universität
wien

MASTERARBEIT / MASTER'S THESIS

Titel der Masterarbeit / Title of the Master's Thesis

Agency Adrift

A Puzzle about Time and the Stability of Intention in Michael E. Bratman's Philosophy of Action

verfasst von / submitted by

Dott. Damiano Ranzenigo

angestrebter akademischer Grad / in partial fulfilment of the requirements for the degree of

Master of Arts (MA)

Wien, 2018 / Vienna 2018

Studienkennzahl lt. Studienblatt / degree programme code as it appears on the student record sheet:	A 066 941
Studienrichtung lt. Studienblatt / degree programme as it appears on the student record sheet:	Masterstudium Philosophie
Betreut von / Supervisor:	Univ.-Prof. Dr. Hans Bernhard Schmid

CONTENTS

ACKNOWLEDGEMENTS

I would like to thank Prof. Hans Bernhard Schmid for his supportive supervision and punctual remarks on my ideas and Prof. Herlinde Pauer-Studer for the supportive feedback. Special thanks to Prof. Michael E. Bratman and also to Prof. Elijha Millgram for their illuminating suggestions for relevant literature on the topic. In addition, Prof. Wolfgang Huemer has given me valuable moral support.

I have profited much from the feedback of my student colleagues at the University of Vienna. In particular, I would like to thank Paul Tucek, Franz Urban Altner and Carlos Nunez, who have offered their constructive criticism and pushed me to develop and refine my ideas. Chiara Dankl has also done an incredible job in revising my work, and helped me improve my style with wit and precision.

I am mostly indebted to the endless discussions about planning agency with my flat mate in Vienna, Jani Rauschenberger, whose insightful patience has been of great philosophical and psychological help. My friend Jesse De Pagter commented the first draft of the thesis and never failed to encourage me, along with Rumi Trendafilova, Paul Seiler, Ana Useros, Mike Eugster, and many others in Vienna. I am also thankful for the enriching discussions with Sofia Ascione, Sara Beretta, Francesca Maffezzoni and Silvia Renica between Brescia and Bologna, and am grateful to all my Italian friends.

Finally, I would like to thank my brother Matteo, for showing me the philosophical equilibrium I sometimes lack, my sister Annalisa, for giving me a valuable insight during our reunion in Italy, and my parents, for sustaining my studies in Vienna and allowing me to cultivate my interests with generosity.

Vienna, 25.10.2018

INTRODUCTION

Michael E. Bratman has written extensively on issues related to our agency over time. How can we control our conduct over time? How can we coordinate earlier actions with later actions? How can we achieve our goals in the future? The key to understanding these issues can be found in the meaning of ‘intention’. Intentions are for Bratman particular mental states, with the function of bringing us to act. One of their most interesting aspects is their role in “[…] extending the influence of Reason on our lives” (Bratman, 1992, p. 2) and controlling our conduct over time. In order for this role to not be undermined by constant reconsideration, intentions, once formed, maintain a certain stability. We do not in fact abandon an intention by chance: we need good reasons to do so. Otherwise, we stick with an intention until we fulfill it, hence its stability.

An interesting question about stability is wondering whether the initial rational status of an intention is timeless and remains unaltered, or if it changes with time, before we form a new intention. I defend this latter idea and argue that the rational status of a newly formed intention is not timeless.

Enquiring into the temporal dimension of our intentions’ rationality is not an easy task, for we are talking about agents who are regularly presented with new information, who are highly reflective and form new intentions every time they adjust their conduct to serve their goals. This is why I focus on intentions covering large intervals of time, such as general policies concerned with recurring circumstances over extended periods, and attempt to isolate them from a context of constant cognitive changes.

I will work out this idea in the guise of a puzzle about stability with two focal points: on the one hand, I assume that a planning agent can't have complete access to the stability of a newly formed intention, because of a kind of ‘prospective opacity’; on the other hand, once an intention is formed, in virtue of the structure of diachronic agency, the intention will tend to drift away from the original rational status at its initial formation. In other words, it is precisely by sticking with an intention over time, and by being unable to ‘see clearly’ the stability of that intention at its formation, that the intention slowly drifts away from its initial rational status. This is what I will call Opacity-Drift Puzzle.

The idea may strike one as implausible, but it shouldn't, for it is not aimed at posing an objection to Bratman's philosophy. Indeed, the Opacity-Drift Puzzle is more to be taken as an oddity or a peculiarity of a model of diachronic agency, such as Bratman's, rather than a threatening theoretical challenge. The puzzle makes explicit a certain reading of Bratman's philosophy: an interpretation of agency over time through certain theoretical lenses, that provide us with an intuition about the kind of agents that we are.

I am aware of the many complexities of Bratman's theory of intention and its vastness and, therefore, I focus on a more manageable and controllable version, coinciding roughly with Bratman's early philosophy. I then proceed to gradually expanding the puzzle to explore if it also applies to other developments of his theory.

In the first chapter, I present the puzzle about stability, and its significance for models of diachronic agency such as Bratman's, whereas in the second chapter, dig deeper into the many issues, which are of direct concern for the puzzle or relevantly related to it. In particular, I am interested in better defining the focal points of the puzzle by juxtaposing them to other concepts of the Bratmanian theory. Here it is useful to consult some of Bratman’ critics, and examine how the puzzle can clarify certain disputes or be, in turn, better illustrated through the contributions of other authors.

Finally, in the third chapter, I respond to three main objections to my proposal, leaving a space of indecision and freedom for the reader to further deliberate on the issues raised. The hardest task is, however, to lead the reader to a favorable ‘hermeneutic spot’, and show the genuineness of the puzzle and the intuition about agency that it gives rise to.

CHAPTER ONE

The Puzzle

1.1 Bratman’s Planning Theory of Intention

1.1.1 Future-Directed Intentions, Planning Agency and Stability

In Bratman's early philosophy of action,[1] agents are temporally extended creatures, who coordinate over time by means of their planning activity. *Future-directed intentions* are the building blocks of this planning activity: they help agents control their conduct over time, rule out inconsistent alternatives, solve means-end coherence problems, settle deliberation in advance and eventually lead to action – although not necessarily.[2]

If I intend to eat ice cream later, I form a future-directed intention to eat ice cream later. My planning activity helps me settle relevant means to achieve my end of eating ice cream and rules out inconsistent alternatives (going to have a granita at the same time) or ineffective means to go to the ice cream parlor (such as with a personal jet, that is not even mine).

The temporally extended structure of agency requires intentions to maintain a certain cross-temporal *stability*: intentions are stable in the sense that we are disposed to retain them without reconsideration. Hence, stability can be thought of as the firmness of an intention, as its inertia or resistance to reconsideration. The stability of future-directed intentions enables my intentional activity to extend over time, from its initial formation to its execution at a later time. In Bratman's own words:

> If I now intend to go to Tanner [library] later today, I normally will not continue to deliberate about whether to go. I will normally see (or, anyway, be disposed to see) the question of whether to go as settled and

[1] In this section, I will mostly refer to Bratman's book *Intention, Plans, and Practical Reason* (1987). As far as I know, Bratman has not substantially changed his early view on agency, but he has completed and enriched his theory later on, in ways I will consider in 1.3.1.

[2] Indeed, future-directed intentions are defeasible: you can intend now to do something later and realize later that you can’t actually carry out the intended action, thence stop intending.

> continue so to intend until the time of action. My intention resists reconsideration: it has a characteristic *stability* or *inertia*. (Bratman, 1987, p. 16)

Going back to our example of the ice cream. I go out to eat an ice cream, because that is what I desire and I don’t normally give up on my intention at any time before the execution, given that I am not presented with challenges to it. Other things equal, I go to the ice cream parlor, choose my favorite flavor, pay, and eat the ice cream without reconsideration.

This doesn't mean that we do not reconsider our intentions: we sometimes do. In these cases, in order for reconsideration to be rational, we need to engage in a reasoning process of weighing relevant desires, given our beliefs of the case, and deliberate about our conduct.

At the ice cream parlor, I see that granitas are cheaper and look particularly tasty. I thence form a new desire to eat granita instead of ice cream, and this desire outweighs that of eating ice cream. I reconsider my intention to eat ice cream and I eat granita instead. I have engaged in a process of reflective *rational deliberation* involving the weighing of relevant desires with the aim of satisfying them.

As a reminder, *practical rationality* is here understood as broadly instrumental, i.e. as serving the satisfaction of our desires and the achievement of our goals. Our rational intentions respond to the norms of means-end coherence, consistency of intentions and beliefs, and stability.

1.1.2 Limited Resources, Habits of Reconsideration and Two-Tier Approach

Even if we can reconsider our intentions, an absence of reconsideration is the default case. We resist reconsideration by default for at least two very important reasons: the first is that reconsideration comes with costs in terms of time and cognitive engagement, the second is that we are agents with temporal and cognitive limited resources, who can't constantly engage in reconsideration.

Thus, we often rely on mechanisms to resist reconsideration nonreflectively, mainly because we dispose of *limited resources*: it would be foolish to constantly engage in deliberative reconsideration, for we would end up constantly reconsidering and doing nothing at all – eating neither ice cream, nor granitas. Bratman introduces the role of these mechanisms as follows:

> An agent's habits and dispositions concerning the reconsideration or nonreconsideration of a prior intention or plan determine the stability of that intention or plan. Such dispositions may vary in a wide variety of ways. My disposition to refrain from reconsidering my prior plan may be rather minimal: I might be inclined to reconsider it given only a slight divergence between the way I find the world when I come to act and the way I expected it to be when I first settled on my plan. Or my disposition may involve substantial rigidity, as when I would only reconsider it in the face of some extreme divergence from my expectations – an earthquake for example. (ibid, p. 65)

Thus, stability sometimes requires that we retain a previous intention in such a way that resistance to reconsideration becomes a nonreflective process.

Consider the case in which I desire to taste all different flavors of ice cream the parlor has to offer this summer. I also know that, if I often give in to the temptation of eating granita instead, summer will be over and flavors will change with the season and my desire to try all summer flavors won’t be satisfied. I thence form a stable intention to go for ice cream and often resist the temptation to eat granita instead. I do not weigh the reasons in favor of ice cream against those in favor of granita each time: there is a nonreflective process determining my degree of resistance to reconsidering my intention to try all different flavors of ice cream before summer ends.

These processes determining reconsideration in nonreflective cases are called *habits of (non)reconsideration* and can be more or less reasonable. Their reasonability depends on how well they help our singular intentions to be effective in satisfying the desires they are related to. A desire to taste all different flavors of ice cream is related to an intention to taste all different flavors of ice cream. This intention is accompanied by a habit of reconsideration, which stops reconsideration enough for the desire to be satisfied.

To assess the rationality of these habits, and hence understand when it is rational to reconsider in a nonreflective case, Bratman adopts a *two-tier consequentialist approach*:

> On this account, we first ask, in a consequentialist spirit, what underlying habits of (non)reconsideration are reasonable. Then we ask whether, in the particular case, the nonreflective (non)reconsideration manifests reasonable habits. It is rational of the agent, in that particular case, to reconsider (or not reconsider) just in case she thereby manifests reasonable habits of (non)reconsideration. (ibid, p. 68)

Was the general habit determining reconsideration for the intention to eat all different flavors of ice cream reasonable? The habit is reasonable if it allows me to satisfy my desire to eat all different flavors of ice cream. Then we can ask if it is rational today that I intend to eat ice cream, given that I could opt for granita instead. If, on balance, my habit of reconsideration allows for such reconsideration and hence that I eat granita this time, we should assess its rationality in terms of how well I can still satisfy the desire to eat all different flavors of ice cream by summer’s end. If I have failed to eat all different flavors, that habit of reconsideration will not have been completely reasonable for me to have – relative to my intention of eating all flavors of ice cream.

1.1.3 Policies, Policy-Based Intentions and Blocking Habits

Given the focus on the temporally extended dimension of agency, general intentions concerning "potentially recurring circumstances in the agent's life" assume a central role in Bratman's account of agency. Bratman calls these general intentions *personal policies* (ibid, p. 87).

Examples of these general policies are buckling up one's seat belt when driving a car, refusing second desserts when on a diet, checking brakes every 6000 miles for safety, reading German prose every night before going to bed, and so on.[3] Often we follow these policies without reopening the question of whether it is rational of us to buckle up our seatbelt or to read German prose. Otherwise the general point of having policies rather than single intentions each time would be gone.

Many of our policies require strategies for nonreflective reconsideration, such as the non-deliberative intentions we have recently analyzed. Like retained intentions, personal policies come with respective habits of reconsideration for non-deliberative cases. However, even if a policy of this kind incurs a problem concerning its application, we don't need to go as far as reconsidering the whole policy. Policies are instantiated through *policy-based intentions.* These are neither deliberative, nor nondeliberative intentions: they are not the outcome of a nonreflective process, nor are they simply retained over time. The general policy controls their functioning over time and if we encounter a problem with a specific policy-based intention, appropriate *blocking habits* might settle the issue without reopening reconsideration about the whole policy.

Consider Bratman's example:

> In rushing my wife to the hospital it may be foolish of me to worry about buckling up. But this does not mean that I should abandon my general policy of buckling up. Rather, I should just block its application to my particular case. (ibid, p. 91)

Blocking habits are analogous to habits of reconsideration and their rationality can also be assessed by means of the two-tier approach: we first see if the overall habit was rational and then we assess its rationality in the specific ‘blocking-case’.

[3] See Bratman, 1987, pp. 87-88.

1.2 A Puzzle About Time and the Stability of Policies

1.2.1 Standing the Test of Time

So far, we have seen that intentions, plans and policies come with a certain stability, measured in terms of resistance to reconsideration and resistance to blocking of application. The rationality of reconsideration (and blocking) is assessed by looking at relevant habits of reconsideration and blocking habits. In general, the stability of an intention is a long-term feature of our policies and plans, in the sense that it is not constantly open for adjustments. This general picture will be termed "rough early model", REM, for short.[4]

REM leaves a question unanswered: Does the rational status of a policy stay unaltered over time, other things equal?

Any newly formed policy comes with a certain stability. According to REM, this stability is determined by habits of reconsideration that indicate when to reconsider and the rationality of which can be assessed by a two-tier approach.

Following problem arises here: these habits may well be timeless mechanisms regulating our conduct over time, in the guise of specific programs controlling the activities of automatons or robots. But it seems that human beings do not know in advance to what degree of stability a newly formed policy, endowed with its nonreflective habit, will lead. We do not ordinarily implement our brains with policies we already know are stable: first we need to form intentions, and then see if we can maintain them. Their stability is a feature we can mostly appreciate as time passes by.

What does this imply for our agency over time? It means that a policy (but also an intention or a plan) is something that *proves its stability* over time rather than from the start. Of course, a policy constantly subject to reconsideration may fail to prove its

[4] I call this model 'rough' because it is a simplification of Bratman's own theory exposed in Bratman, 1987. As a matter of fact, I am trying to depict clearly the features of his theory that are relevant for my later discussion.

stability in this sense, but it would also fail to grant the cross-temporal coordinating role that is necessary for temporally extended agents and thus not be a rational policy at all. Yet this doesn't mean that we can clearly discriminate between stable and less stable policies from the beginning of their formation: only time can truly tell how stable they are.

Building on what has been said so far, it seems that the stability of a policy isn't only an outcome of the firmness of our intention,[5] but also depends on being effectively carried out over a significant period of time. This has implications for our conception of agency: policies control our conduct over time, but we don't completely know in advance which policy will effectively fulfill this role. As a consequence, we try to maintain our policies and see if they stand the *test of time*: after a significant amount of time, we might realize that we effectively achieve our policies' aim. We finally know that the policy is stable: it has passed the test of time.

But how do we know? How can we know? Does this mean that the policy has somehow become reliable, in a sense which we couldn't grasp at the beginning of its formation? It seems that something is missing from REM, but before going into the details of how Bratman could help us shed light on this issue, let me focus briefly on an example which should clarify the problem I have raised now.

1.2.2 The Case of Policies Covering Large Temporal Intervals

Consider the following example of the 'standard case', in which we assume that time doesn't influence the rational status of our policies or related habits of reconsideration/blocking:

> *Vienna Marathon Standard*: March is ending and I think that the time has come for me to run my first Vienna marathon this incoming June. Unfortunately, I am not trained enough. But my desire to run the

[5] As far as I know, Bratman never claimed that the stability of policies is reduced to their firmness. Nonetheless, I want to stress the importance of thinking that subjective firmness at the initial formation of a policy is not all that determines its stability.

> marathon is very strong and I am ready to start a hard training during the following two months to reach my goal. All things considered, it would be optimal to go running thrice a week in Prater, and quite strictly not less than thrice a week – on pain of not being able to complete the marathon. Hence, I form the policy to go running in Prater thrice a week, on Monday, Wednesday and Friday for the following two months.[6]

The policy I have formed must have a high degree of stability for me to carry out my plan of running the marathon. If I were to open reconsideration about the policy or to block its application, even seldom, there is a concrete risk of failing to accomplish my goal.

Suppose now that I am confronted with following challenge:

> *Challenge in the Standard case*: on the first Monday of May – after I have been training for a whole month, I am presented with the possibility of reconsidering/blocking the policy-based intention to go running today, and go to the library instead. I am a student and have a minor exam tomorrow (Tuesday), and it would be great (perhaps decisive) for me to read the exam material once more. Still, my reconsideration/blocking habits are rigid enough to rule out the option of going to the library and I stick with my policy-based intention to train for the marathon. The fatidic Tuesday comes, and I fail the exam.

What happens in above mentioned situation needn't be a case of failure of rationality. If I value running the marathon highly, and I could attempt the minor exam again in September, without much loss for my career as a student, why bother? We could see failing the exam as an unavoidable downside for the success of a more valuable goal, i.e. running the Vienna marathon. We could even go as far as claiming that my habit connected to the policy of running thrice a week was completely rational in resisting reconsideration and blocking.

Now, let's consider a case where a considerable amount of time has passed from the first formation of the policy.

[6] The inspiration for this example comes from a similar one made by Prof. Herlinde Pauer-Studer (University of Vienna, 2016).

> *Vienna Marathon Reiterated*: Independently of how I performed at my first Vienna Marathon, I now settle to run at least as successfully next year, by sticking with the same policy, that has been overall sufficiently stable and reasonable, even considering my other commitments – especially as a student. Hence, I stick with the policy to go running in Prater thrice a week, on Monday, Wednesday and Friday, until next year's marathon.

Suppose that my desire to run the marathon stays unaltered for the whole incoming year, and that nothing in my evaluative system and rankings changes in any relevant way. Running the Vienna marathon is something I value highly, along with accomplishing my obligations as a student. The failure of minor exams has no dramatic impact on my academic success.

Now, consider that the challenge of going to the library instead of running the marathon occurs after months of training:

> *Challenge in the Reiterated case*: The first Monday of October, after training regularly for the whole summer, I am faced with the possibility to reconsider/block my policy-based intention to go running today, and go to the library instead. I have a minor exam tomorrow (Tuesday), and it would be great (perhaps decisive) for me to read the exam material once more. Just like in the Standard case, my habits of reconsideration/blocking rule out the option of going to the library. On Tuesday, I fail the exam.

Can we assess this situation as we did in the Standard case? If we consider that nothing relevant has changed in my evaluative system, in my desires and beliefs, then the assessment of rationality should be the same as before. This would mean that time has played no role, after all. However, something has changed or, better, evolved from our former situation. Keeping up with training after the first marathon was a consequence of my policy having proven its stability/stood the test of time. In a certain sense then, I should rely on my policy in October in a way that was irrational in May: I am no longer exposed to the danger of giving up my training because of a lack of commitment – in terms of stability.

What I didn't know in May but should know now in October, is that I am a person who runs in Prater on a regular basis. Hence, a sense of regretting the failed exam in October would be more justified than the regretting it in May. In May, I had no estimation of my diligence in training, whereas in October I should know better.

My intuition is that failing the minor exam in the Reiterated case challenges my rationality in a stronger sense than in the Standard case. As I will argue, this means that the passing of time alone plays a relevant role for the rationality of our policies.

1.2.3 A More General Depiction of the Puzzle

Many of our personal policies can show that stability is something we discover over time and that the achieved reliability of stable policies plays a role in how we assess our rationality. Similar to the Vienna marathon example, we can consider the case of regularly refusing a second glass of wine after dinner in order not to jeopardize one's productivity,[7] or the case of regularly exercising at the gym to keep healthy,[8] or even reading some German prose every evening before going to bed. The nonreflective mechanisms of reconsideration/blocking of these policies are all regulated by appropriate habits of reconsideration/blocking. These habits usually determine responses to challenges to the policy – either to reconsider/block or not to reconsider/not to block.

As the Vienna Marathon Reiterated example has shown, the way in which habits determine such responses toward the initial formation of the policy, could incur a different assessment of rationality, if it were to occur in the same guise after a considerable amount of time.

Even with a stable policy not to drink a second glass of wine after dinner, there are often good reasons for drinking a second glass of wine. These reasons needn't involve

[7] For discussions on this example see esp. Bratman, 1999, Ch. 3,4, pp. 35-90 and Bratman, 2007, Ch. 12, pp. 257-282.

[8] See Bratman, 2007, Ch.3, pp. 64-65.

a case of momentary temptation in terms of a particularly strong conflicting desire. A long forgotten friend might offer one, or one might think that, independently of the strength of the desire to drink another glass, it would be great to take a break from working after dinner today. Keeping in mind that the situation we are describing does not involve deliberation or reflective weighing of desires: our reconsideration/blocking is regulated by our nonreflective habits of reconsideration/blocking.

So far so good: if the habit is overly rigid, one will irrationally stick with the prior intention even if it is rational to allow for an exception. Our model REM would be at peace with this, but that is not the point of interest right now. My attention is focused on when this policy of avoiding a second glass of wine after dinner becomes a well-established policy. Will we say that the habit accompanying the policy at the beginning of its formation can stand the same rational scrutiny, if it doesn't change in the face of the policy having proven its stability/stood the test of time?

The answer I am inclined to give is no. Something has changed. Most importantly, something has changed and one couldn’t know in advance. There is some kind of ‘prospective opacity’ in forming stable policies. Again, we are not implementing our brains with programs we already know to be stable: we need to bring ourselves to act as we intend, and only after that we can truly assess the stability of our intentions. We cannot see clearly the present and future stability of our intentions at the time of their initial formation.

I can more generally define the problem at hand along the following points:

a) *Stability*: Stability is a long-term feature of our policies, and its rationality depends on nonreflective mechanisms of reconsideration, called habits of reconsideration/blocking;
b) *Limited Resources*: Habits of reconsideration/blocking help agents with limited resources to carry out their intentions without constant reconsideration/blocking. The stability they determine is not constantly up for reflective adjustments;

c) *Prospective Opacity*:[9] Human agents do not completely know in advance how stable their policies are at the time of their initial formation. This is in part due to the fact that stability can't be completely assessed in terms of the subjective firmness of intention, but also in relation to effectively carrying out intentions over time;

d) *Test of Time*: It follows from Stability and Prospective Opacity that only the passing of time can truly tell if a policy has been stable all along. If I have been consistently adhering to it over a considerable amount of time, the policy has stood the test of time and has hence proven its stability.

e) *Ageing of Policies*: The very fact that a policy has stood the Test of Time has consequences for the rationality of reconsidering/blocking: what is rational of us to reconsider/block now, i.e. after the policy has stood the Test of Time, differs from what was rational of us to reconsider/block at the initial formation of the policy. This means that the policy has 'aged', even if other things have remained equal.

f) *Drift from Rationality*: Given Limited Resources, Prospective Opacity and Ageing of Policies*, in the absence of reflective adjustments in our strategies of reconsideration*, there is a natural tendency in how we form our policies to drift from their initial rational status, if they were rational at the beginning, precisely and paradoxically by maintaining their initial stability. We can appreciate this tendency particularly in cases of temporally extended policies, whose rational status drifts more visibly (see Vienna Marathon Reiterated example).

From now on, the problem highlighted by points a) to f) will be termed Opacity-Drift Puzzle – ODP, for short. ODP concerns not only policies, but also general plans and intentions, as long as stability and habits of reconsideration are involved.

[9] This opacity is 'prospective' in the sense that one can't see clearly only from the perspective of the policy's initial formation, but not retrospectively. With retrospection one can 'see' the stability of one's own policy. I first called this phenomenon 'perspective blindness', and I owe its refinement to 'prospective opacity' to a talk with Prof. Hans Bernhard Schmid.

1.3 Quick Response, Answer and Bratman Over Time

1.3.1 A Quick Response

A quick response to ODP is to say that Limited Resources doesn't mean that adjustments do not take place from time to time with a reasonable frequency. Indeed, we could consider it rational of us to adjust the stability of our policies from time to time, as a rational feature of policies themselves.

Bratman writes that

> It seems plausible to suppose that it is in the long-run interests of an agent occasionally to reconsider what he is up to, given such opportunities for reflection and given that the stakes are high, as long as the resources used in the process of such reconsideration are themselves modest. The unexamined life may still be worth living; but the occasionally reexamined life seems likely to be superior, even for limited creatures like us. Alongside the presumption in favor of increased stability of one's plans, then, we should place a presumption in favor of occasional reconsideration, given sufficiently high stakes and an appropriate opportunity. Reasonable habits of nonreconsideration should comport with both these presumptions. (Bratman, 1987, p. 67)

Sergio Tenenbaum provides us with a clear example, where we do not even need a presumption in favor of occasional reconsideration to justify it. Consider his example: Tenenbaum has only one minute with his banker to decide whether to invest in bonds or in mutual funds.

> The day before the meeting, I form the intention of buying into an index fund. Now I am waiting for my banker, and I expect that I'll have at least half an hour before she can see me. Would it be irrational for me to reopen the issue? There's really no new information, no unexpected circumstances (suppose I know that I generally have to wait about half an hour before I meet my banker), but I am bored, I don't know what to do, and reconsidering my intention seems to be a perfectly good way to spend this half hour. It couldn't possibly be a failure of rationality to reconsider my intention in these circumstances. (Tenenbaum, 2016, p. 8)

We see two strategies to articulate the quick response: one is appealing to a presumption in favor of occasional reconsideration; the other is pointing out that reconsideration is a perfectly reasonable way to spend time, in case we don't see any better activity (or inactivity) to engage in. In other words, Bratman says that a life where we occasionally reconsider can be more rational than a life without such occasional reconsideration; Tenenbaum's point would be that occasional reconsideration simply occurs and doesn’t necessarily make us less rational.

1.3.2 Preliminary Answer to the Quick Response[10]

I agree with Bratman on the claim that the occasionally reexamined life can be superior to its unexamined counterpart, but not for instrumental reasons. In REM, rationality is considered instrumental in the sense that it aims to satisfy our desires and achieve our ends by appropriate means. In this respect, REM is compatible with the broad understanding of ‘instrumental’ in Bratman's early writings:

> The crucial point is that I am trying to discuss structures of planning agency, in a way that appeals to the nature of such agency and to demands of instrumental reason but does not depend on arguing that practical reason, by itself, mandates certain ends. (Bratman, 1999, Ch. 4, p. 60 n.5)

Instrumental rationality is in fact ‘end-neutral’.[11] On this account, reconsideration serves exclusively the dictates of instrumental rationality, and can't be an ‘end of its own’, in the sense that we reconsider for reconsideration's sake. I think that REM is limited precisely in this sense: part of the job of reflective beings such as human agents, is to reconsider from time to time, because we are precisely those creatures, who reconsider from time to time.[12] The complete absence of reconsideration, the unexamined life, would strike us as an oddity, not because it would be instrumentally irrational, but because human agents aren’t like that. We could think of a perfectly

[10] I develop the answer further in 2.2.1.

[11] See ibid, p. 6.

[12] I do not further justify this assumption. Its scope is merely to serve as a contrast for REM and show that certain ideas are incompatible with REM, but could be plausible with respect to other conceptions of agency.

rational being, say an angel, who would reconsider only when it is strictly rational to reconsider and, at the same time, who is never presented with occasions of rational reconsideration because of divine providence, destiny, or the like. This angel will never reconsider, he is perfectly rational and his unexamined life is superior to any ‘examined life’ – from an instrumentally rational point of view. But he isn’t like us: he is an angel, after all.

If we stay within REM, then, we would need strict instrumental reasons to ground the possibility of reconsidering from time to time, and a presumption in favor of occasional reconsideration should be so grounded. Bratman makes it dependent on "sufficiently high stakes and an appropriate opportunity". This means two things: 1) underlying habits of reconsideration are high-stakes-responsive and trigger reconsideration, and 2) new information or a change in beliefs and desires makes it instrumentally rational of a person to reconsider. Since we focus on the nonreflective case and on those mechanisms that support coordination over time, we are concerned with 1) rather than with 2) and are led back to relying on our nonreflective habits of reconsideration. The question is then: Why reflectively reconsider our plans from time to time and, indirectly, adjust their stability, given that we rely on nonreflective mechanisms in the absence of new information? As Richard Holton has put it:

> In employing a habit of nonreflective nonreconsideration we do not make ourselves unable to reconsider. We still *could* open the question up again, even if circumstances do not change. It is just that we do not. (Holton, 2004, pp. 515-516)

Tenenbaum's example avoids this question: we don't need any instrumental reason for reconsidering from time to time. From this perspective, reconsidering is a form of escapism from boredom and sounds distinctive of the kind of creature that we are – some of us might enjoy pointless reconsideration when it is not strictly irrational to reconsider pointlessly. But this explanation is out of the theoretical reach of REM. Again, we need instrumental reasons to reconsider, where reconsideration enhances our instrumental rationality, rather than being constitutive of the kind of creatures that we are.

I think REM might still allow for the possibility of reconsideration from time to time, but for a very distinct reason. I have outlined in ODP that the tendency of policies to drift from their initial rational status could constitute the true (instrumental) reason for postulating adjustments of stability from time to time for rationality's sake. Otherwise, these adjustments would simply be a waste of time and energy, given that other cognitive factors do not change in any relevant way.

There is a further caveat: even if we were to adjust the stability of our policies, Limited Resources would hinder that we constantly do so, and Prospective Opacity would make it impossible for us to appreciate in advance how stable the adjusted stability is.

We could describe the final version of the quick response in the following terms: if habits of reconsideration monitor the rational stability of our policies, and have a natural tendency to drift from rationality, don't we need ‘adjusting habits’ to monitor them in turn? This formulation uncovers the risk of an infinite regress of habits monitoring habits. That would not only be implausible, it would also be disruptive of any form of effective agency for agents with limited resources like ourselves.

Perhaps the ‘adjusting role’ shouldn't be played by other nonreflective mechanisms, but rather by reflective or, better, self-governing ones. Still, the puzzle about stability stays so far unchallenged and relevant with respect to REM, because nonreflective non-reconsideration is central for our coordination over time – again, given Limited Resources.

If we hold REM as a background, I think that the puzzle I have outlined is genuine, or so I will further argue.

1.3.3 Early Bratman and Later Bratman[13]

In his later work,[14] Michael Bratman has focused increasingly on the relation between self-governance of agents and stability, rather than solely on the stability of intentions. His later accounts maintain the parsimony of his early philosophy, but aim more directly at understanding how autonomy and self-determination work for an agent.

Bratman's later elaborations are interesting for my puzzle. To see in which way, let us return to the risk of a regress of habits monitoring habits implied by the quick response. We could avoid it simply by postulating some sort of meta-agent, over and above our policies and motivational structures. However, Bratman refuses this view on two grounds:

> One problem is that it is difficult to know what it means to say that the agent – as distinct from relevant psychological events, processes, and states – plays such a basic role in the etiology and explanation of action. Second, and relatedly, in seeing the agent as a fundamentally separate and distinct element in the metaphysics of our action we seem to abandon the idea that our agency is as fully embedded in the event causal order as is the agency of purposive agents like dogs and cats. (Bratman, 2007, Ch. 2, pp. 24-25)

The difference between our agency and that of dogs and cats shall be hence found within the same causal order of events that determines their agency.

The strategy adopted by Bratman is that of referring to a broadly Lockean conception of personal identity:

> Locke and today's Lockeans have argued that the identity of a person over time consists primarily in overlapping strands of various kinds of psychological ties [...] Locke, as he is normally interpreted, focused on backward-looking memory. Today's Lockeans introduce, in addition to memory, forward-looking

[13] In a workshop at the University of Parma (2018), Elijah Millgram encouraged me to dedicate particular attention also to Bratman’s later writings, especially the essays of *Structures of Agency* (2007).

[14] See Bratman 2007, 2010, 2012, 2014.

> connections like those between a prior intention and its later execution, and continuities in desires and the like. (ibid, p. 29)

Planning agency and the rational norms of coherence, consistency and stability that we have seen in REM, favor coordination among such Lockean ties and support them in constituting our cross-temporal personal identities. Drawing from Harry Frankfurt's conception of personhood,[15] Bratman connects Lockean identity to reflectiveness, which is a matter of higher order attitudes: desires about desires, policies about policies, etc.

This hierarchy of attitudes alone doesn't grant personal authority over our own desires:

> After all, as Gary Watson noted in response to Harry Frankfurt's early and seminal work on these matters, we have as yet no reason for saying that, in the face of such conflict, I am on the side of my second-order desire, rather than saying that I am on the side of my first-order desire. (ibid, p. 23)

Reflectiveness without personal standing is called *weak reflectiveness*, whereas *strong reflectiveness* is "the capacity to take a stand" (ibid, p. 24) with respect to our lower-order attitudes. Bratman thinks that agents have particular higher order policies that are strongly reflective and can determine the standing point of an agent. These policies are called *self-governing policies*, i.e. policies "explicitly concerned with the functioning of relevant desires generally in one's temporally extended life" (ibid, pp. 33-34):

> Self-governing policies might, so to speak, crystallize pressures from various elements of one's psychic stew into a more decisive attitude that can, in the relevant context, establish where one stands. (ibid, p. 37)

Still, self-governing policies have the same structure of ordinary policies: they are defeasible and they do not block the regress. A second appeal to Frankfurt helps Bratman in solving this issue. Harry Frankfurt formulated the concept of *satisfaction* in

[15] See Frankfurt, 1971.

terms of not-being-moved to change one's hierarchy of desires on reflection.[16] Bratman extends satisfaction to policies:

> [...] let us say that self-governing policy *P** challenges *P* when *P** is in conflict with and, as a result, the presence of *P** tends to undermine the role of *P* (perhaps by leading to a disposition to change *P*) in supporting coordinating, Lockean ties. *For one to be satisfied with one's self-governing policy is*, to a second approximation, *for that policy not to be challenged by one's other self-governing policies*. (ibid, p. 35, added italics)

With this account of strong reflectiveness, Bratman seems to have avoided the regress of quick response by remaining within the framework of REM. But does the addition of self-governing policies spare REM from ODP? I don't think so.

Consider Prospective Opacity, according to which one can't truly know in advance how stable one's policies are at the time of their formation. This seems to clash with the neo-Lockean approach to personal identity, insofar as it recognizes the central role of forward-looking attitudes for our agency. Forward-looking attitudes are attitudes like future-directed intentions, plans and policies. However, even if such policies play a central role in the constitution of our identity and they control our conduct over time, it doesn't yet mean that they give us a privileged access (i.e. temporally advanced) to their own stability or to the stability of other policies. Neo-Lockean agents are still 'short-sighted' with respect to the stability of their policies, at the time of their initial formation.

The point of Prospective Opacity reaches further with respect to these considerations. Indeed, an agent can only be momentarily satisfied by her newly formed self-governing policy: she can see at present that no potentially conflicting challenge lies at the horizon, but she doesn't yet know if the newly formed policy has stood the test of time. And in turn, once the policy has stood the test of time, how does she know that? And doesn't that very fact change the assessment of the rationality of the stability of the policy? And

[16] See Frankfurt, 1992.

doesn't that further imply, other things equal, that the policy tends to naturally drift away from its initial rational status?

Further treatment is needed to shed light on these issues.

CHAPTER TWO

Refinements and Doubts

2.1 Refinements and Doubts Concerning Stability

2.1.1 Opacity and Predictability

ODP starts from three assumptions:

a) *Stability*: agents develop nonreflective mechanisms to control when it is rational to reconsider, i.e. habits of reconsideration;
b) *Limited Resources*: agents need to coordinate over time with limited resources and can't constantly reconsider or adjust the stability of their intentions-plans-policies;
c) *Prospective Opacity*: agents can’t see clearly the stability of their policies at the time of their initial formation.

The first two premises are explicit in Bratman's work on intention and agency, whereas the third is new, and I have termed it Prospective Opacity. This condition seems to clash with an important point in Bratman's writings: the stability of intentions should enable us to better predict our conduct in the future.

> [...] our pursuit of organization and coordination depends on the predictability to us of our actions. Coordinated, organized activity requires that we be able reliably to predict what we will do; and we need to be able to predict this despite both the complexity of the mechanisms underlying our behavior and our cognitive limitations in understanding those mechanisms. In treating prior plans as settling practical questions we make our conduct more predictable to cognitively limited agents like us by simplifying the explanatory structures underlying our actions. (Bratman, 1999, Ch. 4, p. 59)

On the one hand, "treating prior plans as settling practical questions" allows us to be more predictable to ourselves. On the other hand, there seems to be a "complexity of the mechanisms underlying our behavior", which cannot be fully grasped by our cognition. This underlying complexity determines the stability of our plans which, in turn, are treated as settling our future conduct and making it predictable. Where does the opacity lie and where the clairvoyance?

It seems that Bratman himself hinted at an answer in his early work:

> I do not normally decide how stable my plan is to be. Generally I do not, having settled on a plan, begin to reason yet again about how stable my plan should be (though in special cases I might). Rather, the stability of my plan is largely determined by general, underlying dispositions of mine. (Bratman, 1987, p. 65)

In a certain sense, then, by the initial formation of a plan, we do not really reason about the underlying mechanisms determining its stability. Furthermore, even if we were to reason about such mechanisms, we would at best obtain a reliable introspective report about them, but again, stability is an achievement that only time can evaluate. Rather than being introspective, knowing the stability of our plans shall be retrospective. The opacity lies at the initial formation of a policy: habits of reconsideration are for Bratman bound to our underlying psychological structures, whose stability is not transparent when we form a new policy. Even if we were able to look into the nature of those mechanisms, and hence had a reliable epistemic access to them, their practical significance (stability aimed at coordination over time) would still be out of the reach of our (epistemic) knowledge at the formation of the stable policy.

A habit of reconsideration/blocking is a mechanism that a rational agent develops as a remedy to her limited resources and to the impossibility of constant reconsideration. But the adoption of a particular strategy of reconsideration has more to do with deeply rooted psychological mechanisms depending on the kind of person that we are, rather than with ‘strategic choices’ or introspection:

> For example, if I am the sort of person who is constantly on the alert for dangers in my environment, my plans are likely to have a kind of instability they would not have if I were a person too engrossed in my projects to be very sensitive to such dangers. (ibid, pp. 65-66)

Prospective Opacity is an epistemic condition about stability: it tells us that we can't truly know in advance how stable an intention-plan-policy is (and is going to be) by its initial formation. The practical benefit from having this kind of knowledge would be great, because we would predict our reaction to challenges to our intentions over time and consequently adjust our intentions. However, this kind of knowledge is significantly limited: apart from not being able to predict the future, the stability of a newly formed intention (also at present) is opaque, and this opacity is a limitation for the rationality of our agency.

So, why does Bratman stress the predictability that issues from forming plans? In my opinion, the answer is that predictability, in this sense, is practical rather than epistemic. As Bratman writes, we treat our plans as settling practical issues, and from this settling we can further plan. For instance, suppose that I plan to take the bus to the ice cream parlor and buy an ice cream. If I weren't thinking that I am actually bringing myself to take the bus as a consequence of an intention, further planning would be pointless. So, I form an intention, which provides me with a practical predictability of my behavior: since I have formed the intention to take the bus, I need to form an intention to buy a bus ticket, and an intention to check the timetable. In a certain sense, the prediction lies in settling on practical moves, which follow from an initial practical move (intending to eat an ice cream). Moreover, this practical prediction allows me to form coherent and consistent plans and subplans to shape my later conduct.[17]

I still do not see into the future, but I can treat my present mental states as settling what I will do later. The formation of my intentions-plans-policies now, shapes my later conduct and is structured by the rational norms of cross-temporal agency of means-end

[17] This intuition about the practical nature of predictability seems to be in line with Bratman's later writings about the anticipation of future regret: "[...] this Frankfurtian conception of an agent's standpoint, and the associated norm of synchronic standpoint rationality, puts us in a position to articulate a *practical* role to be played by anticipated regret in our account of rational follow-through. The intuitive idea is that, given the anticipated regret, in certain cases the agent's standpoint does not favor the option favored by her shifted evaluative judgment, but favors, rather, following through with her general policy. The direct impact of the anticipated regret is in this way practical, rather than epistemic" (Bratman, 2014, p. 303).

coherence, consistency and stability. We could say that, from a certain perspective, following the norms of cross-temporal agency shapes at present what we will do later and practically predicts our later conduct by means of its settling function.

Practical predictability is predictability in reverse: instead of predicting from the most proximate steps and proceeding along a causal chain until the final achievement, we start by generally intending an end, and then settle on each step necessary to achieve it. We don't treat our future actions as a domino effect of successive states, as if we were to buy the bus ticket, check the timetable, take the bus and then, surprisingly, end up at the ice cream parlor. On the contrary, we act after having settled on a hierarchical structure of plans, where "plans concerning ends embed plans concerning means and preliminary steps" (ibid, p. 29) – it doesn't matter how detailed this settling is at present, for plans are partial and can be filled with new intentions as time passes by.

To sum up, Prospective Opacity is an *epistemic* limitation about knowing in advance the stability of our intentions-plans-policies at the time of their initial formation, whereas predictability is a *practical* feature of our agency, strictly bound to the settling function of our plans.

2.1.2 Drift from Rationality and Snowball Effect

Three consequences follow from the three premises of the puzzle:

d) *Test of Time*: we can appreciate the stability of a policy only after it has stood the test of time,

e) *Ageing of Policies*: as time passes by, we become good at doing what is supported by a rationally stable policy, and this ‘becoming good’ changes when it is rational of us to reconsider, and

f) *Drift from Rationality*: given the absence of adjustments of stability, our policy tends to drift away from its initial rational status.

Test of Time is the direct counterpart to Prospective Opacity: the fact that we can't see clearly the stability of a policy at its initial formation doesn't imply that we also can't see retrospectively how stable a policy has been. We obviously have access to such retrospection and we can assess how stable a policy has been after a sufficient amount of time has passed by.

Ageing of Policies also seems quite plausible: experts are subject to different rational pressures than beginners. At its initial formation, the costs of reconsideration of a policy might seem much higher compared to the costs seen at the time when we are confident about doing what is supported by the policy. From an external perspective,[18] the stability required might always be that of the ‘expert level’, but internally, when we question our ability to adhere to that newly formed policy, it might be rational of us to increase the resistance to reconsideration even to the point of slightly excessive resistance.

The most problematic consequence to accept might be Drift from Rationality, which together with Prospective Opacity lies at the core of the puzzle. Indeed, I have claimed that a policy would be subject, in the long run, to a tendency to naturally drift from its initial rational status, while its stability stays unaltered.

Drift from Rationality follows from three facts about our planning agency: (a) we do not constantly adjust the stability of our intentions given our limited resources, (b) we do not truly know in advance how stable a newly formed policy is, (c) over time, the criteria to assess the rationality of our policies become different from those at the initial formation of the policy. The result is that ‘going into automatic pilot’[19] has a tendency, over time, to stick with its initial stability and not to adapt to the kind of agent that we have become (or that we have been all along).

[18] In Bratman's early terminology, an external perspective implies an assessment of rationality that takes into account only the expected satisfaction of ‘rational desires’, and brackets the influence of prior plans (see Bratman, 1987, p. 43).

[19] See Bratman, 1987, p. 37.

Even if this might strike one as implausible, we need to keep in mind that we are moving within the framework of REM: according to this framework, our coordination over time is granted by policies supported by appropriate nonreflective habits of reconsideration/blocking, and not by means of reflection. If we were to reflectively reconsider, given that the appropriate conditions obtain, we would perhaps be able to sidestep the inconvenience of Drift from Rationality. But again, the threat would remain and it could be defused only by the triggering of reflective countermeasures. In the nonreflective case, on the other hand, our intentions-plans-policies slowly but relentlessly drift away from their initial rational status.

I sense a weakness in my defense of this specific point, but it might be a weakness of REM itself, rather than of ODP. Indeed, REM is a quite minimal theory of planning agency, where all we need for coordination over time are intentions-plans-policies respecting norms of practical rationality. REM enhances the conative control of future-directed intentions for cross-temporal coordination's sake at the expense of our synchronic reflective and cognitive control on what we are doing at present. This model favors sticking with the initial stability of our policies (other things being equal) over being responsive to an alteration over time of what is rationally demanded for our stability.

A related doubt about Drift from Rationality, is its apparent opposition to another basic assumption of Bratman's planning theory of intention: the *snowball effect*. The following is one of Bratman's definitions:

> [...] once one begins to act on an intention the world may, as a result, change in relevant ways; and it may be that these changes make it increasingly sensible to continue to act on that intention. This is the snowball effect. (Bratman, 2012, p. 74)

Like a tiny snowball that assimilates more and more snow as it rolls down the side of a snowy slope, our prior intentions make it increasingly rational to take further means to their completion. Taking the snowball effect seriously is crucial for the rationality of planning agents: once we settle on a certain course of action, and start investing in it by taking appropriate means and steps for its accomplishment, abandoning our project becomes much more costly, whereas holding on more rational.

Here we have our dilemma: on the one hand, we have a tendency to drift from rationality by sticking with our policies as time passes (as the third consequence of ODP); on the other hand, we have a tendency to be more rational if we stick with an intention (and policies are particular kinds of intentions) that we have already been following for a long enough time (because of the snowball effect).

In order to unravel this apparent opposition, I shall better illustrate what is implied by Drift from Rationality. Indeed, it is not that we are less rational by sticking with our policies, for the snowball effect tells us precisely the opposite. The point of the drift is rather that the rational status of our policies becomes, slowly but increasingly, out of sync over time, because policies maintain the same stability of their initial formation. Moreover, Prospective Opacity hinders the possibility to truly know in advance this stability. There we obtain our drift: over time, our policies have a tendency to stick more with their initial stability rather than being responsive to new ‘drifted demands’ of rationality.

Snowball effect and Drift from Rationality take place, so to say, at two different levels: once we have formed an intention, the former at the more particular level of the more rational course of action, the latter at the more general level of monitoring the stability of our intentions as time passes by. In other words, the snowball effect could participate to the process that makes experts out of us, whereas the drift is about sticking with the stability of beginners, while having already become experts.

Again, this might strike one as implausible. It is part of being an expert to know that one is an expert, or at least skilled, after all![20] I agree but, even if for some particularly

[20] I might have been too charitable. I don’t think that it is a requirement of expertise to know that one is an expert. Think about a 9 year-old Alpinist, brought on the mountains by her parents since she was moving her first steps. She might already be an expert Alpinist, but she also might think that being an Alpinist is quite ordinary and doesn’t

careful agent it might be difficult to appreciate its effects, Drift from Rationality is a process that inevitably concerns planning agents in virtue of their relying on the stability of their policies, no matter how careful they are in reflectively checking their level of expertise.

To sum up, Drift from Rationality is to be understood within REM as a natural development of our reliance on the stability of our policies. It regards our tendency to stick with the initial stability of our policies, other things equal, and is compatible with the snowball effect because this latter regards only what is rational of us to intend, once we have already been sticking with an intention with a certain stability.

2.1.3 What Is a Habit's Threshold?[21]

In his early writings, Michael Bratman has mentioned that, in order for our habits of reconsideration to be reasonable, their expected long-term impact should exceed an appropriate threshold.[22] For an agent to be rational, in a broadly instrumental sense, her habit of reconsideration must be reasonable and enable her to satisfy her desires or achieve her aims.[23] The expected long-term impact of a habit is, then, the general impact it has in helping her to meet her desires or ends.

From the agent's perspective, the stability of an intention-plan-policy can't be ideally rational:[24]

require particular skills. Perhaps, all her classmates and friends have experienced Alpinism and they might consider hiking a common way of spending their free time, not involving any form of expertise. They could all be ‘unknowing’ experts.

[21] This section is almost entirely an outcome of discussions in a seminar with Professor Hans Bernhard Schmid (University of Vienna, 2018). I am especially indebted to the remarks of Paul Tucek.

[22] See Bratman, 1987, Ch. 5, pp. 89-105 and Bratman, 1992.

[23] Certain habits supporting the same intentions could be worse than others in satisfying our desires, but none the less bring satisfaction. Such habits do not make us perfectly rational, but reasonable enough.

[24] The term ‘ideally’ refers here to Bratman’s technical expression ‘ideal stability’, assessed from a non-plan-constrained perspective, from which we foresee the costs of reconsideration and if it is worth it. (see Bratman, 1987, pp. 72-73)

> The stability of an intention or plan is *reasonable* if the associated habits of reconsideration are reasonable of the agent to have – if the expected impact of these habits on the agent's long-term interest in getting what she (rationally) wants exceeds an appropriate threshold. The two-tier theory tells us that if the stability of a prior intention or plan is reasonable, then, in the normal case, its nonreflective (non)reconsideration is rational of the agent. (Bratman, 1987, p. 72)

Recall the ‘two-tier account of the rationality of an agent for her nonreflective (non)reconsideration of a prior intention’.[25] According to this account of rationality, we start by seeing how reasonable a habit is in helping us generally to satisfy a given desire and, afterwards, we see if the habit's determination of reconsideration/blocking in a particular case is rational, all things considered.

There are two scales to evaluate the reasonability of a habit. The first one is temporal, judging how strong a habit is in resisting reconsideration as time passes by, and how well it allows reconsideration only when it is strictly rational. The second one is timeless: for that particular habit we settle on a timeless scale, where being reasonable means that the support for the satisfaction of a related desire exceeds a certain threshold. To put it differently: if the expected long-term impact of a habit of reconsideration justifies having it rather than not, the habit counts as reasonable. The value of a habit depends on that habit exceeding a certain threshold about the satisfaction of our desires.

What does this threshold and this timeless scale have to do with ODP? In a certain sense, ODP questions that our habits of reconsideration can respect this timeless scale, *qua* timeless. The tendency to drift from the initial rationality of a policy means that our policy was settled on a determined threshold at the time of the initial formation of our policy. However, as time passes by and we become experts in doing what the policy supports, the threshold might change, while the stability of our policy remains stuck with the initial threshold. This might sound confusing, so I shall better illustrate this threshold.

[25] See 1.1.2 and Bratman, 1987, p. 68.

Bratman doesn't show particular interest in the conceptualization of this threshold: it is simply a theoretical threshold which, if exceeded, tells that a habit of reconsideration is reasonable. But how does this work? Habits of reconsideration shall determine a certain resistance to reconsideration for the sake of achieving what the agent desires. This means that the less one reconsiders in general, the closer one gets to the perfect satisfaction of one’s particular desire. We can then assume that habits exceeding their thresholds are (roughly) those which make us reconsider least. Refining this statement, we could say that the more rationally stable a habit is, the more it will consistently exceed the threshold.

For its habit, this threshold is an absolute and timeless unity of measure: if the habit exceeds it, it is reasonable, otherwise it is not. But, from an agent's perspective, at the initial formation of a policy, the threshold might be quite high: he is exposed to the risk of *brute shuffling*, that is, of abandoning the policy for excessive instability.[26] So it is reasonable of him to have a stronger resistance to reconsideration. However, after he has been sticking with his policy for long enough, the threshold might get lower and allow more reconsideration. After all, he should be confident that he is doing what is supported by the policy consistently: he is no beginner anymore.

But how does he know that the threshold has been lowered? He can't know from a ‘habit-constrained perspective’, for this threshold, just like stability, falls within the blind spot determined by Prospective Opacity. The threshold determining the reasonability of a habit, rather than being absolute or timeless, changes over time and we have a certain delay in appreciating its changes because of Prospective Opacity. Moreover, this delay causes a drift in our intentions-plans-policies’ rational status while we still adhere to the initial threshold. In other words, Drift from Rationality takes place because we adhere to an initial threshold, while we should have adjusted it instead.

[26] See Bratman, 2012, p. 81.

2.1.4 Smart's Problem

ODP seems to hint at an eerie blind determination of our later conduct in virtue of our previous intentions. This sounds quite similar to what Bratman has called ‘Smart's problem’. In an influential paper published in 1956,[27] J.J.C. Smart worked out some aspects and problems of consequentialist theories of morality, in particular utilitarianism. According to such theories, an action is good or bad depending on the respective goodness or badness of its consequences. Smart distinguished between two kinds of utilitarianism: *extreme utilitarianism*, which focuses only on the particular consequences of particular actions, without any general rule or principle to evaluate them; and *restricted utilitarianism*, according to which,

> [...] the rightness of an action is *not* to be tested by evaluating its consequences but only by considering whether or not it falls under a certain rule. Whether the rule is to be considered an acceptable moral rule, is, however, to be decided by considering the consequences of adopting the rule. Broadly, then, actions are to be tested by rules and rules by consequences. (Smart, 1956, p. 344-345)

Bratman sees an analogy between restricted utilitarianism and his two-tier approach: the general habit, like the general utilitarian rule, is to be assessed by how well it brings us to satisfy a desire (its consequences) and, in the particular case, we see if our conduct conforms with a reasonable habit, just like a good action shall conform to the (restricted) utilitarian rule.

Restricted utilitarianism faces a challenge:

> [...] is it not monstrous to suppose that if we *have* worked out the consequences and if we have perfect faith in the impartiality of our calculations, and if we *know* that in this instance to break [rule] *R* will have better results than to keep it, we should nevertheless obey the rule? Is it not to erect *R* into a sort of idol if we keep it when breaking it will prevent, say, some avoidable misery? Is not this a form of superstitious rule-worship (easily explicable psychologically) and not the rational thought of a philosopher? (ibid, p. 349)

[27] See Smart, 1956.

To illustrate this point, Smart makes an example also quoted by Bratman.[28] Suppose you promise a friend dying on a desert island that you will see that his money is given to a jockey club. You follow the utilitarian rule of always keeping promises, because it has better consequences than sometimes breaking promises. At the same time you know that, on consequentialist grounds, giving the money to a hospital, rather than to the jockey club, would be the best thing to do. If you still keep the promise in this particular case, you might irrationally worship your general rule, given that it doesn't have the best utilitarian consequences. To put it with Bratman, "if [one’s] rationale of accepting the rule is consequentialist, how can [one] rationally resist the consequentialist rationale for breaking the promise?" (Bratman, 1992, p. 9).

This is Smart's problem, and it doesn't only apply to restricted utilitarianism, but also to consequentialist theories in general. Bratman's two-tier account of rationality is a consequentialist theory, hence subject to this problem.

Before going into the details of Smart’s problem applied to Bratman’s two-tier theory, recall the distinction between the reflective/deliberative case and the nonreflective case. In the reflective case, a new belief conflicting with a former intention obliges reconsideration of the intention – if I come to believe that I can't intend what I previously intended, I must change my plans. In the nonreflective case, "some cognitive changes, while they do not straightaway oblige me to reconsider [...], do provide *prima facie triggers of reconsideration*" (ibid, p. 5). Habits of reconsideration have to deal with these weaker cognitive changes and track prima facie triggers of reconsideration. In particular, they should allow reconsideration only in cases in which reconsidering would rationally change the initial intention and the costs of the reconsideration (in terms of time, cognitive resources, etc.) would not outweigh the benefits of reconsidering. These are called by Bratman ‘would-change/worth-it’ cases.

[28] See Bratman, 1992, p. 9.

Habits concerned with prima facie triggers of reconsideration are called *prima-facie-trigger habits* (‘pft habits’ for short):

> 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. (ibid, p. 8)

Going back to Smart's problem for restricted utilitarianism, consider an analogous version for the two-tier account of rationality:

> Suppose that pft habits that are reasonable in our consequentialist sense would lead me at *t*2 not to reconsider my prior intention in the face of a *prima facie* trigger. But suppose that at *t*2 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 in belief from *t*1 to *t*2, my desire-belief reasons at *t*2 argue clearly for abandoning my intention to *A* and reconsidering what to do despite costs of reconsideration. In such a case will my two-tier consequentialist approach sanction irrational *habit* worship? (ibid, p. 9)

To put it differently, if my pft habit were leading me not to reconsider when it would be rational of me to reconsider, would the two-tier account allow irrational habit worship, and hence irrational resistance to reconsideration? Just like worshipping the rule of keeping promises in the face of the rationality of breaking one promise is irrational, so is conforming to a habit *against* reconsideration in a would-change/worth-it case. The key to solving this problem can be found in understanding the meaning of "*it is obvious* to me that mine is a would-change/worth-it case".

An easy response would be to add to the two-tier model an escape clause such as "always reconsider when the habit tells you to, except for those cases in which it is obvious to you that yours is a would-change/worth-it case". But, as Bratman notices, a habit endowed with this escape clause might be less rational than a habit that regulates reconsideration on its own. This is because habits play an important role in supporting coordination over time, and the escape clause is too coarse to respect coordination.[29]

[29] See ibid.

Bratman's response to Smart's problem takes a different track. As we have already seen, habits of reconsideration are concerned with nonreflective cases, where cognitive changes do not directly oblige reconsideration. Reconsideration becomes then a matter of tracking prima facie triggers of reconsideration. But if it is obvious to me that I am in a would-change/worth-it case, i.e. a case that directly obliges reconsideration, my habits of reconsideration won't be involved. My habits of reconsideration do not regulate reconsideration when other factors (cognitive changes and my evidence-responsiveness) oblige reconsideration straightaway.[30]

Does this response also apply to ODP? In a certain sense, the puzzle partially reflects Smart's problem: Drift from Rationality implies that we follow an initial stability over time, while later synchronic rationality would require that we be less rigid or just different. So, if it is obvious to us that our case (later in time) is a would-change/worth-it one, and we still stick to our initial stability, wouldn't we be worshipping our initial stability (our habits)?

In a certain sense yes.[31] Drift from Rationality is indeed some kind of worship of a previously developed habit. On the other hand, I am led to answer no, because, when it is obvious to an agent that she is in a would-change/worth-it case, and this triggers reflectiveness and the priority of present deliberation, habits of reconsideration may not apply, as Bratman pointed out. In such cases, the agent overcomes Drift from Rationality, in a moment of insightful reflection, and reconsiders depending on the synchronic rationality of the reflective moment, rather than depending on the (now drifted) rationality of her habits of reconsideration. However, such moments are not the concern of ODP: ODP is a puzzle about the nonreflective case of tracking prima-facie triggers of reconsideration, and not obvious cases.

ODP is both compatible with a version of Smart's problem, where habit worship is meant as a slow drift from rationality, as well as with Bratman's response, for it doesn't

[30] See ibid, pp. 9-10.

[31] I will formulate this possibility as an objection to my proposal in section 3.1.2.

apply to reflective or obvious cases.[32] ODP applies to REM, which remains a coherent theory of planning agency and the puzzle might simply be the natural development of the model.

2.1.5 The Historical Theory[33]

In his 1987 book *Intention, Plans and Practical Reason*, Bratman attempted to give a historical dimension to the rationality of planning agents. The historical pedigree of retained intentions, of deliberation and policies matters for the assessment of their rationality. If I form an *irrational* intention at t_0 to A at t_2 and I retain it through t_0 to t_1, the intention at t_1 to A at t_2 doesn't work properly as a *rational* default for my later rationality at t_2.[34] The initial irrationality at t_0 undermines the later rationality at t_1:

> [...] rational nonreconsideration in the present does not by itself guarantee the present nondeliberative rationality of an agent in intending. Rather, rational nonreconsideration functions as a kind of *rational link* to prior deliberation. Rational reconsideration allows the earlier rationality of the agent in forming the intention to be transmitted to a later time; it allows the agent now to inherit this earlier rationality. But if there was no such rationality in forming the intention at t_0 there is nothing to transmit or inherit. (Bratman, 1987, p. 80)

Here Bratman has formulated a crucial idea for his theory: non-reconsideration is a rational link and its role is to transmit the earlier rationality of an intention, so that the agent can inherit it. Still, if there was no rationality at the beginning, there can't be any transmission *of rationality* then.

An analogy between the historical theory and ODP ignores this case: for the puzzle to work, the initial formation of an intention-plan-policy must have been rational, which also means ‘endowed with reasonable stability’. But, if the initial formation of the intention was rational, and the habit of reconsideration is reasonable, i.e. it exceeds a

[32] Even if the issues here are more complicated (further discussions in 3.1.2 and 3.2.2).

[33] It was Franz U. Altner, who stressed the relevance of Bratman's historical theory for the discussion of my puzzle (University of Vienna, 2018).

[34] t_0, t_1, t_2,... are points in time such that $t_0 \leq t_1 \leq t_2 \leq \ldots$; ‘A’ is an intended action.

certain threshold, shouldn't I be automatically rational then, and contrarily to what the puzzle holds?

Consider Bratman's historical principle for policy-based rationality:

> *Historical principle for policy-based rationality*
>
> In the policy-based extension of the basic case it is rational of [agent] *S* at t_1 to have the policy-based intention to *A* at t_2 if and only if:
>
> a) it was rational of *S* at t_0 to form the general intention to *A* when *C*; and
>
> b) it was rational of *S* from t_0 to t_1 not to reconsider this general intention; and
>
> c) it was rational of *S* not to block the application of his general intention to this particular case. (ibid, p. 91)

Point a) is quite unproblematic, for it is also assumed by ODP. Point b) however, already shows a discrepancy between Bratman's assumptions and the assumptions in the puzzle. According to what standard of rationality was it rational of S not to reconsider from t_0 to t_1? The rationality bound to the initial stability of the policy or later synchronic rationality, tracking what I have called the 'Ageing of Policies'?[35]

This distinction becomes clearer if put in terms of internal, plan-constrained assessment of rationality as opposed to external, non-plan-constrained assessment of rationality. The 'drifted rationality' is the inherited rationality of the agent, who can assess her rationality only internally, whereas synchronic rationality at later times, tracking the ageing of a policy, can be assessed only externally. Consider the Vienna Marathon Reiterated example. If, after months of training, I am presented with a challenge to my training-policy and I react as if the challenge were to take place towards the beginning of the formation of the policy, something has gone astray. But it has gone astray with respect to what? With respect to synchronic rationality tracking the ageing of the policy, which can be also grasped from a non-plan-constrained point of view, that is external.

[35] See 1.2.3.

But why do I insist that we are facing a *drift* from the internal perspective, if we can't assess it from within, but only externally? After all, Bratman claimed that

> [...] it may be rational of an agent nondeliberatively to intend to *A* even though the recommendation from the external, non-plan-constrained perspective would be to reconsider that intention and decide differently. (ibid, p. 80)

However, he also adds that,

> for this to be true, however, it is not enough that it now be rational of the agent not to reconsider this intention. Rather it must have been rational of the agent initially to have formed that intention, and also rational of her throughout not to have reconsidered. (ibid)

Again, according to what standards of rationality, must it have been rational of the agent not to have reconsidered? From the internal perspective, the only standards are those of the initial formation of the policy, but this is doomed to slowly drift from its initial rational status. So it is true that we are not necessarily irrational if the internal assessment of our rationality differs from the external assessment. But still, our policy's rational status is subject to Drift from Rationality, given that stability doesn't adapt over time to the evolution of the status of the policy. Paradoxically, it is by sticking with our earlier rationality, that our intentions-plans-policies slowly drift from their original rational status.

How does this "sticking with our earlier rationality" work? Through the rational link and the transmitting function of rational non-reconsideration. In other words, we inherit our earlier rationality, but, in the meanwhile, we can't be sure that it has not become out of sync. If the initial formation of the policy was rational, and the non-reconsideration was rational, preserving the initial rationality and transmitting it to a later time, we are not yet fully-blown rational, as Bratman seems to hold, and as REM definitely holds: if we transmit the initial rationality to a later time, but the criteria of assessing rationality have been altered by the mere passing of time, our intentions-plans-policies are drifting away from their initial rational status. This drift can't be appreciated with short-term retained intentions, but becomes clear with largely extended policies, such as the Vienna Marathon Reiterated example's training policy.

Here one might protest: Aren't we reflective agents after all? Of course we keep track of the alteration of the kind of agents that we are as time passes by! That is what it means to be reflective agents and if I want to claim that in the nonreflective case we are slowly drifting from rationality, I must be holding excessively high standards of rationality for limited agents such as ourselves. We aren't only reflective enough, we are also rational enough.

Well, I agree on both points: we are reflective enough, even if we need nonreflective mechanisms to support our cross-temporal coordination, and we are rational enough, even if we have a tendency to slowly drift from how rational we were at the beginning of the formation of a new policy. This doesn't yet mean that we are often irrational. Indeed, ODP doesn't undermine a commonsensical understanding of our agency as reflective and rational: it is more about the theoretical consequences of REM, rather than about phenomena that are easily tracked in our day-to-day life. Remember that the puzzle works under the condition 'other things equal', which is a quite strong condition for real life, because we incur constant cognitive changes and the more time passes by, the more it is probable that we adjust our policies for an infinite number of reasons.[36] However, the puzzle is a valuable theoretical development as a test and, why not, an oddity of a model of planning agency such as REM.

Let's now move to point c) of Bratman's historical principle for policy-based rationality, concerning blocking habits rather than habits of reconsideration.[37] The Vienna Marathon Reiterated example can be seen both as a case of non-reconsideration and as a case of default absence of blocking. Since we go for the training session and we stop before ever reconsidering or blocking, we might never know if we would have

[36] I formulate this point as an objection to my proposal in 3.1.1.

[37] The following observations emerged in conversation with colleagues, and especially Professor Hans Bernhard Schmid (University of Vienna, 2018).

reconsidered rather than blocked the policy. However the most rational thing to do would have been blocking and not reconsidering, given that we do not want to give up the training-policy. I think that the caveat about blocking habits is useful for Bratman to distinguish between merely retained intentions and policy-based intentions (which needn't be reconsidered back to the level of the general policy, but could simply be blocked for a particular case), but is not as interesting for ODP, because ODP works for habits of reconsideration as well as for blocking habits.

2.1.6 Stability, Inertia and Plasticity

ODP is a puzzle concerning those mechanisms that support our coordination over time nonreflectively. It is a puzzle about the exertion of control over our own conduct over time, when this control can be exerted only from a previous point in time and with the resources of that previous point in time. Thus, to a certain extent, it is more a puzzle about the inertial component of stability, rather than about its active and reflective part.

Piotr Tomasz Makowski has tried to work out the difference between inertia and stability, which were interchangeably used in Bratman's early writings.[38] Afterwards, ‘inertia’ has been replaced almost completely by ‘stability’, because, according to Makowski, Bratman wanted to stress the rational and normative character of nonreflective non-reconsideration, which inertia can't supply.

I think that Makowski's challenge to Bratman is based on a misconception of Bratman's work, but his positive attempt to shed light on the concepts of inertia and stability can still pose a significant threat to ODP: Does ODP concern inertia alone or does it concern authentic Bratmanian stability? Before answering this question, I have to first illustrate Makowski's proposal, and second show why I think it is misguided.

Intention inertia is, for Makowski, the possibility of an intention to be retained without reconsideration *simpliciter*. It is its bare defaultness. The inertia of an intention "can be described roughly as its 'managing to survive until the time of action'" (Makowski,

[38] See Makowski, 2016.

2016, p. 1048), it is completely passive and rooted in the structure of our agential psychology.

Stability, on the other hand, has more actively to do with the mechanisms of rational nonreflective non-reconsideration. Its rationality can be assessed with a two-tier model, where the higher tier concerns the overall rational gain given by a habit of reconsideration, whereas the lower tier concerns the rational gain of a particular action depending on the habit of reconsideration. For Makowski, stability, contrarily to inertia, can be *reasonable* in the sense that it must monitor cases of reconsideration in a reasonable way:

> [...] the stability of future-directed intentions involves both active and passive aspects of our psychology – a default tendency toward nonreconsideration combined with a readiness to reassess intentions when needed, as in the face of new information or a change in belief, making intentions *defeasible*. In other words, the reasonable stability of future-directed intentions is a *non-inertial stability*. It is context-dependent, and the degree to which each of these two aspects influences action will depend on the specific circumstances. A reasonable agent will always have an available disposition to reconsider or to refrain from reconsidering, creating a sensible equilibrium. (ibid)

Although Makowski deems stability as an authentic feature of intentions, he is doubtful as to whether it applies also to plans. Indeed, he criticizes Bratman's lack of a clear distinction between the two concepts of intentions and plans. Makowski seems to stress that plans, contrarily to intentions, "should be understood as complex, more or less coherent clusters or chains of intentions" (Makowski, 2016, p. 1049), which are subject to an exclusive demand of rationality called *planning plasticity*.

Planning plasticity is supposed to account for the possibility of abandoning intentions without reconsideration, which Bratman allegedly overlooks, and the ability of agents to rationally respond to unstable environments, while maintaining psychological stability. Makowski defines plasticity with two points:

Planning Plasticity:

> 1) Plans should be flexible. That is, the complex chains of intentions which constitute a plan should be considered alongside the demands of the environment of the planning agent. (ibid)
>
> 2) [Plans manifest a] readiness of an agent to modify a plan, according to the new information about the planning environment. (ibid, p. 1051)

In support of plasticity, Makowski offers two main examples. The first is about learning of a natural disaster in Japan, which he had planned to travel to. Since this plan was under the higher-order plan to visit Asia, planning plasticity allows him to 'reframe'[39] the plan and start planning to travel to Singapore instead. This capacity to reframe plans shows, according to Makowski, the ability of agents to retain "psychological stability despite environmental fluctuations" (ibid).

The second example tries to show that a planning agent not responding to Planning Plasticity is less rational than one who does. Suppose that businessman Mark plans to merge his company A with company B. Mark's planning is rigid and employs all his resources. However, the CEO of company B suddenly retreats and Mark, whose planning wasn't plastic, incurs huge losses and rational failure in achieving what he was aiming at, for he invested all his resources in a merger that won't take place anymore.[40]

For my current purposes, we could summarize Makowski's proposal in three points:

a) Intentions are not the same as plans;
b) Inertia is not the same as Stability;
c) Stability is not the same as Plasticity;

Starting with point a), I think that Makowski's support for the distinction of plans and intentions is much weaker than he assumes it to be. Indeed, in a footnote, he writes that "future-directed intentions probably *always* enter, sooner or later, some planning

[39] I consider the term 'redirect' more appropriate than 'reframe' in this context, but I stick with Makowski's terminology none the less, because the issue is secondary.

[40] See ibid, p. 1051.

structure in which they make sense" (ibid, p. 1055, footnote 1). So what is the difference? As we have seen, Makowski underlines that plans are more complex, and they embed intentions, but this is also what Bratman thinks. Bratman argues that what is peculiar of plans are three features: 1) their being partial, which means that they can be filled-in with new intentions as time passes by, that they are defeasible and that they are not completely settled in the details; 2) their hierarchical structure, which roughly means that plans concerning general ends (such as travelling to Japan) embed plans concerning specific means (like buying the plane tickets); 3) their hybrid character, which consists of deliberative and non-deliberative aspects.[41] Still, this doesn't allow Bratman to draw a neat distinction between intentions and plans precisely because, to a certain extent, intentions have a planning structure, and if we want to understand their role for an agent's practical rationality, we need to refer to such planning structure. On the other hand, Bratmanian plans *are* Bratmanian intentions, if they are to be effective and bring us to action. I think this is enough evidence to deem Makowski's distinction as less sophisticated than Bratman's, and less able to capture important aspects of our agency.

Point b) of Makowski's account is also problematic. In a more recent paper, Bratman has explicitly defended a version of *practical conservativism*: "The view of prior intentions as rational defaults is a kind of practical conservativism" (Bratman, 2010, p. 21). That means that there is rational pressure, other things equal, to follow through with an intention, once it has been formed and it respects the demands of coherence and consistency. It seems to me that Makowski partially or completely ignores the condition ‘other things equal’. Of course, Bratman is not defending the rationality of stubbornness and the rationality of sticking with an intention no-matter-what, once new information or cognitive changes speak clearly for abandoning it.[42] All that Bratman says is that, *other things equal*, once you form an intention, that intention is subject to a rational demand for stability, that is, if you abandon it without good reason, you might be

[41] see Bratman, 1987, pp. 29-30.

[42] See Bratman, 1987, p. 69.

irrational. Hence, intentions become *rational* defaults for our further intending and planning, *unless they are challenged*.

There isn't truly any tension between inertia and stability in Bratman's early writings, even if Makowski is right in noting that stability becomes the preferred term. Arguably, stability provides us with a sense of rational control and can be more easily linked with those mechanisms of our underlying psychology such as habits of reconsideration, that regulate somewhat actively our conduct over time and not just as a matter of inertia. Still, I think that inertia is a part of stability rather than an aspect on its own, and reasonable stability definitely has an inertial component, rather than being non-inertial. As we have already seen, reasonable stability is the stability determined by reasonable habits of reconsideration, that is, habits of reconsideration exceeding a certain threshold in their long-term impact.[43] Reasonable habits are habits that overall help our effective achieving of what they have been formed for, with the least reconsideration possible. Thus, the inertial component of sticking with an intention without reconsideration is always present, and the absence of reconsideration remains the default aspect of intentions. Without this inertial component, it would be almost impossible to coordinate over time without constant reconsideration or deliberation, and stability embraces rather than overcomes this aspect.

Makowski doesn't reject this point: he thinks that stability in general is a mixture of inertia and reasonable stability. What he negates, is that *reasonable stability* is inertial and that inertia has anything to do with rationality. This, I think, is in contrast with Bratman: inertia as the *rational defaultness* of intentions is rational, and reasonable stability consists *also* of some inertial feature of intentions. Again, Bratman's account seems more detailed and careful than Makowski's distinction between inertia and stability. This is the reason why Bratman can't draw the neat distinction between the two concepts that Makowski expects.

[43] See 2.1.3.

Finally, and in line with the recent considerations, there seems to be no need for further demands on plans such as plasticity, and here we get to point c). The main reason why we do not need plasticity is that the rational demand on intentions is not just consistency, but *strong consistency*:

> To coordinate my activities over time a plan should be, other things equal, *internally consistent*. Roughly, it should be possible for my entire plan to be successfully executed. Further, a good coordinating plan is a plan for the world I find myself in. So, assuming my beliefs are consistent, such a plan should be consistent with my beliefs, other things equal. Roughly, it should be possible for my entire plan to be successfully executed given that my beliefs are true. This is a demand that my plans be *strongly consistent*, *relative to my beliefs*. (Bratman, 1987, p. 31)

When Makowski writes that Bratman's "account of stability disregards the nuances of the environment as an additional element that influences planning" (Makowski, 2016, p. 1050) he seems to have forgotten the demand for strong consistency, i.e. that our planning is situated in a context and that it must be consistent with respective beliefs to be rational. If the environment is unstable, I will form beliefs about such difficulties and plan accordingly. Bratman's theory is able to account for such cases, independently of how nuanced they are.

Secondly, and most importantly, Bratman doesn't give up the principle of the *rational priority of present evaluation*, according to which, when you are presented with new information and relevant cognitive changes, your present deliberation makes you rational, and not stability or habits of reconsideration.[44] The issues become more complex when there are cognitive changes or new information that do not trigger reflective control in this direct way, but are rather under the administration of habits of reconsideration. We have already considered these problems and Bratman offers a useful reminder:

[44] See Bratman, 2014, p. 297.

> [...] I distinguished three different kinds of cognitive change. First, in some cases the cognitive change straightaway obliges reconsideration. My example here was the case in which one newly comes to believe that one cannot *A* at *t*2. 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. (Bratman, 1992, pp. 9-10)

Bratman has offered a very inclusive schema to assess a vast range of cases. Abandoning an intention doesn’t fall out of this schema. To see this, let's return to the example of travelling to Japan. If my beliefs change to the extent that it is not worth travelling to Japan anymore, we are in the full-blown deliberative case, which is the first one considered by Bratman. We don't need anything like planning plasticity to abandon the plan to travel to Japan and form another to travel to Singapore: if there is inconsistency between our previous plan and our beliefs, we *ought to* abandon that plan, and if our desires still support planning of another kind, we just go for it. In my opinion, Bratman allows for such abandoning of plans without reconsideration, given that strong consistency is violated: when one is provided with new information which clearly contradicts the feasibility of one’s plan, the plan is defeated and ought to be abandoned without reconsideration, for there is no need to spend resources in pointless reconsideration.

Planning plasticity is about the "flexibility of planning" and the "readiness of changing a plan". These aspects are already included in Bratman's theory, in the guise of strong consistency and in the defeasibility of planning. One of Bratman’s interesting messages is precisely that, despite the priority of present evaluation and of present reflectiveness, if ever we want to achieve our goals, we need to coordinate over time and stick with our intentions without constant reconsideration. This is far from sanctioning excessive rigidity. As a matter of fact, Mark the businessman is not only violating plasticity, but also fails to be a good planner because of the excessive rigidity, stubbornness or wishful thinking of his plan to merge the two companies.

Now, going back to my initial question: Is ODP only a puzzle about inertia and innocuous for stability? I don't think so by following three reasons:

1. Inertia is a constitutive part of stability, hence, if ODP is understood as being about the inertial part of stability, it becomes per definition a puzzle about stability;
2. Adding the demand of Planning Plasticity to plans doesn't overcome Prospective Opacity, unless one is to abandon a planning theory of agency such as REM. This is because intentions, plans and policies are situated in time, and their stability is something that can be appreciated only after they have stood the Test of Time. Being more flexible and more ready for change could reduce the weight of the temporal distortion of our rationality due to Drift from Rationality, but comes with costs and still can't solve the puzzle;
3. One could call ODP a ‘puzzle about intention inertia’, but it would be misleading. Indeed, ODP is supposed to uncover an inertial aspect of stability that is not the ‘rational defaultness’ Bratman talks about. This inertial aspect is given by the fact that our later conduct is anchored to the rationality of our previous formation of an intention-plan-policy and can't but refer back to that rationality. Hence, it is an inertia that implies a slow drift from being rational, in virtue of the passing of time. Bratman's inertia is rational, whereas my (supposed) inertia consists in a dated rationality that is transmitted to later times. Hence, ODP is a puzzle about the inertia of ‘Bratman's inertia’ or, better, a puzzle about an inertial aspect of stability.

2.2 Refinements and Doubts Concerning Authority

2.2.1 Self-Governance and Regress

It is time to go back to some issues raised in the first chapter, concerning Bratman's later philosophy. We introduced self-governing policies as those higher-order policies concerning general desires in our lives, which stop the regress of orders of policies by being satisfied, that is, by not being challenged in their role of supporting cross-temporal Lockean ties – especially forward-looking ones, such as intentions, plans and policies. Recall that the argument of regress came from an attempt to solve ODP by postulating higher-order attitudes, which would have been able to overcome Prospective Opacity and monitor Drift from Rationality. The preliminary response I gave was that Prospective Opacity is a fundamental assumption about planning agents as described by REM and that, once we assume Prospective Opacity together with the other relevant Bratmanian assumptions, Drift from Rationality of intentions-plans-policies follows.

Now, one could doubt that Prospective Opacity is central for the kind of agents that we are. The Vienna Marathon Reiterated example might simply be a case of conflicting self-governing policies that do not challenge one another in Bratman's technical sense of ‘challenging’ – i.e. undermining the role of a policy in supporting coordinating Lockean ties. The first self-governing policy is the one concerned with the desire of running another Vienna marathon, the second self-governing policy is a policy about being a good student which involves passing minor exams. If I fail a minor exam, however, the self-governing policy about being a good student is arguably not challenged. Both self-governing policies are satisfied and nothing of particular interest seems to happen. However, the interesting point is to understand *why* and *how* something has gone astray in the reiterated case, while it was acceptable in the standard case. ODP attempts to give a meaning to this ‘going astray’, by saying that, over a considerable amount of time, we have a tendency to drift from how rational we were at the initial formation of our policies, in part because of Prospective Opacity, that is the difficulty to appreciate how stable our policies are (or are to be) at their initial formation.

However, someone might still stress that nothing has truly gone astray. If self-governing policies, the spearhead of our reflectiveness, do not track problems for our rationality, why shouldn't we be just rational then? My answer is that self-governing policies are not ‘almighty’.[45] They are defeasible, but could still be almighty in the sense that they monitor everything that lies under their scope, in a perfectly rational way from the internal perspective of an agent. I question this point: Self-governing policies do not have such an almighty monitoring role, because if they had it, agency would fall back into a regress of policies governing policies. Indeed, every time we face something conflicting with our policy or a Drift from Rationality, we would need a yet higher-order policy to govern that conflict or track that drift, and this *ad infinitum*.

Giving up this almighty monitoring role, on the other hand, allows for self-governing policies to be subject to ODP, for Prospective Opacity applies and self-governing policies incur Drift from Rationality.

We can formulate this point as a dilemma: the first horn is when self-governing policies are almighty in the sense of monitoring the internal rationality of the agent so that the policy doesn't incur Drift from Rationality; the second horn is when self-governing policies are nothing more than quite general policies that are satisfied in Bratman's technical sense, and ODP applies.

The first horn is vulnerable to the argument of regress. I deem this to be the main critique of Robert J. Muckle to Bratman:

> [...] on Bratman's view, when an agent is going to "take a stand" it is the self-governing policy that speaks on her behalf. However, self-governing policies are subject to rational reflection. He cannot account for who (or what) is doing the reflecting in that sense. In a more concrete sense, Bratman cannot account for what endorses a self-governing policy. Merely calling them "self-governing" does not exempt them from a need for an endorsement of their own. By this I mean that self-governing policies cannot really *be* self-governing in precisely the sense that as an action needs endorsing, so too do self-governing policies.

[45] The reason for this terminology is given by the connection between might and control/monitoring. An all-monitoring or all-controlling policy is, in a technical sense, an ‘almighty policy’.

> There needs to be some point at which the higher-order self-governing policies can come to a stop, or it is a regress. (Muckle, 2010, p. 159)

Even the appeal to satisfaction wouldn't save self-governing policies from regress:

> [...] in the process of rational reflection, satisfaction becomes irrelevant. if I am to revise my self-governing policy to wean myself from coffee then I must bring in another self-governing policy to conflict with it, and decide between the two. Satisfaction then, is left behind in the process of reflection; reflection will bring the agent to be satisfied with the policy she chooses. [...]
>
> If I am right about the regress, then when an agent attempts to adjust a self-governing policy she cannot do so since higher order policies must also be endorsed. Of course he [Bratman] could accept that agents decide those things, but he is openly against agent causation. If in light of the regress, we decide to give up that self-governing policies can be reflected upon then the whole idea of an agent changing what she wants over time is lost, since she cannot change her long-term wants. (ibid, p. 160)

Bratman would probably answer that Muckle is looking for an agent which he has already found: a ‘package of self-governing policies’ is what determines an agent's standpoint, and what the agent rejects or endorses is determined with respect to this background framework of policies. Still, Muckle is right on at least one point: if we maintain strong reflectiveness of self-governing policies in terms of *solely* determining at each point in time, what our relevant standpoint is, we give up an important aspect of our synchronic reflectiveness, that of taking control, synchronically, on what conduct to undertake from now on. Self-governing policies can't be synchronic and diachronic at the same time, they are not almighty and if they were, they would have to constantly govern themselves, falling into the regress pointed out by Muckle.

To save both aspects of our agency, namely the strong reflectiveness of self-governing policies and the importance of synchronic reflection, we shall interpret the role of self-governing policies as more limited than they emerge from Muckle’s criticism. Self-governing policies are limited in their ‘self-adjusting role’; they cannot constitute the whole of an agent's reflectiveness (for there is also synchronic reflective control), and their own stability is not transparent at the time of their initial formation.

Here we find the second horn of the dilemma: we accept that Prospective Opacity can't be overcome by self-governing policies and that it also affects their functioning. This means that the puzzle also applies to Bratman's later considerations about authority and self-determination.

I suspect that accepting the second horn is more in line with Bratman's philosophy, but I don't want to attribute to Bratman more than what I take him to write: accepting the second horn is definitely coherent with REM and that should be enough to appreciate the genuineness of ODP.

2.2.2 Temptation and Regret

Let's linger a little more on the tension between diachronic and synchronic agency. Much of Bratman's later work is devolved to exploring cases of temptation, i.e. situations in which we experience a temporary (synchronic) shift in our evaluative judgments conflicting with the diachronic functioning of our relevant intentions. Suppose that I have a policy of avoiding a second glass of wine after dinner for the sake of practicing the piano afterwards. This can't be done properly if I drink the second glass of wine. In a particular occasion, dinner is over and I have already drunk the first glass of wine and I now experience a shift in what I value most: now I value the second glass of wine more than playing the piano. This shift is temporary: if I go on and drink the second glass of wine, I will regret not having played the piano properly afterwards.

In cases of temptation, what seems to be rational at present contrasts with what is otherwise rational in virtue of a previously formed intention. If we have a presumption in favor of reasonable stability, in terms of resistance to reconsideration and commitment to a prior intention, and we want to respect the rational priority of present evaluation, at the same time, we encounter a tension between the diachronic dimension of agency as opposed to its synchronic counterpart.

There is a direct analogy between temptation cases and cases affected by ODP: both deal with a conflict between the stability of an intention-plan-policy over time and our later rationality. The former in terms of synchronic versus diachronic rationality, the latter in terms of initial rationality versus Drift from Rationality over time.

Bratman has developed two strategies to assess temptation, which roughly correspond to what I have been calling Bratman's early philosophy and later philosophy:[46] the ‘intention stability strategy’ and the ‘agential authority strategy’. According to the *intention stability strategy*,[47] we have:

1) Rational priority of present evaluation;
2) Prior intentions grounded in one's evaluative rankings, which offer an anchor for instrumental reasoning; and
3) Reasonable stability, making it rational of an agent to sometimes stick with the prior intention against a temporary shift in evaluation, hence overcoming 1).

The risk of 2) and 3) is *irrational bootstrapping*: a form of double counting a reason to A, the first time because one desires (values) it, the second time because one intended it. Intentions do not bootstrap reasons in this way: rather they are ‘framework reasons’, which help settle further intentions and practical reasoning, but do not themselves provide an agent with new reasons for action.[48] In this sense, reasonable stability supports

> [...] some sort of defeasible default presumption in favor of following through with one's prior intentions and policies [...]
>
> To grant this pragmatically grounded default in favor of the prior intention or policy does not, however, amount to seeing that intention or policy as providing a further (bootstrapping) reason in deliberation. What it amounts

[46] These two strategies can be theoretically distinguished but should be seen as intertwined to obtain a more exhaustive picture of Bratman's treatment of temptation. (see Bratman, 2007, Ch. 12, pp. 257-282)

[47] See Bratman, 2007, Ch. 12, p. 264, 274-.

[48] See Bratman, 1987, p. 34.

> to is, rather, seeing it as establishing a certain burden of proof on a challenge to that intention or policy. (Bratman, 2007, Ch. 12, p. 276).

Still, in temptation cases, if the shift is strictly contrary and outweighs one's previous evaluation, it seems difficult to maintain the presumption in favor of stability. To grant that still we sometimes might resist temptation, Bratman adds a ‘no-regret condition’: when we can anticipate the regret of giving in to temptation, we raise the bar of the burden of proof needed for temptation to be effective. Anticipated future regret ensures that, at the time of the initial formation of an intention, you know that

> if you were to be guided by your evaluative ranking in favor of the second glass, you would later regret that, and if instead you were to stick with your one-glass policy, you would later be glad that you did. (ibid, p. 277)

This explanation works in cases of no-unanticipated-information:

> [...] if one's plan was rational when formed, then surely it would be rational, barring relevant unanticipated information or change in basic desires or values, to execute it in those circumstances for which one specifically planned. (Bratman, 1999, Ch. 4, p. 62)

Anticipated temptation challenges this picture, but anticipated future regret can re-establish the rational default in favor of the prior intention or policy:

> For a planning agent – one who projects her agency into the future in ways that involve plan-type attitudes – such anticipated regret will normally have a prima facie significance. Our planning agency is future-oriented in a way that normally brings with it a present identification with how one will see matters – including one's now present actions – in the relevant future. This is part of the characteristic future-oriented focus of plans and policies. (Bratman, 2007, Ch. 12, p. 277)

If we have a case of no-unanticipated-information and our agency is future-directed, i.e. a situation addressed by ODP, why doesn't Bratman's ‘intention stability strategy’ already solve the puzzle? Why should we ever drift from rationality if nothing was unanticipated and we met the no-regret condition? Again, because of two main factors. The first one is Prospective Opacity, according to which we can't truly anticipate the stability of our intentions and policies. This opacity also concerns our anticipation of temptation and future regret. The second one is that, over time, we tend to stick with the kind of stability that we had at the initial formation of our intentions and policies, and we do not see that temptation should be considered differently after a considerable amount of time has passed.

A case of temptation in the near future as opposed to the very same case of temptation in the distant future, might incur different assessments depending on the modification of stability over time – something we can't transparently anticipate at present due to Prospective Opacity. True, we could always be incredibly brainy and flesh out all possibilities in the future, but not the stability of our intentions and policies, for that is something that only the Test of Time can show. One knows one’s intentions are stable only when they have already proven their stability, never in advance. Hence, one can't predict a drift from rationality and plan accordingly, for one can only anticipate and assess future temptation and regret with one’s *present* rationality.

The second strategy to assess temptation cases is the *agential authority strategy*:[49] certain attitudes *anchor* one’s instrumental reasoning providing one with a framework determining one’s own standpoint with respect to temptation cases. The anchor in question is provided by an attitude determining one’s evaluative ranking and hence standpoint as an agent. What kind of attitude? For Bratman it has to be a particular policy:

> I value X when I have a policy of treating X as a justifying consideration in my motivationally effective practical reasoning. (ibid, p. 269)

The second strategy to assess temptation then appeals to higher-order policies, constituting, for example, our valuing course of action A over course B. But what confers an authority over our synchronic valuing of temptation to such higher-order policies? Why should a higher-order policy determine where the agent stands with

[49] See Bratman, 2007, Ch. 12, p. 265-.

respect to temptation, more than the present shift in evaluative judgement determined by temptation? If temptation makes one value B over A just this one time, how can the policy of valuing A over B have the last word?

> A policy of practical reasoning is a general intention. Your valuing of two glasses this one time is an intention about practical reasoning, but it is only an intention about present practical reasoning. It is a singular commitment concerning relevant motivationally effective practical reasoning. It is a singular commitment to give relatively more justifying weight on this occasion to a second glass. [...] such a singular valuing may have only an attenuated claim to constitute where you stand on this occasion. It is a singular commitment: its role is not to structure your ongoing practical reasoning and action but only to structure your present reasoning and action. In contrast, your general action policy of only having one glass of wine at dinner does have the role of organizing thought and action over time, in part by way of associated continuities and connections. So there is a case for saying that this action policy, in contrast with your singular valuing, has the stronger claim to authority to constitute where you stand. (ibid, p. 272)

In cases of temptation, the authority of a previously held policy might simply be greater than that of a present evaluation, and hence have a stronger claim in shaping our practical standpoint.

Does this appeal to authority solve ODP? Here we meet the dilemma of self-governing policies again. If the authority of higher-order policies amounts to almightiness (in the sense of monitoring every single case of temptation under their scope, by keeping track of Drift from Rationality and adjusting accordingly), they would solve the puzzle. However, almighty policies fall into an implausible regress, for every time there is a conflict or Drift from Rationality applies, there should be a yet higher-order policy to solve that conflict or track that drift. This is the first horn of the dilemma, whereas the second horn accepts that higher-order policies (even self-governing ones) are relevantly similar to ordinary, lower-order policies and the puzzle applies to them, for they are subject to Prospective Opacity and can't track Drift from Rationality.

2.3 Refinements and Doubts Concerning Readjustments

2.3.1 The Puzzle of the Self-Torturer

The Vienna Marathon Reiterated example showed that the regret of sticking to a plan, when challenged after a considerable amount of time, could have been greater than the regret about the same challenge toward the initial formation of the policy, other things equal. Why? Because, as time passes by, we become better (or less-beginners or more experts) in doing what the policy supports us in achieving and the criteria to assess challenges to the policy might change. In particular, if we do not allow for an exception to a plan when we are already experts and do not incur the risk of brute shuffling, something has gone astray. Due to Prospective Opacity, we can't assess these changes in advance, hence the slow drift from rationality over time.

This example shows that retrospective regret comes in different degrees and can help us assess how rational we have been in the past. However, Bratman focuses on another notion of regret, which is anticipated rather than retrospective:

> Anticipated future regret does not simply provide evidence concerning one's standpoint; it helps shape the contours of one's standpoint. (Bratman, 2007, Ch. 12, p. 304)

The appeal to anticipated future regret also helps Bratman in answering Warren Quinn's *self-torturer puzzle*.[50] I take Quinn's puzzle to work partially in an analogous way to ODP; hence Bratman's reply to the former might turn out to be of interest for the latter.

Quinn's self-torturer puzzle is about rational choice: suppose that you have been offered to try a new medical device, which transmits a painful electric current at each of its settings. On a weekly base, you are allowed to either move from one setting to the next, or stop where you are. For each increment in setting, you will receive $10,000. The device starts from setting 0, to setting 1 up to 1000. For each new setting, the voltage of the electric current is unnoticeably increased, so that you won't experience any

[50] See Quinn, 1990.

difference in pain between voltage at 0 and at 1, at 1 as compared to 2, ... at setting s as compared to s+1. But the difference in pain between sufficiently distant settings will be clear: setting 1000 is so painful that you prefer to give up the whole fortune and go back to 0. Would it be most rational of you to accept or decline the offer? If you accept, at what setting should you stop? How can you ensure that you won't be tempted to go down the slippery-slope of moving to a new setting, given the unnoticeable difference in pain between adjacent settings, each rewarded with $10,000?

Quinn's puzzle is a puzzle about *slippery-slope intransitivity*: for each setting 0, 1, 2,... s, one prefers its adjacent next setting s+1, but this preference, though going down the slippery slope of preferring to constantly add a new setting, is not transitive because given settings x<y<z, it might not be true that if one prefers y to x and z to y, one also prefers z to x, as in the case of setting 1000 compared to setting 0.[51]

Why does the puzzle pose a problem specifically for Bratman's theory of intention? Because one could try to reply, in line with Bratman, that, given a prior intention to stop at a reasonable setting, the self-torturer sticks with this intention and resists the temptation of moving to yet a further step when the time comes. But how can he resist temptation if, when he reaches the setting he planned to stop at, he experiences a reversal of evaluative judgment? Suppose he planned to stop at setting 15, because, say, he prefers 15 to 0 but 0 to 16, hence he overall prefers 15 to 16. However, when 15 comes, he prefers 16 to 15, as illustrated by the puzzle, so what can prevent the self-torturer from going down the slippery slope?

> [...] at the time of his choice between 15 and 16, he can ask: "If I abandon my prior decision to stop at 15, what will then transpire?" and it seems he may reasonably answer: "I would then follow the slippery slope all the way to 1,000." His prior decision to stop at 15 was his best shot at playing the game without going all the way; if he does not stick with that decision, there is little reason to think he would stick with any other decision short of the bottom of the slippery slope. (Bratman, 1999, Ch. 4, p. 81)

[51] I owe this explanation of intransitivity to Andreou, 2014, p. 278.

This is not yet enough for Bratman to ground resistance to the temptation of moving to 16 once the self-torturer has reached 15. Why would moving to 16 mean that you go all the way down the slippery slope? Why moving to 16 is 'evidence' that you reach 1000 because of the 'slippery-slope-nature' of the puzzle?

> The answer seems to be that there is a kind of regret that is grounded in ranking of what would have resulted from certain past conduct as compared to with what has actually transpired. At the end of the day the self-torturer sees that, indeed, after choosing 16 he did go all the way to 1,000, and he sees that that would not have happened if he had stuck with his intention to stop at 15. If he had stopped at 15, he would, as a result, not have ended up at 1,000. He therefore regrets not having stuck with his intention to stop at 15. [...] It is the potential self-torturer's anticipation of such later regret that supports the argument that it may be rational for him to stick with his intention to stop at 15. (ibid, pp. 84-85)

Bratman's answer is not too far away from Quinn's own assessment of his puzzle:

> [The self-torturer] should be stopped by the principle that a reasonable strategy that correctly anticipates all later facts (including facts about preferences) still binds. On such a theory of rationality some contexts of choice fall under the authority of past decisions. In these contexts the Principle of Strategic Readjustment is suspended. *An agent is not rationally permitted to change course even if doing so would better serve his preferences*. (Quinn, 1990, p. 87)

Quinn doesn't take his puzzle to uncover a limitation of human rationality, but rather as calling for a revision of synchronic theories of rationality, based on the Principle of Strategic Readjustment. According to this principle, our strategies and intentions continue to have authority only insofar as they consistently support what is preferred overall by the agent. However, if our conception of rationality is reduced to this principle, it would be impossible to refrain from going down the slippery slope of the self-torturer puzzle, for each setting would compel us to readjust our conduct and move to a further setting.

It seems that Quinn's intents would be compatible with Bratman’s response to the puzzle, given his theory of rationality, which is grounded on the temporally extended nature of agents. Still, this might not be the whole story, for Quinn further writes that

> The self-torturer's predicament thus reveals a quasi-deontological aspect to a fully adequate theory of rational choice. (ibid)

What does this quasi-deontological aspect consist of? To my understanding, it is the possibility to bind our later conduct with a prior intention, even in cases of a reversal in preference. This quasi-deontological binding is not only an outcome of the ‘bare stability’ of our intentions,[52] but also our anticipation of later facts and of later regret. In a certain sense, we cast a project, a model or a pattern on our later conduct, able to bind us (or release us, if necessary).

However, notice that this quasi-deontological aspect comes from the perspective of a previous point in time and withstands rational scrutiny from that previous perspective. It doesn't refer to any timeless standard of rationality. That previous point in time is subject to Prospective Opacity: we weren’t sure about how stable our intention-plan-policy would have turned out to be (and always was), no matter how detailed our anticipation of facts and regret was. If this quasi-deontological aspect of our agency refers back to standards in the past, it will slowly drift, as time passes by, from how we would have acted if we had had the possibility at each point in time to readjust our intention.

Quinn’s puzzle shows precisely why readjusting at each point in time could have pretty nasty consequences. The self-torturer might end up at setting 1000 and strongly regret having even started the whole experiment. At the same time, a diachronic conception of practical rationality can't overcome Prospective Opacity and is subject to Drift from Rationality.

[52] With ‘bare stability’ I mean the minimal degree of stability necessary for an intention to coordinate our conduct over time, without adding further instrumental considerations for sticking with that intention.

In what respect is the self-torturer puzzle similar to ODP? I think that we could adapt the structure of Quinn's puzzle to the puzzle about stability. Suppose that each setting is actually a point in time. For every point in time that passes by, an agent increases her expertise in doing what is supported by her policy to an unnoticeable degree. The gain is only in terms of expertise and it is unescapable, once she starts complying with a policy over a considerable amount of time. While the change in expertise between t_0 and t_1 and of every t_n and t_{n+1} is unnoticeable, the expertise shift between two distant points in time might be quite remarkable. The level of expertise of an agent influences when it is rational to reconsider, and thus it influences the stability of her policy. Still, the time for becoming an expert in doing whatever one’s policy helps one to do is even more complicated to track than the right proportion between pain and monetary reward in the self-torturer puzzle.

In a certain sense, one can't plan to adjust one’s policy at t_{15}, for one would have no reason to choose that point in time over any other point at the time of the policy’s formation. At the same time, Quinn has stressed that synchronic adjustments are central for our intentions to continue to serve our preferences as time passes by:

> Policies need to be monitored to see if they are still serving our preferences. And when they are not, they need to be adjusted. (ibid, p. 85)

This is true, and it might seem to run against the whole idea underlining ODP. But this is simply because ODP is not directed at undermining any conception of diachronic rationality, nor at claiming that we can't ever readjust our policies. ODP is much more limited in scope and its point in much thinner. It says that our planning can't be as precise as synchronic planning: not because we can't foresee future facts and future regret, but rather because we are bound to the rationality at the time of the initial formation of our policy and we can't truly know in advance how stable it will turn out to be (and is at present). Of course we readjust our policies, and obviously we can monitor to a certain extent when our policy is going astray. I am not making the point that we act rationally by chance or not rationally at all.

All I claim is that our planning agency seems to be a matter of casting of patterns into the future from previous points in time in a way which is constrained by this previous perspective, with its respective, non-timeless rationality. Then we stick with this previous rationality, inescapably incurring a slow (almost unnoticeable) Drift from Rationality as time passes by. This might sound obvious, but exploring the details of this proposal is of interest for a planning theory of intention.

2.3.2 Two Concepts of 'Anchor'

Chrisoula Andreou has raised some doubts about Bratman's no-regret condition and his solution to the self-torturer puzzle. In a similar example, an agent can eat as many fun-size cakes as he wants, but he doesn't want to make a pig of himself. Eating one more fun-size cake won't make him a pig, but eating all of them will. Suppose further that the agent resolves to stop at the fifth cake. At this point, Bratman would say that "[...] if he does not stick with that decision, there is little reason to think he would stick with any other decision short of the bottom of the slippery slope" (Bratman, 1999, Ch. 4, p. 81). But is this the right conclusion to draw? Andreou doesn't think so:

> Bratman's reasoning about cases like the fun-size cake case suggests that, if the agent described abandons her plan, it is to be expected that she will proceed to make a pig of herself and end up wishing she had stuck to her plan. But this is questionable. [...] the agent in the fun-size cake case need not expect that, if she abandons her plan, she will proceed to make a pig of herself. There is, after all, the option and genuine possibility that, if she abandons her plan, she will – as seems rationally required, particularly if there is a defeasible requirement to avoid regret – seize one of the perfectly good upcoming opportunities to stop in good time. (Andreou, 2014, p. 279)

There are many such good opportunities, and previously deciding to stop at five can be seen more as a guideline, rather than as an imperative. As Andreou further argues, if the agent goes on and makes a pig of herself, she won't truly regret not having stopped at five cakes as she previously planned, but she will rather regret not having stopped "in the ballpark of five cakes". If she stopped at six cakes, everything would have been fine:

> [...] there is no simple route from the purported (defeasible) requirement that agents stick to intentions that satisfy the no-regret condition to the purportedly desirable verdict that an agent that finds herself in a case like the fun-size cake case should stick with her prior intention. (ibid, p. 280)

The interesting question is how agents can abandon their previous intentions but still act in a reasonable way, say by stopping before going down the slippery slope and not regretting. Again, Andreou provides us with an interesting response:

> [...] there need be no irrationality in the following situation: the agent forms a plan to stop at point *n*; she abandons her plan (even though there is nothing wrong with it); but she still stops in good time. My suggestion is based on the idea that, in at least some cases like the fun-size cake case, forming a specific plan can provide the agent with an effective anchor point, and so need not be seen as a waste even if the plan is abandoned. (ibid, p. 287)

We already encountered the notion of anchor with Bratman and his 'agential authority strategy' to assess temptation.[53] But I think Andreou and Bratman mean something quite different with this terminology. Bratman argues for higher-order prior policies, constituting our valuing, to function as an anchor for instrumental reasoning, i.e. as determining a certain ranking of values which is to a certain extent effective even in cases of temptation and reversal of ranking at a later time. When a policy constitutes such an anchor, it has agential authority, it shapes our standpoint as agents and the practical framework for our conduct over time. This also means that agential authority and support of coordination over time are strictly connected features of our agency for Bratman.

Andreou doesn't employ this meaning for her concept of anchor. Andreou's 'anchoring function' seems to be more a kind of parameter or criterion for our later conduct, rather

[53] See 2.2.2.

than ‘constitutive’[54] of our standpoint in time. When we settle on a policy, say, of always refraining from eating more than five fun-size cakes, but we sometimes eat six, sometimes seven and sometimes four, that is "always in the ballpark of five cakes", this only means that we see that policy as a point in the past that gives us a parameter, an anchor indeed, for our later conduct, without need for constantly reopening the issue and reconsidering how many cakes we should allow ourselves to eat.

Let's call Bratman's concept of anchor ‘forward-dropping anchor’ and Andreou's ‘backward-dropping anchor’. An intention-plan-policy works as a forward-dropping anchor when it gives a framework and shapes the standpoint an agent is to assume over time. A backward-dropping anchor, on the other hand, is the intention-plan-policy seen retrospectively, giving us criteria from the past to act at a later time.

These two concepts of anchor are not incompatible, they are rather two distinct aspects of our agency: the forward-dropping function fulfills the role of agential authority over time, whereas the backward-dropping function seems related to how the rational formation of our policies in the past constrains our later conduct in a non-inflexible way.

Now, these two concepts of anchor have different relationships to ODP. When one drops the forward-dropping anchor, one can do it with one’s present rationality and this anchor is going to shape one’s later conduct to a certain extent, given the rationality at the dropping time. The forward-dropping anchor is a practical anchor, shaping our conduct over time, but it is still subject to the epistemic limitation of Prospective Opacity and eventually incurs Drift from Rationality.

The backward-dropping anchor, on the other hand, is ambiguous. It could be both a practical anchor, in the sense that the dropping point coincides with the initial formation of a policy and it indirectly constrains our later conduct as a consequence of this initial

[54] Bratman himself uses the term: "For valuing to be such an anchor, it needs to *constitute* (at least in part) the agent's practical framework [...]". (Bratman, 2007, Ch. 12, p. 270, added italics)

dropping. Or it might be an epistemic anchor, in the sense of offering us a remarkable point in the past, from which we slowly drift and whose epistemic assessment can be made only retrospectively. Andreou seems to hint at a practical reading rather than at an epistemic one:

> [...] the plan helps by (1) controlling the agent's conduct so long as the plan remains in place and comes to mind at critical junctures, and (2) by creating an anchor point that can continue exerting influence even if the plan is abandoned. (ibid, p. 289)

The influence most likely has a practical nature, and so it seems that her backward-dropping anchor is also a practical anchor.

It is interesting to note that the lack of inflexibility of anchor points do not only serve the rational purpose of letting the agent off the hook of irrationality if they are not rigid enough. They are also compatible with ODP's Drift from Rationality, because a backward-dropping anchor might let one float adrift, in good time for readjusting one's policies to better suit what is rationally expected of one as time passes by.

CHAPTER THREE

Three Objections

In the first chapter I have given a general outline of my reading of Bratman's philosophical understanding of stability of intentions, plans and policies, and proposed a puzzle about this interpretation. In the second chapter I have gone through the details of my proposal and tried to grasp its extent and its significance with respect to Bratman's theory and later developments, as well as with respect to some of his critics' remarks. So far I have not considered serious objections to my reading of Bratman. That is the task of this third chapter.

Criticism to my account can be divided into three main lines: 1) against the plausibility of the Vienna Marathon Reiterated example and against the compatibility of the condition 'other things equal' with the condition that considerable amount of time has passed; 2) against my interpretation of Bratman's conception of stability, which could sound like irrational habit worship; 3) against the overall picture of agents that seems to emerge from my discussions, that is of inertial agents, who have a hard time putting together their diachronic and synchronic rational control on their conduct.

I consider these three lines of criticism to be very effective and cannot ultimately settle all the issues raised. However, I will still respond to each of them, at least in the attempt to better illustrate my point.

3.1 Three Objections

3.1.1 Omission

Recall Vienna Marathon Reiterated. The example is based on two crucial points: 1) a considerable amount of time has passed from the initial formation of the policy of running thrice a week to train for the marathon; 2) no relevant cognitive change has taken place from the initial formation of the policy. With this example, I wanted to show that something must change in the assessment of its rationality, if the same challenge occurs to the very same policy at its initial formation as opposed to when it has become a well-established policy. However, other things equal, nothing changes apart from our uninterrupted sticking with that policy. Hence, so I argued, without further resources of introspection, self-knowledge and so on, we end up following the policy as if it were to incur the same assessment of rationality that took place toward its initial formation. This would have meant that we slowly drift from rationality, because something has changed in the meanwhile without us noticing. I have explained this drift in terms of a puzzle about the stability of intentions, plans and policies (ODP), for models of planning agency such as REM, limited to a broadly instrumental conception of rationality and norms such as means-end coherence, consistency of intentions and beliefs, and stability over time.

The Vienna Marathon Reiterated example illustrated ODP on a macroscopic level, for it would otherwise be quite difficult to notice the almost unnoticeable Drift from Rationality – recall the analogy with the self-torturer puzzle.[55] Nonetheless, there can be a strong impulse to question the validity of this example. Is it plausible that a considerable amount of time by itself doesn't imply changes also on the cognitive level? Are we sure that we are not presented with new information? Is the condition 'other things equal' genuine?

Someone might think that a drift from rationality occurs only if my relevant beliefs and desires have changed without respective adjustments in my policy. Then, if I think that after a considerable amount of time an agent drifts from rationality, I am omitting that something has changed in her beliefs and desires and the condition 'other things equal' is *ad hoc* and forceful, if not directly implausible. We might collect all these doubts and formulate them with the following objection:

[55] See 2.3.1.

> *Omission:*
>
> *The condition 'other things equal' in the Vienna Marathon Reiterated example is not genuine: considerable amounts of time imply by themselves changes other than the passing of time* simpliciter. *In my example, I would be* ad hoc *omitting relevant cognitive changes that would be the true cause of a change in assessment of the rationality of the policy at stake, rather than the purported Drift from Rationality.*

Omission questions three things: 1) the genuineness of the Vienna Marathon Reiterated example, 2) the explanatory power of ODP, 3) the relevance both of the example and of ODP in saying anything interesting about stability.

Before going into the next objection, I anticipate that I accept part of this criticism. I am inclined to consider the Vienna Marathon Reiterated example as a quite artificial example. At the same time, I do not hold its artificiality to undermine the theoretical purposes of the puzzle it is supposed to illustrate. Moreover, both points 2) and 3) objected by Omission are based on a general misunderstanding of the meaning of ODP. Thus, I will try to show that Omission partially misfires and doesn't really pose a threat to my account.

3.1.2 Habit Worship

Another problem with my way of illustrating ODP is that it could be based on a misconception of the rational demand of stability. It could be objected that the agents I am describing are bound to the stability of their prior intentions in such a way that their later reflectiveness would only exceptionally decide their rational course of action. On the contrary, agents appear to be much more in control of themselves: we are self-aware and do not function only on the underlying mechanisms of certain habits of reconsideration/blocking. The excessive weight I would be granting to stability also goes against Bratman's own purposes, for he often remarks the priority of present deliberation over nonreflective mechanisms. This second objection can hence be articulated as a form of habit worship:

Habit Worship:

> *For the puzzle to work, I am giving priority to the stability of plans and policies over the rationality of present deliberation. This is contrary to the spirit of Bratman's planning theory of intention, which is not a theory of irrational stubbornness, but a theory of cross-temporal rational agency.*

My reply to this objection will be based on Edward F. McClennen's criticism of Bratman.[56] Bratman’s attempt to save his account of stability from the possibility of habit worship should accept that a planning conception of agency is bound, to a certain extent, to rely on previous *resolutions* and that, other things equal, those resolutions determine our conduct independently of our active, constant control. This means that the accusation of habit worship can only be resisted to a certain point, and that the practical rationality of planning agents grounds its strength on certain commitments, with the downside of some kind of worship.

3.1.3 Fatalism

An interpretation of ODP might imply that our agency over time is some sort of inertial side-effect of certain initial dispositions. We are destined to act only upon approximations, for we’ll never truly know in advance how stable our own intentions are. If this interpretation of ODP is sound, it would not produce a very reassuring picture of agency. Like self-blind programs,[57] our mental states would bring us to action and our reflectiveness would just be popping up here and there to restart the programs so that we can go on with this passive self-determination. In other words, our synchronic reflectiveness would ensure authority and monitoring but be at the same time quite ineffective, whereas our intentions, plans and policies, while bringing us to effective action, would be almost inertial states of mind, supported by nonreflective mechanisms

[56] See DeHelian, McClennen, 1993.

[57] I use here the term ‘self-blind’ as introduced by Sydney Shoemaker (in Shoemaker, 1996), which is roughly meant as the impossibility to have direct access (being self-acquainted with) one’s mental states. (see in particular Shoemaker, 1996, pp. 30-31)

such as habits of (non)reconsideration. Is this a plausible picture? Isn't it in contrast with Bratman's philosophy?

This interpretation counts as the following objection:

> *Fatalism:*
>
> *If the puzzle is pervasive for rational agency (in REM), it could mean that agents with limited resources inescapably (fatally) tend to drift from rationality over time, precisely because they stick with their long-term intentions, plans and policies. This would also mean that we do not truly control our conduct over time: we implement in our brains programs of which we lack prior understanding, and are then determined by conative mechanisms we can only assess retrospectively. This seems to partially reduce agents to mindless, zombie-like creatures. We arguably have good reasons not to understand agency in this way.*

I take this objection to be of great interest for understanding both the conception of agency emerging from REM and that emerging from Bratman's more articulated model. As a preliminary remark, I shall admit that Bratman's philosophy doesn't seem to be committed to a single and unitary depiction of what an agent is to be. Bratman is more concerned with working out the details of those structures of our mental states that make us agents endowed with practical rationality. The explanatory role of these structures encompasses working out the many sides of intentions, plans and policies, and being of use for the widest range of cases possible. Bratman's account, with its various developments, has remarkably succeeded in overall conferring to his conception of planning agency such multi-faceted role, by maintaining a quite parsimonious core at the same time.

ODP attempts to ask whether, after all, there might still be some kind of unitary idea of agency underlying Bratman's philosophy. This would be particularly elusive, and it is easy to improperly attribute to Bratman ideas that are not implied by his theory. At this point, I developed REM, to better control the general idea of what I have called his ‘early philosophy’. ODP has tried to show that, given REM, our practical rationality appears to be (at least in part) of the inertial kind, and that there is an important difference between reflective-synchronic control and nonreflective-diachronic control. I have further tried to show that this puzzle can apply to later developments of Bratman's philosophy at different degrees. But have I found a unitary idea of Bratmanian agency? And if so, doesn't this idea face the objection of Fatalism?

While I don't think I have been able to draw any such unitary picture of agency, I have done something interesting, as I will argue. I think that my reading of Bratman, through the puzzle’s lenses, can offer us an intuition about how our practical rationality works. I take this intuition to already be present in Bratman, and my work has aimed at making it explicit. Still, I do not deem ODP to go any further than grounding this intuition, and thus I will argue that Fatalism is not the correct way to interpret my puzzle and is hence to be rejected.

3.2 Three Responses

3.2.1 The Usefulness of an Artificial Example

The Vienna Marathon Reiterated example is to a certain extent an artificial example. It assumes that a considerable amount of time has passed from the initial formation of a policy without relevant cognitive changes. It is artificial because an agent who starts complying with a policy to train for a marathon and goes running weekly for a long period is most probably self-conscious of his improvements, by forming new beliefs and checking his condition. Becoming better at running is his purpose after all, isn't it? He might also experience an increased or decreased desire to run the marathon depending on his training or some random occurrence in his life. The more time passes by, the more probable it becomes for something to happen in his life and influence his policy in unpredictable ways.

I take these considerations to be sound and draw from them the conclusion that the example is artificial. However, this doesn't mean that it doesn't illustrate the general point. In fact, we could imagine such an agent, who forms a policy and sticks with it for a considerable amount of time and add the artificial condition that other things stay equal. The question becomes the following, given that the agent is not presented with relevant new information or cognitive changes: Will the rationality of the stability of his policy in terms of (reasonable) resistance to reconsideration stay the same over time?

The answer I gave was "no, it changes". But someone might resist this reply. An agent complying with the policy of running thrice a week can't be blind to his progress. He must know what he is up to and he monitors his activities and intentions. He is self-aware and reflective. Yes, but what is the object of his reflection? And what are the limits of this reflectiveness? It seems to me that this reflectiveness can only be at present and concerned with weighing desires given certain beliefs, but can't exert the same control over the stability of an intention, especially if that intention comes in the form of a policy covering a long period. Moreover, in REM at least, once we reflectively form a policy, given no unanticipated information, we need very good reasons to trigger reconsideration and hence seriously reflect on that particular policy. In a model such as REM, we can't reflect randomly or constantly on that policy, but only when it is time for us to reflect. Here it could be helpful to distinguish two meanings of reflectiveness: the first one is that of having direct access to one's mental states, in terms of avoiding self-blindness[58] – call this *basic reflectiveness*; the second one is that of having mental states about mental states, that is higher-order mental states (in the two Bratmanian variations of weak and strong reflectiveness)[59] – call this *complex reflectiveness*. If basic reflectiveness is about something constitutive of every instant of the reflective life of an agent, complex reflectiveness describes only some of our mental states. If we give priority to this second meaning, as I am currently doing, agents are not constantly (complexly) reflective, but only at salient points, for instance when forming a new intention. Afterwards, nonreflective mechanisms support our following through with

[58] I employ Shoemaker’s use of ‘self-blindness’ again (see previous footnote).

[59] See 1.3.3 and Bratman, 2007, pp. 23-24.

what we intend, even if we remain self-aware (basically reflective) in the sense of having constant access to our psychic life.

Then, other things equal, with a model of planning agency like REM, we can obtain policies alone and ask ourselves the question of whether the assessment of the rationality of their stability stays the same or incurs alterations as time passes by. Even if we could never be estranged from our mental states (be self-blind), we slowly drift from the rationality we had at the initial formation of a policy, for we can't but stick with our initial stability as time passes by. If we have mechanisms monitoring reconsideration, those mechanisms do not change and do not adapt to the ageing of the policy: if other things are equal, the ageing of the policy is determined precisely by the absence of change in those underlying mechanisms when the policy itself should instead change as the level of expertise changes. But isn't this absurd? Of course we adjust our policies if we spot a difference in how we comply with them!

The issues at stake here are more complex. What is meant by "complying with a policy", is the stability of that policy, and Prospective Opacity conflicts with its accessibility. In this sense we can't spot a difference in how we comply with a policy, when that means realizing that the stability of our policies shall be updated without any new relevant belief. We lack such transparent epistemic grasp of stability. Secondly, adjusting a policy is something more serious than toying with a policy: it involves rationally weighing desires with respect to relevant beliefs, costs and benefits of different sorts, and so on. Other things equal, we do not adjust a policy if we are not provided with good reasons to do so.

The artificiality of the Vienna Marathon Reiterated example is supposed to abstract policies from their highly complicated and enmeshed context, and ask: If other things such as beliefs, desires, and so on, were equal, would policies be also equal? Does ‘other things equal’ mean ‘policies equal’? Are policies timeless when not defeated by other factors? My answer is no. And I have tried to show that even under the condition ‘other things equal’, policies tend to deviate from how rational they were at the beginning of their formation, if ‘left on their own’. Of course reality is more complicated, but sometimes understanding reality requires useful abstractions and artificial simplifications.

An interesting variation of the Vienna Marathon Reiterated example is to suppose that, instead of being as rigid as required by the stability at the initial formation of the policy, after a considerable amount of time, the agent allowed the exception and went to the library, passed the minor exam, and did not incur significant losses in his marathon-related objectives.[60] Allowing this exception would mean manifesting habits of reconsideration, which track a relevant trigger of blocking of the policy for that particular occasion. However, diachronic stability implies that those habits were present all along from the initial formation of the policy. That is, the marathon runner would also have allowed for the same exception at the beginning of the policy’s formation. But this would have exposed him to the risk of brute shuffling and thus to irrationality.

Recall that, in the Vienna Marathon Standard case, it could have been completely rational of the agent to resist the challenge and go running instead of going to the library: not passing a minor exam was, after all, not as costly as risking to give up the policy of training thrice a week. That risk was real and threatening at the initial formation of the policy, and a reasonable mechanism of reconsideration would have resisted the challenge. After a considerable amount of time, however, we change our practical standing with respect to the challenge and can allow for an exception. But how, if the mechanism of reconsideration is the same regulating the Standard case? Other things equal, it seems that we just stick with that mechanism and do not reconsider/block. On the other hand, if those mechanisms were to allow an exception in the Vienna Marathon Reiterated case, then they would be bound to allow the same exception in the Standard case, and expose the agent to the risk of brute shuffling. Perhaps there would have been absolutely no Reiterated case if those habits of reconsideration/blocking had allowed the exception at the initial formation of the

[60] My sister Annalisa brought the interest of this variation to my attention.

policy, for the agent would have given up the training by giving in to too many such minor challenges.

3.2.2 Toxin and Resolute Choice

While addressing Bratman’s treatment of Smart's problem, I claimed that ODP could have been understood as compatible with both a form of habit worship and Bratman's final assessment of the issue. Smart's problem is about the dilemma of sticking with the prescriptions of a general habit when it would be rationally required to act differently in a particular case. This is a form of irrational habit worship, but Bratman claims that it doesn't apply to his two-tier account, for if it is obvious to you that in your case you should reconsider your plan and not act as prescribed by the habit, then the habit shouldn't ‘stand in the way’ of reconsideration, and the two-tier account shouldn't apply.

The puzzle about stability is compatible with Bratman's answer for it applies only to cases of nonreflective non-reconsideration. At the same time, it shows that we are bound to the initial stability of our intentions-plans-policies (determined by respective habits of reconsideration), which is some kind of habit worship. This sounds like an odd story. Accepting habit worship, even partially, appears to be in contrast with Bratman's conception of rational agency. Bratman writes extensively about his worry that agents might be excessively rigid and negate the priority of maximizing utility at present.[61] Instead, in order to be rational, agents should always be able to maximize utility and, if they sometimes stick with their prior intentions against a temporary evaluative shift, that must also serve the purpose of utility maximization – mostly in terms of satisfaction of desires.

Some authors do not share Bratman's worries.[62] They argue for resolutely sticking with one's prior intentions other things equal and they are not concerned with the possibility

[61] See for instance Bratman's writings on temptation in Bratman, 1999 and Bratman, 2014. Here I am specifically referring to Bratman, 1992, but we could even deem part of Bratman's work on authority and self-governance (see esp. Bratman, 2007) as an attempt of defusing the threat of irrational habit worship.

[62] I have in mind especially DeHelian, McClennen, 1993, Holton, 2004, Mintoff, 2004.

of maximizing utility temporarily: what counts is the overall rational gain, including that of intrapersonal coordination.[63] To put it differently, if an agent is confident at t_0 that sticking with his resolution formed at t_0 will lead to the overall better outcome, and at t_1 he is not presented with new relevant information, even if it turns out to be temporarily rational for him to give up the resolution, the agent should still stick with it, other things equal. This view, however, encounters a problem illustrated by Gregory Kavka's ‘toxin puzzle’.[64]

Suppose an eccentric billionaire offers you one million dollars if you form the intention before midnight today, to drink a toxin tomorrow afternoon. It will make you painfully sick for a whole day, but won't threaten your life or have lasting consequences. The money will be transferred to your bank account tomorrow morning, that is, after you have formed the intention to drink the toxin, but before you actually drink it. All this information is very reliable, and the *bona fide* of the billionaire is beyond doubt.

Suppose further, that at midnight you will be tested by a brain scanner, able to track with absolute precision your relevant intention to drink the toxin. Depending on the result of the test, you will either receive the million dollars or fail. Drinking the toxin "[...] will not be pleasant, but it is sure worth a day of suffering to become a millionaire" (Kavka, 1983, p. 34), or so you think.

What the puzzle illustrates, is that you might truly be tempted to form the intention to drink the toxin tomorrow afternoon, but not to actually drink it. After all, the billionaire doesn't expect you to drink the toxin tomorrow, but only *to intend today* to drink it tomorrow. So, if you were able to intend to drink the toxin without actually drinking it, you would receive the money you desire and avoid drinking the painful toxin. But can you? Of course, if you aim to intend today to drink the toxin and to stop intending tomorrow to drink the toxin once the money are already on your bank account, you

[63] Along with interpersonal coordination, but we leave this issue aside, for it is not of concern to our current discussion.

[64] See Kavka, 1983.

wouldn't be truly intending to drink the toxin. You would fail the test of the brain scanner and the million dollars would remain a dream.

Kavka's toxin puzzle rules out the possibility of finding external incentives to bring oneself to drink the toxin tomorrow, such as binding oneself with a promise or a legal agreement, or even hiring a hitman with the order to kill in case the toxin is not drunk. Hypnosis, forgetfulness and self-promising are also possibilities out of reach.

It seems that one must bring oneself spontaneously and clearheadedly to drink the toxin tomorrow, but that task is far from easy:

> You are asked to form a simple intention to perform an act that is well within your power. This is the kind of thing we all do many times every day. You are provided with an overwhelming incentive for doing so. Yet you cannot do so (or have extreme difficulty doing so) without resorting to exotic tricks involving hypnosis, hired killers, etc. Nor are your difficulties traceable to an uncontrollable fear of the negative consequences of the act in question – you would be perfectly willing to undergo the after-effects of the toxin to earn the million. (ibid, p. 35)

Bratman gives an original solution to this puzzle.[65] Recall that for Bratman, when forming an intention, it can be within your power to anticipate the future regret of your actions. Hence, if you think that intending to drink the toxin is tightly linked to regretting having drunk the toxin in the future, it can't be rational of you to intend now to drink the toxin later. Bratman is worried that sticking with one's prior intention might be irrational of the agent, when the time comes to drink the toxin: it would be a form of irrational intention worship.

Tomorrow morning, the million dollars will have already been transferred to your bank account and you will lack an important reason to drink the toxin in the afternoon. How could you then drink the toxin if you have no reason to drink it anymore? The crucial point for Bratman seems to be that, if your initial intention to drink the toxin was to be taken deliberatively and be rational, you would see that you can't rationally intend to

[65] See especially, Bratman, 1987, pp. 101-106, and Bratman, 1999, pp 58-90.

drink the toxin because you will lack a reason to drink it afterwards. Moreover, if you were to drink it, you would regret having intended to drink it in the first place. Thence, you shouldn't accept the billionaire's offer.

McClennen doesn’t agree.[66] If the toxin example is a case of no-unanticipated-information, then you should stick to your guns and drink the toxin – if that is the most rational thing to do in order to secure the greatest overall rational gain, without implying any form of irrational intention worship. Paraphrasing Menahem E. Yaari, McClennen writes:

> [...] in the presence of a disparity between one's preferences at different times, if one chooses so as to maximize with respect to whatever preferences one has at each point in time, one will end up doing less well, as measured by the satisfaction of those very same time-defined preferences, than if one had managed to effect some sort of coordination of one's temporally disparate choices. (DeHelian, McClennen, 1993, p. 325)

In other words, McClennen maintains that there is a rational priority of the prior intention, if sticking with it would mean securing major overall gains. This is an account of *resolute choice* that, according to McClennen, doesn't involve irrational intention worship or habit worship. Indeed, McClennen claims that his own account is similar to Bratman's account. The main difference is that in obvious would-change/worth-it cases, where it is clear that one should reconsider an intention, habits of reconsideration could still apply, contrarily to what Bratman holds:

> [...] habits or dispositions of non-reconsideration can be controlling even in what is obviously (from the perspective of an incremental assessment of consequences) a would-change/worth-it case, if it is clear that the gains to be secured by reconsideration would not have been possible if the agent had not been clearly disposed not to reconsider. (ibid, p. 330)

On the one hand, Bratman argues that McClennen's account is vulnerable to Smart's problem and, with it, to the threat of habit worship; on the other hand, McClennen

[66] The paper I am referring to here was written by Laura DeHelian and Edward F. McClennen (1993). For simplicity's sake, however, I will refer to the author as McClennen, for he is the main target of Bratman's criticism.

claims that his account is similar to Bratman's and that, given no-unanticipated-information, one shall stick to one’s guns if that means securing the greater rational gain.

As far as I can see, both authors are right and this dispute can be settled from the perspective of ODP. What I find convincing in McClennen is the idea that one isn’t suddenly tempted at t_1 to reconsider one’s resolute choice taken at t_0, *other things equal*. If getting the million dollars is indeed what one desires overall the most, one should resolutely intend to drink the toxin. Joe Mintoff has put this idea quite clearly:

> [...] it is plausible to claim that persons are being irrational if they persist with a rule that they *come to believe* would have them perform a non-maximizing action, but it does not follow from this that persons are irrational if they persist with a rule that they *believe* (right from the start) would have them to do this. (Mintoff, 2004, p. 411)

Drinking the toxin remains terrible and, luckily, we do not ordinarily experience toxin puzzles. But if we were to, it could even become rational to drink pointless toxins. This is Richard Holton’s argument:

> Suppose that we lived in an environment in which almost every decision had the form of the toxin case. Suppose that, for his own mysterious ends, a perverse god arranged things so that the necessities of life were distributed to those who intended to endure subsequent (and by then pointless) suffering. Imagine how we would bring up our children. If resolute commitment to such intentions were really the only way to form them, that is just what we would encourage. We would inculcate habits of nonreconsideration of resolutions even when their benefit had already been gained, and there was only avoidable cost to come. Such habits would, I suggest, be perfectly rational, since we would go on benefitting from them. (Holton, 2004, p. 529)

I side with McClennen, Mintoff and Holton on this issue, especially with respect to the basic intuition that, contrarily to what Bratman holds, the two-tier model can be extended to resolutions. However, tomorrow morning one will see that drinking the toxin is a difficult task to accomplish, given that one has already received the money.

Bratman has a correct suspicion about the resolute-choice account: there is something odd in sticking with the rationality at the beginning of the formation of a new intention-plan-policy over time.

This thought is captured by the toxin puzzle in a straightforward way, but also by ODP from another point of view. Indeed, once one drops the anchor of a new intention, one's later conduct is bound to the rationality of that prior decision, other things equal. But time flows and, with models like REM or some version of the resolute-choice account, a kind of habit worship is inevitable. At the same time, I would be inclined to say that this particular kind of habit worship is not necessarily irrational: it only implies a slow drift from how rational we were at the initial formation of a new intention or policy. It is precisely by sticking with the previous stability that we drift from rationality, and here is the puzzle.

3.2.3 Interpretation of the Puzzle

Fatalism is an intriguing objection. It asks whether I think that an agent works like a train or an ice skater, with an initial impulse (the intention), and then by inertially following through. In the meanwhile, however, we wouldn't be following the laws of mechanics or physics, but the norms of our own practical rationality: means-end coherence, consistency and stability. Contrarily to the laws of mechanics and physics, which regulate the whole inertial movement of a train or the leap of an ice skater, this rationality is limited to the initial impulse, i.e. to the initial formation of a new intention. We slowly drift from initial rational standards and norms as time passes by, because complying with an intention (but especially with a policy) over a considerable amount of time, means complying with its original rules independently of the passing of time, other things equal.

But again, wouldn't this idea transform us into derailing trains or mindless zombies with sudden moments of clear-headedness? If 'derailing' means generally limited, we might well be slowly derailing, but that is also quite unsurprising. If by 'derailing' we mean less rational, I would not be sure how to answer. Human agents are agents endowed with practical rationality, and this rationality often seems to be reliable. I definitely do not question this. What I have questioned is that the same expectations about synchronic rationality can also be held with respect to the diachronic case. Here we should acknowledge that the passing of time alone plays a crucial role for the assessment of the rationality of our mental states. I have claimed that there is a tendency to drift from the initial reflective impulse. On the other hand, this is far from saying that we are mindless zombies. When we follow through with a policy, we are not giving up our mindedness to become unconscious automatons or blind executors of intentions. We simply rely on certain underlying mechanisms, without constant reconsideration.

ODP doesn't tell a story about our mindless moments. On the contrary, it offers an intuition on how to understand our mindedness: it is not a timeless moment of reflectiveness or perfect retention of rationality, but a complicated mixture of impulses, inertial mechanisms, anchors, and moments of reflectiveness, at different degrees. Our practical rationality is bound to all these mental features, it is constrained by their temporal nature, and can also become effective in virtue of their automation.

Still, one might find the idea of a drift from rationality too counterintuitive for our planning activity. Makowski makes this point quite straightforwardly:

> Our mental resources are of course limited, but it does not mean that we should not be able to use our inventiveness and cognitive flexibility when dealing with demanding situations. It's true that some of us do it better and others do it worse. Nonetheless, our mental economy is generally adequate to enable us to respond to such difficulties. Otherwise, we would have to treat our plans as *traps* whenever the planning environment fluctuates in an unpredictable way. (Makowski, 2016, p. 1052)

It is true that we are very flexible agents, and that our practical rationality is more than adequate to respond to ordinary difficulties. It is also true that we do not ordinarily treat our plans as traps. On the other hand, however, a theory of rational agency over time has made manifest the importance of those underlying mechanisms, which sustain our cross-temporal coordination in a nonreflective way. Thanks to these mechanisms, we are able to inherit the rationality of the past, and our intentions maintain a certain stability, allowing us in achieving effectively our goals and what we desire.

This implies something interesting. Perhaps, plans can't straightforwardly be understood as traps, but there are particularly sophisticated traps, which indeed offer an analogy for the nondeliberative, reason-preserving aspect of plans: spiderwebs. Spiderwebs are not passive traps, like leg-hold traps, body gripping traps, deadfall traps or snares. These latter passive traps are triggered by certain conditions and they passively react. On the other hand, spiderwebs can be thought of as extensions of the spider, and directly connected to its monitoring of what is happening on the web. When an insect gets tangled in the web, the spider is suddenly informed and acts accordingly. Depending on the kind of vibration that is transmitted to it, the spider will either rush to catch its prey or ignore the signal, for it might just be a dry leaf brought by the wind and moving to ‘go and see’ would be a waste of energy at best.

Our mechanisms of retention work similarly to spiderwebs, and the poor vision of many web-weaver spiders may offer an analogy for Prospective Opacity. Spiders rely mostly on their responsiveness to vibrations coming from their webs, rather than on direct visual observation to catch their prey, just like our nonreflective habits of reconsideration track triggers of reconsideration without our direct reflective observation. Our habits trigger reconsideration like when the spider is triggered by an actual prey, or they can sustain resistance to reconsideration, like when nothing at all happens, or in the case of a false alarm, and the spider can keep waiting. Furthermore, spiderwebs are defeasible; they are built up and broken down on a regular basis, depending on their state of deterioration, just like we sometimes abandon certain plans and form new ones, thanks to their defeasibility. We might also draw a parallel between the impact of relevant cognitive changes or substantive new information on our plans and the destruction of the spiderweb because of a distracted passer-by or particularly strong wind.

In all these cases, the spider is always present and the spiderweb is never a passive trap. The spider (together with the spiderweb) represents our mindedness, even if not our full-blown reflectiveness, which is reached only in the deliberative synchronic case. It is a mindedness made of basic reflectiveness, responsiveness to triggers and mechanisms to control our conduct, but in no way it resembles the mindlessness of zombies.

To push the analogy between nonreflective non-reconsideration and spiderwebs a bit further, think about the first spiderweb constructed by a young spider. In its first hunting days, the spider receives a signal from the web, but it is unsure as to whether it is an insect or a dry leaf. This time, it will pass on, for it prefers not to risk making mistakes while still an inexperienced hunter. For our analogy's sake, this would equal sticking with the policy-based intention to go running in Prater when faced with the possibility to go to the library instead. Just like the spider resists going to check what has landed on its web, so our marathon runner resists reconsidering/blocking his policy-based intention.

Suppose now that the same spider, after months of training, is presented with the very same challenge. If it has survived so far, we can assume that it wasn't by chance, but because it has proven its hunting ability. What will it do now? Will it act in the same way as in the first case? How can we assess its situation or the practical rationality that would be required to act?

If these questions are meaningful, I think that my puzzle has been successful in tracking an interesting perspective on planning agency for models like REM.

CONCLUSION

My aim has been to give a reading of Bratman's planning theory of intention through the lenses of a puzzle about intention stability, ODP. I have shown how ODP arises within this theory and captures a subtle aspect of planning agency, that is, its inertial side, implying a slow drift from rationality over time. I have achieved this objective by employing the Vienna Marathon Reiterated example, which can be criticized for its artificiality, but still serves its purpose. In depth attention has been paid to Smart's problem and the risk of understanding our agency as a sequence of habit worship acts. My final assessment of this dispute is that a specific kind of habit worship is unavoidable, but not the kind, which is rejected by Bratman. Instead, I have claimed the compatibility of ODP with his theory. Problems of authority and self-governance have also been of interest for assessing ODP and vice versa: the puzzle could also apply to self-governing policies themselves, if we do not understand them as almighty in my technical sense.

The central aim of these pages has been to disclose a particular perspective on planning agency. We are practically rational agents, who attempt and often succeed in being efficient despite our limited resources. However, a particular sense of transience comes with our intentions: we form intentions and then follow through, not necessarily in a reflective way. Efficiency requires that we form further intentions, make other plans, develop new habits. The role intentions play in settling further planning is the reason why forming an intention often means that we do not linger there, staring at it unfolding, but rather already form new intentions and not look back. The new intention starts shaping our conduct from the point of its formation, and we move on. Meanwhile, the intention is being sustained by underlying mechanisms extending our rational control over time, but, as Bratman asks, “is this thought a way of clothing the dead hand of the past with the cloak of purported rationality?” (Bratman, 2012, p.74).

The matter is more complicated and the problems more subtle than they might appear at first. Earlier intentions influence our conduct as dead hands from the past, and yet they seem to be rational. However, if we shift our point of view on agency, we can find a coherent picture again, by unlocking interesting aspects of agency, its opaque spots, and its drifts.

BIBLIOGRAPHY

Andreou, Chrisoula (2014) ‘Temptation, Resolution, and Regret’, *Inquiry*, 57:3, pp. 275-292.

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

Bratman Michael E. (1992) ‘Planning and the Stability of Intention’, *Minds and Machines*, 2, pp. 1-16.

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

Bratman, Michael E. (2007) *Structures of Agency*, New York: Oxford University Press.

Bratman, Michael E. (2010) ‘Agency, Time and Sociality’, *Proceedings and Addresses of the American Philosophical Association*, Vol. 84, No. 2, pp. 7-26.

Bratman, Michael E. (2012) ‘Time, Rationality and Self-Governance’, *Philosophical Issues*, 22, pp. 73- 88.

Bratman, Michael E. (2014) ‘Temptation and the Agent’s Standpoint’, *Inquiry*, 57:3, pp. 293-310.

DeHelian, Laura, McClennen, Edward F. (1993) ‘Planning and the Stability of Intention: A Comment’, *Minds and Machines*, 3, pp. 319-333.

Frankfurt, Harry (1971) ‘Freedom of the Will and the Concept of a Person’, *The Journal of Philosophy*, 68:1, pp. 5-20.

Frankfurt, Harry (1992) ‘The Faintest Passion’, *Proceedings and Addresses of the American Philosophical Association*, 66:3, pp. 5-16.

Holton, Richard (2004) ‘Rational Resolve’, *The Philosophical Review*, 113:4, pp. 507-535.

Kavka, Gregory S. (1983) ‘The Toxin Puzzle’, *Analysis*, 43:1, pp. 33-36.

Makowski, Piotr Tomasz (2016) ‘Intention Inertia and the Plasticity of Planning’, *Philosophical Psychology*, 29:7, pp. 1045-1056

Mintoff, Joe (2004) ‘Rule Worship and the Stability of Intention’, *Philosophia*, 31, pp. 401-426.

Muckle, Robert J. (2010) ‘Self-Governing Policies: A Critique of Bratman’, *Res Cogitans*, Vol. 1: Iss. 1, Article 19, pp. 156-162.

Quinn, Warren S. (1990) ‘The Puzzle of the Self-Torturer’, *Philosophical Studies*, 59, pp. 79-90.

Shoemaker, Sydney (1996) *The First Person Perspective and Other Essays*, Cambridge: Cambridge University Press.

Smart, J. J. C. (1956) ‘Extreme and Restricted Utilitarianism’, *The Philosophical Quarterly* (1950-), Vol. 6, No. 25, pp. 344-354.

Tenenbaum, Sergio (2016) ‘Reconsidering Intentions’, *Noûs*, 10.1111/nous.12160, pp. 1-30.

ABSTRACT

English

Michael E. Bratman offers an account of the stability of intentions in terms of their reasonable resistance to reconsideration. The stability of our intentions allows coordination over time in a rationally effective way to fulfill our desires and achieve our goals. The condition for our diachronic coordination is that prior intentions transmit our earlier rationality to our later conduct. This understanding of intention stability raises following problem: Is the rationality of prior intentions timeless and transmitted without alterations or does it slowly 'weaken' as we retain an intention formed in the past? I defend this latter idea on two main grounds: 1) due to Prospective Opacity we can't see clearly how stable a newly formed intention is and will continue to be at the time of its initial formation; 2) because of the Drift from Rationality, other things equal, we tend to stick with the rationality of our prior intentions over time, even if synchronic rationality at later times would have suggested acting differently. These two issues cumulate in the Opacity-Drift Puzzle, which offers a specific reading of Bratman's planning theory of intention.

ABSTRACT

Deutsch

Michael E. Bratman erklärt die Stabilität von Absichten als einen vernünftigen Widerstand gegen Neuüberlegen. Die Stabilität unserer Absichten erlaubt es uns, unsere Handlungen über die Zeit hinweg in einer rational wirksamen Weise zu koordinieren, um unsere Wünsche zu erfüllen und unsere Ziele zu erreichen. Voraussetzung für unsere diachronische Koordination ist, dass vorherige Absichten unserer früheren Rationalität auf unser späteres Verhalten übertragen werden. Diese Auffassung von der Stabilität von Absichten führt zu folgendem Problem: Ist die Rationalität von vorherigen Absichten zeitlos und unverändert übertragen oder wird sie ‚schwächer' mit der Zeit, während wir unsere Absichten beibehalten? Ich verteidige letztere Idee aus zwei Gründen: 1) wegen Prospective Opacity können wir die Stabilität von gefassten Absichten zur Zeit ihrer Erstellung nicht genau einschätzen; 2) wegen Drift from Rationality, haben wir eine Tendenz über die Zeit hinweg, *ceteris paribus,* die vorherige Stabilität unserer Absichten beizubehalten, auch wenn spätere synchronische Rationalität uns anweist, anders zu handeln. Diese zwei Punkte erfassen den Kern jener Problemstellung, welche ich im ‚Opacity-Drift Puzzle' darstelle, um eine bestimmte Interpretation Bratmans anzubieten und ausführlich zu analysieren.