Bratman on Intending

Joseph Alberts

In his 1999 paper, "Toxin, Temptation, and the Stability of Intention" Michael Bratman argues that two prevailing descriptions of the nature of intentions, Resolution and Sophistication, are inadequate. By way of a series of examples Bratman attempts to demonstrate that Resolution theory is too restrictive, creating situations in which actors are not capable of altering their choices based on a new evaluative ranking of their options. Sophistication theory, according to Bratman, is unable to explain our intuitions in cases in which the actor should not change her intentions, despite a current evaluative ranking to the contrary. In order to address these weaknesses, Bratman introduces his No-Regret theory, which he believes resolves the problems faced by Resolution and Sophistication. In this paper I will reconstruct Bratman's arguments, demonstrating how he believes that the No-Regret theory improves upon the other two views. I will then demonstrate a crucial error in Bratman's reasoning, to wit: in his development of No-Regret theory he ceases to address intentions, according to most accepted definitions of the term. Finally I will introduce a means of classifying intentions—call it *Intention Indefeasibility*. Appeals to this definition of intention will jibe with the more traditional conceptions of intention, and will, by comparison, demonstrate the flaw in Bratman's argument.

Bratman's article investigates rational agency. That is, he aims to discover what would be rational for a planning agent to do in a given situation. Planning agents "settle in advance on prior, partial plans for future action, fill them in as time goes by, and execute them when the time comes."[1] In his attempt to formulate a theory to explain rational planning, Bratman recognizes an inherent tension. That is, while planning agents formulate

---

Joseph Alberts is a student at Rice University. His major fields of interest are Philosophy of Mind and Biology.

Joseph Alberts:

their plans in advance, they remain temporal agents, retaining control over what they do at the moment. Any theory must accommodate the agent's ability to alter her plans, but at the same time retain the importance of past plans in current decisions. Bratman stresses the importance of plan stability:

> By settling now what she will do later a planning agent puts herself in a position to plan appropriate preliminary steps and means…This will work only if her plans are to some extent stable…not constantly starting from scratch…A theory of instrumentally rational planning agency is in part a theory of intention and plan stability: a theory of when an instrumentally rational planning agent should or should not reconsider and abandon a prior intention.[2]

It is this aspect of the theory that most concerns this paper, i.e. intention stability. When should a rational agent alter her intentions? It seems reasonable to grant that one's intentions might change due to the receipt of previously unforeseen information. For example, one's intention to see a new movie might be changed by discovering that the movie theatre has closed. However, it is key for Bratman's theory that it is possible for intentions to change, even without the intervention of unforeseen events. This will be discussed at greater length later in the paper.

It is helpful here to discuss the two planning strategies that Bratman is criticizing: Sophistication and Resolution. A Sophisticated agent, "Adjusts her prior plans to insure that what she plans to do will be, at the time of action, favored by her then-present evaluative rankings."[3] The Sophisticated planner operates based on the conjunction of the linking principle with the standard view:

Linking Principle: States that if a rational agent has formed a plan (at time t_1) to perform action A at a later time (t2) given a series of events, and those events transpire, then the agent

should not suppose at time t1 that at t2 they will not A Standard View: Instrumental rationality depends on the agent's evaluative rankings at the time of the action.

The alternate strategy, Resolution, operates based on the linking principle alone, ignoring the standard view all together. Essentially Resolution states that, "If it was best in prospect to settle on a prior plan, and if there is no unanticipated information or change in basic values, then it is rational to follow through with the plan."[4]

Bratman is seeking to demonstrate the inability of these two prevailing planning strategies to properly handle instrumentally rational planning agency. In doing this, Bratman introduces a class of examples that he believes will exceed the bounds of any existing planning strategy:

> In all… cases there is a prior plan or policy settling on which is best in prospect. And in [these] cases the agent knows that when the occasion for action arrives her rankings of then-present options will argue against following through.[5]

As stated in the above quote, in each example an agent will form a plan of action, and without any unforeseen events taking place, find themselves in a position in which their current evaluative rankings favor some different action. Bratman holds that a proper strategy model should return results in keeping with our intuitions, that is, in situations in which a rational agent should alter their plans, the model should give that result. Conversely, when the plans should be carried out, despite current rankings, the model should give that result.

A class of cases in which a rational agent should change their plans is identified by Bratman as cases of "Autonomous Benefit", in which the reward for an action is separated from the action itself. Simply forming the intention to perform said action brings about the benefit. It is in this section that Bratman's "Toxin Case" is introduced. A quick sketch of the case:

> A billionaire… offers to give me a lot of money on Tuesday if I form an intention on Monday to drink a disgusting but non-lethal toxin on Wednesday. I would be more than willing to drink the toxin to get the money. However, to get the money I do not need to drink the toxin; I just need to intend on Monday to drink it.[6]

Questions abound. What would an instrumentally rational planning agent do? Can one form an intention on Monday to drink the toxin, knowing full well that on Wednesday one will have no reason to, and indeed will not do so? Bratman maintains that an ideally rational planning agent should, on Monday, form the intention to drink the toxin. For on Monday, the benefits from intending to drink the toxin outweigh the costs. However, that same rational agent, after the money has been exchanged, would then decide not to drink the toxin since there would no longer be any reason to. Bratman feels that a proper theory should return this result.

The Sophisticated planner would be unable to form the intention in the first place. Because they subscribe to both the linking principle and the standard view, and since in this example there can be no unforeseen events, the agent would be incapable of intending to do something that would contradict her predictable evaluative ranking at the time of action. Hence, the Sophisticated planner, being unable to formulate the intention to drink the toxin would not obtain the autonomous benefit.

The planner using the Resolution method will, on Monday, be able to form the intention to drink the toxin, but Resolution requires that, on Wednesday the agent go through with their plan and drink the toxin. This is true regardless of the fact that there is now no benefit brought about by drinking the toxin. Bratman holds that, while this is still and improvement over Sophistication, there remains significant problems associated with Resolution's handling of the Toxin case:

> [Resolution], in seeking a strong role for planning in achieving the benefits of coordination over time and across agents, seems not to do justice to the basic fact that as agents we are temporally and causally located.[7]

So, in cases of Autonomous Benefits, the Resolution method allows the benefit to be achieved but does not seem to comport with our understanding of the temporal element of agency; i.e. that an agent should be capable of altering plans which are in no way beneficial (following through with the intention to drink the toxin) regardless of past preference.

The second class of cases introduced by Bratman is known as "Temporary Reversal in Rankings", or Temptation examples. In these cases the agent will form a plan at time t1, then without any unforeseen information, at t2 the evaluative rankings of the agent will change. However, the change in rankings is only temporary and the former rankings will return at a time in the future t3. Ann's case is representative of this class:

> Ann enjoys a good read after dinner but also loves fine beer at dinner. However, she knows that if she has more than one beer at dinner she cannot concentrate on her book after dinner. Prior to dinner Ann prefers an evening of one beer plus a good book to an evening with more than one beer but no book… [However] each evening at dinner, having drunk her first [beer]…she prefers a second beer to her after dinner read… As she knows all along this change in ranking will be short-lived: after dinner she will return to her preference for a good read.[8]

How would our two conventional strategies handle Ann's case? Sophistication suggests that, because at dinner Ann always prefers a second beer, she could never have settled on the one-beer plan in the first place (remember, sophistication re-

quires the agent to adhere to both the linking principle and the standard view). Resolution however, allows that Ann could very well form the intention to drink only one beer at dinner. Ann would then be required, (due to the linking principle) of carrying out her intention in the event of no unforeseen information. In this case Ann is aware of all relevant information at the time of her plan formation, and as such she should drink only one beer in accordance with her original plan. Bratman holds that this is the desired outcome for a rational agent, to be able to form plans, aware of the fact that at some point in the future the ranking of the options would switch, and still carry out one's original plan.

The contrasting results in the Toxin case and Ann's case trouble Bratman. In each of the situations, it seems that Resolution is better able to explain the situation than Sophistication, enabling the receiving of the autonomous reward and requiring Ann to drink only one beer. However, Bratman holds that the Toxin case demonstrates the true weakness of Resolution, its failure to account for the temporal status of the agent. It seems that a truly rational agent, having received the reward, would not drink the toxin; there would simply be no benefit to be had. Why is this the case? Why in the Toxin case do we expect a rational agent not to follow through with her intentions, but in Ann's case we do? Resolution and Sophistication are incapable of returning results in keeping with our intuitions as to the behavior of a rational agent. A new rational planning strategy is required.

It is in this pursuit that Bratman introduces the No-Regret planning condition that he feels, "avoid[s] both extremes [Resolution and Sophistication]."[9] The No-Regret condition deals with the agent's decision at the time of the action, on Wednesday for the Toxin case and during dinner for Ann. Bratman carefully dissects the argument for action based on future regret:

> One should act in accord with prior intention in the event that:
> Upon sticking with your prior intention, you will

be glad you did.
Upon failing to stick with your prior intention,
you will wish you had.
So, other things being equal,
Though you now prefer to abandon your prior intention, you should nevertheless stick with it.[10]

In this way the subject retains the ability to act as a temporal causal agent. At the time of action the agent anticipates the consequences of her following through with previously held plans. If she decides that at an "appropriate later time" following through with her intentions would lead to regret, then she abandons her plan.[11] On the other hand, if she decides that acting in accord with her previous plan would not bring about regret, she acts consistent with her plan. How does this divide up the cases with which we were dealing? In the Toxin case the No-Regret condition stipulates that we deal with the agent's decision at the time of action, on Wednesday. Just before drinking the toxin the subject must consider her future regret (i.e. Will she regret not drinking the toxin?) It seems that the answer is obviously no. In fact it is much more likely that a rational agent, faced with the choice of drinking a revolting concoction with no possibility of benefit, would certainly regret drinking said mixture. This seems the proper response, allowing the agent to make a decision based on their current temporal state.

This leaves us with Ann's case. Evaluating Ann's situation based on the No-Regret principle requires us to evaluate her decisions during dinner, at a point when she genuinely prefers a second beer to reading her book. Here again we are given the desired response. Ann, understanding that her preference of a second beer over "a good read" is temporary, would, at some future point, regret having consumed the second beer. She will regret having abandoned her single-beer policy.

This seems a victory for Bratman and his No-Regret condition. In both situations of Autonomous Benefits and Temptation we are given the response understood to be that of a rational

agent. In this way we avoid the difficulties of both Sophistication and Resolution and in the process, "arrive at a view of instrumentally rational planning agency that does justice both to the fact that we are planners and to the fact that we are temporarily and causally located agents."[12]

Bratman's investigation into instrumentally rational planning agency is well thought out in its handling of rational planning. It seems that Bratman's No-Regret model does provide results that correspond to our intuitions. Neither the Resolution nor Sophistication models were able to give the correct responses in both of the above situations. In this way Bratman's No-Regret model is a success. However, I hold that his criticism of both theories (which provided the impetus for positing the No-Regret theory) rested on an interpretation of intention that is in conflict with all prevailing definitions offered by psychologists and philosophers engaged in the study of intention. Here I will attempt to give a brief definition of intentions as they are commonly thought of. As the definition becomes more explicit, it will become ever clearer that Bratman's concept of intention is at odds with the field's view in general.

An intention is typically thought of as a pro-attitude (its content is the desired alteration of the external world) and consists of a desire and a belief. While desires themselves are also pro-attitudes, there are several distinctions. Because intentions are the conjunction of a desire and a belief, they unlike desires, must be consistent with our set of beliefs.[13] Agent S could safely say that she desires to jump over the Empire State building in a single bound. However, should S inform her friends that she *intends* to perform the same task, she will be thought of as very odd indeed. Louis J. Moses states that in order for S to intend to A then S must believe at least one of the following:

> I can-A
> There is some chance that I will A
> I probably will A
> I will A[14]

Moses feels that option (a) is likely too weak, and option (d) is likely too strong. He concludes that, "intending to A requires a relatively strong belief that one will A."[15] In addition to consistency with beliefs, an intention carries a greater commitment to action than a desire. When an agent is said to intend, they are said to have "actually decid[ed] to perform the action in question," this is not the case with desire.[16] It seems that intentions commit us to action in a way that desires do not. While we have no strict consensus on a definition of intention (and a growing segment of philosophers believe such a definition will never exist), there remains some general consensus; that is, an intention must be consistent with one's beliefs and it commits one (in some degree) to action. As will be demonstrated in the remainder of this paper, Bratman's use of "intention" in no way meets the above conditions.

In his cases of Autonomous Benefit, Bratman describes a case in which an intention must be formed. However, when the time comes to perform the action, that was the content of the original intention, the agent may then decide whether to follow through. This seems correct, at least it seems in keeping with our experience of everyday rational planning. Bratman uses the example of preparing for a job interview with Jones, but upon arriving, Smith has taken her place. In this situation we must, as temporally located agents, be understood to be capable of reconsidering and perhaps changing our plans. However, in the paper under examination, Bratman is not interested in these everyday vicissitudes, he is interested in cases in which, "one's circumstances are, in all relevant aspects, those for which one has specifically planned."[17] Here Bratman's theory as to instrumentally rational planning faces a problem. In the Toxin case, the agent must be aware of "all relevant aspects", and it seems that this must include the agent's foreseeable disposition not to drink on Wednesday. The agent is aware that as of Wednesday the money will have already been distributed, or not, and will understand that there will then be no incentive to actually drink the toxin. This is certainly a relevant aspect of the case, and the agent (being ideally rational) must know that on Wednesday she

will not drink the toxin. This puts the agent in the precarious position of intending to do something he knows that he will not do. This is absolutely in conflict with the accepted definitions of intention given above. Bratman is in violation of the condition of belief consistency. One cannot intend to do what one does not believe one will do. In this case the agent, being aware of all relevant aspects, does not believe that he will drink the toxin, therefore, he cannot have intended to drink. What the above definitions imply, and what Bratman fails to acknowledge, is that intention, in the way it is understood in folk psychology and by experts in the field, takes on an indefeasible quality.

Indefeasibility is a concept most commonly associated with Epistemology. Briefly, in any theory of knowledge subscribing to an indefeasible standard, one is said to know a proposition P if and only if there is no undefeated defeater of your justification for knowing P. To clarify, in order for subject S to know P, there can exist no evidence E that would undermine S's justification for knowing P. Such evidence is referred to as a "defeater." However, S can still be said to know P, if S can present evidence D which undermines E. In this case evidence D is referred to as a "defeater-defeater". If S is to know P there can be no piece of evidence for which there is no "defeater-defeater", such irrefutable evidence is known as an "Undefeated Defeater."[18] This typically epistemological notion can also be understood as applying to intention. Just as knowledge is not granted in the event of an undefeated defeater of justification, intention status may be withheld due to the existence of an undefeated defeater of the agent's belief. In the case of intentions, a defeater can be understood as a piece of information that, if known, would defeat the belief that one could or would perform the action. For example, S might intend to join the army. However, person X might inform S that in order to join the Armed Forces, one must be at least 5'5" tall, which S is not. Once understood by S, this piece of evidence could be said to successfully defeats S's intention to join the Army. This is not necessarily an undefeated defeater however, if S knows that X is a liar (or simply as poorly informed as the author of this paper) and the actual height re-

quirement is 5'0", which S easily clears, then S could still be said to intend to joint the army.[19]

We can now formulate a definition of intention as characterized by indefeasibility; we will call it *Intention-Indefeasibility:*

> Subject S, at time t1, can be said to intend to A, at time t2, if there exist no undefeated defeaters of S's intention to A

This can be understood as strong *Intention-Indefeasibility*, and it does not seem to correspond to our intuitions as to intentions. For example, in the above case, before it was explained to S that he was too short to join the army, one would say that S intended to join the army. It seems that S must be **aware** of the defeater of his intention in order to be said not to intend. This leads to the positing of weak *Intention-Indefeasibility:*

> Subject S, at time t1, can be said to intend to A, at time t2, if there exist no undefeated defeaters of S's intention to A of which S is aware.

This seems to correspond nicely with our understanding of intentions. In fact, it serves a similar function to the first of our original conditions, that S's intention be consistent with his beliefs.

When examining intentions as used by Bratman through the lens of *Intention-Indefeasibility* it becomes clear that Bratman has a substantial problem. In the case of the toxin, subject S, at time t1 is being asked to form an intention to drink a toxin at t3. However, as subject S is aware of all relevant aspects of the case, S understands that at t3 she will already have acquired the money (or not) at t2, and as such will have no reason to drink the toxin. Being a rational agent, S will understand that at t3 S will not drink the toxin. S's knowledge that she will not drink is a clear "undefeated-defeater" of her ability to form an intention to drink. As S is clearly aware of her belief that she will not drink the toxin, weak *Intention-Indefeasibility* tells us that at time t1, S

cannot be said to intend to drink the toxin. In Bratman's example, the billionaire possesses an intention detector and will discover this lack of intention; ergo, S will not receive the benefit. Only by committing in a binding way (indicating that S truly believes that at t3 she will drink the toxin) will S receive the benefit. If she were aware of all relevant information, she would have to be committed to following through with her plan. Any deviation from the plan could only be explained by unforeseen events, which are explicitly ruled out of the equation. Bratman has a problem.

On the other hand, Bratman's examination of the Temptation case seems to pass the test of weak *Intention-Indefeasibility*. There does not exist an undefeated-defeater, that Ann is aware of, which will preclude her from forming her intention. Ann's temporary ranking of a second beer over a night with a book might be considered a defeater of her intention. However, one might posit that Ann's understanding of her long-term evaluative ranking (reading over beer) might fill the role of defeater-defeater. Weak *Intention-Indefeasibility* concludes that Ann can intend to have only one beer and barring unforeseen information (here an impossibility) will follow through with her original plan.

Bratman's investigation into instrumentally rational planning does elucidate the likely actions of a perfectly rational planning agent. In both the Autonomous Benefit and Temptation cases, we would like to say that a rational agent would decline to drink the toxin or the second beer. However, closer examination has given us reason to doubt the applicability of Bratman's No-Regret theory to his "Toxin Case". Simply put, Bratman just isn't talking about intentions as we know them.

Notes

1. Michael Bratman, *Faces of Intention* (Cambridge: Cambridge University Press, 1999), 58-59.
2. Ibid., 61.

3. Ibid., 69.
4. Ibid., 70.
5. Ibid., 76.
6. Ibid., 62.
7. Ibid., 73.
8. Ibid., 74.
9. Ibid., 89.
10. Ibid., 79.
11. Ibid., 80.
12. Ibid., 89.
13. Louis J. Moses "Complex Intentional Concepts and Young Children," in *Intention and Intentionality: Foundations of Social Cognition,* ed. Bertram F. Malle, Louis J. Moses, and Dare A. Baldwin (Cambridge: MIT Press, 2001), 72.
14. Ibid., 73.
15. Ibid.
16. Ibid., 46.
17. Ibid., 62.
18. Gilbert Harman, "Selections from *Thought*," in *Epistemology: An Anthology*, ed. Ernest Sosa and Jaegwon Kim (Malden: Blackwell Publishers, 2001), 72.
19. Information that casts doubt upon the intention, but does not eliminate the possibility is less clear. For example, if S were to learn that the only way to join the army was by way of lottery, and in that drawing he would have a 50% chance of being accepted, would we still say he has an intention to join the army? If so, would he retain his intention if he stood a 10% chance of being admitted? It is possible that in such a situation S could simply be said to intend to attempt to join the army. However, this seems to imply that S would succeed in his intention regardless of the outcome of the drawing, a sentiment that would probably not be echoed by S upon failing to have his number drawn. These are problematic cases, and as such are worthy of a lengthy inquiry, that we will be unable to examine here. However, in dealing with Bratman's examples we can restrict ourselves to cases in which the defeater precludes the possibility of performing the action.

Works Cited

Bratman, Michael. *Faces of Intention*. Cambridge: Cambridge University Press, 1999.

Harman, Gilbert. "Selections from *Thought*." In *Epistemology: An Anthology*, edited by Ernest Sosa and Jaegwon Kim. Malden: Blackwell Publishers, 2000.

Moses, Louis J. "Complex Intentional Concepts and Young Children." In *Intention and Intentionality: Foundations of Social Cognition*, edited by Bertram F. Malle, Louis J. Moses, and Dare A. Baldwin. Cambridge: MIT Press, 2001.