Review of Frank Jackson, From Metaphysics to Ethics: A Defense of Conceptual Analysis

for Philosophical Books

Stephen Yablo, MIT

first draft, June 30, 1999

I. Introduction

I would start by saying that every analytic philosopher should buy this book, but I suspect that most of them already have. They got their money's worth. The book operates at two levels, developing a grand-scale methodological hypothesis while trying the hypothesis out on some ground-level disputes. The grand-scale hypothesis is that conceptual analysis, properly understood, is just as crucial to analytic philosophy as in glorious days gone by. The ground-level disputes are in various areas but especially mind, color, and morality. A theme throughout is that physicalists are committed to the a priori deducibility of truths about -- well, anything you like -- from physical premises.

A rough and selective guide to the book's contents: "Serious metaphysics" seeks to weave "a limited number of …basic notions" (4) together into a complete story -- a story implicitly containing (read: entailing) any truth you care to mention. Once that story is found, we have a

hoop we can ask candidate features of reality to jump through; for nothing is real unless the story entails it. This is the "entry by entailment" solution to the "location problem" for the given feature of reality. Physicalists think they can tell the story in physical terms, so they must hold that nothing is real unless its presence is demanded by what goes on physically.

Chapter 2 finds a place for conceptual analysis in all this. A concern with the location problem is a concern about "entailment theses between matters described in some preferred vocabulary and matters described in other vocabularies" (28). And what is conceptual analysis if not the study of inter-vocabulary entailments? Suppose that a physicalist wants to defend freedom of the will. Her job is to show that "it was done freely" is entailed by true physical premises. To do that, though, she will need to get clear on the conditions under which an action is to be described as free.

Described by whom? Not just anyone; the topic is "free action according to our ordinary conception" (31). This conception, a.k.a. the "folk theory" of freedom, needs to be teased out by consideration of possible cases. The teasing out is a delicate business. One-off intuitions about which things are in fact K have got to be weighed against other evidence, including "the theoretical role [subjects] give K-hood, signs of confused thinking, …, their readiness to back off under questioning, and the like…" (35) Like the empirical scientist, we seek "the hypothesis that best makes sense of [subjects'] responses" (36).

The exercise is not to be conceived as obnoxiously prescriptive. If our ordinary conception takes us in unhappy directions, we may seek to tweak it a little, as the compatibilist perhaps does with the folk conception of freedom. (This is "the sense in which we pay due homage to Quine's critique

of analyticity" (44).) But that tweaking is necessary just underscores the fact that we are dealing with sentences that "that could only come false by virtue of meaning change" (46). A sentence whose meaning guarantees its truth -- so that it is true on any hypothesis about which world is actual -- deserves to be called a priori. So there is also a sense in which Quine's critique is shrugged off.

Chapter 3 asks: How did we get from the idea that all truths have got to be entailed by "basic" truths to a conclusion about what is knowable a priori? Your average physicalist would much rather see the entailment relations as necessary a posteriori. And given the work of Kripke, there would seem to be nothing to stop him.

This is to read Kripke in an unduly mystery-mongering way, Jackson thinks. There is only one kind of necessity in Kripke; what there are two of is propositions expressed by a given sentence. One, the C-proposition, is the set of worlds, considered as counterfactual, that the sentence truly describes. ($||\text{water} = H_20||_C$ is the set of all worlds, or all containing water.) The other, the A-proposition, is the set of worlds such that the sentence is true if they are actual. ($||\text{water} = H_20||_A$ contains the worlds in which H_20 is the clear drinkable stuff of our acquaintance.) The alleged distinction between metaphysical and epistemic necessity is best understood as the distinction between the necessity (simpliciter) of the C-proposition and that of the A-proposition.

All right, but why can't our average physicalist say that the C-proposition expressed by [physical statement] P entails that expressed by [mentalistic statement] M, but the A-propositions do not relate in this way? She can, but this only buys her a little time. The reason has to do with the

nature of understanding. To understand a sentence is to appreciate how its C-proposition depends on context. If a speaker who understands $\underline{P}$ and $\underline{M}$ fails to see that $\underline{P}$ entails $\underline{M}$, that is because she doesn't don't know enough about context to figure out which C-propositions they express. When the gaps in his contextual information are filled -- as according to the physicalist can be done with a physical sentence $\underline{P}^*$ -- the speaker is out of excuses. Armed with $\underline{P}^*$, he can identify $||\underline{P}||_C$ and $||\underline{M}||_C$, whereupon she will see that every world in the one belongs also to the other. Hence the physicalist's commitment to the "in principle a priori deducibility of the psychological from the physical" (83)

Chapter 4 takes up the location problem for color. A "subject-determining platitude" about color is that "Yellowness is the property of objects putatively presented to subjects when those objects look yellow" (89). Since presentation requires causation, and dispositions are causally inert, dispositional theories of color -- including those that identify colors with reflectance-dispositions -- are dead on arrival. As we all know, "the only causes… of objects' looking yellow are complexes of physical properties" (93). And so yellowness must be a complex of physical properties.

A perhaps surprising consequence is that "the way [colors] look does not reveal their essential nature" (102). But "revelation" was never part of our folk theory of color anyway. It must be conceded too that, insofar as different physical properties make things look yellow on different occasions, there is no single property of yellowness shared by everything yellow. But on the one hand, not-excessively-disjunctive properties can be causes, so a common physical cause can be found more often than you might think. And on the other, it does no great violence to our folk conception of yellowness if it splits jadeishly into a small number of disparate types. If (but only

if) there were no "interesting distinctive distal commonalities underlying similarities of apparent colour" (112), color would have to be regarded as an out and out illusion.

The last two Chapters concern the location problem for ethics. Jackson assumes cognitivism as against expressivism: ethical sentences are in the business of "saying how things are," even if some of them for reasons of vagueness (or whatever) lack truth-value. He concludes that "ethical properties are descriptive properties" (117). It's a subject-determining platitude that nothing can be called "rightness" that fails to supervene on the descriptive. But then the disjunction of the total descriptive natures of all possible right actions is necessarily equivalent to rightness, and so, fussiness about property individuation aside, identical to it.

Which descriptive property is it? The one playing the rightness role in (ramsified) folk morality -- with the twist that it is not current folk morality that calls the shots but a "maturer" version. Reference is pegged to the place "where folk morality will [would] end up after it has been exposed to debate and critical reflection" (133). Jackson agrees with the Cornell realists that ethical properties are descriptive properties, and agrees (or can) that the identities here are necessary a posteriori. But he rejects their skepticism about the possibility of analyzing ethical language descriptively. It is analytic that an action is right iff it has whatever property D meets descriptive condition M, a condition extracted by the Ramsey-Lewis method from mature folk morality.

This might seem to make Jackson easy prey for the open question argument: "I see that keeping my promise has the M-property, but is it right?" He has two main replies. (1) Since "what matters is

the nature of mature folk morality, there will, here and now, inevitably be a substantial degree of 'openness' induced by the very fact that the rightness role is currently under investigation" (151).

(2) If the question still seems open when "all the negotiation is over and we have arrived at mature folk morality, we…are entitled to dig in our heels and insist that the idea that what fits the bill that well might fail to be rightness is nothing more than a hangover from the platonist conception that the meaning of 'right' is somehow a matter of its being mysteriously attached to the form of the right" (152).

II. Ethics

It says a lot for the book's unity and integrity that later chapters raise similar issues to earlier ones. Let me start with Jackson's reply to the open question argument and work backwards. (I'll be brief with the chapter on colour; see "Singling out Properties.")

If we accept reply (1), then we ground the openness of "is that right?" in the openness (for us today) of "is that mature folk morality?" Jackson occasonally talks as though this latter openness were a matter of ordinary descriptive ignorance about the future; we haven't got there yet, so we don't know. But of course morality could take a direction that we would regard as quite misguided, and for good and specific reasons; tomorrow's moralists might end up resting a lot on tendentious analogies, or giving in to self-interest, or etc. The point is just that it is no part of current folk morality to defer to whatever comes along. If deference must be paid, it is owed to those (probably hypothetical) future populations who had thought things through carefully, reasonably, and with

due concern for all. So the best (1) can hope to accomplish is to ground the openness of "is it right" in that of "is that the most careful, reasonable, etc. way of developing the theory from here?"

But "careful," "reasonable," “due concern” and so on are themselves evaluative terms. So Jackson faces a choice. Either he maintains about reasonableness too -- let it stand in for the rest -- that it is analyzable in descriptive terms, or he treats it as irreducibly evaluative.

Suppose that he sticks with his earlier descriptivism. "Reasonable" picks out whatever property plays the reasonableness-role in mature folk reason-theory, the theory we would arrive at if existing folk reason-theory were developed along the most reasonable lines. But this is circular; "the most reasonable lines" are the lines that would be identified as such by the theory lying at their terminus.[1]

Does it go better if we treat "reasonable" as <u>not</u> picking out a descriptive property? That might look like an impossible position; wouldn’t the same supervenience considerations that led Jackson to his descriptivism about the ethical apply here too? But the most that comes out of supervenience is that any worthwhile <u>candidate</u> for the role of being an evaluative property has to got to be a descriptive property. It's compatible with this that there are and will continue to be many

[1] Could it be that Jackson thinks that a mature theory of reasons is not so far out of reach? It's not clear, but his one comment on the issue sounds a (to me) surprisingly optimistic note: "it is certainly true that we cannot, as of now, write down in a natural language necessary and sufficient conditions for being rational. (Though we can say something useful and to the point – whatever the defects of the inductive logic sections of textbooks and extant discussions of experimental design, they are very far from useless.) What would be incredible…would be if there were no story to be told constructible from our folk-classificatory practice: we are finite beings; we do not work by magic; we give useful information to each other by means of the word "ratonal." There must,

candidates for the reasonableness-role, so that there is no particular descriptive property that deserves to be called "the" property of reasonableness.[2]

And now someone might say: Why should it make our blood race if "reasonable" -- and by implication "right" -- is determined to pick out some descriptive property or other, if the question of which one, and hence of the word's application to particular actions, remains as contestable as ever? A finding of truth-aptitude begins to seem like a pretty superficial victory for the cognitivist. He is right about the technical semantical point: evaluative predicates, to the extent that they refer, refer to descriptive properties. But as for there being a truth of the matter about what is right (reasonable), matters are as described by the non-cognitivist..

How would Jackson rule out this weakened cognitivism? He maintains that "it is part of current folk morality that convergence will or would occur. We have some kind of commitment to the idea that moral disagreements can be resolved by sufficient critical reflection which is why we bother to engage in moral debate" (137). But commitment to moral debate is one thing, commitment to convergence is another. It would sufficiently explain the first commitment that we always hope to find a basis for agreement; that we expect that agreement will (or would) come seems like a hypothesis running far ahead of the data.

****

---

therefore, be a story to be told (extracted). And when it is told (extracted), rationality will have been codified" (67).

[2] This would not necessarily show that "reasonable" was vague. How far semantic precision requires definiteness of descriptive extension is a hard question, about which theorists may be expected to disagree. Indefiniteness of extension is the natural model to the possible worlds semanticist; but that may be a reason for looking beyond possible worlds semantics. See also note 4.

A different reason for imputing a commitment to convergence has to do with moral debate. Suppose that we were to arrive not at "a single mature folk morality but rather different mature folk moralities for different groups in the community" (137). Then, Jackson thinks, "the adherents of the different mature folk moralities will mean something different by the moral vocabulary because [their] moral terms...will be located in significantly different networks" (137). One assumes that current speakers too, to the extent that they belong to precursors of Jackson's "different groups in the community," will have meant something different by their moral vocabulary.

But then, insofar as we see today's factions as in danger of evolving into "different groups in the community," we see them as in danger of not really communicating, on account of their similar-sounding words having different semantic values. Contraposing, we have got to see ourselves as converging if we want to see ourselves as communicating.

I don't know whether Jackson is attracted to this reasoning or not. But it looks to be in tension with a point he makes in his discussion of color. He says that the folk might well take it for granted that there is "a kind, indeed a natural kind, distinctive of the exemplars of water and gold. [But] the folk are too sensible to have presupposed something as risky as that there is a distinctive kind in common to things we call 'red'" (108).[3] The presupposition would be "risky" because it would put the legitimacy of our practice with "red" at the mercy of developments that we're not in a position to prejudge. Presupposing moral convergence would be risky in the same sense. If the folk don't do it with "red," it stands to reason that they wouldn't do it with "right" either.

Then what does entitle moral disputants to regard themselves as communicating? Not enough attention has been paid to the cognitivist who answers like this: If something ought to be written into our folk theory, it's not that risky conditions ABC are met – given which such and such a semantic mechanism sees to it that we attribute the same property by "right" -- but simply that you must mean something different by "right" is a diagnosis of the very last resort. The project of using evaluative language together in figuring out how to act is so overwhelmingly important that we do not allow others to opt out with this facile semantic excuse. Our primary commitment in other words is that "right" and similar action-guiding terms should stand for the same or similar properties in all our mouths.

The commitment comes at a price: a certain kind of anti-dogmatism. The readier we are to claim conceptual authority for our own moral views – to say they follow from what "right" means in our mouths -- the harder it becomes to hold onto the idea of coreference as between disputing parties. But then, rather than its being a condition of moral communication that we expect to arrive at a single moral truth, the proper condition is that nothing will ever be regarded as the point of arrival: the point at which reference is finally fixed and moral theory acquires a conceptual imprimatur. This is what makes Jackson's second response to the open question so disturbing. He says that

> the idea that what fits the bill that well might fail to be rightness is nothing more than a hangover from the platonist conception that the meaning of "right" is somehow a matter of its being mysteriously attached to the form of the right (152).

[3] I have my doubts that they presupposed it for "water," as opposed to wanting the extension to be

I would have thought it was part of the bill that what satisfied it might still fail to be rightness. This is not because of platonism but the opposite: we refuse to attach "right" to any particular property so tightly that moral dissidents, even hypothetical ones, come out as simply misusing "right." Similarly, it seems part of what we have in mind by "reasonable" that the door is left open to the brilliant iconoclast who gets us to see that we have all along been acting contrary to reason. To dismiss such a person as meaning something else by "reasonable" strikes us as dogmatic. I doubt that as folk reason-theory matures it will take a kinder view of dogmatism than we do today.[4]

III. Color

A couple of words about Jackson's defense of physicalism against "revelation" – which says that "colour experience is transparent in the sense of revealing the essential nature of colour" -- and "unity" – which says that "redness (e.g.) is the property common to all red things."

The objection from unity is that redness-qua-common-to-red-things will have to be a disjunction of microphysical surface properties; but disjunctive properties can't be causes; so, given the role of causation in representation, redness is not the property presented in red experience. Jackson replies that disjunctive properties can be causes provided they're not excessively disjunctive.

---

as natural-kindy as possible. If for "water," why not also "earth" and "air"?

[4] More needs to be said about how the envisaged referential indeterminacy relates to garden-variety vagueness. They are certainly not the same, since garden-variety vagueness is tolerant of brute, "no-fault," disagreement, whereas disagreements about what is right are felt -- or at least hoped -- to reflect muddled reasoning (or etc.) on one side or the other.

Someone might ask: what forced us to admit that redness was disjunctive in the first place? That it is (up to necessary equivalence) a disjunction of other properties? No, for that much can be said of every property. That it is a disjunction of natural properties. No, because natural (and so presumably nondisjunctive) properties can be that too: charge is the disjunction of positive charge with negative. A better idea (borrowed from Lewis) is that

> a property is disjunctive to the extent that it is a disjunction of properties than which it is much less natural.

The question now is: why think that redness (qua common to red things) is much less natural than its microphysical disjuncts? One reason would be that they are physical, and it is not physical. (This would seem to follow from the discussion of physicality on pp. 6-7.) I don't know whether this is the reason Jackson would give, but if so he may be assuming a stronger form of physicalism than he officially espouses. This stronger physicalism maintains not just that the physical story is complete – all other truths are entailed by it -- but that it is (far and away) the most natural – all other taxonomies introduce an element of the arbitrary.

I don't say that this stronger physicalism is wrong. What I do want to point out is that someone who disagrees with it – call him the taxonomical pluralist -- can avoid the unity problem entirely; he can deny that redness is much less natural than its disjuncts, thus denying that it's in any important sense disjunctive.

There is a connection with revelation as well. That color experience doesn't reveal anything microphysical about redness no longer means that it does not reveal the "essential nature" of redness. Because while it may be of the essence of Jackson's disjunctive redness to be built on microphysical disjuncts, redness conceived as nondisjunctive is no more microphysical than, say, the property of being (approximately) round. Both can be implemented in lots of microphysical ways, but since neither is a disjunction -- in the relevant sense -- of these implementations, it seems gratuitous to write them into the essence. Of course, if all necessary properties are written into essence, then color experience does indeed become "non-transparent." But then so does the experience of roundness.

IV. Necessity

Not many eyebrows will be raised by Jackson's view that metaphysics is committed to "entry by entailment" theses. The hubbub concerns his claim that (if physicalism is true), physicalistic statement P necessitates mentalistic statement M only if M is a priori necessitated by P & P*. His argument looks straightforward:

> The crucial point is the way that the contextual information -- by virtue of telling us in principle the [C-]propositions expressed by [P and M] -- enables us to move a priori from the [physical] way things are to the [mental] way they are. But if physicalism is true, all the information needed to yield the propositions being

> expressed about what the actual world is like in various physical sentences can be given in physical terms [P*], for the actual context is givable in physical terms according to physicalism (83).

An immediate puzzle is that Jackson is attempting here to ground an a priori entailment in C-propositions, when his official account has it that X a priori entails Y iff a certain A-proposition holds in all worlds, viz. the A-proposition expressed by "if X then Y." To square his apriority claim with the official account, he needs to show that the result the speaker allegedly can establish,

> (i) assuming that this is a P*-world, all worlds that are P are also M

can be parlayed into the result that (officially) makes for an priori entailment, viz.

> (ii) whatever world this may be, if P&P* then M.

But this is easily done. From (i) we see that if this is a P*-world, then if it is P it is M. Clearly then if this is a P*&P-world, it also an M-world -- which is just what (ii) says. The problem lies more with (i). How exactly does the speaker establish it?

If Jackson is right about what it is to understand, then the speaker knows (based on her knowledge of context) what it takes to for a world to be P, what to be M. Why must she therefore know that that the one set of worlds is a subset of the other?

The result would indeed follow if (*) to know what an F was, and what a G, you had to know the inclusion relations among the Fs and the Gs. And I don't deny that there's be a sense of "knowing-which" that makes (*) true. The problem is that it's a "high" sense; it sets the bar higher than users of the phrase ordinarily do. Speaking in accordance with (*), even trained mathematicians don't know what an even number is, and/or a sum of two primes. (If they did, they'd be able to figure out the truth-value of Goldbach's conjecture.) The question for us is whether understanding a sentence, and knowing the context, confers knowledge in a high sense of what an S-world is, or knowledge in a more ordinary sense?

For me to know what an S-world was, I need a way of picking out the S-worlds in thought; and not any old way will do. But this is a far sight from (*). I know what a "there is pain"-world is by knowing that it is a world in which there is pain. I know what a "things are physically thus-and-so"-world is by knowing that it's a world in which matters are physically thus-and-so -- and here I might be able to reel off some specific physical requirements. Obviously though to know in these sorts of ways what the P- and M-worlds are does not put me in a position to tell whether M is true in every P-world, even if in fact it is.

One response might be to insist that understanding a sentence is matter of knowing which set of worlds it expresses in a special canonical way: a way that better responds to what worlds in their innermost nature are. Since the physicalist thinks that worlds are in their innermost nature physical, he will presumably insist on a physical specification. It can't be that speakers "miss" the fact that any world physically like ours is a pain-world simply through failing to think of the pain-worlds in

physical terms. Thinking of them in physical terms is a condition of understanding, and we are talking about speakers who understand.

The suggestion is that if physicalism is true, then to understand S one must be able to decide (i) on the basis of physical information (ii) how to make the cut between S- and non-S-world in physical terms. (If physicalism is true, then understanding is "physical" understanding.) This plugs the gap in Jackson's argument, and his conclusion is reinstated. Whatever physical premises necessitate at all, an expanded set of physical premises conceptually necessitates. Merely to understand the sentences is to appreciate their truth-relations.

But this is a result we can happily accept. It would indeed come as a surprise if a normal understanding of "things are physically like so" and of "there is pain" sufficed for knowledge of the conditional. That a physical understanding of the same sentences should have this result is not surprising at all. A physical understanding of "there is pain" is by definition an ability to tell whether worlds presented in physical terms do or do not contain pain.

Everything here goes back to the idea that the physicalist will insist on a physical specification of the verifying worlds. Why should she? Physicalism was supposed to be an ontological theory, not a theory of understanding. This distinction is trampled on when understanding is equated with canonical grasp of truth conditions. It now becomes a "consequence" of physicalism that typical speakers, to the extent that they can't deduce pain from physics, don't understand "there is pain"! The physicalist presumably finds this as bizarre as anyone else.

But let that worry be waived. Suppose we have for each possible world w a complete physical specification P_w in physical terms; and suppose that my understanding of a sentence S is given by the truth values I assign to instances of the schema:

(S_{wv}) if P_w is true in actual fact, then S would have been true had it been that P_v.

Now another problem arises. One cannot decide the truth-value of arbitrary statements – "there is pain," for example – with respect to physically specified worlds without asking what it would be reasonable to say about these worlds. What is the the best explanation, for example, of the fact that there is weeping and wailing and gnashing of teeth? That the weepers are in in pain, or that they are in such and such brain states but feel nothing? If you and I have sufficiently different ideas about this – because of our larger theoretical and practical projects, how anxious we are to avoid multiplying entities, etc. – we may be expected to assign different truth-values to S_{vw}. But then, just because of our methodological differences, we wind up meaning different things by "there is pain"!

Another example: Suppose that you and I are confronted with a world in which events with the characteristic physical manifestations of pain occur on all the same occasions as events that we agree deserve to be called c-fiber-firings. You decide for ontological economy reasons that "pain" and "c-fiber firing" pick out one and the same type in this world. I am not too worried about ontological economy; I decide on the basis of my attachment to psychological autonomy and modal

intuition that they pick out different though de facto correlated types. What does the Jacksonian physicalist say about this case? Again, that you and I mean different things by "pain."

That is not how ordinary speakers see it. Remember the great identity debates of the 1950s, when it was assumed that mental/physical correlations would soon be found and the question was what ontological conclusions to draw. The disputants didn't think of these debates as driven by differences about the meaning of "pain"; they thought they were arguing about the metaphysics of pain. Of course, everyone is entitled to use words however they like. But if Jackson is using "conceptual" in a special sense, to describe differences which others would classify as doctrinal, then that bears on the interpretation of his claim that physicalists are committed to the conceptual entailment of the psychological by the physical. It's not clear – I don't say it isn't true, just that it isn't clear – that the claim comes to more than this: those who find it on balance reasonable to apply mentalistic description to a physically given world can portray those who disagree with them as "meaning" something different by the mentalistic description.

V. Summing up

Quine speaks somewhere of the tendency among philosophers "to seek the gist of every statement in objects that it is about." Jackson does something related; he seeks the gist of every (truth-apt) statement in worlds that we would/should regard as verifying it. Add to this that understanding is "getting" the gist, and Jackson's main claim – that, physicalism granted, merely to understand a statement S is to be in a position to deduce it from physical statements – more or less falls out.

The question I have been pressing is: why should we chalk it up to conceptual competence that a speaker regards "there is pain" ("that was a decent thing to do," "he's being illogical") as an appropriate thing to say about a descriptively given world? Some of the credit surely goes to the speaker's sense of what is or is not a sensible thing to think in light of the available physical evidence. If that is right, then the physical/mental (etc.) conditionals that Jackson would call conceptual truths, or truths of reason, might better be described as truths of reasonableness. They do not compel the assent of every rational person who speaks the language; they record the epistemic intuitions of every speaker who shares those intuitions.