Ethics and Rationality

Philip Clayton and Steven Knapp

Despite their differences in other respects, a number of recent philosophical accounts of ethics have been marked by a shared skepticism toward attempts to derive ethical obligations from the requirements of rational agency. Thinkers such as Alasdair MacIntyre and Martha Nussbaum have argued influentially that *ta ethika* are not a matter of universal principles of reasoning but of individual character, social standards, or the ideals that define a particular community. Similarly, for Bernard Williams ethical reasons are strictly "internal" to each agent's set of motivations, unlike the "external reasons" involved, for instance, in scientific inquiry.[1] Hence, according to Williams, there is no way to get from the requirements of rational agency to any specifically *ethical* requirements.

In part at least, the anti-rationalist tendency has been motivated by a sense that rationalist approaches are incompatible with recognizing the diversity of human ethical practices and the extent to which such practices are inseparable from the particular historical and cultural contexts from which they emerge. The underlying assumption of the rationalist seems to be that, if a commitment to certain ethical norms is already built into rational agency in general, then it must be possible to deduce a full range of ethical prescriptions from a context-independent analysis of what rationality entails. But such a deduction seems, to current theorists, neither possible nor desirable; hence, they think, the rationalist project itself should be abandoned.

In this paper, we argue that it is a mistake for those who hold a contextualist or dispositionalist view of agency to reject a rationalist derivation of ethics.[2] Indeed, we argue, the rationalist and the contextualist/dispositionalist accounts, rightly understood, are not only mutually compatible but inseparable. Our argument should be interpreted, then, not just as an updating of the Kantian project (though it is that) but also as an attempt to reveal what is necessarily involved in contextualist or dispositionalist accounts of ethics if they wish to do justice to a (rational) agent's own desire to be rational. At the same time, we show why the right sort of rationalist derivation of ethics does not, in fact, entail the ethical consequences that traditional rationalists hoped for or that current anti-rationalists fear.

The difficulties attending previous attempts to derive ethical principles from the requirements of rationality stem, in our view, from an inadequate notion of what rationality actually requires. Hence we will begin by developing two theses regarding the nature of rationality: (1) that rationality in general, properly understood, involves evaluation against the standards of a community of inquiry (and there are multiple such communities); (2) that *practical* rationality involves an ongoing evaluation of an agent's reasons for action, reasons that can only be evaluated against the standards implicit in each agent's particular self-conception.

(1) The standards for what count as good reasons are not independent of one's social and cultural context. Far from reflecting timeless methods and criteria, human thought and evaluation are pervasively influenced by the values of one's reference group, whose standards reflect a particular construal of what counts as reasonable. We can no longer suppose, in ethical or aesthetic debates, that human beings are in a position to read the best justified theory off the face of nature. On the other hand, developments in the philosophy of science make it equally groundless to suppose that ethical and aesthetic debates, because of their dependence on a certain context of discourse, are more "subjective" than empirical ones. The challenge faced by a rationalist account of ethics is not to separate "purely objective" factual questions from the "purely subjective" ethical ones, but rather to substantiate the claim of some obligations to be binding on rational agents as a consequence of their commitment to rationality.

The best way to spell out this commitment is to formulate an account of rationality in pragmatic or procedural terms: the rationality of a given claim lies in its relation to an ongoing process of collective assessment.[3] Put minimally, a necessary condition for my claiming that a belief is rational is that it has been subjected to (or is genuinely open to) criticism by what I take to be the relevant community of inquiry.[4] Similarly, a necessary condition for my claiming that an *action* is rational is that it has been subjected to assessment by the relevant community.

While rationality necessarily entails an openness to input from the relevant community of inquiry or assessment--so that one cannot rationally go about ignoring such input--there is no reason to hold that any *presently existing* community fully instantiates the agent's sense of what the relevant community is. (From this point of view, the prophet or revolutionary who relies almost wholly on an appeal to the standards of an as-yet-non-existent community is just a limiting case of rational agency in general.) For in general it makes sense to suppose that the relevant community itself will continue to be transformed by ongoing rational discussion, at least until (as C. S. Peirce imagines) all possible grounds of doubt are eliminated (if that state can ever be reached). So a rational agent, in being open to the views of what she takes to be the relevant community, has to be open not just to whatever conclusion that community reaches at any given moment in its history but to the future (properly-related) views of that community (or, indeed, that community's properly-related descendant communities). From this point of view, it would seem that a rational agent has to have in mind not just some criterion for identifying what she takes to be a relevant existing community but at least a partial image of that community's normative *future* state--which is to say, an *ideal* of what that community will become if it goes on properly pursuing its course of inquiry.

(2) Now to practical reasoning. Very few theorists are still inclined to defend a context-free account of human rationality. Similarly, the Kantian (and neo-Kantian) picture of *agency*, with its tendency to abstract from individual motivation, is no longer tenable: individuals are not separable from their particular sets of desires, wants, and other dispositions. Since we are animals who are characterized by (some degree of) self-awareness, these sets of motivations are inevitably accompanied by--indeed, often directed by--some sense of who this "I" is who wishes, wants, and tends to act in a certain way. We will call this the individual's *self-conception*; our thesis is that the notion of an individual's self-conception provides the indispensable starting point for an

account of ethics while at the same time showing its intrinsic connection to rationality.

It is not hard to show that every rational agent needs to have some self-conception (SC), that is, an image, however ill-defined, of the self she wants to be or become. For in the first place, a theory of practical rationality presupposes a notion of what it is to perform an action rationally. This involves the question, What are the minimal conditions for reasoning about action? Acting rationally means, at least, giving (or being able to give) a rationale for one's proposed or past actions. And (given thesis (1) above) a rationale is a reason that would be taken to be adequate in the right social--or, as we shall say, *intersubjective*--context. Now in order that my reason for action be available (in principle) for assessment by a relevant community of other persons, it has to be possible to pose the question of whether the action I contemplate performing makes sense for *me* to do, that is, *given the kind of agent I am*. For only if I can represent (or imagine representing) my actions as the actions of a certain *kind* of agent can I coherently engage (or imagine engaging) in a discussion with others about what it makes sense for *me* to do. No one, then, can deliberate rationally about his or her possible actions without relating those possible actions to one or more conceptions of the kind of agent she takes herself to be. It follows that every rational agent who wants to deliberate rationally is actually required to have (at least one) SC.[5]

The fact that there is a rational imperative to *have* a SC does not mean, however, that all the contents of an agent's SC are derived from the very idea of rational agency. There may be many things that an agent considers worth doing that are nonetheless neutral with regard to (intersubjective) rational evaluation. Still, whatever other projects may belong to a person's set of motivations, the project of realizing a SC is mandated by her being a rational agent.[6]

Finally, it seems clear that the intersubjective assessment to which a rational agent necessarily remains open is not something that can take place all at once. Intersubjective assessment is processual, and consequently so is an agent's rational justification in holding a particular conception of herself and a particular account of how to go about realizing that conception. In general, to hold that *belief B is rationally justified* is, in our view, to hold that *B* would be accepted in the long run by the relevant experts, after sufficient time had elapsed for relevant testing and intellectual scrutiny. We do not share Peirce's contention that the term *truth* can simply be defined as the opinion that the ideal scientific community would settle on given a sufficiently long (possibly infinite) period of inquiry. But Peirce was correct in maintaining that the conclusions of such a community are the model for the *ideally* rational. Rationality is both an intersubjective and a normative notion; a given belief counts as rational to the extent that the believer has reason to think it meets or could meet what the believer takes to be the standards of the ideal discursive community.

This brings us to our main argument. Kantian moralists claim that a given obligation *O* must pertain to all members of some specified group, or perhaps to all persons in general, on the basis of the requirements of reason alone. But making the right sort of connection between rationality and ethics does not require us to suppose that obligations are deduced directly from reason. If, as we argue, the question of what actions are appropriate for a given agent to perform is relative to the requirements of that agent's SC, and if we are right in denying that particular SCs can be grounded directly in reason, then the strong Kantian account of obligation cannot be correct. In contrast to the Kantian position, our argument for connecting ethics and rationality starts only with the following:

> There is at least one general obligation *O* such that, if a SC is to be rational, it must contain *O*. That is, one cannot rationally develop a SC and at the same time omit *O*, for no SC that contains not-*O* is consistent.

Our claim is that all SCs must contain the relevant obligation; persons who wish to be rational but whose SC contains not-*O* rather than *O* are guilty either of an existential contradiction or of holding an inconsistent set of beliefs. In either case, they are behaving in a way that contradicts their own desire to be rational. Hence the individual can rationally pursue an SC containing *O* but not one omitting *O*.

Our task, then, is to show that there is at least one such general obligation that is binding on all individuals or, as we parse it, on all who seek to develop and maintain a SC in a rational manner.[7] Presumably, most theorists who reject the notion of a rationally inescapable obligation will grant that there are at least some general prescriptions (say, a prescription to refrain from randomly slaying passers-by) such that it is better if most agents follow them. But why suppose, these theorists will ask, that any obligation is such that it would be *irrational* for a given individual to make herself an exception by disregarding it?

Given the intersubjective model of rationality sketched above, we suggest that there exists a necessary feedback relationship between the individual who forms a SC and her social world. (Let us call this *the Feedback Principle*.) A rational agent is one who attempts to acquire beliefs, including beliefs about herself, such that she is rationally justified in believing them to be accurate. It is not simply that people would *prefer* their view of themselves to be accurate; they can't even count as holding a SC in a rational manner unless they are interested in rationally determining whether their account of its intentional content, as well as of the degree to which their actions fit the conception in question, is accurate. But in order to determine (rationally) whether my SC is accurate, I require feedback from others. Martha believes she is a successful philosopher, but her belief is rational only to the extent that she expects it to be confirmed by evidence (public acclaim, book sales, lecture invitations) from those whom she takes to constitute the relevant community of inquiry. (For reasons mentioned earlier, that community *may* be, but need not necessarily be, some presently existing community, such as a specialist subcommunity of the community of professional philosophers.) In other words, Martha only counts as holding her belief *rationally* if she is open in principle to (the right sort of) feedback.

The epistemic picture behind the Feedback Principle can be phrased in a sort of practical syllogism for SC-success:

(P1) Martha is rationally justified in believing that S_1 - S_n are the standards for being a successful philosopher.

(P2) Martha is rationally justified in believing that she meets S_1 - S_n.

(C) Martha is rationally justified in believing herself to be a successful philosopher.

The standards mentioned in (P1), of course, already involve some reference to other people. In the *Gorgias*, the hedonism with which Callicles concludes may be more value-free than his earlier defense of heroic amorality; yet it is false to suppose that Callicles can even conceive something as an object of pleasure without having gotten that notion, directly or indirectly, from others. Essential to an object's rationally counting as desirable for an agent is that agent's sense that (at least certain) others would also find it so--the others whose views the agent takes, implicitly or explicitly, as his standards of what is desirable.[8]

But we can show that (P2) requires feedback in an even stronger sense. The individual cannot rationally claim that she meets the general standards (P1) unless she is prepared to take into account others' evaluation of the actions she has performed and/or contemplates performing (P2). SCs presuppose intersubjective standards, and their application to oneself is rational only when subjected to the (actual or rightly imagined) test of intersubjective debate.[9]

The Feedback Principle has an important entailment: since feedback is a necessary condition for SC-holding, it is not rational to hold a SC that conflicts with (or disregards) the feedback process itself. For I cannot claim to be interested in knowing something and at the same time disregard the conditions that make it possible to know it. The anti-rationalist need not resist this point. Yet, she retorts, surely one cannot base any universal moral obligations upon such a foundation. Suppose someone's self-conception is that of a great composer, a sort of cross between a Wagnerian aesthete and a Nietzschean superman (he seeks to emulate Wagner as Nietzsche wanted him to be). And suppose that our composer (call him Rick) wishes (perhaps surprisingly) to hold this SC rationally. Given the Feedback Principle, he will note the intersubjective standards for being a great composer and will seek to verify from the relevant set of composers and critics that he has indeed met them. But in this particular case, the anti-rationalist asks, couldn't all the relevant standards be aesthetic? For surely this is a case in which someone holds a SC in the rationally appropriate way--only the SC happens to be such that it does not contain *any* general requirements on his behavior. Why then is Rick under any sort of *ethical* obligation? Put differently, how does he betray any *rational* obligation in acting as ruthlessly as he might wish in attaining his goal?

Note that the standard Kantian response to this kind of objection is inadequate. The Kantian might complain that Rick's stance fails the universalizability test: *to attain one's life ambition by any possible means* is not a principle of action that can be generalized. The very structures for distinguishing between great and not-so-great composers would be endangered by Rick's principle, such that, if it were universalized, he could not have the one thing he wants from life. Moreover, if all would-be great musicians cheated, lied, and murdered to win public acclaim, the field of music as we know it, and perhaps the fabric of society itself, would be destroyed. But this response begs the question against the anti-rationalist. The universalizability theorist notes correctly (a) that not all persons can lie and murder at the same time--not all can assume that the basic mores of social interaction do not apply to themselves--and concludes (b) that the *individual* amoralist cannot rationally declare herself an exception. But surely this is wrong. For Rick can freely grant point (a) while denying (b): let the masses keep the musical establishment running, so that I can more easily rise to greatness. But why should my dependence on *others'* behaving in this manner require *me* to do so? Let *them* provide what I need for my success; the fact that they follow the norms I need them to follow in no way rationally compels me to follow the same norms![10]

Interestingly, however, despite its successful rejoinder to the Kantian rationalist, this plausible objection still fails. It fails because even a would-be amoralist has to accept some general ethical obligations in order to evaluate her own SC and the means of realizing it. Recall that an agent can only count as rationally pursuing her SC if she is open to intersubjective feedback; it follows that she is committed, if she wants to be rational, to whatever principles are entailed by the feedback process itself. First among these principles is the need to be truthful. For no agent can rationally attempt to realize her SC while lying to all others about her actions. Unless I remain open to feedback from the relevant communities, I cannot rationally determine whether my conception of myself as a great fly fisherman or as a model philanthropist is accurate; hence I cannot rationally take appropriate steps to bring my actions into line with my SC. Yet the feedback I require is obviously worthless to me (rationally speaking) if my reports of what I caught or what I gave are completely fabricated. Only by being truthful in my dealings with the relevant community of evaluation can I rationally (attempt to) determine whether my actions and my SC are consistent.[11]

But surely, someone will object, there is something wrong with treating this pragmatic need to be truthful as a general obligation. Perhaps an agent who wants to be rational is constrained from *always* lying, since otherwise she cannot hope to receive the critical feedback required by her own rational assessment of her actions. But why should we view this constraint as as an actual ethical obligation, rather than as principle of expediency that merely *resembles* one? How is it different, for instance, from any case in which someone refrains from lying out of mere pragmatic necessity? Imagine a drug dealer (and habitual liar) who seeks out a doctor after being wounded in his latest gunfight with police. Suppose the drug dealer can't survive--let alone realize his self-conception as the perfect drug dealer--unless he refrains from lying to the doctor, for instance about where it hurts. His truth-telling, on this occasion, hardly seems ethical; its only motivation is a combination of raw self-interest and an ongoing commitment to a life of crime. Why isn't truth-telling in such cases merely an expedient to which the agent resorts for the sake of a SC that acknowledges no general obligations whatsoever?

If the kind of honesty required by an agent's rational pursuit of her SC could be *localized* in the way suggested by the incident of the wounded drug dealer, the objection would be unanswerable. But the honesty required by rational agency cannot, in fact, be confined to a local episode of truth-telling, a mere exception to the general rule of an agent's dishonesty. For a *rational* interest in one's SC, and therefore in one's success or failure at realizing it, entails an openness, in principle at least, to ongoing discussion of a wide range of one's actions. Indeed, since the agent is rational *to the degree that* she is open to such feedback, the fully rational agent will be one whose actions do not block but encourage continued evaluation of the total set of (actual or contemplated) actions relevant to the realization of her SC. Now the objection requires us to suppose that an agent could be committed to such long-term and wide-ranging honesty and nevertheless "really" be a liar, someone who only resorted to truth-telling as a mere pragmatic necessity. Yet it is hard to see on what basis we could decide that someone who showed an ongoing commitment to a certain kind of behavior--in this case, honesty--wasn't *really* the kind of person that behavior implied. Above all, it is hard to see how *the agent herself* could rationally interpret her own behavior in a way that so drastically disregarded the evidence of her own dispositions. What alternative evidence would she have? And how can an agent be said to hold or pursue a SC rationally if she is precluded from rationally interpreting her own behavior and therefore from rationally assessing her own success or failure?

It follows from this argument, then, that at least one kind of SC has to be excluded from the set of SCs that an agent can rationally hold: an agent cannot rationally hold the SC of a consummate liar. For, as we have seen, an agent can only be said to have a SC if she can interpret her own behavior. She can't do *that* unless she can identify her own dispositions; and she can't separate her sense of her own dispositions from an account of her behavior over time and across a range of actions. Since the rational possession of any SC entails a broad and continuous commitment to honesty, it also entails that the agent *interpret* herself as (broadly and continuously) committed to honesty. Obviously, an agent cannot rationally view herself as holding a SC--and therefore cannot rationally hold one!--that contradicts the agent's self-interpretation. Thus it seems that no one can be a rational agent without having a commitment to honesty as a principled feature of her SC (and not just a local or occasional convenience).

At this point the anti-rationalist may be willing to concede that a commitment to honesty, once given a central place in an agent's project, will indeed be difficult to reduce to a narrowly pragmatic (and hence expedient or "merely internal") principle. But the anti-rationalist may still want to raise a more fundamental objection to our whole picture, an objection that reaches all the way back to our Feedback Principle itself. She may want to protest that our prohibition against lying depends, after all, on a mistaken account of the sort of information a rational agent actually needs. Because I am committed to consider the assessments of my actions that I would expect to receive from what I take to be the right community of inquiry, must it follow that I need to receive *actual* assessments from *actual* other persons (and hence am prohibited from depriving them of information they need to form their assessments)? Do I really need to know, for instance, whether any actual person thinks me a great composer? Or do I merely need to know what the standards for a great composer are and what music I have written, so that I can I can extrapolate from these to an account of what well-informed persons *would say*? And even if I need some *initial* input from others to help me form my self-conception, why suppose that I need *continuing* feedback from actual other persons? Isn't it enough to learn what the standards are and, once one has internalized them, to go on trying to meet them, without needing constantly to return for fresh information to the community from which one received them in the first place?

This objection is intuitively rather compelling. For it is clearly implausible to suppose that I can be acting rationally as I attempt to realize my self-conception only if I can check my performance against the views of some *actually existing* community of inquiry. What if I am the last survivor of what I take to have been the relevant community of inquiry; am I rationally required, in that case, to abandon my project of realizing my self-conception, just because there is no one left with whom I can discuss my progress? Do we really want to say, for instance, that Sir Bedivere, the last survivor of the Knights of the Round Table, is behaving *irrationally* if he goes on trying to follow the chivalric code?

It seems to us that someone in Sir Bedivere's situation would not be behaving irrationally in relying on the views of his community as he *remembers* it.[12] But this only shows that, in cases where an actual (relevant) community of inquiry is unavailable, an agent is rationally justified in falling back on a remembered community. It does not show that an agent can rationally act in such a way as to deprive himself of information from what he takes to be a relevant *existing* community of inquiry. For to be rational, on our account, is precisely to be open to rational discussion. In relying on a remembered image of a community when no actual relevant community is available, an agent is not *closing* himself to the views of what he takes to be the relevant community of inquiry. But an agent who took no interest in whether or not such a community existed, or who acted in such a way as to deprive himself of relevant information from what he took to be the relevant community, could hardly be said to be open to the rational assessment of his actions (and hence of the degree to which he was or was not succeeding in realizing his SC). Thus Sir Bedivere *would* be behaving irrationally--closing himself to the relevant information--if he discovered that a new chivalric community was emerging but either took no interest in its emergence or decided to lie to it about his exploits.

Suppose, however, that the trouble with the agent's reference group (i.e., what the agent takes to be the relevant community of inquiry) is not that it *no longer* exists but that it has *never* existed. Surely we don't want to say that a prophet or a revolutionary can only be rational if he or she remains bound by the standards of some past or present community of inquiry; for to do so would be to exclude from our account some of the most interesting and important forms of ethical practice. Here again it seems to us that the present unavailability of the relevant community rationally justifies the agent's reliance on an internal vision of that community--this time an *imagined* community rather than a remembered one. And here again, while it would be perfectly rational for the agent to rely on an image of the community until the actual one arrived, it would not be rational for the agent to take no interest in the conditions of possibility of this community's arrival or, when it did arrive, to prevent it from acquiring the information it needed to carry on its inquiry.

The conclusion is clear: if one is involved in formulating a SC (as all rational agents are), *one cannot rationally enter into pervasively dishonest relations with others*. Even if the right sort of discussion community is unavailable (because it no longer exists or does not yet exist), rational agency entails a certain *dispositional* honesty: the agent must be ready to communicate honestly with the right community should it arrive or return. To count as having a rational interest in realizing one's SC as a saint or a philosopher--or even a successful bank robber--an agent is necessarily committed to the prospect (even if remote) of an appropriate assessment of her actions by others. In short, *truthfulness is rationally indicated*.

Of course, one might complain that an injunction against lying--a requirement of truthfulness at least in one's dealings with some persons on some occasions--does not amount to a full moral philosophy. It does not apply to all other persons without exception (and hence fails to be Kantian); it neither has a metaphysical basis nor entails moral realism; and it is not a virtue that derives its force from a "thick" description of the moral life. Indeed, the ethics entailed by rationality seems compatible with some rather *im*moral character ideals, including for instance the goal of *being a successful Jack the Ripper*. The objector is right: given what we have argued so far, the would-be Jack can ply his trade on the streets of London, secure in the knowledge that he is not directly betraying any universal rational obligation by aspiring to realize his criminal SC.

Still, in the present context of "internalist" ethics and deep skepticism about whether there are *any* general obligations, it is already a rather significant development to discover that an (adequately contextualized) notion of rationality does give rise, unexpectedly, to a set of obligations. Even Jack will need his club at the end of the day, where he can report his day's accomplishments and verify his evaluation of their significance. Should his reports be dishonest (he actually spent the day helping orphans and widows in their distress), or should he fail to exclude his peers from his activities (he murders *all* persons upon sight), he cannot count as acting in a rational manner, since he is acting in ways that deprive him of the relevant means of determining whether his actions meet or fail to meet the standards set by his own character ideal.

In fact, the constraints on the rational agent may be even more extensive than this. We have concentrated up to this point on a rational agent's initial obligation to be truthful. But upon closer examination, it turns out that the implications of our argument extend well beyond this minimal requirement. Consider again our would-be Jack the Ripper. We found that Jack faced immediate limitations on his actions: for one, he couldn't kill those on whom he relied for SC-assessment. But Jack needs more than a few judges of his performance; he also needs enough social stability to make the formulation and communication of their assessments possible in the first place. This commits him to supporting (or at least not actively endangering) the sorts of social institutions and practices that are required to mediate this feedback: certain media of communication, certain basic societal mores, just enough centralized authority to ensure that society does not disintegrate into "the war of all against all." Likewise, it commits him to fostering, or at least refraining from destroying, whatever he takes to be the conditions that will enable such social institutions to emerge and endure. To the extent, in other words, that the Feedback Principle itself presupposes an adequate social system--that is, a social system adequate to support the ongoing intersubjective assessment of individual self-conceptions and the means of realizing them--the rational agent as such must share a commitment toward fostering such a system. Even Jack needs his club (or something like it).[13]

Once again, however, the anti-rationalist may resist. It may well be that Jack the Ripper wants the police to capture *other* Jack's, so that his continues to be a society in which rational evaluation is possible and he can obtain the feedback (not to mention the unsuspecting victims) he requires. But why should he not wish *himself* to be the one exception to his general wish that the police successfully maintain law and order? Jack may need to hope for a just aristocracy, or even a stable democracy, in which enough fair debate occurs for him to meet with and receive feedback from his peers. But why should *he* be just? Even if honesty is entailed by rational agency (as a necessary condition for the agents' realization of his own SC), why suppose that *justice* is entailed by it?

By now our response to this sort of objection may be obvious. I cannot consistently seek to realize a SC and at the same time engage in actions that, if successful, would make it impossible for me to attain my goal. Thus Jack cannot consistently desire the relative social stability required by the Feedback Principle and at the same time act in such a way as to bring about complete anarchy; for if he did so, the rationally indicated interpretation of his actions (indeed, his *own* rationally indicated interpretation!) would be that he is not, in fact, committed to the Feedback Principle required by his own rationality. To be rational--to count as rational even in his own terms--he must act in such a way that makes it possible to infer his commitment to continuing feedback; and this means acting in a manner that promotes adequately just social and political conditions. Now the extent of the just conditions to which our argument commits an agent may be initially somewhat limited (the agent need not be committed to a full-blown Rawlsian theory of justice, or may find it consistent to engage in petty crimes). Still, at minimum it seems that a rational agent must favor a political system in which it is possible that (at least some) rational conversations will occur and (some) decisions will be based on the force of the better argument.

Further, it may turn out in the long run that certain self-conceptions are by their very nature incompatible with the requirements of the Feedback Principle. After all, even a would-be Hitler is committed to at least some standards of justice if he wishes *rationally* to pursue his project of becoming a Hitler. And surely, at some point in the rational assessment of his actions, his own desire for consistency (based on his need to form a unified and hence intersubjectively discussable SC) should lead him to ask whether the project of becoming a Hitler is *itself* consistent with the commitment to justice into which he is compelled by his very desire to realize his fascist SC in a rational manner.[14]

Thus it seems that, in the long run, a commitment to standards of justice and truthfulness--motivated at first, perhaps, by nothing more than a desire to find out how best to be evil--will, if carried through consistently, finally feed back to one's assessment of one's SC itself. If this is true from the agent's perspective, it is also true from the perspective of anyone who inquires into the relation between particular SCs and the conditions that enable their assessment. Hence it should be possible in principle for ethical inquirers to evaluate alternative SCs (not in Kantian abstraction from their social and personal context but using "thick" descriptions, etc.) in order to determine which are or are not consistent with the conditions of their being known and evaluated.

Indeed, what we now discover, at day's end, is that the rationalist argument is not just about individual ethical principles such as honesty, but is fundamentally about the demand of consistency on rational behavior. To be to any extent inconsistent is to deprive oneself, to that extent, of the possibility of carrying out one's own projects in a rational manner. Hence the significance of our argument lies not (just) in our showing that even Jack the Ripper must sometimes listen to rather than kill his friends. It also lies in the discovery that rational agency, because it involves a general commitment to reflecting on one's actions, already contains within itself a movement toward comprehensiveness, toward that account that gives an explanation of *all* that one is as an agent (call it one's *comprehensive self-conception*). For it seems that the rational evaluation of any particular action involves, at least potentially, an assessment of that action in the light of a comprehensive account of the agent's behavior. Hence even a rational agent who wants to be a successful amoralist cannot in principle avoid (in the long run) a requirement to provide a comprehensive theory of her own behavior that, if successful, would have to explain how her necessary commitment to certain standards of honesty and justice was consistent with her amoralist SC. If her amoralist SC prevents her from meeting this requirement (because she cannot account for the split between her own altruistic and egoistic dispositions), then she cannot pursue her amoralist SC without abandoning the claim that she is *rationally* pursuing it.

In conclusion, then, we have shown that certain general ethical dispositions are implied by the long-term requirements of rational agency as such: (1) a commitment to honesty or truthfulness, and (2) a commitment to the emergence and/or preservation of a certain type of social order. Not *all* of a rational agent's behavior must be ethical; ours is not an exceptionless derivation of ethics based on rationality. Still, at least *some* ethical and political commitments do follow from the structure of rationality, including a commitment to honesty and a commitment to fostering the kind of society that makes feedback possible. Rationality, then, does entail a certain unavoidable investment in the interests of others, though this is an implication of rationality only as mediated through the requirements of individual SCs. Although arising in each agent's case out of that agent's particular SC, the requirements themselves are universal.

We argued that one necessarily faces these obligations *if* (1) one wishes to be rational *and if* (2) one chooses to develop a self-conception. But we also argued that (2) is not a matter of choice but rather a requirement derived from (1). Hence, we discovered that rationality does in fact entail both the core of an ethics and the kernel of a social philosophy. We also discovered, however, that rationality may in the long run entail more than that: for the rational requirement of consistency and therefore of comprehensiveness may eventually entail the rejection of some SCs that at first seem rationally (if not ethically) permissible. At the very least, there is no way to limit in advance the degree to which the ongoing process of rational assessment may determine, even if it cannot fully dictate, the set of concrete self-conceptions an agent can rationally attempt to realize, and hence the set of concrete actions the agent can rationally perform.

ENDNOTES

. Alasdair MacIntyre, *After Virtue: A Study in Moral Theory*, 2nd ed. (Notre Dame: University of Notre Dame Press, 1984); Martha C. Nussbaum, *The Fragility of Goodness* (Cambridge: Cambridge University Press, 1986); Bernard Williams, "Internal and External Reasons," in *Moral Luck: Philosophical Papers 1973-1982* (New York: Cambridge, 1981), pp. 101-113; Williams, *Ethics and the Limits of Philosophy* (Cambridge: Harvard University Press, 1985).

. We use the terms *ethics* and *ethical* in the broadest possible sense, referring to actions an agent *ought* to perform (and attitudes or dispositions regarding these actions), without prejudging the question of whether such obligations are merely internal or external as well. Where we wish to imply the latter, we do so with adjectives (*general* obligations or *universal* ethical principles). We use *moral* as a synonym for *ethical.*

. There may of course still be formal requirements for rational judgments, but these can presumably be derived as entailments of rational assessment. Take, for example, consistency: it is irrational for me to contradict myself in defending a position not because noncontradiction is a law of nature but because, if I do so, you will not be able to understand, let alone evaluate, my claims. Cf. Nicholas Rescher, *Rationality: A Philosophical Inquiry into the Nature and the Rationale of Reason* (Oxford: Clarendon, 1988), esp. chap. 8.

. The notion of a "community of inquiry" is derived, of course, from the pragmatism of C. S. Peirce.

. Note that we do not argue that every human being is obligated to be rational. Clearly, no rational argument will influence an individual, if one exists, who has no interest in whether or not her actions are rational.

Further, we are not supposing that an agent ever *begins* with a fully-formed SC, any more than we are supposing that an agent begins with a fully-formed image of the relevant community of inquiry.

Presumably an agent starts with motivations that point in the direction of various SCs and that provide hints as to the sort of inquirers whose opinions would be relevant to the rational assessment of each of those self-conceptions. Isolating one or more self-conceptions and defining the relevant community or communities is likely to be a long-term, if not indeed a life-long, process.

. Presumably the agent's set of motivations also includes a disposition to assess the possible means of achieving her SC rationally. Even if one can sometimes act on particular desires in complete abstraction from the views of a wider community, the project of realizing one's SC requires at least a degree of willingness to consider other people's accounts of what that conception involves and consequently of what it takes to realize it. Indeed, the very fact that an agent *has* a SC shows that she has already taken some others' views into account, since no one can invent a SC that is not based at least in part on images and reasons derived from others.

. Insisting on this *O* clearly separates us from those who would relativize all ethics to one's SC (or to other contextual parameters). Conversely, the fact that we derive obligations from the agent's practical interest in assessing her own actions separates us from universalizability theorists such as Alan Gewirth, who argues that ethical obligations spring directly from the definition of an agent as such. According to Gewirth, "the agent's description of himself as a prospective purposive agent is both a necessary and a sufficient condition" for his claim to have certain "generic rights" that he must also extend others (Gewirth, *Reason and Morality* (Chicago: Univ. of Chicago Press, 1978), p. 109; cf. pp. 64ff). Gewirth hopes to avoid "the variabilities as to content" that have crippled previous universalizability arguments; by admitting only the categories of purposiveness and voluntariness, he "substitutes rational necessities for these contingent contents" (Gewirth, *Moral Rationality*, the 1972 Lindley Lecture (n.p.: The University of Kansas, 1972), p. 28).

We grant that rights would have to be extended to all agents simply as such *if* they really were built into the definition of an agent. But Gewirth hasn't shown why the universal logical conditions of agency *must* be taken as rights or universal rules in the first place. He overlooks the possibility that an amoral agent might just go around behaving purposively and voluntarily without ever imagining that it was his *right* to do so, let alone caring whether anyone else was free to behave in the same way (or whether, in fact, any other agents survived at all). Bernard Williams correctly observes that "the argument needs to tell us what it is about rational agents that requires them to form this conception of themselves" as legislating moral rules or acting on the basis of rights (*Ethics and the Limits of Philosophy*, chap. 4, quote p. 63). And such questions clearly go beyond the logical definition of human agency as such.

. Of course, Callicles can avoid charges of inconsistency if he drops his claim that his hedonistic stance is rational. Although rational discourse about desires (e.g., defending my having these particular desires or my acting on them) presupposes input from others, the mere having of desires need not do so.

9. The role of the feedback process is not surprising given the fact that self-conceptions are generally socially defined. As Hegel argued, even the master only knows himself to be a master with reference to his slave. In addition to the references that SCs typically make to other persons, there are also, of course, essential social influences on individual identity formation. As G. H. Mead argued in detail, one's introspection and "internal dialogue" will take on the characteristics of the external dialogue to which one has been exposed; see Mead, *Mind, Self, and Society*, ed. Charles Morris (Chicago: Univ. of Chicago Press, 1934), e.g. chap. 3. Hence what I find subjectively convincing is strongly dependent on what I have found to be accepted, or can imagine myself as winning acceptance for, in an intersubjective context.

. Our version of this objection (apart from narrative details) is indebted to discussions with Bernard Williams.

. Suppose, however, that Rick doesn't want to *be* a great composer but only to be *perceived* as one; why does he have to be truthful if *that* is his ambition? The answer is that he cannot check his own progress even in realizing this ambition unless he is willing to disclose, without distorting, sufficient information to insure, for instance, that the mistaken judgments he relishes actually refer to *him* and are actually produced by the strategies he employs to produce them. (We are indebted for the counterexample to Charles Altieri.)

. Presumably this intuition is in part what underlies Alasdair MacIntyre's attempt to retrieve Aristotelianism and his defense of a Thomistic community in an age in which, as he sees it, virtue is no longer possible.

. Our argument in this paper leaves open the question of how far one can go in specifying the necessary features of such a community. For instance, it might or might not prove necessary to adopt the specific requirements proposed by Jürgen Habermas in his theory of "ideal speech situations"; see e.g. *A Theory of Communicative Action*, 2 vols., trans. Thomas McCarthy (Boston: Beacon, 1986), e.g. Intro. and chap. 1; or "Toward a Universal Pragmatics," in *Communication and the Evolution of Society*, trans. Thomas McCarthy (Boston: Beacon, 1979), pp. 1-68.

. Someone might object that we have skewed the picture by including only positive examples of feedback, that is, cases where most people would share the intuition that the agent is better (or better off) after the feedback than before. But surely, the objector will point out, there are cases where the results of feedback are destructive rather than helpful--cases where adjusting one's SC in the light of rational requirements will weaken or even eliminate the qualities that made the SC interesting or attractive in the first place. One can easily imagine situations in which, for instance, an attitude of egalitarian toleration, reflecting a rationally well-founded fear of giving offense and hence making discussion impossible, might supplant such virtues as courage, fortitude, loyalty, punctilious honor, etc. There is no reason to think that such changes are intrinsically valuable, and from an aesthetic standpoint they may seem positively reprehensible. In short, there is no reason to assume that what is good for the rational criticism of a particular human practice is therefore good *for the practice itself*. Thus, even if our account succeeds in deriving certain ethical values from rationality, doesn't it equally show that rationality and at least some important values are actually incompatible? If so, the victory for a rationalist account of ethics would seem to be a Pyrrhic one.

We grant that the costs of rational feedback to an agent's pursuit of the values implicit in her SC can sometimes outweigh its benefits. But no agent with an interest in consistently and deliberately pursuing her own projects can suppose that this is so *in general* or *in the long run*. For only if she supposes that, in general, her chances of success are better if she knows what she's doing can her activities take the form of projects she can consistently *try* to carry out.