Comment on Thomas Hill, Jr.: JUSTIFYING TO ONESELF*
Christine M. Korsgaard

I find myself in close agreement with Professor Hill about the issues discussed in his paper. In my comments, therefore, I will summarize and supplement rather than criticize Hill's argument. I will review its main points, and occasionally contribute some further considerations which I think lend it support.

In this paper and other recent work, Hill has been concerned to develop a Kantian model of practical reasoning, in opposition to some other models current on the intellectual scene. Among Hill's primary targets are those which we might call "instrumental-maximizing" theories because they have the following two features: First, some selected end or ends – say, pleasure, the satisfaction of desire, or intrinsically valuable states of affairs – are treated as automatically or necessarily reason-providing (or, we may say more simply, as good). I will call these things "material ends." The claim is that we have reason to do whatever is instrumental to bringing about these ends. The reasons provided by any given end are usually regarded as *prima facie*, and may be outweighed by reasons provided by other good ends. But there is always at least a *prima facie* reason to bring about a pleasure, avoid a pain, satisfy a desire, or what have you. Second, according to these theories, the rational thing to do is to maximize the net balance of these good ends over time, usually, treating all parts of time alike.

The instrumental-maximizing model is generally thought to be much less controversial for what we might call personal reasons, or reasons of private interest, than for moral reasons. One might conceivably endorse, say, a universalizability criterion for

---

* Alternative title: PAINS AND PROJECTS

moral reasons alongside an instrumental-maximizing model of reasons of private interest. Then you would think that when an agent is deliberating only about what concerns herself, and other persons' interests are not involved, she should aim to maximize these material ends over the course of her life. But Hill argues that even in this less controversial setting, an instrumental-maximizing theory, or indeed any theory which involves calculation with a fixed set of *prima facie* reasons, is unacceptable.

In opposition to such theories, Hill argues that no such material end is automatically or necessarily reason-providing. Material ends are always open to assessment, and to re-assessment, by the rational agents whose ends they primarily are. This means that we cannot take it for granted that a desire or a pleasure provides a reason for acting. And it means that we *need not* take it for granted that the only way to avoid the reason-providing force of a desire or a pleasure is to have it outweighed. One immediate benefit of this view is that it avoids some of the odder consequences of theories of *prima facie* reasons. One such consequence is pointed out by Thomas Nagel in a footnote in *The Possibility of Altruism*. Nagel notes that we should not ordinarily say that "to someone driving a severely injured person to the hospital, the beauty of the scenery along a considerably slower alternative route provides a reason to take it, although that reason is outweighed by the urgency of the circumstances …" And yet (Nagel observes somewhat resignedly) this is what his account of reasons implies, since the reason provided by the beautiful scenery is "derivable from appropriate general reasons which apply in the case." (PA, p. 51) By denying that appropriate general reasons can be specified by their material content, Hill avoids results like these.

This is one of the ways in which Hill's theory is Kantian. Instead of the having a theory of *prima facie* reasons which are weighed up, Kant has a theory of incentives (*Triebfeder*) which the agent may or may not adopt as reasons. The objective laws of practical reason, the imperatives, in a sense can be regarded as testing the status of an incentive as a reason in a given case. Incentives are like prima facie reasons in that they

always provide reasons when they are the only incentives in the case. In *Religion within the Limits of Reason Alone*, Kant says that it is not open to human beings to altogether ignore the incentives. When duty does not forbid it, we will treat our desires and inclinations as reasons. Our freedom consists in our ability to determine the order in which we adopt them into our maxims. Kant means that we can decide to rank the incentives of inclination below moral incentives. But incentives are unlike *prima facie* reasons in that they are tested rather than weighed. When rejected, they are not regarded as reasons to act.

Yet Hill's view gives rise to two somewhat disturbing consequences of its own, which he tries to answer. One is that no future material ends are automatically or necessarily reason-providing, so that, as in present-aim theories, concern for the future itself does not appear to be rationally required. The other, really a more general version of the first, is that it is not clear that there are any material constraints on rational conduct. If we cannot say, with Nagel, that pain is obviously bad and there is obviously a reason to avoid it, then neither can we say that someone who chooses pain is obviously irrational.

Hill's view also leaves him with a project, which is to say how ends *are* assessed by rational agents. What provides the terms of assessment? If we leave some ends fixed, of course, we can assess others in terms of them. If I take it as a given that, for instance, I want to advance my career, then I can consider in light of that whether I can really afford to take a long vacation, have a child, or publish a controversial piece of work. I can ask whether these other ends fit in with my general project of career advancement. But Hill is concerned with an earlier question. What occurred in the deliberation in which I arrived at the decision to advance my career? Hill wants to discuss a kind of practical reasoning which he calls deep deliberative reflection, in which *all* of one's material ends are open to question. But of course we must say something about what question, exactly, we should ask about them.

When considering an ultimate end, we should not simply ask: "will it maximize my pleasure?" for several reasons. The main reason, which is brought out by Hill in his paper,

is that in deep deliberative reflection we do not take it for granted that we should maximize pleasure, or that seeking pleasure is all that we should do. This, like any other material end, is open for assessment.

There is another important reason which I think supports Hill's contention. It is true that I often adopt something as an end because it I foresee that it will give me pleasure. The reason it will give me pleasure is, schematically put, because of my nature. There is, as Bishop Butler says a "prior suitableness between the object and the passion." (Sermon XI, Library of Liberal Arts, p. 5) But not everyone agrees with Butler that the suitableness has to be all that prior. I can take pleasure in the achievement of an end *because it is my end*, even if I did not make it my end because I foresaw that it would give me pleasure. It doesn't really matter why you make something your end: to the extent that it really is an *ultimate* end for you, and not a means to something else, the successful pursuit of it will normally be pleasant.

You may be tempted to think that I am guilty of a commonplace error in making these remarks. You may think I must be confusing pleasure with satisfaction or gratification: once I've made up my mind to do or achieve something, I am of course gratified to see it done. But that point applies even to projects undertaken for the most purely instrumental reasons, projects that have not come to be ultimate ends. The point I want to make is a different one, having to do with what it *means* to say that you have made something your end. My claim is that if you succeed in making something an ultimate end, you will to take pleasure in its successful pursuit. This is a claim that Kant was prepared to make in his later ethical writings. The passage that brings it out most clearly is in the *Metaphysical Principles of Virtue*. Kant is explaining the duty of beneficence, and he says:

> Beneficence is a duty. Whoever often exercises this and sees his beneficent purpose succeed comes at last really to love him whom he has benefited. When therefore it is said, "Thou shalt love thy neighbor as

> thyself," this does not mean you should directly (at first) love and through this love (subsequently) benefit him; but rather, "Do good to your neighbor," and this beneficence will produce in you the love of mankind (as a readiness of inclination toward beneficence in general. (MM 402/61; Ellington).

Kant doesn't just mean that you will inevitably be gratified by succeeding where you've tried. He means that, although you make the good of others your end for moral reasons, you nevertheless come to enjoy pursuing the good of others because it is your end. We can change our nature, at least to some extent. You will notice that Kant does not think that adopting something as an ultimate end is the work of a moment, achieved by a simple act of mental resolution. The theory here is like Aristotle's theory of habituation: you come to value certain sorts of activities for their own sake – that is, you become the sort of person who values them for their own sake – as a result of practicing them. In *Religion within the Limits of Reason Alone*, Kant even says that unless the virtuous man has a joyous frame of mind, he is never really certain of having "attained a love for the good." (R 21n/19n)

Let me get back to the point, which is this. If what you take pleasure in depends on what ends you have, then you should not settle the question which ends you will have by asking which ones will give you pleasure. Kant's example is of someone adopting an end for a moral reason, but you might adopt ends for a variety of reasons, admirable and contemptible. You might adopt ends out of personal ambition or to please people you love or for moral reasons or to fit in with your neighbors. These are all different kinds of reasons for adopting ends, and nothing is gained by treating them all as various ways of trying to achieve pleasure (which they are not) or as various species of a monolithic motive called "desire." But achieving any of the ends to which these various motives prompt you might give you pleasure. All that is required is that it has really become an ultimate end for you. Although Hill's point, that deep deliberation shouldn't presuppose the value of

maximizing pleasures, makes this observation superfluous, it is still perhaps worth pointing out that even if you did want to maximize pleasure it wouldn't make sense to adopt only ends from which you anticipate pleasure. Anticipation of pleasure *is* one reason for adopting an end, but there is no particular reason to believe that the achievement of ends adopted for the sake of pleasure actually affords more pleasure than the achievement of ends adopted for other reasons.

All of that was by way of supporting Hill's contention that the right question to ask when deliberating about your final ends is not "Will they maximize my pleasures?" Similar considerations support the contention that you shouldn't ask, "Is it what I really want?" This question falsifies the deliberative situation. Deliberation is a practical enterprise, an enterprise of decision and construction, not a theoretical inquiry into your nature. Who after all is this "I" into whose real desires you would inquire? Is it your true self? There is something romantic and misleadingly metaphysical about asking what you really want. It is as if you had an essential self that was also your best and happiest self, and the business of deliberation was to uncover it. Of course we talk this way. We say, "here's what you really want" when what we mean is "here's a better thing for you to want." But this is probably a holdover from a view of human nature which few of us now accept, a view in which metaphysics and psychology somehow coincide. It's like St. Augustine claiming that what we really love about anything that we love is God. (Well, yes and no.) If the question "what do I really want?" is a metaphysical one, then *all there is for it to mean* is "what should I want?" If the question "what do I really want?" is a psychological or psychoanalytical one, the answer is probably something dreadful, which you had better not do.

Hill's proposed alternative question, one that *is* appropriate to deep deliberation, is something like "What can I justify to myself?" (I could not find a favored formulation of his deliberative question, so I take this one from his title.) Hill describes this as "a Kantian supplementary principle that requires rational concern for one's future and one's self-

respect." The principle requires that you be satisfied with what your choice makes or reveals of you. It requires that you in effect acknowledge the self-constituting character of the choice of ultimate ends, by considering whether you are and will be satisfied to be the person who has made these choices. As Hill nicely puts it, you are to choose with awareness that the choice itself, and not just its costs and benefits, is yours.

This leaves us with two further questions. What does the principle require of us, and how is it justified? I start with the first question. The principle seems primarily to restrict, rather than determine, the content of choice. Hill says it is "just the procedural, second-order concern that one's choices, whatever their content, be capable of surviving a deeply reflective scrutiny of and by oneself." (pp. 10-11) The principle does not most immediately suggest ends, but rather provides a perspective from which we can review the candidate ends that are proposed by various natural human incentives: desire, the prospect of pleasure, the interests of loved ones who have a stake in what we are, pride, ambition, and social pressure. In Hill's examples, the important thing seems to be avoiding either objects of choice or grounds of choice which are in one way or another contemptible. Presumably objects and grounds could be rejected separately. For instance, you might be tempted to pursue a project in itself admirable, but be aware that you are interested in it for a reason that makes you uneasy: perhaps you are trying to impress someone. Or you might have a ground of choice in itself unexceptionable, and yet be prompted by it to an action you don't think so well of: say, accepting an assignment you don't wholeheartedly approve of, because you are asked by someone to whom you are grateful.

Although Hill reminds us early in his paper that a practically rational choice must be informed by what we know about our motives rather than by what is true about them, his principle does seem to dictate that we try to achieve self-knowledge. A commitment to being responsible for the choices you make includes wanting to know what they really are.

My examples, like Hill's, are rather intuitive, and his principle would be strengthened if we were able to say something more determinate about what sorts of grounds *are* acceptable to a rational agent who respects herself as such. One definite thing can be said. You must regard yourself as choosing your own ultimate ends in the first place because in deliberation you must act under the idea of freedom. You must choose as if you are free to do and to be whatever your choice directs. If the rational agent as such respects herself as free, this is at least reason to avoid choice on overtly heteronomous grounds, such as feeling yourself trapped, bribed, or intimidated. One may say this and still stop short of Kant's early view that desire in general is a heteronomous ground of choice. On the other hand, however, if one endorses Kant's view that a free will and a will under moral laws are the same, or anything like it, then at this stage in the argument we are on an express train into moral territory, and the distinction between private interest and morality is about to become unclear.

Hill also argues that the principle of justification to oneself will lead us to concern for the future. This resolves one of the two problems with the view mentioned earlier. If the rational agent is not bound to accept any material ends as necessarily reason-providing, then *ipso facto* he is not bound to accept any future ends as necessarily reason-providing. But Hill argues that if a person regards himself as being the same rational agent over time, he has the same reasons of self-respect to be responsible to his future self that he has to be responsible to himself now. The reason is that it is him, and he is answerable to himself. I should note that this is consistent with the general defense of future concern advanced by Nagel in *The Possibility of Altruism*. Nagel's argument is to the effect that one must be motivationally responsive to one's future *reasons*, whatever they might be. Nagel treats desire as a plausible example of a reason, but the argument does not depend on that. So Hill could accept Nagel's general account of the basis of future concern. There is still a difference, for Hill's argument makes it harder, if not impossible, to anticipate what my future reasons will be. This doesn't mean that one cannot be prudent:

it means that prudence will consist to a great extent in keeping certain possibilities open. But this will not be for the common skeptical reason, that my future interests are hard to predict. Rather, being responsible to my future self means in part that I must not allow myself to reach the point where I feel trapped, bribed, or intimidated by my own past actions.

Hill proposes that we switch the focus of the thought "I will be the same person" from "I will be the same subject of experiences" to "I will be the same rational agent." This proposal has another advantage, which Hill just touches on. "I will be the same rational agent" is not a metaphysical claim about the continuity of consciousness or the existence of Cartesian egos. In *Reasons and Persons*, Derek Parfit may have shown that I have no special relation to the subject of experiences that will occupy my body in the future, especially the far future. But there is a reason – not metaphysical but practical – for regarding myself as the same rational agent as she will be.

To see this, forget about the problem of identity over time for a moment, and think about the problem of identity at any given time. Why do you think of yourself as one person? Hume believed that the self is a bundle. Various thoughts, feelings, desires, and other psychological paraphernalia are bundled together. When Hume asked what bundles them, he could not get an answer. In the *Treatise of Human Nature,* Hume thought he had a solution, but in the appendix added later he confessed that his solution didn't work, and gave the problem up as too hard for him. There is a problem about what makes you one person *now*, and this problem will seem especially pressing if Parfit has convinced you that you are not identified by a single continuing Cartesian Ego. You are even at a given moment just a conglomerate of conscious and unconscious psychological and physical functions. What make you one?

In the third part of his book *On the Soul*, Aristotle says that the practical faculty of the soul must be one thing. We can tell that it has parts, of course, because we sometimes have appetites that are contrary to practical reason, or experience conflict

among our various desires. Still the faculty that originates motion must be regarded as one thing, because we do act. Somehow, the conflicts are resolved, and no matter how many different things you want to do, you in fact do one rather than another. It may be that in actual fact the strongest one somehow wins. (Whatever that means.) But that isn't the way you think of it when you deliberate. When you deliberate, you think that there is something over and above all of these desires, something that is you, and that decides what to do. This is a conception you hold of yourself as a deliberating agent.

This conception of yourself as a single unified agent is not based on facts or metaphysical theories, but on sheer practical necessity. In advancing his arguments, Parfit makes a great deal of cases in which the two hemispheres of the brain function separately, and are unconscious of each other's activities. When their line of communication, the corpus callosum, is cut, they become separate agents. These cases suggest that the two hemispheres of the brain are not related in any metaphysically deeper sense than, say, two people who are married. They share the same quarters and, with luck, are in regular communication. (Even their characteristic division of labor turns out to be largely conventional.) Now, imagine that the right and the left half of your brain disagree about what to do. Suppose that they do not try to resolve their differences, but each merely sends motor orders, by way of the nervous system, to your limbs. Since the orders are contradictory, your limbs do not get clear signals. They start to do one thing and then start to do the opposite. Unless they can come to an agreement, both hemispheres of your brain are ineffectual. Like parties in the original position, they must come to a unanimous decision somehow (even if it is merely by way of the strongest one winning). You are a single person at any given time because you must act, and you only have one human body to act with. This is not a deep metaphysical fact, but a simple necessity.

Now let's see if we can extend this necessity to unity over time. Many considerations suggest that we can. First of all, most of the things we do take up time. Some of the things we do are projects that extend over long periods. This is especially true

of the pursuit of our ultimate ends. It is also true that many of us think of our various activities and pursuits as interconnected in various ways. We think that we are the authors of rational plans of life. To carry out a rational plan of life, you need to be one continuing person. You normally think you lead one life because you are one person, but according to this argument the truth is the reverse. You are one person because you have one life to lead.

Parfit might reply that this concedes his point about the insignificance of personal identity. If personal identity is just a matter of effectiveness in action and the necessities of cooperation, individual human beings do not have to be its possessors. We could, for instance, always act in groups. The answer to this is surely that for many purposes we do. A person is, on the view I am proposing, an agent, and an agent must be one unit in order to act and plan. Whenever some group wants or needs to act as a unit, it must form itself into a person: a legal person, a society, or a corporation. There are agents of differing sizes in the world. But this doesn't show that these agents aren't necessary. When a group of human beings occupy the same territory, we have an imperative need to form such a unit, and become a single society. When a group of psychological functions occupy a single human body, they have an imperative need to form a unit, and become a single person. This is why the human body must be conceived as a single agent. As things stand, it is the basic kind of agent.

Of course if the technology were different, individual human bodies might not be the basic kind of agent. My argument supports a physical criterion of identity, but it is more conditional than the versions of that criterion which Parfit challenges. Given the technology we have now, the unit of action is a single human body. The fact that the unit of action might be different if the technology were different is neither here nor there. The relevant necessity is the necessity of acting, and it remains. The main point of the argument is this: a focus on agency makes more sense of the notion of personal identity than a focus on

experience, for there is a natural connection between agency and unity which requires no metaphysical support.

Let me return to the other question I wanted to take up, the question of how a principle like the one Hill proposes might be justified. The argument that Hill suggests is, if I have understood him rightly, negative in character: since the requirements of deep deliberation forbid us to take any material end as a given, inevitable source of reasons, what is left is that any end we do treat as a source of reasons be able to stand up to deep and persistent reflection. Hill says that for the deep deliberator "nothing is beyond doubt -- unless seriously doubting it proves to be incompatible with the very undertaking the deliberator has set about." (p. 5) In deep deliberation we cannot doubt that we are responsible for what we treat as a reason. We must justify our reasons to ourselves.

I would like to sketch a slightly different line of argument, leading to a similar spot, which I believe to be (for what it's worth) closer to what Kant had in mind. I will start from a remark Hill himself makes when discussing the possibility of assessing our ends. Hill says:

> To say that we can reflect in this way does not mean that we will find the action-guiding answers we might hope for; for some contend that such questions are in principle unanswerable. On this view, rational assessment of ends can only be made relative to other ends which are taken for granted but not rationally required. The choice of ultimate ends is thus constrained only by human nature, not by reason; that is, though there would be natural limits to what creatures like us, on reflection, will choose as ends, nothing we can so choose will be irrational to choose. (pp. 6-7)

Hill describes this as a skeptical position. But there is a way to move from this thought to a more positive Kantian position. We can reason this way: nothing about my pleasures or the satisfaction of my desires themselves guarantees that they are good. To simply believe them good - where that means reason-providing - is just dogmatic metaphysics. Yet,

human beings, who must think of themselves as rational agents, do choose and pursue the objects of their pleasures and desires, and so must regard them as good. If the goodness of these things is relative to human nature, then human nature must be regarded as a source of value. Humanity must be regarded as an end in itself.

For this argument to land where us Hill's did, we need only remind ourselves that for Kant, rationality is the distinguishing characteristic of humanity, and the power to choose our own ends is the distinguishing characteristic of rationality. Kant thinks that all of our desires beyond the primitive instinctual drives for nourishment and sex are produced by reason through its operation of comparison: first comparison of natural objects of instinct with others like them, which Kant describes in *Conjectural Beginning of Human History*, then later comparison of our own situation with that of others like us, which Kant describes in *Religion within the Limits of Reason Alone*. (CB 111-112/55-56; R 22) Comparison proliferates the objects among which we choose our ends and the grounds on which we might choose them. Like Hill's deep deliberator, we find ourselves responsible for our own choice of ends, but without clear grounds on which to choose them. Describing this position as if it were an historical event, Kant wrote:

> Until that moment instinct had directed [man] toward specific objects of desire. But from these there now opened up an infinity of such objects, and he did not yet know how to choose between them. On the other hand, it was impossible for him to return to the state of servitude (i.e. subjection to instinct) from the state of freedom, once he had tasted the latter. (CB 112/56)

But one restriction on choice does offer itself, which is that if any of these ends are good at all, it must be because of the humanity, or rationality, which chooses them. Rational choice must be consistent with the value of humanity.

In a simple sense, this Kantian argument is transcendental. It asks, "how is it possible for human choice and action to be rational?" and then, given the relativity of

human practical reasons to human nature, it answers: "it is only possible if humanity itself is taken to be unconditionally good." Since the deliberative standpoint requires that we regard our actions and ends as rationally justified, it demands that we regard humanity as unconditionally good. And from this, substantive principles of action do follow.

This transcendental argument, like others, is intended to avoid dogmatic metaphysical claims. In particular, it avoids metaphysical claims like the one that Hill begins his paper by protesting: that pains are intrinsically bad. On Kant's argument, physical pains will be bad in exactly in the way that grief, frustration, and failed projects are bad: because they are contrary to the human choice that gives things value. And this in turn means that we cannot, without further argument, deem someone irrational merely because of the content of his choices. If someone chooses to undergo pains in order to carry out his projects, we must regard his pains as at least worth it. If someone chooses to abandon his project to avoid pains, that is what we must want for him. As Hill says, if someone sees *no reason at all* to avoid pain, or seeks it out, that is a sign of disorder, but the disorder is not a mistake about whether pains are intrinsically bad. The Kantian analysis of the disorder is closer to that we might give from a psychological point of view: the person does not have a perverse attitude about the value of pain, but rather about his own value. In some sense he surely does not desire the pain, so in some sense he declares his desires and so his humanity worthless. And severe physical pain reduces, enfeebles, and incapacitates us. Seeking it out would therefore express a deep kind of *disrespect* for oneself as a rational agent.