Mill's Proof of the Principle of Utility*

Elijah Millgram

I

In a famous, or infamous, paragraph or so early on in chapter 4 of *Utilitarianism,* Mill provides his argument for the Principle of Utility. I will first quote the passage at some length and rehearse two very familiar objections to it. Then I will go on to say what I intend to do with the material I will have introduced.

> The utilitarian doctrine is, that happiness is desirable, and the only thing desirable, as an end; all other things being only desirable as means to that end. What ought to be required of this doctrine—what conditions is it requisite that the doctrine should fulfil—to make good its claim to be believed?
>
> The only proof capable of being given that an object is visible, is that people actually see it. The only proof that a sound is audible, is that people hear it: and so of the other sources of our experience. In like manner, I apprehend, the sole evidence it is possible to produce that anything is desirable, is that people do actually desire it. If the end which the utilitarian doctrine proposes to itself were not, in theory and in practice, acknowledged to be an end, nothing could ever convince any person that it was so. No reason can be given why the general happiness is desirable, except that each person, so far as he believes it to be attainable, desires his own happiness. This, however, being a fact, we have not only all the proof which the case admits of, but all which it is possible to require, that happiness is a good: that each person's happiness is a good to that

* Thanks to Alyssa Bernstein, Sarah Buss, Roger Crisp, and Gideon Yaffe for commenting on earlier drafts of this paper. I'm also grateful to Brian Hirsch, Brian Lee, Christian Piller, Debra Satz, Candace Vogler, and audiences at Vanderbilt University, Carnegie Mellon University, Reed College, Ohio State University, Stanford University, the University of California, Davis, the University of Miami, Arizona State University, Wake Forest University, Syracuse University, the University of Illinois at Chicago, the University of Florida at Gainesville, the University of Utah, and the University of Mississippi for helpful discussion. An ancestor of this material benefited from comments from Hilary Bok, Alice Crary, Christoph Fehige, Bill Haines, Robert Nozick, Hilary Putnam, and Tim Scanlon. Work on this paper was supported by a fellowship from the National Endowment for the Humanities.

Ethics 110 (January 2000): 282–310
 0014-1704/2000/11002-0002 $02.00

person, and the general happiness, therefore, a good to the aggregate of all persons. (234/4:2–3)[1]

Almost from the moment it went into print, the passage that I have just reproduced was notorious for the two fallacies it was alleged to contain, one for each of the two stretches of argument that make it up.[2] The first stretch of argument (which is what I will call it from here on in) moves from the premise that each person desires his own happiness to the conclusion that each person's happiness is desirable for him. The second stretch of argument moves from the conclusion of the previous stretch, that each person's happiness is desirable for him, to the further conclusion, that "the general happiness . . . [is] a good to the aggregate of all persons."

The first stretch of argument consists, more or less, of an explanation of the legitimacy of the transition from its premise to its conclusion; the explanation adduces analogous arguments that are clearly in order. You show that something is visible by showing that it is seen; you show that something is audible by showing that it is heard. Similarly, you show that something is desirable by showing that it is desired. It is standardly pointed out, however, that Mill's use of his model arguments seems to turn on an equivocation. 'Audible' means "can be heard," and 'visible' means "can be seen." But 'desirable', in the sense required by the argument's conclusion, does not mean "*can* be desired," but something along the lines of "*should* be desired" or "*worth* desiring." The first stretch of argument is apparently supported by no more than a bad pun.

The second stretch of argument looks just as bad, on a par with taking what is good for each of Boeing's workers—say, a pay hike—to be good for Boeing, only with the added obstacle that "the aggregate of all persons" is not a corporate person in the way that Boeing is, and so there is no clear sense in which something can be good for it.

I am going to demonstrate that each stretch of argument is, not fallacious, but deductively valid. I will show that the conclusion of Mill's

1. References to *Utilitarianism* are by page number in the standard edition of John Stuart Mill's *Collected Works* (Toronto and London: University of Toronto Press, Routledge & Kegan Paul, 1967–89), vol. X, followed by chapter and paragraph for the convenience of readers to whom the standard edition is not easily accessible; so the reference above is to p. 234, chap. 4, pars. 2 and 3. Other references in the text will be to the standard edition by volume and page number alone; vols. VII and VIII, which I will refer to frequently, are Mill's *System of Logic*.

The quoted passage is presented as only half of the argument: it is supposed to show that happiness is desirable, but not yet that happiness is the only thing desirable. It will turn out, however, that the apparent division of labor between the two halves of the argument is at least misleading.

2. For a discussion of the early history of such criticism, see Steve Gerard, "Desire and Desirability: Bradley, Russell and Moore versus Mill," in *Early Analytic Philosophy: Frege, Russell, Wittgenstein*, ed. W. W. Tait (Chicago: Open Court Press, 1997), pp. 37–74.

argument has been badly misunderstood. And, in the course of so doing, I will show that *Utilitarianism* is a much more tightly constructed text than it is generally taken for.

The exercise is intended to be not merely of textual or historical interest. I mean, first, to use the passage to investigate the relations between theories of practical reasoning and moral or ethical theories. Practical reasoning is reasoning aimed at figuring out what to do; a theory of practical reasoning is accordingly a theory of *how* one should go about figuring out what to do. A moral or ethical theory (I will use these terms interchangeably here) is, on one fairly standard construction of the notion, a general and systematic theory of *what* one should do; that is, it can be thought of as a compendium or summary of the results of practical reasoning. If that is right, we might expect a philosopher's theory of practical reasoning to determine the architecture of his moral theory. In the course of reconstructing Mill's arguments, we will see that he does not disappoint us on this score: Mill's utilitarianism drops directly out of his instrumentalist theory of practical reasoning. To be sure, Mill's is only a single case. But making out the claim regarding Mill will lend plausibility and substance to the more general suggestion regarding the dependence of substantive moral theory on the theory of practical reasoning. If taken up, this suggestion would motivate a shift in focus in ethics, away from the moral intuitions which are today routinely treated as the touchstone of moral theorizing, and toward the views of practical reasoning that underwrite particular substantive moral or ethical theories.

Second, while I have said that I will show each stretch of argument to be deductively valid, I do not think that they are unproblematic when put side by side. They do not in fact sit at all well together, and responsibility for the incoherence of the argument that they jointly constitute rests, I will argue, with an incoherence in the underlying theory of practical reasoning. Now the instrumentalist account of practical reasoning that I will attribute to Mill is a very close relative of views still widely accepted today. So the incoherence that our discussion of Mill will expose in the instrumentalist theory is of current philosophical interest. Moreover, if Mill's reasons are still among the best reasons we have for accepting utilitarianism, and if they turn out to invoke an incoherent account of practical reasoning, we will have a reason for balking at utilitarianism. To the extent that utilitarianism is still a live moral theory, this result is also of current philosophical interest.

Finally, I will consider how Mill found himself saddled with the instrumentalist theory that was the source of the difficulty I will identify in his argument. The explanation I will offer, a failure of nerve in Mill's attempt at a thoroughgoing empiricism, can serve as an object lesson for contemporary philosophy of an empiricist bent.

II

If, as I have suggested, philosophers have misread Mill's arguments for over a century, there are reasons. One is Mill's practice, necessary in an age when philosophy was much less professionalized than it now is, of writing prose that could be read and at least partially appreciated by an intelligent lay audience (and in particular by authors of public policy whom Mill hoped to influence).[3] The elegance of Mill's writing has obscured the way in which he deploys a technical philosophical vocabulary.[4]

By way of example, consider the glosses Mill provides for the words attached to the utilitarian conception of the good. In the course of explaining what the Principle of Utility is supposed to amount to, Mill tells us that "by happiness is *intended* pleasure" (210/2:2).[5] Shortly before, he has stated that "every writer . . . who maintained the theory of utility, *meant by it* . . . pleasure . . ." (209/2:1).[6] And elsewhere, as one might by now expect, he uses 'utility' as a synonym for 'happiness', as in, "a perfectly just conception of Utility or Happiness" (213/2:9). Mill overtly introduces 'utility', 'happiness', and 'pleasure' as synonymous terms.

He is equally explicit in connecting these terms, almost as closely, to the notion of desiring. "Desiring a thing and finding it pleasant . . . are . . . *in strictness of language,* two different modes of *naming* the same psychological fact" (237/4:10, my emphasis). "Almost as closely," because there is an important distinction to be made between items desired only as means to satisfying further desires and items desired on their own account: the terminological identity is meant to apply only to the latter. The restriction of the scope of the identification to things desired on

3. I'm grateful to John Rawls for emphasizing to me Mill's interest in reaching the politicians.

4. Compare Skorupski's recent description of *Utilitarianism:* "It is written for the general reader. It is not a technical treatise of philosophy, like the *System of Logic;* neither is it a carefully polished piece of political argument, as *On Liberty* is" (John Skorupski, *John Stuart Mill* [New York: Routledge, 1989], p. 283). Skorupski takes the second sentence to follow from, or at any rate to be of a piece with, the first. This, I will show, is a mistake. *Utilitarianism* is written for the general reader, *and* it is a "technical treatise of philosophy."

This attempt to write for both audiences at the same time is not unique. The preface to Mill's *Principles of Political Economy* concludes with the following announcement: "Although [the writer's] object is practical, and, as far as the nature of the subject admits, popular, he has not attempted to purchase either of those advantages by the sacrifice of strict scientific reasoning. Though he desires that his treatise should be more than a mere exposition of the abstract doctrines of Political Economy, he is also desirous that such an exposition should be found in it" (II:xcii).

5. My emphasis. As per usual, Mill adds a clause about the absence of pain; to keep the writing manageable, I'm going to suppress this rider from here on in.

6. Again, my emphasis. Mill is being disingenuous here: as he well knew, in the writing of earlier British moral philosophers, such as David Hume and Adam Smith, 'utility' had meant something very different, and the misunderstanding Mill is trying to correct was in fact quite natural. See Geoff Sayre-McCord, "Hume and the Bauhaus Theory of Ethics," *Midwest Studies in Philosophy* 20 (1996): 280–98.

their own account is important, and I will return to it shortly. With this restriction in place, 'desired' and 'found pleasant' are also synonymous terms, and consequently, "to desire anything, except in proportion as the idea of it is pleasant, is a physical and *metaphysical* impossibility" (238/4:10, my emphasis).

If the object of desire is, as a matter of mere nomenclature, pleasure, and if 'pleasure' is just a synonym for 'happiness', then happiness should be, again as a matter of mere nomenclature, the object, and the sole object, of desire.[7] And so it is: Mill follows the argument we are trying to unravel with a further argument, purporting to show that people desire only happiness. The argument proceeds by insisting that "whatever is desired otherwise than as a means to some end beyond itself . . . is desired as itself a part of happiness" (237/4:8); inspecting the argument shows the point to be that, if something is desired (again, with the scope of the claim restricted to desires whose objects are not merely desired as means to further ends), it thereby *counts* as part of happiness.

With this, we have arrived at a view as to the status of the premise of the first stretch of Mill's argument, that each person desires his own happiness. The premise is analytic, true by virtue of the meanings of the words. (As is the further claim, that people desire only happiness noninstrumentally.) It means no more and no less than: people desire what they desire. This should be puzzling; it is hard to see how any substantive conclusion that Mill wishes to draw could be derived from an empty tautology, and it might also seem that this reading of the premise does not match the advertising. Mill, after all, describes the claim that people desire only happiness as "a question of fact and experience, dependent, like all similar questions, upon evidence" (237/4:10). I will return to the puzzles later on; for the present, I want to let this reading of the premise stand, and note only that it allows us to shelve, at any rate temporarily, doubts we may have had about the premise's truth. With those doubts out of the way, we can go on to consider the validity of the first stretch of Mill's "proof."

III

Mill introduces his argument by stating "that questions of ultimate ends do not admit of proof, in the ordinary acceptation of the term" (234/4:1). This passage is commonly used as an excuse for holding Mill to lower standards of argumentation than usual; if an interpretation of the argument makes it come out loose, or shoddy, or invalid, or simply less than an argument, the interpretation can be defended by pointing out that it was not meant to be a *real* argument, anyway.[8] But another look at the passage shows it to be a reminder of an earlier discussion: "*[i]t has*

7. I should perhaps note that I am not here taking over Mill's technical sense of "nomenclature" (VIII:704–5).

8. Even as astute a reader of Mill as Skorupski falls into this trap; see n. 16 below.

already been remarked, that questions of ultimate ends do not admit of proof, in the ordinary acceptation of the term"; this is in virtue of features that they share, as "the first premises . . . of our conduct," with "the first premises of our knowledge" (234/4:1). So before we decide that Mill is telling us that we are about to get hand-waving instead of an argument, we need to look back at this earlier discussion, spell out the shared features, and see exactly why it is they preclude "proof." Now in chapter 1, Mill tells us that "such proof as [the Utilitarian or Happiness theory] is susceptible of . . . cannot be proof in the ordinary and popular meaning of the term" (207/1:5). This is evidently the earlier discussion to which Mill is referring, and it is here we should look to fill out the analogy between the first premises of knowledge and the first premises of conduct on which the inference from desire to desirability is supposed to turn.

Description of something as a "first premise" implies a pattern of inference in which it occupies this position. So what we need, first of all, from this earlier discussion is the pattern of inference in which this characterization has its home. Mill obligingly identifies it for us immediately after the passage I have just quoted. "Whatever can be proved to be good, must be so by being shown to be a means to something admitted to be good without proof" (207–8/1:5). A contemporary way of putting Mill's view might be to say that all practical reasoning is means-end reasoning; or, to use another bit of current vocabulary, that Mill is an instrumentalist.[9]

9. Mill's instrumentalism seems to have been inherited; in an editorial footnote, he approvingly quotes his father's "Fragment on Mackintosh": "All action, as Aristotle says, (and all mankind agree with him) is for an end. Actions are essentially means" (James Mill, *Analysis of the Phenomena of the Human Mind* [London: Longmans, 1878]; "New Ed. with Notes Illustrative and Critical by Alexander Bain, Andrew Findlater, and George Grote, Edited with Additional Notes by John Stuart Mill"; originally published in 1829; vol. 2, p. 262n.).

For a display of instrumentalism in the younger Mill's own writing, see the last chapter of the *System of Logic,* evidently intended by Mill as stage-setting for *Utilitarianism* (cf. VIII: 951n.), where Mill describes the relations between Science and Art. "Art" is Mill's label for the domain of practice and action, the area of which "the imperative mood is . . . characteristic" (VIII:943). Instrumental reasoning is not only the sole form of practical reasoning mentioned in the course of Mill's discussion; it is held entirely responsible for the structure of the domain. Art supplies ends, and the role of Science is to determine the means to those ends. Not surprisingly, Mill ends up identifying "the general principles of . . . Teleology, or the Doctrine of Ends" with "the principles of Practical Reason" (VIII:949–50).

There is, however, a caveat required here: while Mill's official doctrine remained unflinchingly instrumentalist until the day of his death, Vogler has pointed out that in his own argumentative practice he helped himself to other and richer forms of practical reasoning. What is more, he produced a striking indictment of instrumentalism in his retrospective diagnosis of his "mental crisis": in a strange and touching passage in his *Autobiography,* Mill argued that if you are an instrumentalist, you ought to find life unlivable (I:141–43). For discussion, see Candace Vogler, *The Deliberative Landscape* (New York: Garland Press, in press).

We can now make out the analogy Mill has in mind, between the 'first premises' of our knowledge and of our conduct. On the instrumentalist view, a practical justification consists in showing that a proposed end is a means to a further end. (More Millian terminology: "Questions about ends are, *in other words,* questions what things are desirable" [234/4:2, my emphasis].) That further end may be justified in turn, that is, shown to be desirable, by showing it to be a means to a still further end. Iterating justifications in this way induces larger patterns of justification that can fall into three general types: chains of justification can continue backward forever; they can contain cycles; or they can terminate in ends that are not themselves further justified.

The choice between these patterns should sound familiar; it was certainly familiar to Mill. Beliefs can be justified by further beliefs, and when they are, it is usually thought, the same choices arise: between an infinite regress, circularity (so-called if you don't like it) or coherence (so-called if you do), and foundational beliefs, that is, beliefs that terminate chains of justification and are not themselves further justified. In a survey of fallacies in *A System of Logic,* Mill rejects the first two options, for reasons that are general enough to apply to practical justifications as well: on the one hand, "there cannot be an infinite series of proof, a chain suspended from nothing" (VIII:746); on the other, "Reasoning in a Circle" is in Mill's view simply a "more complex and not uncommon variety of" "Petitio Principii, or begging the question" (VIII:820). Mill opts for foundationalism, the view that there are "propositions which may reasonably be received without proof" (VIII:746).

I have so far identified the class of foundational beliefs, and the class of foundational ends, simply in terms of structural features of the patterns of justifications in which they appear: foundational beliefs are those beliefs which are not further justified; foundational ends are those ends that are not merely means to further ends. But it is typical of philosophizing in epistemology and ethics to look for a further and independent way of identifying these classes, for example, by taking them to be, or to be associated with, an especially immediate *feeling.* Not surprisingly, given the tradition of British Empiricism in which Mill stood, he took the foundational beliefs to correspond to "sensations."[10] Mill similarly takes the foundational ends to be objects of especially immediate feeling, and it is this presumption that we may suppose licenses Mill's introduction of vocabulary that, we saw, identifies the objects of (non-instrumental) desires with pleasure and happiness, states naturally described as states of feeling.

Now consider what attempts at justification look like at terminal and

10. "The truths known by intuition," he says, "are the original premises from which all others are inferred"—and by 'intuition' Mill does not mean what a late-twentieth-century philosopher would: he uses the word as a synonym for 'consciousness', i.e., for "what one sees and feels" (VII:6–7).

nonterminal locations in the respective foundationalist structures we have just described. A belief that is not simply the correlate of a sensation ought to come with a further justification: other propositions that you believe. So when someone asks you why you believe it, you can, if the belief is not unjustified, adduce these further propositions. But when you are asked why you hold a terminal belief (that is, a belief that is simply the correlate of a sensation), there is no further proposition to adduce. All you can do is reiterate your claim to be sensing what you are sensing, or feeling (since Mill takes sensation to be a species of feeling) what you are feeling.[11]

This is the point of Mill's remark that "the only proof capable of being given that an object is visible, is that people actually see it." Philosophers have misread the passage by overemphasizing the modal dimension of 'visible' (and, in the following sentence, of 'audible'), taking 'visible' to mean "*capable* of being seen." This is a mistake. 'Visible' here means, if you like, "can be seen"; but only in the sense in which a pilot might report that he can see the target: he *sees* the target. 'Audible' is likewise meant in a similar sense, the sense in which the petrified victim in the monster movie whispers, "It's audible now—I can hear it moving down below." This means, not that in some other possible world, the monster is heard moving down below, nor that in the future, with some extra effort, the monster might be heard, but that in *this,* actual world, the monster is *now* heard moving down below. At the terminus of the chain of justification, when we have gotten back to the sensations, the only "proof" that an object is visible, that is, *is seen,* is to gesture once again at its being seen. The only "proof" that an object is audible, that is, *is heard,* is to gesture at its being heard.

This is Mill's model for "desire as proof of desirability."[12] At a nonterminal location in a chain of practical justification, the end occupying that location and the justification for it are distinct; whether something is desired, and whether it is desirable, are two different questions. Asked why you want to do such-and-such, you can adduce the further ends to which it is a means. But at the terminus of the chain, when we have come back to the feeling of pleasure or happiness, or the desire whose satisfaction is that feeling, there is nothing further to be said. All there is to the desirability of the object of the desire is that it is desired, and all there is left, at the end of the chain, for 'desirable' to *mean* is "desired."

11. It might be thought that, at the terminus of the chain of justification, belief, feeling and justification all collapse into one. But Mill does in fact continue to distinguish the feeling from the belief, for reasons having to do with the role of general terms in observation. These issues can be put aside here.

12. The phrase is borrowed from Norman Kretzmann, "Desire as Proof of Desirability," in *Utilitarianism with Critical Essays,* ed. Samuel Gorowitz (Indianapolis: Bobbs-Merrill, 1971), pp. 231–41 (originally published in *Philosophical Quarterly* 8 [1958]: 246–58).

We are now in a position to give the first of Mill's two reasons for withholding the title of "proof" from his argument for the Principle of Utility. Recall again that Mill was also the author of *A System of Logic:* "The proper subject . . . of Logic is Proof" (VII:157), and so we should not be surprised to find that 'proof' is, in Mill, a technical term.[13] Proof is inference, the movement from evidence to conclusion: "We say of a fact or statement, that it is proved, when we believe its truth by reason of some other fact or statement from which it is said to *follow*" (VII:158).

> In so far as belief professes to be founded on proof, the office of logic is to supply a test of ascertaining whether or not the belief is well grounded. With the claims which any proposition has to belief on the evidence of consciousness, that is, without evidence in the proper sense of the word, logic has nothing to do. (VII:9)

Sensations, and the beliefs that correspond to them, are not inferred, and so when we get back to sensations, we have left the domain of properly so-called "proof." Likewise, the "first principles" of conduct are ultimate ends, desires that are, for the instrumentalist, ex hypothesi not inferred from further premises. The only question that arises regarding these desires is whether one actually has them, and that is a matter of "consciousness" or "intuition." Since they are not inferred, they too lie outside the domain of proof.

Let's return from this lengthy excursion to the first stretch of Mill's argument. Its conclusion was that each person's happiness is desirable for him. Let us keep in place Mill's restriction to ends that are not desired merely as means to further ends, but on their own account. Then happiness, recall, is, simply as a matter of terminology, what one desires. Again, with the restriction in place—that is, when we are at the termini of the agent's chains of means-end justification—'desirable' just means "desired": there is nothing else left for it to mean. So the conclusion of the argument just means that what one desires (noninstrumentally), one (noninstrumentally) desires. Recall further that the premise of the argument turned out to mean only that each person desires what he desires. We can now see why the first stretch of argument is so short: the inference rule being used is $p \rightarrow p$. And it is evidently not fallacious after all, contrary to the views of Moore, Sidgwick, and many others: if *any* argument is deductively valid, it is this one.[14]

13. As far as the point at hand goes, Mill's use of 'proof' as a technical term matches, and justifies his invoking, "the ordinary and popular meaning of the term" (207/1:5). But there is a further point, which we will soon come to, that stretches what is today the ordinary meaning a good deal.

14. Deductive validity notwithstanding, there are a couple of difficulties that are worth highlighting. In working the finger-exercise version of the first stretch of argument, the claim that happiness is the aggregate of the objects of noninstrumental desire is going to play a pivotal role, and the question is how we come by it. There seem to be two ways of

IV

Before moving on to the second stretch of Mill's argument, we need to pause briefly to consider what we are to make of the first. The stretch of argument we have been looking at consists entirely in the repetition of the trivial truth that people want what they want, and seems to be a matter of laying out a series of verbal equivalences. This might still seem hard to believe, but any residual incredulity can be met by Mill's own redescription of his thesis: "what is the principle of utility," he asks, "if it be not that 'happiness' and 'desirable' are synonymous terms?" (258n./5:35n.). Still, if that is all there is to the argument, what is it doing here? How could Mill, normally so impatient with "reasoning [that] consists in the mere substitution of one set of arbitrary signs for another," have thought it to be worth presenting?[15]

We can begin by giving the second reason that Mill withholds the title of 'proof' from this stretch of argument. Commentators have generally assumed that if Mill is unwilling to call his argument a proof, he must mean at least that the argument is not deductively valid.[16] But to take Mill this way is to get him exactly backward. With the stipulated meanings of the words put in place, the first stretch of Mill's argument turns out to be the repetition of a tautology. Now, in *A System of Logic*, Mill "exclude[s] from the province of Reasoning or Inference properly so called, the cases in which the progression from one truth to another is only apparent, the logical consequent being a mere repetition of the

---

getting there. The first is to read it off the subsequent argument, by noticing that Mill takes being an object of noninstrumental desire to be equivalent to being a component of happiness. But if this is what Mill is doing, then he seems to be entirely ignoring questions of the *organization* of the parts of a person's happiness into a whole. And I think that this is a genuine problem in Mill's view.

The second way of getting the pivotal claim is to use the verbal equivalence of 'happiness' and 'pleasure', and then that of 'desiring a thing' and 'finding it pleasant'. The problem here is that pleasure has to come out being the *object* of desire; but as we ordinarily use the words, to find something pleasant is not the same as that thing's *being* pleasure. By our lights, there is a slide from treating pleasure as a propositional attitude to treating pleasure as the object of the attitude. Now, by the end of the passage that introduces the equivalence, it is clear that Mill takes himself to have introduced the terms in a way that bridges the apparent gap: having said that "to desire anything, except in proportion as the idea of it is pleasant, is a physical and metaphysical impossibility," he then says, by way of repetition, "that desire can [not] possibly be directed to anything ultimately except pleasure." But why would Mill want to use his words this way? I will towards the end of the argument return to the question of what substantive view underlies this aspect of Mill's vocabulary.

15. VII:176; compare VIII:760–61.

16. For instance: "Mill stands accused of committing glaring logical blunders. . . . such accusations . . . are not justified. . . . he explicitly states that a *proof*, in the commonly understood sense of the word, of the Utility Principle . . . cannot be given. So he is not claiming to present a deductively valid argument—which is what one must assume him to be doing to pin on him the familiar fallacies" (Skorupski, pp. 285–86).

logical antecedent."[17] Proof is inference, and to count as an inference, Mill holds, there must be more to the conclusion than to the premises. So neither a tautology nor its repetition can count as a proof. This is not, it needs to be emphasized, a minor or dispensable Millian doctrine. Mill was embarked on the radical empiricist project of showing that there is *no such thing* as deductive inference; it is one of the more important burdens of the *System* to argue that all inference (and consequently all proof) is inductive.[18] Because the position is so alien to contemporary sensibilities, commentators tend to leave it behind when they read Mill's political and ethical writings; as a result, they badly misinterpret the passage we are reconstructing. Mill is denying that his argument is a "proof," *not* because it is not deductively valid, but precisely because it *is*.

We have two reasons on the table for Mill's first stretch of argument not being a proof. One is its subject matter; ultimate ends are a category of items picked out as beyond the reach of inference or proof. The other is the argument's form: the argument is (we would say) deductively valid and (Mill would have insisted) merely "apparent, not real" (VII:158). We can now see that the subject matter explains the form. When inference is no longer available, all Mill can do is remind us of something we already know, "not proving the proposition, but only appealing to another mode of wording it, which may or may not be more readily comprehensible by the hearer."[19] Mill is not in a position to argue that happiness is what we are really after, but he can put it to his reader in a way that will remind the reader that what is at stake for him is that he get what he wants, and that this is what he calls "happiness." "If the end which the utilitarian doctrine proposes to itself were not, in theory and in practice, acknowledged to be an end, nothing could ever convince

17. VII:162; the full discussion appears in the previous section, at VII:158–62. See n. 48, below.

18. VII:186–93. This may also seem hard to believe; after all, Mill would not have denied, of a syllogism like, 'All men are mortal; the Duke of Wellington is a man; so the Duke of Wellington is mortal,' that when the major premise is true, all its instances are true also. However, we now have a widely familiar analog of syllogistic reasoning as Mill understands it, and highlighting the analogy may make Mill's claim seem more reasonable.

Think of the file compression utility (e.g., Zip or Compress) on your computer. When you compress and reexpand a file, you do not think of yourself as deriving new results, but as merely reformatting the file for more convenient storage or access. Now, in Mill's view, the actual inference to the Duke of Wellington's mortality is from particulars to particulars: from 'A was a man, and he died', and 'B was a man, and he died too', and so on, and 'The Duke of Wellington is a man', to 'The Duke of Wellington will die'. The major premise of our sample syllogism is not part of the inference proper, but is rather a way of *storing* the observed evidence in a way that makes it easy to keep track of and use; the syllogism is a method of *extracting* already available information, not of inferring new information.

19. VII:158–59; compare VII:66, or again, the *Analysis:* "We can afford . . . no aid to the reader in distinguishing [a pleasurable or painful sensation], otherwise than by using such expressions as seem calculated to fix his attention upon it" (James Mill, vol. 2, p. 190).

any person that it was so" (234/4:3). Now if what Mill is trying to elicit is an acknowledgment of what the reader already has as an end, Mill's logical views make syllogistic deduction an appropriate vehicle. Syllogisms are a means of "deciphering our own notes" (VII:187), and the point of deductively valid "argument" is to remind us to what we are already committed.[20]

What is doing the work in Mill's argument? The first stretch of argument reduces to a series of interlocking definitions, and interlocking definitions of themselves only produce empty tautologies. Now, as Mill himself remarked, "most questions of naming have questions of fact lying underneath them."[21] We have already seen one instance of this phenomenon, in the view that feelings, in Mill's sense, terminate chains of justification; this allowed Mill to choose terminology that identified pleasure, happiness, and the objects of noninstrumental desire. I want to suggest that Mill's interlocking definitions are underwritten by his instrumentalist understanding of practical reasoning, and that this instrumentalism is the motor of the first stretch of argument, and of Mill's utilitarianism more broadly.

This is an occasion to say more precisely what I mean by 'instrumentalism'. I am using it as the label for the still widely shared view that there is only one kind of practical inference, that which takes you from an end to the means to that end; or, in psychological vocabulary, that practical reasoning consists in moving from a desire for one object to a desire for a way of bringing about or attaining that object. Grant that justifying an end amounts to recapitulating an inference by which that end could be arrived at. Every instrumentalist inference proceeds from a desire. So instrumentalism guarantees that when a chain of practical justification is followed back to its terminus, there will be a desire at the terminus.

It is this that licenses "desire as proof of desirability." Suppose for a moment that instrumentalism were false, and that practical inferences could proceed from premises that were not desires. Then "desire as proof of desirability" would have to be given up, and replaced with the much less suggestive "whatever premises are used to demonstrate desirability are, taken jointly, proof of desirability." The identification of those premises with a class of feelings that cannot be argued about, but simply gestured at, would be suddenly implausible. 'Happiness' would

20. So while Mill did not think that deduction was properly understood as inference, he did not deny, but rather insisted on, the heuristic importance of deduction, and syllogistic deduction in particular. See, e.g., VII:196–99; VIII:665.

21. VII:15n.; this quotation is from the 1856 edition of *The System of Logic*, and was omitted in subsequent editions. Compare his criticism, at VII:93–97, of views of naming on which all truth comes out merely verbal, and his discussion "Of Propositions Merely Verbal" where he claims that "when any important consequences seem to follow . . . from a proposition involved in the meaning of a name, what they really flow from is the tacit assumption of the real existence of the objects so named" (VII:113).

no longer be a reasonable label for the aggregate of those premises. Mill's precision-tooled vocabulary would become unusable, and his argument would cease to make sense. Instrumentalism is the substantive premise lying beneath the first stretch of argument, and expressed through the terminological stipulations that make it up.

Many philosophers think that there must be a close and intrinsic connection between rational motivation and what we can for the moment call the Good. Instrumentalism makes it hard to see the dependence as running from the Good to the motivation: if you don't happen to have the right desires, the knowledge that such-and-such is the Good will not amount to a reason to act. So it is natural to reverse the direction of dependence: the Good is analyzed as a construction out of one's preferences or desires or ultimate ends—that is, as something very much like Mill's sophisticated notion of utility. So we should expect to find the philosophical motivation for utilitarianism and its relatives to be instrumentalism, in cases other than Mill's.

V

It's time to turn our attention to the second stretch of Mill's argument: the move from "each person's happiness is a good to that person" to "the general happiness, therefore, [is] a good to the aggregate of all persons" (234/4:3). The objection, recall, was that the conclusion just doesn't seem to follow from the premise. Rather than repeat my phrasing of the complaint, I'll let Sidgwick give his version:

> even if we grant that what is actually desired may legitimately be inferred to be . . . desirable . . . an aggregate of actual desires, each directed towards a different part of the general happiness, does not constitute an actual desire for the general happiness, existing in any individual; and Mill would certainly not contend that a desire which does not exist in any individual could possibly exist in an aggregate of individuals. There being therefore no actual desire—so far as this reasoning goes—for the general happiness, the proposition that the general happiness is desirable cannot be in this way established: so that there is a gap in the expressed argument . . .[22]

If the gap is genuine, a further premise will be needed to fill it. (Sidgwick volunteered "the intuition of Rational Benevolence.") But it is worth noticing that Mill himself had heard similar complaints, and thought there was no gap. Here is Mill responding to a criticism of Spencer's that was evidently very close to Sidgwick's:

> [Spencer] says . . . the principle of utility presupposes the anterior principle, that everyone has an equal right to happiness. It may be

22. Henry Sidgwick, *The Methods of Ethics* (Indianapolis: Hackett, 1981), p. 388.

> more correctly described as supposing that equal amounts of happiness are equally desirable, whether felt by the same or by different persons. This, however, is not a *pre*supposition; not a premise needful to support the principle of utility, but the very principle itself . . . (258n./5:35n.)

This is an obscure passage, and it is still too early to try to say what it means; it is clear enough, however, that Mill does not agree that there is a missing premise. So we should try to find an interpretation of the argument on which all the pieces are in place. I propose to do that in what may seem like an especially roundabout way, by returning to the model of practical reasoning that, I claimed a moment ago, drives the first stretch of Mill's argument.

I introduced instrumentalism as the doctrine that all practical reasoning consists in finding means to ends, or, equivalently, finding ways to satisfy the desires one has. But that is simply not enough to count as a complete theory of practical reasoning. First of all, this very minimal instrumentalist theory, or rather, theory fragment, takes no account of a very basic fact of human life, that you can get what you wanted and still be disappointed. (This fact has a less depressing flip side, that you can be pleasantly surprised by something you had not desired.) In Mill's vocabulary, the problem can be located in the ambiguity of the phrase 'find pleasant', which, it will be recalled, is introduced as equivalent to 'desire': in desiring something you may find—that is, *expect*—it to be pleasant, but that does not mean that when the object of your desire is obtained, it will be found—that is, *turn out*—to be pleasant. The possibility of having the thought that you were wrong about what you had wanted has got to be accommodated in one way or another.

Second, the theory fragment says nothing at all about choice in the face of competing desires. Almost every decision involves competing priorities, and so a putative theory of practical reasoning that failed to speak to the issue at all would not be much of a theory of practical reasoning: it would not explain how you figure out what to do.[23] So the minimal theory fragment must come supplemented in a way that enables it to address these requirements.

Instrumentalism is an exclusionary view: *only* means-end reasoning counts as practical inference. So the supplement must consist not of further rules of inference, but of a penumbra, around the one legitimate rule, of claims about the material to which it might apply. The Benthamite penumbra consisted in the view that pleasures could be felt to be stronger or weaker; that the stronger pleasure was to be preferred; and that pleasure itself, as opposed to the more or less reliable means through which it might be obtained, would never disappoint. This is not the place to reconstruct Mill's reasons for rejecting Bentham's views;

23. Mill registers his awareness of this demand at VIII:951.

suffice it for now that Mill found implausible the notion that pleasures could differ only in strength: "Neither pains nor pleasures are homogeneous, and pain is always heterogeneous with pleasure" (213/2:8). Mill's more sophisticated alternative is presented in chapter 2 of *Utilitarianism*, in the course of a discussion whose ostensible purpose is to meet the objection that utilitarianism is a moral theory unable to accommodate our interests in the finer things in life. The relevant passages are these:

> Of two pleasures, if there be one to which all or almost all who have experience of both give a decided preference . . . that is the more desirable pleasure. (211/2:5)[24]
>
> From this verdict of the only competent judges, I apprehend there can be no appeal. On a question which is the best worth having of two pleasures, or which of two modes of existence is the most grateful to the feelings . . . the judgment of those who are qualified by knowledge of both, or, if they differ, that of the majority among them, must be admitted as final. And there needs be the less hesitation to accept this judgment respecting the quality of pleasures, since there is no other tribunal to be referred to even on the question of quantity. What means are there of determining which is the acutest of two pains, or the intensest of two pleasurable sensations, except the general suffrage of those who are familiar with both?
>
> What is there to decide whether a particular pleasure is worth purchasing at the cost of a particular pain, except the feelings and judgment of the experienced? (213/2:8)
>
> The test of quality, and the rule for measuring it against quantity, [is] the preference felt by those who, in their opportunities of experience, to which must be added their habits of self-consciousness and self-observation, are best furnished with the means of comparison. (214/2:10)

Mill's proposal, often referred to as the "decided preference criterion," addresses the second lacuna in the instrumentalist theory fragment directly: competing desires are to be referred to the preferences, over their objects, of the majority of the experienced. And the first lacuna is indirectly addressed; if we think of what will be a disappointing object of desire as in competition with the contrasting state of not having it, the preference over these options can be corrected by appealing to the preferences of the already-disappointed.

Mill is providing a logically decisive criterion of desirability. It is "final," a "verdict . . . [from which] there can be no appeal"; "there is no

24. The omitted phrase is a qualification designed to avoid circularity in the establishment of a specifically moral criterion. This omission is repeated in the next passage.

The passage is introduced as an explanation of difference in quality in pleasures, as opposed to quantity. But since, as we will see in a moment, quantitative comparisons work no differently, we can disregard the apparent restriction in scope.

other tribunal." The experientially privileged (as I will call them henceforth) are not being used as reliable witnesses of some independent matter of fact; no matter how reliable, it is always possible that all the witnesses are wrong; but Mill's experientially privileged judges are infallible. (The metaphor is instructive: a judicial verdict constitutes the legal fact.) Like a contemporary decision theorist, Mill is not taking the utilities to be reflected (possibly inaccurately) in the preferences; rather, he is taking them to be constructions from the preferences.[25]

Contemporary instrumentalists need to supplement the instrumentalist theory fragment, for the same reasons that Mill must, and it's worth pausing for a moment to point out the similarities and the differences in their respective solutions to the problem. Brandt, in this regard an entirely typical moral philosopher of the last generation, "call[s] a person's desire, aversion, or pleasure 'rational' if it *would* survive or be produced by" a process he labels "cognitive psychotherapy." Rawls defines "a person's plan of life [as] rational if" it satisfies two conditions, one of which is that it "*would* be chosen by him with full deliberative rationality, that is, with full awareness of the relevant facts and after a careful consideration of the consequences."[26] As these examples suggest, the fashionable way of doing these things nowadays is to invoke a counterfactual self: what is decisive is what a better-informed, more carefully deliberative, and otherwise improved version of yourself *would* think or want. Mill manages to avoid the many problems that come with relying on counterfactual selves[27] by appealing to the *actual* preferences of a majority of

25. Mill does differ from contemporary decision theorists in how he thinks of preferences. The current economist's notion of preference ties it very closely to overt behavior: preference is "revealed" in choice. Mill's notion of preference is closer to that of a comparative evaluation: one such preference is described as "a full appreciation of the intrinsic superiority of the higher [pleasure]." The description appears in the course of Mill's acknowledgment that "the influence of temptation" and "infirmity of character" may produce the opposite "election" (212/2:7)—"election" being much nearer to what we would call "preference" today.

26. Richard Brandt, *A Theory of the Good and the Right* (Oxford: Clarendon Press, 1979), p. 113; John Rawls, *A Theory of Justice* (Cambridge, Mass.: Harvard University Press, 1971), pp. 408–9; cf. also pp. 417, 421–22; emphases mine.

27. Mill's take on what these problems are is sufficiently nonstandard to be worth spelling out. First, appeal to the preferences of counterfactual selves involves making psychological predictions (about what you would prefer if . . .). But psychological laws fall into two categories. On the one hand, there are what Mill calls "Empirical Laws," that is, uniformities that are observed to hold, but whose underlying reasons are not understood. Mill holds that these cannot be relied upon "beyond the limits of actual experience"; varying the "collocations" (Mill's term for boundary conditions) may disrupt the uniformity (VII: 516, 519, 548; for 'collocation', VII:465). But assessing the truth of a psychological counterfactual normally involves "varying" the collocations "from those which have actually been observed" (VII:516). So Empirical Laws cannot be relied upon to assess psychological counterfactuals. On the other hand, there are psychological predictions derived from the fundamental laws of the human mind. These are ironclad, but, unfortunately, they cannot normally be applied. What a person will do in response to a given stimulus is a matter of his

(*other*) people.[28] But the devices have a good deal in common. Each uses suitably chosen persons as what engineers call a black box. We can vary the box's inputs, and see what outputs it produces. We are instructed to use the outputs in a certain manner. But we are given no account whatsoever of what goes on inside the box. I will return to this point shortly; but it is now time, finally, to reconstruct the second stretch of Mill's proof of the Principle of Utility.

VI

Mill's object is to show that the general happiness is desirable to everybody (or, "to the aggregate of all persons"—a turn of phrase I'll gloss a few steps down the road). Consider a particular person, say John Doe. The problem was that the desires he currently has do not necessarily involve a desire for the general happiness; and that even if Doe does have such a desire, it might be insufficiently weighty in his scheme of things. But we now have Mill's technique for correcting desires, and for showing that something that someone does not in fact desire may be desirable for him nonetheless. If an overriding preference for the general happiness can be shown to be the preference of the majority of the experientially privileged, then it will have turned out to be desirable for John Doe, even if John Doe himself does not actually desire it.[29] And since John Doe is

---

character; his character is shaped by his entire history; and this is not something we can observe (VII:865). Consequently, a criterion that appealed to psychological counterfactuals would be unusable: there is simply no way to determine what you would prefer, if . . .

Mill's second problem is that there are many counterfactual selves—some dulled, some especially nasty, some overly tremulous, etc.—whose judgments are going to have to be ruled out as defective. It's hard to believe that this can be done without importing into the selection of counterfactual selves the full and robust set of evaluations that the use of the counterfactual selves was meant to certify. But if this is what is going on, this use of counterfactual selves is viciously circular: "to explain away the numerous instances of divergence from their assumed [moral] standard, by representing them as cases in which the perceptions are unhealthy" is "[a] striking instance of reasoning in a circle" (VIII:826).

Notice that this is also a reason for not looking to the preferences of a carefully chosen coterie of right-thinking persons: appealing to the judgments of the *phronimoi* will not do. This is why Mill usually requires only experience of the options being compared. And I am accordingly inclined to think that his sole mention of "habits of self-consciousness and self-observation" is a slip on Mill's part.

28. A quick textual point, for those who are used to reading the passages in question as invoking counterfactuals. In English, counterfactuals are generally (although not necessarily) signaled by the subjunctive mood. These passages, amounting to about half a page of text, contain not even a single subjunctive, and the explanation cannot be stylistic: Mill's discussion of "Permanent Possibilities of Sensation" uses the subjunctive without hesitation (e.g., IX:184).

29. If this sounds implausible—why should I care about what other people prefer, if their preferences turn out not to match mine?—notice that the currently popular counterfactual technique has the same problem. Why should I care about what a counterfactual self would prefer, if his preferences turn out not to match mine?

The difficulty is, however, perhaps especially acute for Mill, who in *On Liberty* had

simply an arbitrary person, to show that the general happiness is desirable for John Doe is to have shown it to be desirable for everybody.

A first glance at the experientially privileged might make this seem an unpromising strategy. For present purposes, the relevant group can be taken to be, very roughly, people who have both experienced an improvement in the general happiness at some cost to themselves, and, on some other occasion, experienced an improvement in their own happiness at the expense of the general happiness. When we look around at such people, it is hard to say offhand that one preference predominates; and certainly the cynical among us will suspect that those with the egoistic preference are bound to outvote the altruists.

But when votes don't go the way you like, there's a standard solution: gerrymander the voting districts. Mill, the radical utilitarian activist, would have thought of the response in more noble-sounding terms; the obstacle to enacting the utilitarian political program, he was convinced, was the restriction of the franchise, and the first item on the agenda was always to distribute the right to vote more broadly. What populations should we survey, when we are looking for the experientially privileged? All those now living, certainly; but there is no obvious principled reason for excluding the judgments of persons past and future, and if there is no reason to exclude them, they must be taken account of as well; the group to be surveyed is, in Mill's phrase, "the aggregate of all persons" (234/4:3).[30] This bit of redistricting has the potential drastically to alter the outcome of any poll: the human race has had a short and underpopulated past, and Mill would have expected it to have a long and populous future. So the votes of the future experientially privileged will outweigh the votes of their predecessors; what the future sees to be desirable, will (by the decided preference criterion) *be* desirable. So we need to determine what the experientially privileged of the future will prefer.

Mill himself raises the problem we have been trying to solve for him: "why am I bound to promote the general happiness? If my own happi-

---

pointed out that others' experience may be an "unsuitable" guide if one has an "uncustomary" character, that "human beings are not . . . undistinguishably alike," "that people have diversities of taste, . . . [and] different conditions for their spiritual development," and that there "are . . . differences among human beings in their sources of pleasure, their susceptibilities of pain, and the operation on them of different physical and moral agencies" (XVIII:262, 270).

I will get around to discussing the more general version of this problem in due course.

30. Kretzmann notices and highlights this move, but worries that he has "extended [Mill] and even departed from him" (Kretzmann, p. 241).

One other worry that might arise at this point is whether the decided preference criterion is really usable: do we have to wait until all the votes are in before we can make up our minds? Notice, however, that this problem is very like another that utilitarianism already has. When assessing an action or rule, how far into the future do we need to look for the effects it will have on overall utility? The in-principle answer is, of course, to the end of time; but that cannot be the normal procedure.

ness lies in something else, why may I not give that the preference?" (227/3:1). The occasion for asking the question is his discussion of "The Ultimate Sanction of the Principle of Utility," in chapter 3 of *Utilitarianism*. Chapter 3 is conventionally taken to have nothing to do with the subsequent "proof" in chapter 4: the former, it is said, deals with questions of motivation, while the latter has to do with questions of justification, and the consequentialist form of Mill's theory makes these entirely separate questions. As we are already in a position to see, however, the criterion of desirability introduced in chapter 2 entails that these are not separate questions, after all. And as we should by now expect, chapter 3 contains the last missing premise: Mill's view, and the argument for it, as to what the experientially privileged of the future will prefer with regard to the general happiness.

The chapter contains a number of intertwined arguments, and in the interest of brevity I will here pick out just one strand. Because each individual has an interest in *others'* altruism, individuals, whether or not altruistic themselves, will, "in proportion to the amount of general intelligence" (228/3:3), support measures that will induce utilitarian attitudes in their fellows:

> whatever amount of this feeling [for the good of others] a person has, he is urged by the strongest motives . . . of interest . . . to the utmost of his power to encourage it in others; and even if he has none of it himself, he is as greatly interested as any one else that others should have it. Consequently, the smallest germs of the feeling are laid hold of and nourished by . . . the influences of education; and a complete web of corroborative association is woven round it, by the powerful agency of external sanctions.
>
> . . . the influences are constantly on the increase, which tend to generate in each individual a feeling of unity with all the rest; which feeling, if perfect, would make him never think of, or desire, any beneficial condition for himself, in the benefits of which they are not included. If we now suppose this feeling of unity to be taught as a religion, and the whole force of education, of institutions, and of opinion, directed, as it once was in the case of religion, to make every person grow up from infancy surrounded on all sides both by the profession and by the practice of it, I think that no one, who can realize this conception, will feel any misgiving about the sufficiency of the ultimate sanction for the Happiness morality. (232/3:9)[31]

31. Mill has just given a prediction of the future state of society. Now we saw that Mill does not think that individual psychologies are predictable (n. 27, above). But he thinks that problems of individual psychological prediction can be sidestepped when large groups of people are being studied. The political scientist "can get on well enough with approximate generalizations on human nature, since what is true approximately of all individuals is true absolutely of all masses" (VII:603). Mill had been impressed by the new science of statistics (VIII:932) and was quite willing to believe that even if you could not predict the

Sooner or later, Mill thinks, self-interested motives make it almost inevitable that institutions guaranteeing a society of natural utilitarians be put in place. The persons created by these institutions will prefer the general happiness to any other alternative.[32] Because these institutions will be stable, the persons shaped by them will outnumber those living in previous historical periods. The preferences of the majority of the experientially privileged are decisive with respect to desirability of the objects of one's desires.[33] And consequently, the general happiness is desirable for each individual, no matter what his desires, and so is desirable for everyone. The second stretch of Mill's argument is deductively valid as well.

VII

Before considering how well the two stretches of Mill's argument sit together, I want to make a couple of remarks about the stretch of argument that we have just seen.

---

counter-factual preferences of individuals, under certain circumstances predicting the preferences of the majority of the experientially privileged was well within the realm of possibility (VIII:846–47; cf. VIII:873, 890).

In that case, the reader might be wondering, why not appeal to the counterfactual preferences of the experientially privileged as a group? Why the reliance on actual preferences? Mill does not actually discuss this option, but he would have had reasons for not availing himself of it. First of all, the desired social science (the science of character, which Mill dubbed "Ethology") was never more than a gleam in his eye; barring special circumstances, such as the argument we now have on hand, we don't for the present have any way of predicting the preferences of a group. Second, prediction in the social sciences has its limits, and of precise predictions of the history of society, extended arbitrarily far into the future (similar to those we have come to expect from astronomy), Mill thought, "there is . . . no hope" (VIII:877). Small errors snowball, and so the predictions of a "Deductive" science like Ethology would need to be continually checked against and corrected by observation—in this case, observation of the actual preferences of the experientially privileged. Since the actual preferences are not, on Mill's account of the social sciences, dispensable, he may have preferred to cut out, as so much wasted motion, the detour through counterfactuals.

32. Mill sometimes suggests that the economic arrangements of the future will also work to mute the conflict between the alternatives; e.g., in the *Principles of Political Economy*, he speaks of a time when "civilization and improvement have so far advanced, that what is a benefit to the whole shall be a benefit to each individual composing it," having in mind in this case the more equitable distribution of gains in productivity (III:768).

33. There's a nicety here that deserves pointing out. The criterion of desirability introduced in chap. 2 avoids circularity by specifying that the preferences consulted shall be those given "irrespective of any feeling of moral obligation to prefer" one item rather than another (211/2:5). (See n. 24, above.) In order to prevent this condition from excluding the preferences of the acculturated utilitarian majority for the general happiness over all else, Mill argues that the "strengthening of social ties . . . leads [each individual] to identify his *feelings* more and more with [others'] good . . . He comes, as though instinctively, to be conscious of himself as a being who *of course* pays regard to others. The good of others becomes to him a thing naturally and necessarily to be attended to" (231–32/3:9, emphases Mill's). Future utilitarians, thinks Mill, will not want one thing but feel constrained by the moral law to prefer another. Their desire for the general happiness will be a native preference, one that can thus serve as grist for the "decided preference" criterion.

First, the argument is valid, that is, its conclusions follow from its premises. But it is not, I think, sound. It was reasonable in Mill's day to believe that entire populations could be made over into the natural adherents of any ideology whatsoever. In addition to the considerations we have just laid out, Mill had a number of other reasons for thinking that utilitarian indoctrination would be possible. Some are mentioned in the same chapter, for example, the receptiveness of normal individuals to a humane ideology appealing to the sympathetic feelings. Others, such as his associationist psychology, are left in the background. But Mill's reasons are now beside the point: today we know better. The twentieth century has been a large-scale test of the effectiveness of ideological indoctrination, and it has turned out not to work. Mill thought that Comte had "superabundantly shown the possibility of giving to the service of humanity . . . the psychical power and the social efficacy of a religion" (232/3:9). But Mill and Comte were, as a matter of empirical fact, mistaken.[34]

Contemporary moral philosophers will tend to find the use of an empirical premise of this kind unsettling: if we have gotten the argument right, evaluative issues, such as whether the general happiness is a good, or the most important good, are *contingent.* This is easily overlooked when we are considering the long-term social processes that Mill thought would result in a society of utilitarians. But the desirability of the general happiness depends not just on the sociological laws that make those processes seem inevitable; there must be a sufficiently long run in which they are able to do their work. If human history is prematurely truncated by natural catastrophe, then utilitarianism will be made false: an asteroid striking the earth can, on this view, destroy not only things of value, but the values themselves.[35] Assessing the plausibility of this view is a topic for some other occasion; for now, it suffices that this upshot is not a reason to think we have misread Mill. Mill's official view is that *no* body of theory—logic and the mathematical sciences included—is a priori true or necessary;[36] so moral facts should be no less contingent than any other

34. For a lengthier and much more thorough rendition of Mill's views on this point, see his essay, "Auguste Comte and Positivism" (X:263–368).

35. There is in fact a more likely way for the contingency of value to exhibit itself, and one which I suspect worried Mill a good deal. If *A*'s being preferable to *B* is a matter of the majority of the experientially privileged preferring *A* to *B*, then *A*'s being preferable to *B* depends on there being individuals having experience of the different options: if there are no such individuals, or perhaps not enough of them, then there is no fact of the matter as to which is preferable. The applicability of the decided preference criterion requires that there be people who try things out—who conduct, in Mill's phrase, "experiments of living" (XVIII:261)—and I am inclined to think that this was one of Mill's reasons for insisting so strongly on the liberty to try things out. (Notice the awkward tradeoff: in order for there to be a standard of preferability, people have to, and hence have to be *encouraged* to [XVIII: 269], do what will prove to be the *in*correct thing.)

36. See VII:227, 231, 237–61, 277.

facts. Whether Mill managed to live up to his official view is a question that I will address when the time comes.

Second, the conclusion of Mill's argument was that the general happiness is "a good to the aggregate of all persons" (234/4:3). And we saw Sidgwick, not atypically, trying to read Mill as meaning that the general happiness is desirable for the aggregate: that it is a good of, as it were, the collective entity made up of all persons. But this, we can now see, is to misconstrue the sentence, and in particular, to mistake the force of the preposition. Mill is not committed to anything along the lines of a collective good. The "aggregate of all persons" is not some metaphysically dubious creature whose good the general happiness is, but the population from which the experientially privileged who provide the test of desirability are drawn. "The general happiness [is] . . . a good *to* the aggregate of all persons" means, roughly: among all persons, those capable of judging from experience, or at any rate a majority of them, find (or will find) the general happiness to be a good.[37]

Before proceeding to its problems, let me pause to recapitulate the merits of the argument as I have reconstructed it. It is deductively valid. That it is deductively valid explains Mill's unwillingness to call it a proof, and does so without uncharitably construing Mill as attempting to excuse his own sloppiness in advance. The premises of the argument are contributed by the previous chapters of *Utilitarianism;* on the reading just advanced, the book is a single argument, each carefully machined component of which is necessary for the conclusion finally drawn in chapter 4. And that conclusion proves to be metaphysically sane, in not presupposing the existence of collective social organisms with desires and interests.

If the reading I have given is on target, the text proves to be much more tightly written, and in some ways much more difficult to swallow, than it is usually taken to be. Mill is often pressed into service as the spokesman of a vague and well-meaning lowest common denominator of liberalism. But *Utilitarianism* is not a quotable collection of inoffensive platitudes. It is philosophical writing of the most muscular kind.

37. The sentence can read on the model of sentences like: "*To* the scientific community, the theory of relativity is true." Norman Kretzmann takes Mill's argument to lean in a pragmatist direction (Kretzmann, p. 239), and it is in this vicinity that I find his suggestion most evocative.

At this point we are able to say what Mill has in mind in his response to Spencer. That "the truths of arithmetic are applicable to the valuation of happiness" (258n./5:35n.) is to be read off the preferences of the experientially privileged. (They are not always applicable: higher and lower pleasures are incommensurable, and this fact is also to be read off the preferences of the experientially privileged.) The processes that will make future persons into natural utilitarians will, Mill further argues, also create persons who take it for granted "that the interests of all are to be regarded equally" (231/3:9).

VIII

We have seen that Mill's argument is valid, but it is not for that reason unproblematic. Its problems are worth exploring, and it is now time to take a closer look at them.

The first stretch of Mill's argument was driven by the common core of instrumentalist accounts of practical reason: the view, which I earlier pointed out is somewhat less than a theory of practical reasoning, that all practical inference is means-end inference. Since any chain of practical inferences will terminate in a desire that one just *has,* and since, at the end of the chain, there is no further argument to be given for the desirability of the final end, desirability can be identified with being desired. This identification underwrites Mill's treatment of felt desire "proving" desirability on the model of felt sensation "proving" the basic premises of one's system of beliefs.

The second stretch of Mill's argument exploits a device introduced to handle the shortcomings of the theory fragment that is the common core. These shortcomings, recall, were that the desires one actually has cannot be made the measure of desirability after all, due to the possibility of disappointment, and that the need to choose between competing desires must be accommodated in one's theory of practical reasoning. The device Mill uses, a standard made up of the preferences of other more experienced persons, adheres to the letter of the instrumentalist common core, in that it does not require any further forms of practical inference, over and above means-end reasoning. The inputs to one's practical inferences are indeed being corrected, but not inferentially. The corrected preferences are generated by a black box, a device whose inner workings are invisible and irrelevant.

The problem, at the level of the underlying account of practical reasoning, is this. An instrumentalist theorist who has gotten to this point is not in a position to justify the use (or choice of) a black box; he cannot say why what it produces are *corrected* preferences. For suppose there were an argument that showed the method employing the black box to be a method of correction. That argument could with minimal effort be turned into an argument that the desires or preferences produced by the method were *correct.* But that argument would be an instance of practical reasoning; that practical reasoning would be noninstrumental; and so acknowledging it would mean abandoning the instrumentalist common core.[38] The incoherence is, as far as I can see, inescapable: the prob-

38. It might look like there are a couple of places to get off the boat here: you might object that the reasoning in question isn't actually practical, or, alternatively, that, although practical, it is merely means-end reasoning. But while one can always get out of the boat, it turns out that, in this case, there's no pier handy to step onto.

Notice, first, that the two objections are not as different in spirit as they might at first glance seem to be. Since the reasoning was identified as practical by identifying its conclusion as practical, the first objection requires insisting that that conclusion does not give the

lem is a problem not only for Mill, but for contemporary instrumentalists who appeal to counterfactually informed desires. Any instrumentalist must use some such device; but justifying the device means abandoning instrumentalism. Instrumentalism is consequently an insupportable view of practical reasoning.[39]

The incoherence near the surface of Mill's argument can be summarized as follows. The first stretch of the argument makes sense on the supposition that, once you have gotten through the instrumental justifications, there is nothing to be said for an object of desire being desirable, other than that you desire it. The second stretch of argument makes sense on the supposition that there *is* something more to be said about whether an object of desire is desirable, namely, that certain other people, better situated than you are, desire it (or that they don't). Mill is careful to avoid actual contradiction; the decided preference criterion can be inserted parenthetically at the appropriate points in the first

---

agent a reason for action. Now this latter thought is normally underwritten by the idea that sentences like "I ought to do this" or "This is the correct preference" do not express reasons for action because they do not express desires. But the requirement that reasons for action be desires is in turn normally underwritten by instrumentalism.

Because the point of correcting the agent's preferences was to pick out his reasons for action, let's abandon the first objection in favor of the second: the reasoning is conceded to be practical, but is entirely means-end, that is, correcting one's preferences is shown to be a way of satisfying some further desire. But now the reasoning so construed faces an uncomfortable dilemma. Either the desires from which it proceeds are corrected, or they are not. If they are uncorrected, then we have the appearance of a pragmatic contradiction: the conclusion of the reasoning, recall, was that one should use in one's practical reasoning, not one's uncorrected, but *corrected* desires or preferences. If, on the other hand, they are corrected, then we have the question blatantly begged. To see that, consider the "black box" that corrects your preferences to mine. Asked why you should regard my preferences as those you should act on, rather than your own, I would not convince you by answering: because I would prefer it. The cleanest way out of the dilemma would be to invoke desires or preferences that one was sure would not need correction, and so for which the uncomfortable choice does not arise. But the confidence required here would be inappropriate for human beings: to be sure (ahead of time, of *any* desires) that one will not be disappointed—so sure, as to be willing to make those desires a pivot of one's practical logic—is just to show a failure of imagination in the course of setting oneself up for a fall. Human beings can end being disappointed in anything.

39. The point is not that appeals to informed desires are in principle dispensable: taking the argument in that direction would bring us around to a conclusion I would not wish to endorse, that there are no properties that can be understood only via the reactions of their perceivers. (Such properties have traditionally been called "secondary qualities.") The point is, rather, that the explanation for taking the perception of some secondary quality to provide a reason for action will involve a noninstrumental pattern of practical inference.

The problem of informed desires is an interesting and important topic in its own right, not merely an unworkable fix for instrumentalism. I do not here mean to be endorsing one or another approach to it, and in particular I should not be taken as recommending the decided preference criterion as a replacement for contemporary counterfactual-based devices.

stretch of argument without aborting it.[40] But the two stretches of argument cannot both make sense together.

This near-surface incoherence is, we now see, the immediate expression of a similarly shaped incoherence in the underlying view of practical reasoning. Instrumentalism is the engine of Mill's utilitarianism, and instrumentalism is not a sustainable view of practical rationality. Mill's ingenious proof of the Principle of Utility does not commit the crude mistakes for which it is known. It commits instead a deep and philosophically interesting mistake, and one from which we have much to learn.

IX

I have argued that Mill's instrumentalism gets him into trouble; so we now need to explain Mill's being an instrumentalist. I'll begin by explaining why an explanation is called for.

Instrumentalism, once again, is the view that all practical inference is means-end inference. So it is a view about what patterns of inference are legitimate, which is to say that it is a logical thesis, in Mill's sense of "Logic." Now Mill was not in the business of accepting received views about inference just because they were received: his *System of Logic*, I have already mentioned, argued for the heretical view that deductive inference is not actually inference at all. Mill was not shy when it came to arguing for philosophical views that he needed. So if instrumentalism is a view in Logic, if it is explicitly acknowledged, and if it is doing the work for him that I have argued it is, then Mill, of all people, should be providing us with an argument for it.

But the expected argument is conspicuously lacking. The closest Mill comes is a discussion which perhaps can be made to speak to a puzzle we have had on the queue for a while. Recall that I suggested that, on Mill's view, the claim that people desire their own happiness amounts to a tautology, roughly, that people desire what they desire; Mill, how-

40. Very quickly: To desire something, noninstrumentally, is to find the idea of it pleasurable, that is, to expect it to be pleasurable; and since the "decided preference criterion" is being used as a way of correcting desires so their objects *are* found pleasurable, we can read 'desired' as 'desired, subject to correction by the judgments of the experientially privileged'. Happiness means pleasure, and so now means something like: the aggregate of the objects of one's desires, as corrected by the preferences of the experientially privileged. So the premise of the stretch of argument that we are now marvering comes out as: the objects of desires corrected by the preferences of the experientially privileged are desired, when those desires are corrected by the preferences of the experientially privileged. Once again, the premise of the argument is a tautology. Now one's noninstrumental desires, corrected by the preferences of the experientially privileged, are what is desirable (by the "decided preference criterion"). So the conclusion, that each person's happiness is desirable for him, comes out meaning: the objects of one's desires, as corrected by the preferences of the experientially privileged (that is, when attained, happiness), are the objects of one's desires, as corrected by the preferences of the experientially privileged (that is, what is desirable). The conclusion is also, once again, a tautology; and is, once again, the premise of the argument repeated.

ever, describes the claim, not as true by terminological stipulation, but as "psychologically true" (237/4:9). Now this might of course simply be a reflexive expression of Mill's doctrine that the truth of *any* claim is an empirical question, but there is likely to be more to it than that: tautologies need not be "unmeaning" if they bring to the fore the presuppositions that make the vocabulary in them usable.[41]

For Mill's psychological views, the best source is his father's *Analysis of the Phenomena of the Human Mind,* to which the younger Mill, as editor of a second edition, added extensive commentary. That commentary consists quite often of rather blunt correction; we may take it that where the editor's notes do not indicate disagreement or qualification the views expressed are John Stuart Mill's as well. So we should look to the *Analysis* for the question of psychological fact that Mill takes to lie beneath the questions of naming.[42] Now what remains, in 'people desire what they desire', to be underwritten by experimental psychology, is the nonredundant assertion that people *desire:* that human beings are wanting creatures. And the *Analysis* does indeed contain an extended exposition of just this view. Desires are analyzed (as ideas of pleasure); commonplace desires (e.g., for "Wealth, Power, and Dignity") are accounted for; and their role in producing action is explained.[43]

That human beings are desiring creatures, and that their desires move them to action, is not itself an argument for instrumentalism. Mill was quite clear that whether something is an inference is a matter not of what one *does,* but of what one *should* do. The *Analysis* does contain an associationist reconstruction of the psychological movements involved in means-end reasoning; and it defines the Will in such a way that "in all cases in which the action is said to be Willed, it is desired, as a means to

41. IX:468–69; the passage provides two examples.

42. See n. 21 above.

43. Mill, *Analysis,* vol. 2, pp. 191–92, for the analysis; notice the younger Mill's modification to the definition in the editorial footnote at pp. 194–95. For the desires for wealth, etc., see pp. 207–14. For the role of desires in action, see pp. 256–59, 262n.–64, 327–403.

We can now tie up another loose end: the analysis of desire accounts for what we would see as the terminological slide that I highlighted in n. 14, above. Like his son, James Mill identifies, as a matter of terminology, pleasure (or more carefully, the representation of pleasure) with desire: "The terms . . . 'idea of pleasure,' and 'desire,' are but two names; the thing named, the state of consciousness, is one and the same" (pp. 191–92; cf. p. 327). A desire just is an idea of the sensation of pleasure, which idea may be associated with other ideas, most prominently, of causes of the sensation. It follows that understanding desires as (what we would nowadays call) propositional attitudes, i.e., attitudes whose objects are as it were logically integral to the mental state, is a *mistake:* "when the word desire is applied to the cause of a sensation . . . it is employed in a figurative, or metaphorical, not in a direct sense" (p. 351). What we would call the objects of desire are actually just the causes of the real objects, pain and pleasure: "The illusion is merely that of a very close association" (p. 192). This view, quite startling by today's lights, accounts for Mill's treating the moves from desiring *X,* to finding *X* pleasurable, to desiring pleasure (which happens to be associated with *X*) as traversing a series of innocuous synonymies.

an end."[44] But there is no argument to the effect that associations that proceed in these ways (and only in these ways) are *correct*, and the *Analysis* similarly reconstructs other patterns of association that clearly are not understood inferentially, for example, "proneness to sympathize with the Rich and Great."[45] Mill's psychological views go a great distance toward explaining why he chose the terminology he did, and why he regarded it as expressing interesting empirical facts; but it does not help at all in providing Mill with grounds for his instrumentalist take on practical reasoning.

Beyond these considerations, Mill's corpus contains no treatment of the problem whatsoever.[46] One possible explanation is that the thought that practical inference might have a place in an exhaustive treatment of the theory of inference eluded him, because "Logic, as [Mill] conceive[d] it, is the entire theory of the ascertainment of reasoned or inferred truth" (VII:206)—and practical inference is naturally seen as having something other than truth as its formal object. Since what we are trying to explain is Mill's silence on the matter, any line we take will be going out on a limb. But as long as we are being speculative, there is a way of playing out this possibility that I would like to explore.

Let us consider for a moment what a Millian investigation of the forms of practical inference would have looked like. Mill had definite views about how to argue about logic: it is an empirical science, and one establishes what the correct forms of inference are inductively.[47] To show that a candidate pattern of inference is legitimate, that is, is a pattern of *inference*, is to show, among other things,[48] that it works; and the demonstration must be an induction from particular instances.

Now verification of this kind ultimately comes down to observation of particular cases.[49] Assessing the effectiveness of a particular inference will involve assessing the starting points and terminus of that inference. You must determine that, for instance, the premises of an argument being inspected were true, and the conclusion was also true; and doing this will require observing matters of particular fact. So the empirical investigation of inference requires the ability to make observations in the per-

44. Ibid., vol. 2, pp. 357n., 378.

45. Ibid., vol. 2, pp. 209–11.

46. To be sure, one can see how the fact that human beings are desiring creatures might be made the starting point of an argument for the view we are considering. But the argument is not in Mill, and the mode of argument, as we will see in a moment, would not be one that Mill could admit.

47. See VII:567–77; also VII:306–14; VIII:833.

48. Foremost among further criteria that need to be satisfied in order for a mental transition to amount to an inference is, recall, that "the conclusion is . . . wider than the premises from which it is drawn" (VII:288). Mill never seems to have asked whether means-end reasoning, which he endorses, satisfies this condition, and it is unclear to me, in view of his discussion of parallel questions, how he could have given a satisfactory answer.

49. See, e.g., VII:193.

tinent domain. Were the putative inferences to be practical, that would mean making observations whose content was practical; that is to say, some statements along the lines of "Such-and-such is desirable" would have to count as *observations*.

Mill was a British Empiricist—the fourth of the great British empiricists, and the most thoroughly radical. Even Hume had allowed logic to be a priori; Mill was determined to be empiricist about logic as well. But even Mill was not an empiricist to the end. Like his predecessors, he took the class of basic observations for granted: they were sensations, the descendents of (most recently) Humean "impressions of sensation."[50] Because he was not empiricist enough to allow what an observation is to be an empirical question, there was no room in his systematic view for practical observation. Without practical observation, there was no room for the empirical investigation of practical inference. And so, despite his avowed empiricism about logic, Mill allowed himself to treat his views about practical inference as true a priori.

Mill has recently gotten a certain amount of publicity for having taken an empiricist approach in moral matters.[51] But, if I am right, the

50. Even though sense-data are now out of style, taking a class of basic observations for granted is not. A familiar recent example is van Fraassen's identification of observation with "what the unaided eye discerns" (*The Scientific Image* [Oxford: Clarendon Press, 1980]), p. 59; see also pp. 16 ff., 56 ff., 214; for criticism, see essays in *Images of Science: Essays on Realism and Empiricism*, ed. Paul Churchland and Clifford Hooker [Chicago: University of Chicago Press, 1985]).

51. Anderson, for instance, adduces under this heading Mill's diagnosis of his nervous breakdown as an "experiment in living" (Elizabeth Anderson, "John Stuart Mill and Experiments in Living," *Ethics* 102 [1991]: 4–26). I think that it can be so understood, but not the way Anderson would like to: Mill himself was surprisingly unsuccessful in learning the lessons of his breakdown and partial recovery. He was never able to provide a satisfactory account of the importance in his own life of Romantic poetry; willing to surrender neither his instrumentalism nor his associationism, he was forced to treat the resistance of associations formed under the influence of poetry to analytical corrosion as a brute psychological fact. And while Mill is quite clear about his own motivational structure up until his breakdown, his view of himself after his recovery fails, quite strikingly, to match the ways he thought, felt, and lived. I hope to discuss these issues further elsewhere.

There is another confusion in Anderson's paper that is common enough to be worth remark. Anderson thinks that "Mill's emphasis on the intrinsic desirability of gratifying the [higher] sentiments strongly suggests that he believed that dignity, beauty, honor, and so forth are values distinct from pleasure," and she takes Mill to be departing from "the basic premise of ethical hedonism, that pleasure is the sole respect in which things can be intrinsically valuable." She concludes that "Mill's conception of the good is not hedonistic but pluralistic and hierarchical."

But this is just to lose track of Mill's terminology. For something to be desirable is, as we have seen, just to be desired (noninstrumentally, and subject to the correction of the experientially privileged); this is, merely as a matter of nomenclature, to be found pleasurable. Consequently, being pleasurable, in Mill's sense, is not a *respect* in which things can be found desirable. And if we tie the word 'hedonism' to Mill's word 'pleasure', whether Mill's conception of the good is "pluralistic and hierarchical" has nothing at all to do with whether it is "hedonistic."

publicity is unwarranted: it was precisely here that Mill chickened out. Mill is able to let what people want be a matter of experience,[52] and finding out how to get what one wants is of course a matter of experience. But that what *matters* is people getting what they want is not a matter of experience, because Mill is barred from understanding the preferences of the experientially privileged as observations. This restriction is on display in two particularly prominent places. The first is the infallibility of the collective preferences of the experientially privileged. Mill cannot be providing an empirical test for distinguishing higher from lower pleasures, because empirical methods are fallible. Second, the processes described in chapter 3 of *Utilitarianism,* through which the preferences of the majority of the experientially privileged are to be brought into line, are too clearly manipulative to allow the resulting preferences to be understood as observations. The citizens of Mill's Brave New World are being *conditioned.* The resulting preferences depend on the social pressures that drive the processes of conditioning, and not, in any way appropriate to observation, on the preferences' objects.

And so we have come to the final moral I wish to draw from the story I have now finished recounting. We have already seen that Mill's utilitarianism is very closely tied to his instrumentalism; that his argument for the Principle of Utility, while tight, is deeply incoherent; that the incoherence stems from an incoherence in instrumentalism; and that Mill's instrumentalism turns out to have been an island of apriorism in an otherwise empiricist project. It is tempting to think that if Mill had been willing to look to experience at this point also, his theory of practical reasoning, and consequently his moral theory, would have turned out quite differently—and perhaps less incoherently. So the final moral is that if you're going to be empiricist, be empiricist all the way: about practical reasoning, and about observation.

52. That people want happiness is, we have seen, a verbal matter, and not a matter of experience at all. But for that reason, happiness is not a substantive account of what is wanted; the substantive account is to be given by empirical investigation of what people, and their experientially privileged stand-ins, actually desire.