Planning and the Stability of Intention

MICHAEL E. BRATMAN
Philosophy Department, Stanford University, Stanford, CA 94305, U.S.A.

Abstract. I sketch my general model of the roles of intentions in the planning of agents like us – agents with substantial resource limitations and with important needs for coordination. I then focus on the stability of prior intentions: their rational resistance to reconsideration. I emphasize the importance of cases in which one's nonreconsideration of a prior intention is nondeliberative and is grounded in relevant habits of reconsideration. Concerning such cases I argue for a limited form of two-tier consequentialism, one that is restricted in ways that aim at blocking an analogue of Smart's concerns about "rule-worship". I contrast this with the unrestricted two-tier consequentialism suggested by McClennen. I argue that my restricted approach is superior for a theory of the practical rationality of reflective, planning agents like us. But I also conjecture that an unrestricted two-tier consequentialism may be more appropriate for the AI project of specifying a high level architecture for a resource-bounded planner.

Key words. Intention, planning, practical reasoning.

I believe that the concept of intention stability is important for the theory of action. I have discussed this concept in several earlier publications [Bratman (1983) and (1987); Bratman *et al.* (1988)]. I return to it here to develop several points, to respond to a recent discussion, and to sketch a conjecture about different projects within which issues about reasonable intention stability can arise.

Intention and Plans

To set the stage I need to tell you how I approach the topic of intention. Begin with a point suggested by Anscombe (1963). We use the concept of intention to characterize both our actions and our minds: we characterize actions as done intentionally and with a certain intention; and we attribute mental states of intending, or having an intention to act in certain ways now or later. A standard approach to intention is to begin with intentional action and action done with an intention. When we do this it is tempting to see the intentionality of action as lying in its *relation* to appropriate desires and beliefs. I raise my arm intentionally and with the intention of signalling for a cab. So I must want to signal for a cab and believe I can do this by raising my arm. It is tempting to suppose that for me to raise it with the intention of signalling is just for my raising it to stand in the right relation – perhaps a certain kind of causal relation – to that desire and belief. And it is similarly tempting to see the fact that my arm-raising is intentional as consisting in the fact that it stands in the appropriate relation to some such desire-belief pair. In this way we are led to see intention not as a distinctive state of mind, but as consisting solely in certain facts about the

Minds and Machines **2**: 1–16, 1992.

relations between actions, desires and beliefs. You might call this a "glue theory" of intention in action.

It is then natural to try to extend this reductive approach to intention for the future. I now intend to fly back to San Francisco on Sunday. What sort of state of mind is this? On such a reductive approach we might try to see it as some sort of belief-desire complex [Audi (1973)].

I take a different tack. [See esp. (1987). The present section is a brief summary of ideas I develop more fully there.] I begin with future-directed intention and ask about the roles it plays in our lives, eschewing the assumption that intention must somehow be reducible to desire, belief, causation and action. I try to articulate the systematic relations between such intentions, other psychological states, practical reasoning, and action. I try to describe a network of regularities and norms in terms of which we can understand what it is to have an intention for the future. I take this tack because I believe that future-directed intentions play a central role in our psychology, both individual and social, and that it is a serious error to ignore them in theorizing about intelligent agency. Intentions for the future involve a characteristic and important kind of commitment to action, and we can get at what this commitment is by articulating this nexus of roles and norms.

When you take future-directed intentions seriously in this way you run up against an obvious question: Why bother with intentions for the future? Why don't we just cross our bridges when we come to them? I believe there are two main answers. First, we are not frictionless deliberators. Deliberation is a process that takes time and uses other resources, so there are obvious limits to the extent of deliberation at the time of action. By settling on future-directed intentions we allow present deliberation to shape later conduct, thereby extending the influence of Reason on our lives. Second, we have pressing needs for coordination, both intra-personal and social; and future-directed intentions play a central role in our efforts at achieving such coordination.

Future-directed intentions typically play these roles as elements of larger, partial plans. My intention to fly back to San Francisco this Sunday helps coordinate my various activities for this weekend, and my activities with the activities of others, by entering into a larger plan of action – one that will eventually include specifications of when to leave the hotel and of how to get to the airport, and one that will be coordinated with my spouse's plans for meeting me when I arrive. Such plans will typically be partial and will need to be filled in as time goes by with appropriate specifications of means, preliminary steps, and the like. In filling in such partial plans in stages we engage in a kind of practical reasoning that is distinctive of planning agents like us.

Such reasoning is structured by two major demands on one's intentions and plans. First, there are demands for what I call strong consistency: One's plans taken together need to be both internally consistent and consistent with one's beliefs. Second, though partial, one's plans need to be filled in as time goes by.

Your plans need to be filled in with sub-plans concerning means and the like, sub-plans at least as extensive as you believe is now required to do what you plan. Otherwise your plans will suffer from means-end incoherence.[1]

Associated with these two demands are two roles intentions and plans play as inputs into practical reasoning. Given the demand for means-end coherence prior intentions pose problems for further deliberation, thereby establishing standards of relevance for options considered in deliberation. And given the needs for consistency, prior intentions constrain further intentions. In this way prior intentions provide a filter of admissibility on options that can be considered in deliberation aimed at resolving the problems posed by the incompleteness of the plans.

Consider my intention to fly back to San Francisco on Sunday evening. As part of a partial plan it poses a problem for further deliberation: how am I to get to the airport from the hotel? One solution to this problem might be to take the afternoon limo to the airport. But this solution is inadmissible, given that I am also planning to meet an old friend in the afternoon. Other solutions include taking a cab and taking the bus, in either case after I meet my friend. Both are relevant and admissible options. But which is superior? Here I weigh relevant desire-belief reasons for and against these competing solutions; I weigh, for example, speed and convenience against cost, in order to reach a decision. This deliberation is framed by my prior, partial plan. My prior plan provides a background framework within which such weighing of desire-belief reasons is to be done.

For all this to work well intentions will need to have two further features. First, when the time for action is seen to have arrived one's prior intentions will normally control one's conduct. Suppose I plan to take a cab at 6 p.m., if I see that it is now 6 p.m. then in the normal course of events I will at least endeavor to take one. The second feature brings me to the main topic of this paper. Prior intentions are not irrevocable. If things change in relevant ways it might behoove me to change my plan for returning Sunday night. Still, prior intentions will need to have a certain stability: if we were constantly to be reconsidering the merits of our prior plans they would be of little use in coordination and in helping us cope with our resource limitations. The nonreconsideration of one's prior intentions will typically be the default.

This means two things. First, having settled on an intention to A one will normally be disposed not to reconsider this intention except in the face of some relevant change of belief or desire. Such resistance to reconsideration is a defining feature of intention, one that helps support the role of intention in the kind of planning-by-stages that is so important to agents like us – agents with substantial resource limits and strong needs for coordination. Second, this resistance to reconsideration will to a certain extent extend even to cases in which one acquires new information. Resistance to reconsideration in this latter kind of case will be the focus of the early parts of this essay; but towards the end of this essay I will

return to a special query about reconsideration even in the absence of new information.

Stability of Intention

When we take the idea of intention stability seriously we are led to the following result. It will many times be reasonable of an agent to act on her prior intention to A even though the agent would reasonably have abandoned this prior intention and acted differently had she stopped to reconsider that intention. This result lies at the heart of the idea that prior intentions involve a distinctive commitment to action.

To discuss these issues I need two further ideas. First, I will assume here that practical rationality is, at bottom, a matter of satisfying rational desires.[2] One's prior intentions do not provide reasons for action in the basic way in which one's belief-desire reasons do. Still, for the reasons just emphasized, prior intentions and partial plans play central roles in the normal functioning of agents like us, agents with substantial resource limitations and with basic interests in coordination. Prior intentions and plans provide the background framework within which most deliberation takes place; and these prior intentions shape such deliberation by determining, in part, which options are relevant and admissible. In this way intentions provide *framework reasons* – reasons that shape what it is rational to decide to do, but reasons whose ultimate rational force rests on the overall contribution of this planning system to the satisfaction of rational desire [Bratman (1987), Section 3.3]. This mixed status of intention-based framework reasons drives my treatment of intention stability.

Second, it is important to note that in most cases the reconsideration or nonreconsideration of a prior intention is not itself the product of deliberation. Typically, we cannot or do not take the time to deliberate about whether to reconsider a prior intention. Instead, we either proceed in a non-deliberative fashion not to reconsider our prior intention, or we proceed directly to reconsideration. In either case the (non)reconsideration[3] is itself a product not of explicit deliberation about whether to reconsider but rather of other kinds of psychological processes, processes grounded in general habits and other non-deliberative mechanisms.

Consider now cases in which I form a future-directed intention but then have some relevant change in belief. At $t1$ I form the intention to A at $t2$. I form this intention on the basis of my relevant desire-belief reasons for and against A at $t2$ and for and against its relevant and admissible alternatives. And I see the desire-belief reasons I thereby have in favor of my A-ing as superior to those I have in favor of the relevant alternatives to A. Essential to these reasons at $t1$ are relevant intrinsic desires and beliefs about the circumstances I will face at $t2$. Let us suppose that when $t2$ arrives there is no change in my relevant intrinsic desires, but there are differences between what I expected at $t1$ to be the case at $t2$ and

what I now, at $t2$, believe to be the circumstances at $t2$. Sometimes my new belief at $t2$ will simply be an addition to my earlier beliefs. Sometimes my cognitive change will essentially involve the rejection of some earlier belief.[4] In either case we can ask whether such a cognitive change should lead me to reconsider my prior intention to A.

On the one hand, some kinds of divergence between my earlier expectations and my later, updated beliefs about my circumstances at $t2$ will straightway oblige me to reconsider. I cannot rationally intend to A at $t2$ and also believe that I cannot A at $t2$. So if I newly come to believe that I cannot A at $t2$ then I am rationally obliged to reconsider. That I am obliged to reconsider in such a case is a consequence of a basic structural constraint on rational intention.[5] On the other hand, some kinds of divergence between earlier and later beliefs will normally have no tendency at all to trigger reconsideration. If I discover yet another reason for A-ing, or yet another reason against one of A's alternatives, I need normally have no inclination to reconsider. In such cases the cognitive change will normally simply reinforce my earlier decision at $t1$ to A at $t2$.

In contrast with both such cases, some cognitive changes, while they do not straightway oblige me to reconsider in the way in which a new belief that I cannot A would, do provide *prima facie triggers of reconsideration*. Suppose for example that I have earlier decided to go to the theater instead of the piano concert. I arrive at the theater and discover that the tickets are more expensive than I had earlier anticipated (though I do have enough money with me to pay for them). Or, instead, as I approach the theater I newly learn that there is also a string quartet concert that I could attend instead. In the first case my cognitive change somewhat weakens my desire-belief reasons for going to the theater. (In a variant of this case I discover instead that the tickets for the piano concert are less expensive than I had thought and so my desire-belief reasons in favor of an alternative to the theater are strengthened.) In the second case my cognitive change introduces a new and attractive alternative to my going to the theater, an alternative not previously considered. Such cognitive changes potentially threaten my earlier decision even though they do not change my belief that I can do what I earlier decided to do. But in response to such cognitive changes should I always reconsider?

It may seem that once my beliefs change in such ways there is no presumption at all in favor of my intention to A. This sentiment is captured nicely by Donald Davidson:

> A present intention with respect to the future is in itself like an interim report: given what I now know and believe, here is my judgment of what kind of action is desirable, . . . My intention is based on my present view of the situation; there is no reason in general why I should act as I now intend if my present view turns out to be wrong. (Davidson, 1980, p. 100)[6]

I believe, however, that it can be misleading to see intention in this way as "like

an interim report," for this may tempt us to ignore important pressures in the direction of stability.

Suppose the change in my beliefs from $t1$ to $t2$ provides a *prima facie* trigger of reconsideration. We can go on to ask [Bratman *et al.* (1988)]: if at $t2$ I were to stop and reconsider my prior intention to A in the light of this change in my beliefs would I still see my A-ing as best supported by my desire-belief reasons, or would I instead decide that a relevant alternative to A is superior and so abandon my intention to A? Let us call a case in which such reconsideration *would* lead to a rational change of intention a *would-change* case, and a case in which there would be *no* such change, despite the change in belief, a *wouldn't-change* case. In wouldn't-change cases the desire-belief reasons I had at $t1$ for A-ing at $t2$ are themselves resilient in the face of my change in belief: I continue to have sufficient desire-belief support for my intention to A. In this case my intention would not reasonably change even if I were to go ahead and reconsider. So my intention is stable in the face of the changes in belief.[7] This stability derives directly from the resilience of my desire-belief reasons for A-ing. In contrast, the intention stability that is of most interest here is not derivative in this way from the agent's desire-belief reasons for the intended action.

Reconsidering a prior intention is an activity that uses up time and other limited resources; while engaged in reconsideration I am unable to do other valuable things. To appreciate these potential costs of reconsideration note a further fact: prior intentions tend to become *enmeshed* in our various plans. Given the prior intention to A at $t2$ I will typically have gone some way towards settling on how to A; and I will typically have adjusted my various other plans so that they are compatible with my A-ing at $t2$. When I reconsider whether to A at $t2$, then, I will typically need to consider issues about the compatibility with my other plans of alternatives to A; and I will also need to be prepared to trace out means to these alternatives to A. As a result, reconsideration of my prior intention can become complex, involve significant further reasoning and planning, and risk undermining coordination with other plans.[8]

These observations lead to a distinction between two species of would-change cases. In a would-change case, recall, if I were to reconsider my prior intention to A at $t2$ in the light of my beliefs at $t2$ – beliefs that diverge in some relevant way from those on the basis of which I formed my intention at $t1$ – I would reasonably abandon my intention to A. We have seen, however, that this reconsideration will itself have costs. Consider the benefits I would achieve by abandoning my intention to A in favor of some alternative to A – benefits assessed in terms of the satisfaction of my rational desires. If I could reflect on the matter would I see such benefits as outweighing the costs of the reconsideration itself? If the answer is "yes" I will call the case a *would-change/worth-it* case (sometimes, for short, a *worth-it* case); if the answer is "no" a *would-change/not-worth-it* case (sometimes, for short, a *not-worth-it* case).[9]

It is only in would-change/worth-it cases that reconsideration of my prior

intention, given my change in beliefs, can be directly recommended to my concern with the satisfaction of rational desire. So we have discovered a limited way in which the presence of my prior intention to *A* makes a difference to what it is reasonable for me to do at *t*2. Given this prior intention it might be reasonable for me to *A* at *t*2 even though, had I started from scratch at *t*2 and tried to determine what to do in the light of my desire-belief reasons, I would reasonably have plumped instead for some alternative to *A*. My case might be a would-change/*not*-worth-it case.

We now need to come to terms with the fact that (non)reconsideration of a prior intention is typically not the result of explicit deliberation about whether or not to reconsider. Instead, my (non)reconsideration will typically be the result of various underlying habits – what David Velleman (1989) once called "delicate mechanism[s]." One cannot regularly stop and deliberate with care about whether or not to reconsider without getting hopelessly tangled up. Sometimes such second-order deliberation may be in order; but more typically we rely on background habits, strategies and policies. We rely on psychological mechanisms of salience, problem detection, and the like. So that we have a single term with which to work I will lump these mechanisms together under the heading: *habits of reconsideration*.

Of particular interest here are those habits of reconsideration which concern cases in which the agent's cognitive change, though it does not straightway oblige her to reconsider, does present her with a *prima facie* trigger of reconsideration. Let us call such habits of reconsideration *prima-facie-trigger habits* – for short, *pft habits*. Pft habits shape an agent's non-deliberative responses to circumstances in which she is not straightway obliged to reconsider but is presented with a *prima facie* trigger of reconsideration. Recognizing the relevance of my pft habits to my non-deliberative (non)reconsideration of prior intentions, we may ask about the expected impact of those habits on the long-term satisfaction of my rational desires. Taking a broadly pragmatic approach, we can say that these pft habits are reasonable when this expected long-term impact exceeds an appropriate threshold.

Suppose now that I refrain at *t*2 from reconsidering my prior intention despite being faced with a *prima facie* trigger of reconsideration; and suppose this is an upshot of my relevant pft habits. Suppose, further, that these pft habits are themselves reasonable: their expected long-term impact on my rational desire satisfaction exceeds an appropriate threshold. In such a case I think we should say that it was reasonable of me at *t*2 not to reconsider my prior intention; for in retaining my prior intention and not subjecting it to reconsideration I am functioning in a way that has an appropriate pragmatic rationale. This is a *two-tier consequentialist* account of reasonable non-deliberative nonreconsideration in the face of a *prima facie* trigger of reconsideration. This account ramifies within our treatment of reasonable intention and action. In particular, if it was reasonable of me to form my intention in the first place, and it has since then been reasonable

of me not to reconsider, then it is at $t2$ reasonable of me to intend to A at $t2$. And if it is reasonable of me so to intend then it is reasonable of me to execute that intention by A-ing at $t2$. [Bratman (1987), Chs. 4–6]

How does this two-tier consequentialist account fit with our earlier classification in terms of worth-it cases, not-worth-it cases, and wouldn't-change cases? Well, one thing we can now say is this: Other things equal, we want pft habits that issue in reconsideration only in worth-it cases, though, of course, we cannot expect perfect fine-tuning in such habits and strategies. [Bratman *et al.* (1988)]

But we do not want to stop here. I have emphasized the role of plans in helping us coordinate our activities, both intra-personally and socially. When we assess pft habits we need to keep track not only of their impact on the extent to which the agent reconsiders only in worth-it cases; we also need to keep track of their impact on the agent's general ability to benefit from forms of coordination. I have already suggested that these two desiderata are linked: a characteristic risk of the reconsideration of one's intention to A is the potential for undermining coordination with other plans that have earlier been adjusted to cohere with one's A-ing; and if I spend too much time reconsidering my intention I may be late for our planned, coordinated activity. But one's pft habits will also have impacts on coordination that are not so directly linked to the costs and risks of the activities of reconsideration that directly result from these habits.

Suppose you and I plan to meet today for lunch. It will be important to me to know how reliable you are about such things. If you are rather resistant to reconsidering such prior intentions, and I know this, I will be somewhat more willing to make such plans with you and to go out of my way to keep such appointments with you. My knowledge of your habits of reconsideration will directly affect the extent to which I am willing to be a partner with you in mutually beneficial coordinating schemes. And this will lead to social pressure in the direction of pft habits that support increased stability of intentions. Again, in embarking on a long and complicated project – such as writing a book – I will be aided by the knowledge that my habits of reconsideration support fairly stable intentions, that I will not treat almost any *prima facie* trigger as an excuse for reconsidering my decision to write the book. Such reliability may help me justify to myself now sacrifices I must now make in the initial stages of working on the book.

Smart's Problem

It is time to get to an important complexity. Two-tier consequentialist approaches have been widely studied in recent moral philosophy, and many believe that they face a deep problem. I will call this *Smart's Problem*, for its formulation as an objection to rule-utilitarianism was initiated by Smart (1967). Rule utilitarianism sanctions two forms of moral reasoning: we are to assess general rules by appeal to their consequences, but we are to assess particular acts by appeal to their fit

with such general rules and not by appeal to *their* consequences. But given its basic concern with the goodness of consequences it is unclear how this conception can non-arbitrarily block consequentialist reasoning concerning particular acts. Suppose that a rule requiring the keeping of promises is justified on consequentialist grounds, but that in a special case my breaking a promise made to a dying person on a desert island would have the best consequences. The rule utilitarian will say that it is nevertheless wrong to break the promise, for I would thereby violate a rule which is itself justified on utilitarian grounds. But Smart would say that this is "superstitious rule-worship" [(1967), p. 177].

The main target of this objection is the agent who justifies a rule to herself on consequentialist grounds, sees that this rule dictates keeping the promise, but also sees that in her special circumstances breaking the promise would have the best consequences. If her rationale for accepting the rule is consequentialist, how can she rationally resist the consequentialist rationale for breaking the promise? For an agent who justifies accepting the rule on consequentialist grounds this will look like irrational rule-worship.

So understood, the objection depends on its being the very same agent who seeks a justification both of the rule and of the particular act. Let us call agents who seek justification both for the rules they accept and for the acts they perform *reflective* agents. Smart's Problem is a problem for any two-tier consequentialist theory whose intended range includes the activities of reflective agents.

My conception of the roles of intentions in practical reasoning is intended to apply, *inter alia*, to reflective agents. So I face an analogous concern about my two-tier consequentialist approach to reasonable non-deliberative nonreconsideration of a prior intention. Suppose that pft habits that are reasonable in our consequentialist sense would lead me at $t2$ not to reconsider my prior intention in the face of a *prima facie* trigger. But suppose that at $t2$ it is obvious to me that mine is a would-change/worth-it case. That is, it is obvious to me that, given my change of belief from $t1$ to $t2$, my desire-belief reasons at $t2$ argue clearly for abandoning my intention to A and reconsidering what to do despite costs of reconsideration.[10] In such a case will my two-tier consequentialist approach sanction irrational *habit* worship?

We might try to reply by insisting that any reasonable habits of reconsideration will have built into them a general escape clause: if it is *obvious* that this is a worth-it case, reconsider what to do.[11] But though this will normally be right, if we stick with our consequentialist approach we cannot be sure that such habits will always be superior to habits without such a general escape clause. Indeed, I will discuss an example below in which it seems that the opposite may be the case.[12] So rather than insist that consequentially justified habits of reconsideration must have such a general escape clause, I propose taking a different tack.

Earlier I distinguished three different kinds of cognitive change. First, in some cases the cognitive change straightway obliges reconsideration. My example here was the case in which one newly comes to believe that one cannot A at $t2$.

Second, there are cases in which the cognitive change will normally exert no rational pressure at all towards reconsideration. Third, there are cognitive changes – such as my new information about the higher cost of theater tickets – that provide *prima facie* triggers. My two-tier consequentialist approach has concerned habits of reconsideration whose targets are cases of this third type which do not also include cognitive changes of the first type. Such habits of reconsideration I have called "pft habits." And now what I want to say is this: *it's being obvious to me that mine is a would-change/worth-it case, in the absence of a special reason to distrust my own judgment, straightway obliges reconsideration.* This means that cases in which it is obvious to me that mine is a would-change/worth-it case are (barring reason to distrust my own judgment) *not* cases to which my pft habits are applicable. In such cases my two-tier consequentialist approach is not engaged: for this approach concerns only cases in which my pft habits determine whether I reconsider. So my two-tier consequentialist theory need not sanction irrational habit-worship. Reasonable stability is not stubborness.[13]

Contrasting Perspectives

It will be useful to locate my view on a (partial) map of contrasting views of the stability of intention. My attention will be confined to views which suppose that, in addition to ordinary desires and beliefs, there really are future-directed intentions. The differences to be traced concern the kind of stability that such an intention should rationally have over time. We can distinguish six different positions:

View #1: The interim-report view. On this view prior intentions have a stability that is totally derivative from the resilience of their underlying desire-belief reasons for the act that is intended. This is the view that seems to be suggested by the remarks of Davidson quoted above.

View #2: Interim report plus costs of reconsideration. A prior intention to *A* has a stability that is not totally derived from the resilience of the underlying desire-belief reasons for *A*-ing, but this is solely because reconsideration of a prior intention has its own costs. Because of the costs of reconsideration it can sometimes be reasonable to retain one's prior intention even though reconsideration, were it to occur, would argue for a different option. This is why our distinction between would-change/worth-it and would-change/not-worth-it cases is important.

View #3: Restricted two-tier approach, focusing on costs of reconsideration. Views 1 and 2 do not yet come to terms with the point that (non)reconsideration is typically non-deliberative and habit-based. View 3 tries to do justice to this point by adopting a two-tier consequentialist strategy. View 3 seeks habits of reconsideration which will, as much as possible, issue in reconsideration in all and only would-change/worth-it cases. In order to side-step Smart's Problem, how-

ever, View 3 limits this two-tier consequentialist strategy in the way described above. In particular, this two-tier consequentialist strategy does not apply to cases in which it is obvious to the agent that his is a would-change/worth-it case.[14] We can call this a *restricted* two-tier theory.

View #4: View 3 plus appeal to benefits of coordination. View 4 accepts the restricted two-tier consequentialist approach of View 3. But it expands the basis for the consequentialist assessment of pft habits. In particular, it explicitly includes in this basis the impact of such habits of reconsideration on an agent's ability to benefit from forms of coordination. This is the view I have defended here.

View #5: Unrestricted two-tier theory. This is View 4 without the restriction motivated by Smart's Problem. On View 5 if a habit or policy of reconsideration has a pragmatic rationale for a certain range of cases then it is reasonable to follow this policy whenever one's case is within that range of cases. The departure from View 4 concerns those cases, if such there be, in which a pragmatically justified habit or policy dictates nonreconsideration even though in the particular case it is obvious to the agent at $t2$ that it would then be better to abandon his prior intention. View 5 supposes that in such a case a reasonable agent should nevertheless stick with his prior intention. This is an *unrestricted* two-tier consequentialist conception.[15]

View #6: Intrinsic stability. View 6 agrees with Views 3–5 that there is a source of the reasonable stability of intention that goes beyond what is envisaged in Views 1 and 2. But on View 6 this source of stability is not to be found in two-tier consequentialist considerations but rather in something intrinsic to what intentions are. Perhaps this is the view of Michael Robins in identifying intentions with "normative commitments to bring about their satisfaction conditions" [Robins (1984), p. 35].

I suspect that View 6 is threatened by what elsewhere [Bratman (1987), Ch. 2] I have called "bootstrapping" problems. However, I will focus here on the contrast between Views 4 and 5, and in particular on a version of View 5 recently suggested by Edward McClennen (1990). This will allow me to arrive at the conjecture promised at the beginning of this essay.

McClennen's Challenge

McClennen argues in favor of a strategy of "resolute choice" for certain special kinds of "dynamic choice" situations. McClennen's subtle discussion is complex and touches on a wide range of issues I cannot discuss here. But we can get at some relevant ideas by reflecting on his version of an example introduced into the literature by Gregory Kavka (1983).[16] There is a non-lethal toxin drink which makes one sick. A billionaire offers you the following deal: See if you can intend at midnight to drink the toxin the next afternoon. If you really do so intend at

midnight you will be, in McClennen's words, "tested by the brain-mind machine to see if you have the *capacity* to act on such a plan" [McClennen (1990), p. 229].[17] If you have both the intention and the "capacity" to act on it your bank account will be credited irrevocably with one million dollars that morning, prior to when you must drink the toxin. And you will know that morning that you have received the money. Now, you clearly prefer getting the million dollars at the cost of some sickness to remaining poor without the sickness. But how can you satisfy the billionaire's conditions? Come the next afternoon you will have no reason at all to drink the toxin, since by then the money will either be in your bank account or not, and you will know either way. So, being a rational sort of person, you will then choose not to drink the toxin. But then even if you could somehow get yourself to forget all this at midnight and get yourself to intend to drink,[18] you will not have the "capacity" to carry out this intention. So, given an accurate "brain-mind machine", you will not get the money.

In our discussion so far of intention stability we have focused on cases in which there is, between the formation of the intention at $t1$ and the time for its execution at $t2$, some change of belief about circumstances at $t2$. Our question has been: when should this change of belief actually trigger reconsideration of the prior intention? As noted earlier, however, when there is *no* change in relevant belief or desire one should normally not reconsider. And in a way the toxin example is such a case. After all, you know, at the time of getting yourself to have the intention to drink, what things will be like the next afternoon. The problem you will have with executing such an intention the next afternoon is not unexpected and does not depend on any correction of belief from $t1$ to $t2$. But the toxin example does present a special problem, for your special reasons at midnight to get yourself to intend to drink the toxin the next day will not be reasons the next day for actually drinking the toxin. Does the presumption in favor of nonreconsideration in the absence of new information nevertheless still apply here?

McClennen thinks it does. He thinks that in a case of this kind one should be a "resolute chooser". A resolute chooser in such a case would form the intention at midnight to drink the toxin the next afternoon and then stick to his guns the next afternoon if in so doing he conforms to a pragmatically justified general policy.[19] In this toxin case a policy of sticking to one's guns would have such a pragmatic justification; for if one internalized such a policy one would be able to win the million dollars. So a resolute chooser would intend to drink and then stick to his guns and drink the toxin the next afternoon; and so he would win the million dollars. The resolute chooser would drink the toxin even though, when it comes time to drink, his ordinary desire-belief reasons then argue overwhelmingly in favor of not drinking.

McClennen correctly anticipates that I would not agree with this endorsement of "resolute choice". As he says, I would suppose (along with Kavka) that "whether the money is in your bank account or not, deliberation the next

afternoon (if it does take place) will take place just by reference to the prospects that still lay open to you: a painless afternoon vs. much pain and discomfort during the afternoon" [McClennen (1990), p. 230][20]. But McClennen wonders whether I have failed to appreciate where my view of intention stability really leads. He writes:

> Reasonable habits of nonreconsideration, according to Bratman, are based on a consideration of the impact of a habit of nonreconsideration on the agent's long-term prospects of getting what he wants. . . . One way in which such a habit can have a favorable impact on long-term prospects concerns savings with respect to decision-making costs. . . . But a habit of nonreconsideration may also be grounded in . . . the need to coordinate. Coordination can bring benefits over and above those associated with reduced decision-making costs. This is true not only with regard to plans that involve a number of agents, but also, as Bratman makes clear, when the single agent must coordinate with his own future selves. Now the modified version of the toxin puzzle presents the agent with just such a coordination problem. This evening it is distinctly in his interests not only to plan to drink the toxin the next afternoon, but to have the capacity to execute that plan. From the perspective of his future self – the self of the next afternoon – however, it is not in his interests to drink the toxin. If he can resolve this coordination problem, he stands to gain . . .
>
> Thus it would appear that Bratman's framework can be employed to rationalize a policy of nonreconsideration that would apply to the modified version of the toxin example. . . . a policy of nonreconsideration can be rationalized by reference to the gains to be had from being able to coordinate the decisions of different time slices of oneself. [McClennen (1990), pp. 230–231]

McClennen's point that "coordination can bring benefits over and above those that are associated with reduced decision-making costs", argues for the transition from View 3 to View 4. Here I agree with McClennen. But McClennen's use of this point, to argue that you should be a resolute chooser and go ahead and drink, depends on an *unrestricted* version of a two-tier consequentialist conception. After all, when the afternoon arrives it will be obvious to you that, considering only the desire-belief reasons you will then have, it would be silly to drink the toxin. If the two-tier consequentialist theory at work were a restricted theory of the sort embraced by View 4 it would not apply to such a case. McClennen needs the unrestricted two-tier consequentialist theory of View 5 to arrive at his endorsement of "resolute choice" in such examples.

McClennen and I agree in rejecting models of practical rationality within which, as he says, "there is no temporal thickness to any commitment to a plan" [McClennen (1990), p. 208]. But we disagree about how to capture this idea. McClennen goes all the way to View 5; given Smart's Problem I think we should stop with View 4. In this way we can take seriously the distinctive stability of prior intentions and plans without sanctioning unreasonable habit worship. Life, after all, is complicated. In working on a book project or planning to meet your friend for lunch it is unusual for it to become obvious that you would do best, from then on, to abandon your prior plan. However, if it does, you should (baring reasons to doubt your own judgment). But so long as your ordinary, reasonable pft habits of reconsideration are engaged they will normally invest your prior intentions and plans with a significant level of stability. So View 4 is not in danger of retreating to a "no temporal thickness" view.

Two-Tier Theories and Artificial Intelligence

There is yet another complexity, however. Suppose we are concerned with, as my friends in AI say, design specifications for constructing an intelligent machine. Suppose that we consult the model I have sketched of intentions and plans, their interaction with desires and beliefs, and their role in action. And suppose that we try to develop this model as a high level architecture for constructing a resource-bounded planner.[21] To do this we will need, *inter alia*, to specify appropriate strategies of reconsideration.[22] For *this* project it seems that we can reasonably proceed in accordance with an *unrestricted* two-tier conception. We can reasonably just decide which general strategies of reconsideration would be close enough to optimal and put them into the architecture. We need not worry that the system we are thereby specifying will run up against a version of Smart's Problem. Our concern is with the long-run effectiveness of the system and we can simply allow that concern to drive our decisions, at the upper tier, about general strategies of reconsideration. In pursuing this project we are not ourselves engaged in two kinds of reasoning, trying to be consequentialists with respect to one and yet to block consequentialist reasoning with respect to the other. We are just designing the system at the upper level, and reasoning in a consequentialist fashion about that. So it may well be that View 5 is appropriate for this AI project.

Of course, even if we endorse an unrestricted two-tier consequentialism, the habits/strategies that are justified on such pragmatic grounds might themselves involve a general escape clause along the lines of: if it is obvious to you that you do best to reconsider then do so. But given an unrestricted two-tier conception, whether or not the justified habits involve such a general escape clause is itself a matter of the relevant consequences. And McClennen may be right that for a certain range of cases the pragmatic argument supports, instead, a strategy of resolute choice. If so, such a strategy might recommend itself for such an AI architecture, even though, as a recommendation to reflective agents, it runs into a version of Smart's Problem.

I have urged that Smart's Problem should lead us to stop at View 4 in our effort to construct a plausible conception of the stability of the intentions of reasonable, reflective agents. If we continue all the way to View 5 we risk putting into the head of one and the same agent the view both that her habits of reconsideration are justified on consequentialist grounds and that she should not reconsider this time despite an overwhelming consequentialist argument for doing so. And that is the kind of situation we have tried to avoid. However, we have just seen reason to suppose that View 5 *is* an appropriate perspective for the corresponding project in AI. If both these conjectures are on the right track we have uncovered an important difference in these projects. In one case we are interested in a conception that can itself be endorsed by a reflective agent, and we seem well advised to stop with View 4. In the other case we are just interested in designing an efficient system, and here the unrestricted two-tier consequentialism of View 5

seems to provide an appropriate perspective. Our view about reasonable stability of intention will need, then, to be relativized to the intellectual project within which that view is to function.[23]

Notes

[1] I have sometimes simply called this "incoherence". See Bratman (1989).

[2] Though other views about practical rationality – for example, views which tie it to a conception of an objectively good life – are compatible with most of what I will say here. For relevant discussions see Brandt (1979) and Parfit (1984).

[3] I use "(non)reconsideration" to abbreviate "reconsideration or nonreconsideration".

[4] A distinction along these lines was emphasized in correspondence by John Pollock.

[5] George Smith's comments on an earlier version of this essay (read at the Greensboro Conference on "Approaches to Cognition") helped me clarify this point.

[6] My concern here is with the conception of intention as an "interim report". I am not here concerned specifically with the idea that intention is a desirability judgment. I discuss this latter idea in Bratman (1985).

[7] In a wouldn't-change case my cognitive change might still have led to a change in my subsidiary intentions about how to do A. Change in subsidiary intentions in response to relevant cognitive changes is characteristic of having an intention to A. If I intend to hit a target and I get new information about where that target is I will normally change my subsidiary intentions about how to hit the target. See Bratman (1987), Ch. 10.

[8] I was helped here by Martha Pollack.

[9] Thanks to Jane Aronson for suggesting this terminology.

[10] Note that this may be true even if it is not obvious which specific alternative should replace A in my plans.

[11] This is close to what I did assume in Bratman (1987), p. 106.

[12] See the section called "McClennen's Challenge".

[13] In this paragraph I have benefited from conversation with Joshua Hoffman.

[14] This view, without the restriction driven by Smart's Problem, is hinted at in Bratman *et al.* (1988), even though it was not my view in Bratman (1987) and is not now my considered view. This discrepency is due to the fact that in the former essay Israel, Pollack and I simply didn't pursue the questions that lead to a distinction between Views 3 and 4.

[15] In Bratman (1987) I did not have this distinction between unrestricted and restricted two-tier theories, though I was explicitly concerned with Smart's Problem. (See p. 69.) Reflection on a discussion of my views by Edward McClennen (1990) – see below – has led me to believe that that this distinction is needed by my theory – though this is not the moral McClennen intended me to draw!

[16] I discuss Kavka's puzzle in (1987), pp. 101–106. McClennen labels his example the "modified version of the toxin example" (p. 229). In discussing McClennen's version of the example I will make some small changes in order to simplify the discussion.

[17] You have this "capacity", in McClennen's sense, only if you will not later reconsider and abandon this intention. This test for the "capacity" to act on the intention is McClennen's main modification to Kavka's original example. It is interesting to note that, so modified, the example is similar to a "Newcomb" example [Nozick (1969)] in which, contrary to the usual story, *both* boxes are transparent. David Gauthier discusses such a modified Newcomb example in Gauthier (1988–89). Gauthier's essay is relevant to present concerns, but I cannot discuss it here.

[18] Given your knowledge of what things will be like the next afternoon, you cannot get yourself to have this intention simply by considering the reasons you will have then for drinking the toxin.

[19] To be more precise, he will stick to his guns when "both the *ex ante* self and the *ex post* self can reasonably expect to benefit from such a policy". McClennen (1990), pp. 212–213.

[20] Note that in the general case the parenthetical qualification is crucial. On my view a planning agent will normally simply not reconsider his prior intention, and this tendency is supported by appropriate habits of reconsideration. However, if the agent *does* reconsider and start his deliberation about what

to do from scratch then he should in such a case (in which there are, for example, no relevant promises) concern himself with what will happen *from then on*. And in the toxin case a reasonable agent would reconsider: or so I argue.

[21] This is what we tried to do in Bratman *et al.* (1988).

[22] In Bratman *et al.* (1988) this problem appears as the problem of specifying the "filter override mechanism".

[23] Ancestors of this essay were presented to the Rational Agency research group at the Center for the Study of Language and Information, and at the conference on "Approaches to Cognition", April 1991, at the University of North Carolina at Greensboro. I benefited from the comments of many of those in attendance on these occasions. I want particularly to thank my commentator at the Greensboro conference, George Smith, whose comments helped me clarify and improve this essay in a number of ways. Work on this essay was supported in part by the Center for the Study of Language and Information and by the Stanford Humanities Center.

References

Anscombe, G. E. M. (1963), *Intention*, (2nd edition). Ithaca: Cornell University Press.

Audi, Robert (1973), 'Intending', *Journal of Philosophy* **70**, pp. 387–403.

Brandt, Richard (1979), *A Theory of the Good and the Right*, Oxford: Oxford University Press.

Bratman, Michael (1983), 'Taking Plans Seriously', *Social Theory and Practice* **9**, pp. 271–287.

Bratman, Michael (1985), 'Davidson's Theory of Intention' in Bruce Vermazen and Merill B. Hintikka, eds., *Essays on Davidson: Actions and Events*, Oxford: Oxford University Press, pp. 13–26.

Bratman, Michael E. (1987), *Intention, Plans, and Practical Reason*, Cambridge: Harvard University Press.

Bratman, Michael E., Israel, David J., and Pollack, Martha E. (1988), 'Plans and Resource-Bounded Practical Reasoning', *Computational Intelligence* **4**, pp. 349–55.

Bratman, Michael E. (1989) 'Intention and Personal Policies', *Philosophical Perspectives* **3**, pp. 443–469.

Davidson, Donald (1980), 'Intending', in Donald Davidson, *Essays on Actions and Events*, New York: Oxford University Press.

Gauthier, David (1988–89), 'In the Neighbourhood of the Newcomb-Predictor (Reflections on Rationality)', *Proceedings of the Aristotelian Society* **89**, pp. 179–194.

Kavka, Gregory (1983), 'The Toxin Puzzle', *Analysis* **43**, pp. 33–36.

McClennen, Edward F. (1990), *Rationality and Dynamic Choice: Foundational Explorations*, Cambridge: Cambridge University Press.

Nozick, Robert (1969), 'Newcomb's Problem and Two Problems of Choice', in Nicholas Rescher, ed., *Essays in Honor of Carl G. Hempel*, Dordrecht: D. Reidel, pp. 114–146.

Parfit, Derek (1984), *Reasons and Persons*, Oxford: Oxford University Press.

Robins, Michael (1984), *Promising, Intending, and Moral Autonomy*, Cambridge: Cambridge University Press.

Smart, J. J. C. (1967), 'Extreme and Restricted Utilitarianism', in Philippa Foot, ed., *Theories of Ethics*, Oxford: Oxford University Press, pp. 171–183.

Velleman, J. David (1989), 'Bratman's Anti-Reduction', presented at the 1989 Meetings of the Central Division of the American Philosophical Association.