VII

CONDITIONS OF PERSONHOOD

DANIEL DENNETT

I am a person, and so are you. That much is beyond doubt. I am a human being, and *probably* you are too. If you take offense at the "probably" you stand accused of a sort of racism, for what is important about us is not that we are of the same biological species, but that we are both persons, and I have not cast doubt on that. One's dignity does not depend on one's parentage even to the extent of having been born of woman or born at all. We normally ignore this and treat humanity as the deciding mark of personhood, no doubt because the terms are locally coextensive or almost coextensive. At this time and place human beings are the only persons we recognize, and we recognize almost all human beings as persons, but on the one hand we can easily contemplate the existence of biologically very different persons—inhabiting other planets, perhaps—and on the other hand we recognize conditions that exempt human beings from personhood, or at least some very important elements of personhood. For instance, infant human beings, mentally defective human beings, and human beings declared insane by licensed psychiatrists are denied personhood, or at any rate crucial elements of personhood.

One might well hope that such an important concept, applied and denied so confidently, would have clearly formulatable necessary and sufficient conditions for ascription, but if it does, we have not yet discovered them. In the end there may be none to discover. In the end we may come to realize that the concept of a person is incoherent and obsolete. Skinner, for one, has suggested this, but the doctrine has not caught on, no doubt in part because it is difficult or even impossible to conceive of what it would be like if we abandoned the concept of a person. The idea that we

might cease to view others and *ourselves* as persons (if it does not mean merely that we might annihilate ourselves, and hence cease to view anything as anything) is arguably self-contradictory.[1] So quite aside from whatever might be right or wrong in Skinner's grounds for his claim, it is hard to see how it could win out in contest with such an intuitively invulnerable notion. If then the concept of a person is in some way an ineliminable part of our conceptual scheme, it might still be in rather worse shape than we would like. It might turn out, for instance, that the concept of a person is only a free-floating honorific that we are all happy to apply to ourselves, and to others as the spirit moves us, guided by our emotions, aesthetic sensibilities, considerations of policy, and the like —just as those who are *chic* are all and only those who can get themselves considered *chic* by others who consider themselves *chic*. Being a person is certainly *something* like that, and if it were no more, we would have to reconsider if we could the importance with which we now endow the concept.

Supposing there *is* something more to being a person, the searcher for necessary and sufficient conditions may still have difficulties if there is more than one concept of a person, and there are grounds for suspecting this. Roughly, there seem to be two notions intertwined here, which we may call the moral notion and the metaphysical notion. Locke says that "person"

> is a forensic term, appropriating actions and their merit; and so belongs only to intelligent agents, capable of a law, and happiness, and misery. This personality extends itself beyond present existence to what is past, only by consciousness—whereby it becomes concerned and accountable. (*Essays*, Book II, Chap. XXVII)

Does the metaphysical notion—roughly, the notion of an intelligent, conscious, feeling agent—*coincide* with the moral notion—roughly, the notion of an agent who is accountable, who has both rights and responsibilities? Or is it merely that being a person in the metaphysical sense is a necessary but not sufficient condition of being a person in the moral sense? Is being an entity to which states of consciousness or self-consciousness are ascribed *the same* as being an end-in-oneself, or is it merely one precondition? In Rawls's theory of justice, should the derivation from the original position be viewed as a demonstration of how metaphysical per-

sons *can become* moral persons, or should it be viewed as a demonstration of why metaphysical persons *must be* moral persons?[2] In less technical surroundings the distinction stands out as clearly: when we declare a man insane we cease treating him as accountable, and we deny him most rights, but still our interactions with him are virtually indistinguishable from normal personal interactions unless he is very far gone in madness indeed. In one sense of "person," it seems, we continue to treat and view him as a person. I claimed at the outset that it was indubitable that you and I are persons. I could not plausibly hope—let alone aver—that all readers of this essay will be legally sane and morally accountable. What—if anything—was beyond all doubt may only have been that anything properly addressed by the opening sentence's personal pronouns, "you" and "I," was a person in the metaphysical sense. If that was all that was beyond doubt, then the metaphysical notion and the moral notion must be distinct. Still, even if we suppose there are these distinct notions, there seems every reason to believe that metaphysical personhood is a necessary condition of moral personhood.[3]

What I wish to do now is consider six familiar themes, each a claim to identify a necessary condition of personhood, and each, I think, a correct claim on some interpretation. What will be at issue here is first, how (on my interpretation) they are dependent on each other; second, why they are necessary conditions of moral personhood, and third, why it is so hard to say whether they are jointly sufficient conditions for moral personhood. The *first* and most obvious theme is that persons are *rational beings.* It figures, for example, in the ethical theories of Kant and Rawls, and in the "metaphysical" theories of Aristotle and Hintikka.[4] The *second* theme is that persons are beings to which states of consciousness are attributed, or to which psychological or mental or *Intentional predicates*, are ascribed. Thus Strawson identifies the concept of a person as "the concept of a type of entity such that *both* predicates ascribing states of consciousness *and* predicates ascribing corporeal characteristics" are applicable.[5] The *third* theme is that whether something counts as a person depends in some way on an *attitude taken* toward it, a *stance adopted* with respect to it. This theme suggests that it is not the case that once we have established the objective fact that something is a person we treat him or her or

it a certain way, but that our treating him or her or it in this certain way is somehow and to some extent constitutive of its being a person. Variations on this theme have been expressed by MacKay, Strawson, Amelie Rorty, Putnam, Sellars, Flew, Thomas Nagel, Dwight Van de Vate, and myself.[6] The *fourth* theme is that the object toward which this personal stance is taken must be capable of *reciprocating* in some way. Very different versions of this are expressed or hinted at by Rawls, MacKay, Strawson, Grice, and others. This reciprocity has sometimes been rather uninformatively expressed by the slogan: to be a person is to treat others as persons, and with this expression has often gone the claim that treating another as a person is treating him morally—perhaps obeying the Golden Rule, but this conflates different sorts of reciprocity. As Nagel says, "extremely hostile behavior toward another is compatible with treating him as a person" (p. 134), and as Van de Vate observes, one of the differences between some forms of manslaughter and murder is that the murderer treats the victim as a person.

The *fifth* theme is that persons must be capable of *verbal communication*. This condition handily excuses nonhuman animals from full personhood and the attendant moral responsibility, and seems at least implicit in all social contract theories of ethics. It is also a theme that has been stressed or presupposed by many writers in philosophy of mind, including myself, where the moral dimension of personhood has not been at issue. The *sixth* theme is that persons are distinguishable from other entities by being *conscious* in some special way: there is a way in which *we* are conscious in which no other species is conscious. Sometimes this is identified as *self*-consciousness of one sort or another. Three philosophers who claim—in very different ways—that a special sort of consciousness is a precondition of being a moral agent are Anscombe, in *Intention*, Sartre, in *The Transcendence of the Ego*, and Harry Frankfurt, in his recent paper, "Freedom of the Will and the Concept of a Person."[7]

I will argue that the order in which I have given these six themes is—with one proviso—the order of their dependence. The proviso is that the first three are mutually interdependent; being rational is being Intentional is being the object of a certain stance. These three together are a necessary but not sufficient condition for

exhibiting the form of reciprocity that is in turn a necessary but not sufficient condition for having the capacity for verbal communication, which is the necessary[8] condition for having a special sort of consciousness, which is, as Anscombe and Frankfurt in their different ways claim,[9] a necessary condition of moral personhood.

I have previously exploited the first three themes, rationality, Intentionality and stance, to define not persons, but the much wider class of what I call *Intentional systems*, and since I intend to build on that notion, a brief résumé is in order. An Intentional system is a system whose behavior can be (at least sometimes) explained and predicted by relying on ascriptions to the system of *beliefs* and *desires* (and other Intentionally characterized features —what I will call *Intentions* here, meaning to include hopes, fears, intentions, perceptions, expectations, etc.). There may *in every case* be other ways of predicting and explaining the behavior of an Intentional system—for instance, mechanistic or physical ways—but the Intentional stance may be the handiest or most effective or in any case *a* successful stance to adopt, which suffices for the object to be an Intentional system. So defined, Intentional systems are obviously not all persons. We ascribe beliefs and desires to dogs and fish and thereby predict their behavior, and we can even use the procedure to predict the behavior of some machines. For instance, it is a good, indeed the only good, strategy to adopt against a good chess-playing computer. By *assuming* the computer has certain beliefs (or information) and desires (or preference functions) dealing with the chess game in progress, I can calculate —under auspicious circumstances—the computer's most likely next move, *provided I assume the computer deals rationally with these beliefs and desires*. The computer is an Intentional system in these instances not because it has any particular intrinsic features, and not because it really and truly has beliefs and desires (whatever that would be), but just because it succumbs to a certain *stance* adopted toward it, namely the Intentional stance, the stance that proceeds by ascribing Intentional predicates under the usual constraints to the computer, the stance that proceeds by considering the computer as a rational practical reasoner.

It is important to recognize how bland this definition of *Intentional system* is, and how correspondingly large the class of Inten-

tional systems can be. If, for instance, I predict that a particular plant—say a potted ivy—will grow around a corner and up into the light because it "seeks" the light and "wants" to get out of the shade it now finds itself in, and "expects" or "hopes" there is light around the corner, I have adopted the Intentional stance toward the plant, and lo and behold, within very narrow limits it works. Since it works, some plants are very low-grade Intentional systems.

The actual utility of adopting the Intentional stance toward plants was brought home to me talking with loggers in the Maine woods. These men invariably call a tree not "it" but "he," and will say of a young spruce "he wants to spread his limbs, but don't let him; then he'll have to stretch up to get his light" or "pines don't like to get their feet wet the way cedars do." You can "trick" an apple tree into "thinking it's spring" by building a small fire under its branches in the late fall; it will blossom. This way of talking is not just picturesque and is not really superstitious at all; it is simply an efficient way of making sense of, controlling, predicting, and explaining the behavior of these plants in a way that nicely circumvents one's ignorance of the controlling mechanisms. More sophisticated biologists may choose to speak of information transmission from the tree's periphery to other locations in the tree. This is less picturesque, but still Intentional. Complete abstention from Intentional talk about trees can become almost as heroic, cumbersome, and pointless as the parallel strict behaviorist taboo when speaking of rats and pigeons. And even when Intentional glosses on (e.g.) tree-activities are of vanishingly small heuristic value, it seems to me wiser to grant that such a tree is a very degenerate, uninteresting, negligible Intentional system than to attempt to draw a line above which Intentional interpretations are "objectively true."

It is obvious, then, that being an Intentional system is not sufficient condition for being a person, but is surely a necessary condition. Nothing to which we could not successfully adopt the Intentional stance, with its presupposition of rationality, could count as a person. Can we then define persons as a subclass of Intentional systems? At first glance it might seem profitable to suppose that persons are just that subclass of Intentional systems that *really* have beliefs, desires, and so forth, and are not merely *supposed* to

have them for the sake of a short-cut prediction. But efforts to say what counts as really having a belief (so that no dog or tree or computer could qualify) all seem to end by putting conditions on genuine belief that (1) are too strong for our intuitions, and (2) allude to distinct conditions of personhood farther down my list. For instance, one might claim that genuine beliefs are necessarily *verbally expressible* by the believer,[10] or the believer must be *conscious* that he has them, but people seem to have many beliefs that they cannot put into words, and many that they are unaware of having—and in any case I hope to show that the capacity for verbal expression, and the capacity for consciousness, find different *loci* in the set of necessary conditions of personhood.

Better progress can be made, I think, if we turn to our fourth theme, reciprocity, to see what kind of definition it could receive in terms of Intentional systems. The theme suggests that a person must be able to reciprocate the stance, which suggests that an Intentional system that itself adopted the Intentional stance toward other objects would meet the test. Let us define a *second-order Intentional system* as one to which we ascribe not only simple beliefs, desires and other Intentions, but beliefs, desires, and other Intentions *about* beliefs, desires, and other Intentions. An Intentional system *S* would be a second-order Intentional system if among the ascriptions we make to it are such as *S believes that T desires that p, S hopes that T fears that q*, and reflexive cases like *S believes that S desires that p.* (The importance of the reflexive cases will loom large, not surprisingly, when we turn to those who interpret our sixth condition as *self*-consciousness. It may seem to some that the reflexive cases make all Intentional systems automatically second-order systems, and even *n*-order systems, on the grounds that believing that *p* implies believing that you believe that *p* and so forth, but this is a fundamental mistake; the iteration of beliefs and other Intentions is never redundant, and hence while some iterations are normal [are to be expected] they are never trivial or automatic.)

Now are human beings the only second-order Intentional systems so far as we know? I take this to be an empirical question. We ascribe beliefs and desires to dogs, cats, lions, birds, and dolphins, for example, and thereby often predict their behavior—when all goes well—but it is hard to think of a case where an ani-

mal's behavior was so sophisticated that we would need to ascribe second-order Intentions to it in order to predict or explain its behavior. Of course if some version of mechanistic physicalism is true (as I believe), we will never *need* absolutely to ascribe any Intentions to anything, but supposing that for heuristic and pragmatic reasons we were to ascribe Intentions to animals, would we ever feel the pragmatic tug to ascribe second-order Intentions to them? Psychologists have often appealed to a principle known as Lloyd Morgan's Canon of Parsimony, which can be viewed as a special case of Occam's Razor; it is the principle that one should attribute to an organism as little intelligence or consciousness or rationality or mind as will suffice to account for its behavior. This principle can be, and has been, interpreted as demanding nothing short of radical behaviorism[11] but I think this is a mistake, and we can interpret it as the principle requiring us when we adopt the Intentional stance toward a thing to ascribe the simplest, least sophisticated, lowest-order beliefs, desires, and so on, that will account for the behavior. Then we will grant, for instance, that Fido *wants* his supper, and *believes* his master will give him his supper if he begs in front of his master, but we need not ascribe to Fido the further *belief* that his begging induces a *belief* in his master that he, Fido, *wants* his supper. Similarly, my *expectation* when I put a dime in the candy machine does not hinge on a further *belief* that inserting the coin induces the machine to *believe* I *want* some candy. That is, while Fido's begging looks very much like true second-order interacting (with Fido treating his master as an Intentional system), if we suppose that to Fido his master is just a supper machine activated by begging, we will have just as good a predictive ascription, more modest but still, of course, Intentional.

Are dogs, then, or chimps or other "higher" animals, incapable of rising to the level of second-order Intentional systems, and if so why? I used to think the answer was Yes, and I thought the reason was that nonhuman animals lack language, and that language was needed to represent second-order Intentions. In other words, I thought condition four might rest on condition five. I was tempted by the hypothesis that animals cannot, for instance, have second-order beliefs, beliefs about beliefs, for the same reason they cannot have beliefs about Friday, or poetry. Some beliefs can only be acquired, and hence represented, via language.[12] But if it is true

that some beliefs cannot be acquired without language, it is false that all second-order beliefs are among them, and it is false that non-humans cannot be second-order Intentional systems. Once I began asking people for examples of non-human second-order Intentional systems, I found some very plausible cases. Consider this from Peter Ashley (in a letter):

> One evening I was sitting in a chair at my home, the *only* chair my dog is allowed to sleep in. The dog was lying in front of me, whimpering. She was getting nowhere in her trying to "convince" me to give up the chair to her. Her next move is the most interesting, nay, the *only* interesting part of the story. She stood up, and went to the front door where I could still easily see her. She scratched the door, giving me the impression that she had given up trying to get the chair and had decided to go out. However as soon as I reached the door to let her out, she ran back across the room and climbed into her chair, the chair she had "forced" me to leave.

Here it seems we must ascribe to the dog the *intention* that her master *believe* she *wants* to go out—not just a second-order, but a third-order Intention. The key to the example, what makes it an example of a higher-order Intentional system at work, is that the belief she intends to induce in her master is false. If we want to discover further examples of animals behaving as second-order Intentional systems it will help to think of cases of deception, where the animal, believing p, tries to get another Intentional system to believe *not-p*. Where an animal is trying to induce behavior in another which *true* beliefs about the other's environment would not induce, we cannot "divide through" and get an explanation that cites only first-level Intentions. We can make this point more general before explaining why it is so: where x is attempting to induce behavior in y which is inappropriate to y's *true* environment and needs but appropriate to y's *perceived* or *believed* environment and needs, we are forced to ascribe second-order Intentions to x. Once in this form the point emerges as a familiar one, often exploited by critics of behaviorism: one can be a behaviorist in explaining and controlling the behavior of laboratory animals only so long as he can rely on there being no serious dislocation between the actual environment of the experiment and the environment perceived by the animals. A tactic for embarrassing behaviorists in the laboratory is to set up experiments that deceive the subjects: if the deception succeeds their behavior is predictable

from their false *beliefs* about the environment, not from the actual environment. Now a first-order Intentional system is a behaviorist; it ascribes no Intentions to anything. So if we are to have good evidence that some system *S* is *not* a behaviorist—is a second-order Intentional system—it will only be in those cases where behaviorist theories are inadequate to the data, only in those cases where behaviorism would not explain system *S*'s success in manipulating another system's behavior.

This suggests that Ashley's example is not so convincing after all, that it can be defeated by supposing his dog is a behaviorist of sorts. She need not believe that scratching on the door will induce Ashley to believe she wants to go out; she may simply believe, as a good behaviorist, that she has conditioned Ashley to go to the door when she scratches. So she applies the usual stimulus, gets the usual response, and that's that. Ashley's case succumbs if this is a *standard* way his dog has of getting the door opened, as it probably is, for then the more modest hypothesis is that the dog believes her master is conditioned to go to the door when she scratches. Had the dog done something *novel* to deceive her master (like running to the window and looking out, growling suspiciously) then we would have to grant that rising from the chair was no mere conditioned response in Ashley, and could not be "viewed" as such by his dog, but then, such virtuosity in a dog would be highly implausible.

Yet what is the difference between the implausible case and the well-attested cases where a low-nesting bird will feign a broken wing to lure a predator away from the nest? The effect achieved is novel, in the sense that the bird in all likelihood has not repeatedly conditioned the predators in the neighborhood with this stimulus, so we seem constrained to explain the ploy as a bit of genuine deception, where the bird *intends* to induce a false *belief* in the predator. Forced to this interpretation of the behavior, we would be mightily impressed with the bird's ingenuity were it not for the fact that we know such behavior is "merely instinctual." But why does it disparage this trick to call it merely instinctual? To claim it is instinctual is to claim that all birds of the species do it; they do it even when circumstances aren't entirely appropriate; they do it when there are better reasons for staying on the nest; the behavior pattern is rigid, a tropism of sorts, and presumably the controls are genetically wired in, not learned or invented.

We must be careful not to carry this disparagement too far; it is not that the bird does this trick "unthinkingly," for while it is no doubt true that she does not in any sense run through an argument or scheme in her head ("Let's see, if I were to flap my wing as if it were broken, the fox would think..."), a man might do something of similar subtlety, and of genuine intelligence, novelty, and appropriateness, and not run through the "conscious thoughts" either. *Thinking the thoughts*, however that is characterized, is not what makes truly intelligent behavior intelligent. Anscombe says at one point "If [such an expression of reasoning] were supposed to describe actual mental processes, it would in general be quite absurd. The interest of the account is that it described an order which is there whenever actions are done with intentions."[13] But the "order is there" in the case of the bird as well as the man. That is, when we ask why birds evolved with this tropism we explain it by noting the utility of having a means of *deceiving* predators, or inducing false beliefs in them; what must be explained is the provenance of the bird's second-order Intentions. I would be the last to deny or dismiss the vast difference between instinctual tropistic behavior and the more versatile, intelligent behavior of humans and others, but what I want to insist on here is that if one is prepared to adopt the Intentional stance without qualms as a tool in predicting and explaining behavior, the bird is as much a second-order Intentional system as any man. Since this is so, we should be particularly suspicious of the argument I was tempted to use, viz., that *representations* of second order Intentions would depend somehow on language.[14] For it is far from clear that all or even any of the beliefs and other Intentions of an Intentional system need be *represented* "within" the system in any way for us to get a purchase on predicting its behavior by *ascribing* such Intentions to it.[15] The situation we elucidate by citing the bird's desire to induce a false belief in the predator seems to have no room or need for a representation of this sophisticated Intention in any entity's "thoughts" or "mind," for neither the bird nor evolutionary history nor Mother Nature need think these thoughts for our explanation to be warranted.

Reciprocity, then, provided we understand by it merely the capacity in Intentional systems to exhibit higher-order Intentions, while it depends on the first three conditions, is independent of the fifth and sixth. Whether this notion does justice to the reciprocity

discussed by other writers will begin to come clear only when we see how it meshes with the last two conditions. For the fifth condition, the capacity for verbal communication, we turn to Grice's theory of meaning. Grice attempts to define what he calls nonnatural meaning, an utterer's meaning something by uttering something, in terms of the *intentions* of the utterer. His initial definition is as follows:[16]

> "U meant something by uttering x" is true if, for some audience A, U uttered x intending
> (1) A to produce a particular response *r*.
> (2) A to think (recognize) that U intends (1).
> (3) A to fulfill (1) on the basis of his fulfillment of (2).

Notice that intention (2) ascribes to U not only a second- but a third-order Intention: U must *intend* that A *recognize* that U *intends* that A produce *r*. It matters not at all that Grice has been forced by a series of counterexamples to move from this initial definition to much more complicated versions, for they all reproduce the third-order Intention of (2). Two points of great importance to us emerge from Grice's analysis of nonnatural meaning. First, since nonnatural meaning, meaning something by saying something, must be a feature of any true verbal communication, and since it depends on third-order Intentions on the part of the utterer, we have our case that condition five rests on condition four and not vice versa. Second, Grice shows us that mere *second*-order Intentions are not enough to provide genuine reciprocity; for that, *third*-order Intentions are needed. Grice introduces condition (2) in order to exclude such cases as this: I leave the china my daughter has broken lying around for my wife to see. This is not a case of meaning something by doing what I do intending, what I intend, for though I am attempting thereby to induce my wife to believe something about our daughter (a second-order Intention on my part), success does not depend on her recognizing this intention of mine, or recognizing my intervention or existence at all. There has been no real *encounter*, to use Erving Goffman's apt term, between us, no *mutual recognition*. There must be an encounter between utterer and audience for utterer to mean anything, but encounters can occur in the absence of non-natural meaning (witness Ashley's dog), and ploys that depend on third-

order Intentions need not involve encounters (e.g., *A* can intend that *B* believe that *C* desires that *p*). So third-order Intentions are a necessary but not sufficient condition for encounters which are a necessary but not sufficient condition for instances of nonnatural meaning, that is, instances of verbal communication.

It is no accident that Grice's cases of nonnatural meaning fall into a class whose other members are cases of deception or manipulation. Consider, for instance, Searle's ingenious counterexample to one of Grice's formulations: the American caught behind enemy lines in World War II Italy who attempts to deceive his Italian captors into concluding he is a German officer by saying the one sentence of German he knows: *"Kennst du das Land, wo die Zitronen blühen?"*[17] As Grice points out, these cases share with cases of nonnatural meaning a reliance on or exploitation of the rationality of the victim. In these cases success hinges on inducing the victim to embark on a chain of reasoning to which one contributes premises directly or indirectly. In deception the premises are disbelieved by the supplier; in normal communication they are believed. Communication, in Gricean guise, appears to be a sort of collaborative manipulation of audience by utterer; it depends, not only on the rationality of the audience who must sort out the utterer's intentions, but on the audience's *trust* in the utterer. Communication, as a sort of manipulation, would not work, given the requisite rationality of the audience, unless the audience's trust in the utterer were *well-grounded* or reasonable. Thus the *norm* for utterance is sincerity; were utterances not normally trustworthy, they would fail of their purpose.[18]

Lying, as a form of deception, can only work against a background of truth-telling, but other forms of deception do not depend on the trust of the victim. In these cases success depends on the victim being *quite* smart, but not quite smart enough. Stupid poker players are the bane of clever poker players, for they fail to see the bluffs and ruses being offered them. Such sophisticated deceptions need not depend on direct encounters. There is a book on how to detect fake antiques (which is also, inevitably, a book on how to *make* fake antiques) which offers this sly advice to those who want to fool the "expert" buyer: once you have completed your table or whatever (having utilized all the usual means of simulating age and wear) take a modern electric drill and drill a

hole right through the piece in some conspicuous but perplexing place. The would-be buyer will argue: no one would drill such a disfiguring hole without a reason (it can't be supposed to look "authentic" in any way) so it must have served a purpose, which means this table must have been in use in someone's home; since it was in use in someone's home, it was not made expressly for sale in this antique shop... therefore it is authentic. Even if this "conclusion" left room for lingering doubts, the buyer will be so preoccupied dreaming up uses for that hole it will be months before the doubts can surface.

What is important about these cases of deception is the fact that just as in the case of the feigning bird, success does not depend on the victim's *consciously entertaining* these chains of reasoning. It does not matter if the buyer just notices the hole and "gets a hunch" the piece is genuine. He *might* later accept the reasoning offered as his "rationale" for finding the piece genuine, but he might deny it, and in denying it, he might be deceiving himself, even though the *thoughts* never went through his head. The chain of reasoning explains why the hole works as it does (if it does), but as Anscombe says, it need not "describe actual mental processes," if we suppose actual mental processes are conscious processes or events. The same, of course, is true of Gricean communications; neither the utterer nor the audience need consciously entertain the complicated intentions he outlines, and what is a bit surprising is that no one has ever used this fact as an objection to Grice. Grice's conditions for meaning have been often criticized for falling short of being sufficient, but there seems to be an argument not yet used to show they are not even necessary. Certainly few people ever consciously framed those ingenious intentions before Grice pointed them out, and yet people had been communicating for years. Before Grice, were one asked: "Did you intend your audience to recognize your intention to provoke that response in him?" one would most likely have retorted: "I intended nothing so devious. I simply intended to inform him that I wouldn't be home for supper" (or whatever). So it seems that if these complicated intentions underlay our communicating all along, they must have been unconscious intentions. Indeed, a perfectly natural way of responding to Grice's papers is to remark that *one was not aware* of doing these things when one communicated. Now

Anscombe has held, very powerfully, that such a response establishes that the action under that description was not intentional.[19] Since one is not *aware* of these intentions in speaking, one cannot be speaking *with* these intentions.

Why has no one used this argument against Grice's theory? Because, I submit, it is just too plain that Grice is on to something, that Grice is giving us necessary conditions for nonnatural meaning. His analysis illuminates so many questions. Do we communicate with computers in Fortran? Fortran seems to be a language; it has a grammar, a vocabulary, a semantics. The transactions in Fortran between man and machine are often viewed as cases of *man communicating with machine*, but such transactions are pale copies of human verbal communication precisely because the Gricean conditions for nonnatural meaning have been bypassed. There is no room for them to apply. Achieving one's ends in transmitting a bit of Fortran to the machine does not hinge on getting the machine to recognize one's intentions. This does not mean that all communications with computers in the future will have this shortcoming (or strength, depending on your purposes), but just that we do not now communicate, in the strong (Gricean) sense, with computers.[20]

If we are not about to abandon the Gricean model, yet are aware of no such intentions in our normal conversation, we shall just have to drive these intentions underground, and call them unconscious or preconscious intentions. They are intentions that exhibit "an order which is there" when people communicate, intentions of which we are not normally aware, and intentions which are a precondition of verbal communication.[21]

We have come this far without having to invoke any sort of consciousness at all, so if there is a dependence between consciousness or self-consciousness and our other conditions, it will have to be consciousness depending on the others. But to show this I must first show how the first five conditions by themselves might play a role in ethics, as suggested by Rawls's theory of justice. Central to Rawls's theory is his setting up of an idealized situation, the "original position," inhabited by idealized persons, and deriving from this idealization the first principles of justice that generate and illuminate the rest of his theory. What I am concerned with now is neither the content of these principles nor the validity of

their derivation, but the nature of Rawls's tactic. Rawls supposes that a group of idealized persons, defined by him as rational, self-interested entities, make calculations under certain constraints about the likely and possible interactive effects of their individual and antagonistic interests (which will require them to frame higher-order Intentions, for example, beliefs about the desires of others, beliefs about the beliefs of others about their own desires, and so forth). Rawls claims these calculations have an optimal "solution" that it would be reasonable for each self-interested person to adopt as an alternative to a Hobbesian state of nature. The solution is to agree with his fellows to abide by the principles of justice Rawls adumbrates. What sort of a proof of the principles of justice would this be? Adopting these principles of justice can be viewed, Rawls claims, as the solution to the "highest order game" or "bargaining problem." It is analogous to derivations of game theory, and to proofs in Hintikka's epistemic logic,[22] and to a "demonstration" that the chess-playing computer will make a certain move because it is the most rational move given its information about the game. All depend on the assumption of ideally rational calculators and hence their outcomes are intrinsically normative. Thus I see the derivations from Rawls's original position as continuous with the deductions and extrapolations encountered in more simple uses of the Intentional stance to understand and control the behavior of simpler entities. Just as truth and consistency are norms for belief,[23] and sincerity is the norm for utterance, so, if Rawls is right, justice as he defines it is the norm for interpersonal interactions. But then, just as part of our warrant for considering an entity to have any beliefs or other Intentions is our ability to construe the entity as *rational*, so our grounds for considering an entity a person include our ability to view him as abiding by the principles of justice. A way of capturing the peculiar status of the concept of a person as I think it is exploited here would be to say that while Rawls does not at all intend to argue that justice is the inevitable result of *human* interaction, he does argue in effect that it is the inevitable result of *personal* interaction. That is, the concept of a person is itself inescapably normative or idealized; to the extent that justice does not reveal itself in the dealings and interactions of creatures, to that extent they are not persons. And once again we can see that there is "an order which is there" in a just society that is independent of any actual

episodes of conscious thought. The existence of just practices and the "acknowledgment" implicit in them does not depend on anyone ever consciously or deliberately going through the calculations of the idealized original position, consciously arriving at the reciprocal agreements, consciously adopting a stance toward others.

> To recognize another as a person one must respond to him and act towards him in certain ways; and these ways are intimately connected with the various prima facie duties. Acknowledging these duties in some degree, and so having the elements of morality, is not a matter of choice or of intuiting moral qualities or a matter of the expression of feelings or attitudes...it is simply the pursuance of one of the forms of conduct in which the recognition of others as persons is manifested.[24]

The importance of Rawls's attempt to derive principles of justice from the "original position" is, of course, that while the outcome is recognizable as a *moral* norm, it is not *derived as* a moral norm. Morality is not presupposed of the parties in the original position. But this means that the derivation of the norm does not in itself give us any answer to the questions of when and why we have the right to hold persons *morally* responsible for deviations from that norm. Here Anscombe provides help and at the same time introduces our sixth condition. *If I am to be held responsible for an action* (a bit of behavior of mine under a particular description), I must have been *aware* of that action under that description.[25] Why? Because only if I was aware of the action can I *say* what I was about, and participate from a privileged position in the question-and-answer game of giving reasons for my actions. (If I am not in a privileged position to answer questions about the reasons for my actions, there is no special reason to ask *me*.) And what is so important about being able to participate in this game is that only those capable of participating in reason-giving can be argued into, or argued out of, courses of action or attitudes, and if one is incapable of "listening to reason" in some matter, one cannot be held responsible for it. The capacities for verbal communication and for awareness of one's actions are thus essential in one who is going to be amenable to argument or persuasion, and such persuasion, such reciprocal adjustment of interests achieved by mutual exploitation of rationality, is a feature of the optimal mode of personal interaction.

This capacity for participation in mutual persuasion provides

the foundation for yet another condition of personhood recently exposed by Harry Frankfurt.[16] Frankfurt claims that persons are the subclass of Intentional systems capable of what he calls *second-order volitions.* Now at first this looks just like the class of second-order Intentional systems, but it is not, as we shall see.

> Besides wanting and choosing and being moved *to do* this or that, men may also want to have (or not to have) certain desires and motives. They are capable of wanting to be different, in their preferences and purposes, from what they are.... No animal other than man, however, appears to have the capacity for reflective self-evaluation that is manifested in the formation of second-order desires. (p. 7)

Frankfurt points out that there are cases in which a person might be said to want to have a particular desire even though he would not want that desire to be effective for him, to be "his will." (One might, for instance, want to desire heroin just to know what it felt like to desire heroin, without at all wanting this desire to become one's effective desire.) In more serious cases one wants to have a desire one currently does not have, and wants this desire to become one's will. These cases Frankfurt calls second-order volitions, and it is having these, he claims, that is "essential to being a person" (p. 10). His argument for this claim, which I will not try to do justice to here, proceeds from an analysis of the distinction between having freedom of action and having freedom of the will. One has freedom of the will, on his analysis, only when one can have the will one wants, when one's second-order volitions can be satisfied. Persons do not always have free will, and under some circumstances can be responsible for actions done in the absence of freedom of the will, but a person always must be an "entity for whom the freedom of its will may be a problem" (p. 14)—that is, one capable of framing second-order volitions, satisfiable or not. Frankfurt introduces the marvelous term "wanton" for those "who have first-order desires but...no second-order volitions." (Second-order volitions for Frankfurt are all, of course, *reflexive* second-order desires.) He claims that our intuitions support the opinion that all nonhuman animals, as well as small children and some mentally defective people, are wantons, and I for one can think of no plausible counterexamples. Indeed, it seems a strength of his theory, as he claims, that human beings—the only persons

we recognize—are distinguished from animals in this regard. But what should be so special about second-order volitions? Why are they, among higher-order Intentions, the peculiar province of persons? Because, I believe, the "reflexive self-evaluation" Frankfurt speaks of is, and must be, genuine self-consciousness, which is achieved only by adopting toward *oneself* the stance not simply of communicator but of Anscombian reason-asker and persuader. As Frankfurt points out, second-order desires are an empty notion unless one can *act* on them, and acting on a second-order desire must be logically distinct from acting on its first-order component. Acting on a second-order desire, doing something to bring it about that one acquires a first-order desire, is acting upon oneself just as one would act upon another person: one *schools* oneself, one offers oneself persuasions, arguments, threats, bribes, in the hopes of inducing oneself to acquire the first-order desire.[27] One's stance toward oneself *and access to oneself* in these cases is essentially the same as one's stance toward and access to another. One must *ask oneself* what one's desires, motives, reasons really are, and only if one can say, can become aware of one's desires, can one be in a position to induce oneself to change.[28] Only here, I think, is it the case that the "order which is there" cannot be there unless it is there in episodes of conscious thought, in a dialogue with oneself.[29]

Now, finally, why are we not in a position to claim that these necessary conditions of moral personhood are also sufficient? Simply because the concept of a person is, I have tried to show, inescapably normative. Human beings or other entities can only aspire to being approximations of the ideal, and there can be no way to set a "passing grade" that is not arbitrary. Were the six conditions (strictly interpreted) considered sufficient they would not ensure that any actual entity was a person, for nothing would ever fulfill them. The moral notion of a person and the metaphysical notion of a person are not separate and distinct concepts but just two different and unstable resting points on the same continuum. This relativity infects the satisfaction of conditions of personhood at every level. There is no objectively satisfiable sufficient condition for an entity's *really* having beliefs, and as we uncover apparent irrationality under an Intentional interpretation of an entity, our grounds for ascribing any beliefs at all wanes,

especially when we have (what we always *can* have in principle) a non-Intentional, mechanistic account of the entity. In just the same way our assumption that an entity is a person is shaken precisely in those cases where it matters: when wrong has been done and the question of responsibility arises. For in these cases the grounds for saying that the person is culpable (the evidence that he did wrong, was aware he was doing wrong, and did wrong of his own free will) are in themselves grounds for doubting that it is a person we are dealing with at all. And if it is asked what could *settle* our doubts, the answer is: nothing. When such problems arise we cannot even tell in our own cases if we are persons.

NOTES

1. See my "Mechanism and Responsibility," in T. Honderich, ed., *Essays on Freedom of Action* (London: Routledge & Kegan Paul, 1973).

2. In "Justice as Reciprocity," a revision of "Justice as Fairness" printed in S. Gorovitz, ed., *Utilitarianism* (Indianapolis: Bobbs Merrill, 1971), Rawls allows that the persons in the original position may include "nations, provinces, business firms, churches, teams, and so on. The principles of justice apply to conflicting claims made by persons of all these separate kinds. There is, perhaps, a certain logical priority to the case of human individuals" (p. 245). In *A Theory of Justice* (Cambridge, Mass.: Harvard University Press, 1971), he acknowledges that parties in the original position may include associations and other entities not human individuals (e.g., p. 146), and the apparent interchangeability of "parties in the original position" and "persons in the original position" suggests that Rawls is claiming that for some moral concept of a person, the moral person is *composed* of metaphysical persons who may or may not themselves be moral persons.

3. Setting aside Rawls's possible compound moral persons. For more on compound persons see Amelie Rorty, "Persons, Policies, and Bodies," *International Philosophical Quarterly*, Vol. XIII, no. 1 (March 1973).

4. J. Hintikka, *Knowledge and Belief* (Ithaca: Cornell University Press, 1962).

5. P. F. Strawson, *Individuals* (London: Methuen, 1959), pp. 101-102. It has often been pointed out that Strawson's definition is obviously much too broad, capturing all sentient, active creatures. See, e.g. H. Frankfurt, "Freedom of the will and the concept of a person," *Journal of Philosophy* (January 14, 1971). It can also be argued (and I would argue) that states of consciousness are only a proper subset of psychological or Intentionally characterized states, but I think it is clear that Strawson here means to cast his net wide enough to include psychological states generally.

6. D.M. MacKay, "The use of behavioral language to refer to mechanical processes," *British Journal of Philosophy of Science* (1962), pp. 89-103; P.F. Strawson, "Freedom and resentment," *Proceedings of the British Academy* (1962), reprinted in Strawson, ed., *Studies in the Philosophy of Thought and Action* (Oxford, 1968); A. Rorty, "Slaves and machines," *Analysis* (1962); H. Putnam, "Robots: machines or artificially created life?" *Journal of Philosophy*

(November 12, 1964); W. Sellars, "Fatalism and determinism," in K. Lehrer, ed., *Freedom and Determinism* (New York: Random House, 1966); A. Flew, "A Rational Animal," in J.R. Smythies, ed., *Brain and Mind* (London: Routledge & Kegan Paul, 1968); T. Nagel, "War and Massacre," *Philosophy and Public Affairs* (Winter 1972); D. Van de Vate, "The problem of robot consciousness," *Philosophy and Phenomenological Research* (December 1971); my "Intentional Systems," *Journal of Philosophy* (February 25, 1971).

7. H. Frankfurt, "Freedom of the will and the concept of a person," op. cit.

8. And sufficient, but I will not argue it here. I argue for this in *Content and Consciousness* (London: Routledge & Kegan Paul, 1969), and more recently and explicitly in my "Reply to Arbib and Gunderson," APA Eastern Division Meetings, December 29, 1972.

9. I will not discuss Sartre's claim here.

10. Cf. Bernard Williams, "Deciding to Believe," in H.E. Kiefer and M.K. Munitz, eds., *Language, Belief and Metaphysics* (New York: New York University Press, 1970).

11. E.g., B.F. Skinner, "Behaviorism at Fifty," in T.W. Wann, ed., *Behaviorism and Phenomenology* (Chicago: University of Chicago Press, 1964).

12. For illuminating suggestions on the relation of language to belief and rationality, see Ronald de Sousa, "How to give a piece of your mind; or, a logic of belief and assent," *Review of Metaphysics* (September 1971).

13. G.E.M. Anscombe, *Intention* (Oxford: Blackwell, 1957), p. 80.

14. Cf. Ronald de Sousa, "Self-Deception," *Inquiry*, 13 (1970), esp. p. 317.

15. I argue this in more detail in "Brain Writing and Mind Reading," in K. Gunderson, ed., *Language, Mind, and Knowledge* (Minneapolis: University of Minnesota Press, 1975), and in my "Reply to Arbib and Gunderson."

16. The key papers are "Meaning," *Philosophical Review* (July 1957), and "Utterer's meaning and intentions," *Philosophical Review* (April 1969). His initial formulation, developed in the first paper, is subjected to a series of revisions in the second paper, from which this formulation is drawn (p. 151).

17. John Searle, "What is a Speech Act?" in Max Black, ed., *Philosphy in America* (London: Allen & Unwin, 1965), discussed by Grice in "Utterer's Meaning and Intentions," p. 160.

18. Cf. "Intentional Systems," pp. 102-103.

19. G.E.M. Anscombe, *Intention*, p. 11.

20. It has been pointed out to me by Howard Friedman that many current Fortran compilers which "correct" operator input by inserting "plus" signs and parentheses, etc., to produce well-formed expressions arguably meet Grice's criteria, since within a very limited sphere, they diagnose the "utterer's" intentions and proceed on the basis of this diagnosis. But first it should be noted that the machines to date can diagnose only what might be called the operator's syntactical intentions, and second, these machines do not seem to meet Grice's subsequent and more elaborate definitions, not that I wish to claim that no computer could.

21. In fact, Grice is describing only a small portion of the order which is there as a precondition of normal personal interaction. An analysis of higher order Intentions on a broader front is to be found in the works of Erving Goffman, especially in *The Presentation of Self in Everyday Life* (Garden City: Doubleday, 1959).

22. See Hintikka, *Knowledge and Belief*, p. 38.

23. See Dennett, "Intentional Systems," pp. 102-103.

24. J. Rawls, "Justice as Reciprocity," p. 259.

25. I can be held responsible for events and states of affairs that I was not aware of and ought to have been aware of, but these are not intentional actions. In these cases I am responsible for these further matters in virtue of being responsible for the foreseeable consequences of actions—including acts of omission—that I was aware of.

26. H. Frankfurt, "Freedom of the will and the concept of a person." Frankfurt does not say whether he conceives his condition to be merely a necessary or also a sufficient condition of moral personhood.

27. It has been brought to my attention that dogs at stud will often engage in masturbation, in order, apparently, to *increase their desire* to copulate. What makes these cases negligible is that even supposing the dog can be said to act on a desire to strengthen a desire, the effect is achieved in a non-Intentional ("purely physiological") way; the dog does not appeal to or exploit his own rationality in achieving his end. (As if the only way a person could act on a second-order volition were by taking a pill or standing on his head, etc.).

28. Margaret Gilbert, in "Vices and self-knowledge," *Journal of Philosophy* (August 5, 1971), p. 452, examines the implications of the fact that "when, and only when, one believes that one has a given trait can one decide to change out of it."

29. Marx, in *The German Ideology*, says: "Language, like consciousness, only arises from the need, the necessity, of intercourse with other men.... Language is as old as consciousness, language is practical consciousness." And Nietzsche, in *The Joyful Wisdom*, says: "For we could in fact think, feel, will, and recollect, we could likewise 'act' in every sense of the term, and nevertheless nothing of it at all need necessarily 'come into consciousness' (as one says metaphorically;.... *What* then is *the purpose* of consciousness generally, when it is in the main *superfluous*?—Now it seems to me, if you will hear my answer and its perhaps extravagant supposition, that the subtlety and strength of consciousness are always in proportion to the *capacity for communication* of a man (or an animal), the capacity for communication in its turn being in proportion to the *necessity for communication*.... In short, the development of speech and the development of consciousness (not of reason, but of reason becoming self-conscious) go hand in hand."