thereby committed to (SQ*i). (SQ*i) is true if and only if

(14) Eating meat is wrong ↔ eating meat is wrong according to my moral standards.

is true in every world, including w. By hypothesis, the right hand side of (14) is true in w. So I am committed to the left hand side's being true in w, too. But this does appear to be a substantive moral commitment. To put the same point another way, consider

(15) If my moral standards prohibited eating meat, then eating meat would be wrong.

(SQ*) entails (15). If you discover that I am committed to (15), you have, intuitively, discovered something substantial about my moral standards. And this fits our formal characterization, for (15) is not true all along the (three dimensional) diagonal of its matrix. It is false, in particular, when the context is a speaker, Gilbert, at a world, the actual one, and it is evaluated at Gilbert's moral system and the actual world. For Gilbert's actual moral system permits eating meat in any nearby world, including those nearby worlds in which he would have a moral system that prohibited eating meat. (He regards such worlds as worlds in which he has a mistaken, or overly fastidious, moral view.)

Notice that (SQ*i) is itself false at some points along its diagonal. To see this, note that it will be false at any context and moral system at which (14) expresses a proposition that is not necessarily true. Take as a context, again, Gilbert at the actual world, and as a moral system Gilbert's actual moral system, MG. Then the proposition expressed by (14) is true at just those worlds, w, in which Gilbert has a moral system that agrees with MG on the morality of eating meat in w. And this means that it is false at any world at which Gilbert has a moral system that forbids the eating of meat. And there are, of course, many of these. So (14) does not express a necessary proposition, relative to the context and moral system in question. And that means that (SQ*i) is false at such a context and system, necessarily false; it expresses a proposition that is false everywhere. So its matrix is false at that spot on its diagonal.
We have just shown that contrary to first impressions, (SQ*) does carry normative commitment (it is morally revealing). But we can now remove that commitment by a simple device. Consider

\[(ASQ) \square (x) (x \text{ is wrong } \leftrightarrow x \text{ is wrong according to my actual moral standards})\]

(ASQ) is the ‘actualized’ version of a Secondary Quality theory. The standards denoted on the right hand side of the biconditional formula are my actual ones. Even when we evaluate instantiations of this formula at other worlds, the standards we use are the ones to which I actually subscribe, not the ones to which I subscribe at the world of evaluation. And now the normative moral commitment of the theory has disappeared.

(ASQ) is true at every point of its matrix’s diagonal. For, when evaluating its truth on the diagonal, we will always select the same moral system for both sides of the biconditional. For the left hand side, we will always use the system of the context (the speaker’s system in the world of the context). And for the right hand side, we are also constrained to use that same system; because of the ‘actual’ we will not go to the world of evaluation to find the speaker’s system at that world. So for each choice of x and each world, the two sides of the biconditional will say the same thing, so the biconditional will be true everywhere, satisfied by every object.

To see in a more intuitive way that an assertion of (ASQ) lacks all moral commitment, ask what commitments are compatible with (ASQ). Take some arbitrary normative commitment I might have, to the effect that \(\phi\)-ing is wrong in some possible world, \(w\). So \(\phi\)-ing is wrong in \(w\), according to my moral standards. I accept (I believe, I am willing to assert) that in \(w\), \(\phi\)-ing is wrong; I believe also that it is wrong according to my moral standards. Since I accept both sides of the biconditional, “\(\phi\)-ing is wrong \(\leftrightarrow\) \(\phi\)-ing is wrong according to my actual moral standards,” I accept the biconditional. Given an arbitrary world, and an arbitrary \(\phi\), then, (ASQ) entails nothing incompatible with my normative view about \(\phi\)-ing in \(w\). But such entailments are the only implications that could conflict with my normative views. So my normative views do not conflict with (ASQ). No matter what my normative views, I should be willing to accept (ASQ). Then my acceptance of (ASQ) reveals nothing about my normative views; it rules nothing out. So it is not morally revealing.

(ASQ) is a standard meta-ethical theory whose assertion is not morally revealing. So, I say, there is a good sense in which it carries no normative moral commitment. True enough, there is also a perfectly good sense in which it has normative moral implications (Karmo’s sense—there are certainly sentences it entails which are moral in this world). Nonetheless, it could not reasonably be called a normative moral theory, just because its assertion carries no normative moral commitment.
A. Dworkin's Mistake

Dworkin points out one fact we have noted above, namely, that (SQ*) does carry moral commitment. But he also insists that even (ASQ) is a substantive moral thesis. In this he is mistaken.

About a theory roughly equivalent to (SQ*), he writes,

But someone who holds that moral properties are secondary properties does take sides in actual or potential substantive disputes. Suppose we discovered that, contrary to our expectations, contemplating genocide does not in fact outrage even most normal people. Genocide would not then be morally wrong on that dispositional account, though, of course, many people would think it was. [101]

We have already seen that Dworkin is right about this. But then he considers a theory roughly equivalent to (ASQ), and he writes,

The dispositional account might, it is true, take a different form. It might hold, for example, that what makes genocide wrong is the reaction, not of whichever kind of people happen to exist from time to time, but of us, that is, of people with the physiological structure, basic interests, and general mental dispositions that people actually have now. In that case, it would no longer follow that genocide would cease being wicked if human beings developed very different general interests or different neural wiring. But some plainly substantive and controversial claims would still follow: for instance, that genocide would not have been wicked if economic or other circumstances had been different as human reactions evolved, so that creatures with our general interests and attitudes had not been revolted by genocide. [102]

But this is wrong. For imagine a possible world in which human beings evolved differently, so that human beings with our general interests and attitudes had not been revolted by genocide. Would it be true, at that world, that genocide tends to cause in us as we actually are a feeling of moral disapproval? Of course. We can even do the experiment ourselves: don't you feel moral disapproval of any genocide that might occur in such a world? Of course; for according to your actual moral standards, genocide is morally wrong.

Dworkin continues,

[Just as any philosophically illuminating account of what the disgustingness of rotten eggs consists in yields counterfactual claims about the circumstances in which rotten eggs would not be or have been disgusting, so any illuminating account of
moral properties as secondary entails counterfactuals that state substantive moral positions. [102–3]

I can’t find any argument for this claim. So I think it is not supposed to be a claim that needs to be argued for at all, but rather a judgment about what would count as an “illuminating account”. Dworkin is of the opinion that any theory that lacked substantive moral entailments would thereby fail to be illuminating. Maybe he can’t quite believe that anyone would offer a theory like (ASQ), maybe he thinks it’s just obviously unilluminating. If so, I think he’s wrong about that.

B. How a Theory Could Be Illuminating

Let’s add a typical clause to (ASQ). It’s a weasely clause, notoriously so, but we’ll add it anyway, and call the result a Weasely Actualized Secondary Quality theory:

(WASQ) An act is wrong iff it tends to cause in standard perceivers as we actually are a feeling of moral disapproval.

The weasel: what makes a perceiver standard? We dodge this question.

Now, someone might say, “Of course, if a perceiver is ‘standard’, then he will certainly disapprove morally of all and only the wrong things. Otherwise he doesn’t count as standard.” And he might facetiously propose

(SQUARE) An object is square iff it tends to cause in standard perceivers as we actually are an impression of squareness.

The idea, of course, is that there is little question but that we can rig ‘standard’ in such a way that (SQUARE) comes out true. But this obviously doesn’t show that squareness is a projected property, that it is a feature contributed by our minds or perceptual abilities rather than by the properties that things have independently of us. (SQUARE) seems to explain nothing about the nature of squareness or squares. So why should (WASQ) be thought to explain anything about the nature of moral wrongness?

We might maintain that there is an important difference between (WASQ) and (SQUARE). For each is a biconditional, but there is a difference in the direction in which the biconditional is to be read; that is, the explanation runs in different directions in each case. In (SQUARE), most of us will agree that it is our being standard perceivers, along with the fact that the jewelbox is square, that explains why it tends to cause in us an impression of squareness. Whereas, it might be maintained, in the case of (WASQ), it is our being standard perceivers, along with extortion’s tendency to cause in us a feeling of moral disapproval, that explains the wrongness of extortion, explains in virtue of what
extortion is wrong. Or we could think of it this way. In the one case, we think there is some independent explanation of what squareness is, and then the notion of a standard perceiver of squares will just follow, empirically, from (SQUARE). Whereas in the other case, we might think that there is some independent explanation of what it is to be a standard moral perceiver, and then the notion of moral wrongness may be read off of (WASQ).

A theory of the nature of moral properties, even one which lacked altogether any substantive moral entailments, could be philosophically illuminating in just this way, that it would explain what moral properties are. It would say that they are dispositions. Moral properties, according to such a theory, would be disposition to affect us as we actually are in certain affective ways.

Dworkin is concerned in “Objectivity and Truth” to show that there are no Archimedean metaethical theories, perspectives from which one might pass unfavorable judgment on the objectivity of moral statements without oneself incurring any startling first order moral commitments. I have argued that he is mistaken. For (ASQ) looks to be a significant and (for all that’s been shown) coherent Archimedean position. No doubt there are some serious objections to (ASQ). And perhaps it is not entirely clear or rigorous exactly in what sense it might be thought to imply that moral judgments are not ‘objective’. But we cannot reject it as a failed attempt to establish an Archimedean theory. (ASQ) seems to say exactly what some Archimedean want to say about moral properties and facts.11

Notes

1. (Dworkin 1996), hereafter cited only with page numbers in the body of the text.
2. See (Dreier 1996), especially comments by Blackburn and Zangwill and Dworkin’s reply.
3. See (Davies 1980), where this sort of theory is discussed but not endorsed.
5. Though note that Humberstone’s taxonomy does not support this supposition. See (Humberstone 1982) and (Humberstone 1996).
6. See (Moore 1942 p. 543), and also (Moore 1962 p. 277).
7. This is Prior’s example.
8. See my (Dreier 1990) for details.
9. Due to (Davies 1980).
10. As I noted above, in the Introduction, Dworkin couldn’t mean all of our attitudes. For one of our attitudes is the attitude of revulsion at genocide. Presumably Dworkin means to be imagining that human beings have the attitude of revulsion, but not toward genocide.
11. Thanks to audiences at Arizona State University and the New School’s “Conference on Methods” where I read early versions of this paper, and also to Ronald Dworkin, who got me interested in the topic, and who will certainly disagree with the spirit of my position.
References


