Personal Rules and Rational Willpower*

MICHAEL E. BRATMAN**

In my remarks on the rationality of rule following I will focus on the case of committing oneself in advance to a personal rule with an eye to resisting certain temptations. I will discuss why, in this case, it may be puzzling how it could be rational to follow the rule. And I will explore two lines of argument for understanding how such rule-following can, nevertheless, be rational.

I. A PROBLEM ABOUT RATIONAL WILLPOWER

Suppose you value both pleasant dinners and productive work after

* A version of this Essay was presented as a talk at the AALS Symposium on "The Rationality of Rule-Following." It was completed while I was a Fellow at the Center for Advanced Study in Behavioral Sciences. I am grateful for financial support provided by The Andrew W. Mellon Foundation.

This Essay is primarily a condensed synopsis of aspects of a longer paper of mine, *Temptation Revisited*. A version of *Temptation Revisited* was presented at the 2002 Amsterdam Workshop on Intention and Rationality. *Temptation Revisited* is to be published in a volume of essays associated with that workshop, edited by Bruno Verbeek and tentatively titled *Reasons and Intentions*. This Essay will also appear in MICHAEL E. BRATMAN, STRUCTURES OF AGENCY: ESSAYS (Oxford University Press, in preparation). See that essay for relevant further developments, qualifications, and references.

** Michael E. Bratman is U.G. and Abbie Birch Durfee Professor in the School of Humanities and Sciences, and Professor of Philosophy at Stanford University. He is the author of INTENTION, PLANS, AND PRACTICAL REASON (Cambridge: Harvard University Press 1987; reissued by CSLI Publications 1999), FACES OF INTENTION: SELECTED ESSAYS ON INTENTION AND AGENCY (Cambridge: Cambridge University Press, 1999), and various articles in philosophy of action and related fields. Some his recent work on agency and self-governance will be collected in his STRUCTURES OF AGENCY: ESSAYS (Oxford University Press, in preparation).

dinner. One pleasant aspect of dinner is a glass of wine, and two glasses would be very nice. However, a second glass of wine undermines your efforts to work after dinner. So, you have a general evaluative ranking concerning dinners: one glass of wine over two glasses. When you are in the middle of dinner, however, you are frequently tempted. For a short period of time, faced with the immediate prospect of a second glass, you value two glasses over one glass just this once.

Because you know that this will happen frequently, you ask yourself: "Can I rationally commit to a general policy—a general intention, a personal rule—of having only one glass of wine at dinner, a policy supported by my stable, evaluative ranking of an overall pattern of one glass over an overall pattern of two glasses, and then, when tempted, rationally follow through with this policy?" What is the answer?

II. INSTRUMENTAL RATIONALITY

How should we understand this talk of rationality? One central idea here is the idea of instrumental rationality. Instrumental rationality, as I understand it here, is rationality relative to the agent's current valuings, policies, cares, commitments, and the like. We may challenge some of these attitudes as failing to track important goods, but such a challenge goes beyond a judgment of instrumental rationality. And we can ask whether it would be instrumentally rational for you to follow through with your one-glass policy in the face of an evaluative ranking in favor of a second glass this once.

Instrumental rationality is rationality relative to an agent's present ends, valuings, and the like. But human agents are complex, and in many cases there is conflict. Our temptation case is one example: relative to your one-glass action policy, it is rational to refrain from the second glass; relative to your present directed valuing, it is rational to drink the second glass this once. It is, however, typically supposed that we can go on to an *on balance* judgment that says what, relative to all your relevant ends and the like, it is instrumentally rational to do. And it is commonly assumed that on balance instrumental rationality is rationality relative to the agent's evaluative ranking at the time of action of relevant options. That is why it is puzzling to suggest that it might be instrumentally rational on balance for you to follow through with your one-glass policy in the envisaged circumstances.

But why think that it is always the present evaluative ranking of present options that grounds on balance judgments of instrumental rationality? We need to consider the roles of evaluative rankings in the psychic economy of our agency in order to see if these roles can justify the normal priority of such rankings for on balance judgments of

instrumental rationality. And we need to see whether this priority can sometimes be defeated by a prior action policy.

III. STABILITY AND AUTHORITY

Let's distinguish two different strategies for developing the suggested connection between instrumental rationality, willpower, and the psychic economy of our agency.

The first strategy grants that on balance judgments of instrumental rationality are normally anchored in the agent's evaluative ranking, at the time of action, of present options. And it grants that intentions to act are normally grounded in one's evaluative rankings. So, to treat such intentions as themselves an independent anchor for on balance judgments of instrumental rationality normally would involve an odd double counting, one that would support an odd bootstrapping.

However, this strategy then goes on to consider the complex, cross-temporal and social coordinating roles of intentions, plans, and policies.[1] Such cross-temporal and social organization is, for agents like us, a means to an enormously wide range of human ends. Intentions, plans, and policies, in playing these cross-temporal and social organizing roles, need to have a certain stability: they need to have a certain resistance to reconsideration and revision. So we can ask: given a prior intention or policy, when would it be reasonable for the agent to reconsider or change it or both, and when, in contrast, would it be reasonable to stick with it? In particular, might a one-glass action policy reasonably have a kind of stability such that it can be instrumentally rational to stick with it even in the face of a temporary evaluative ranking to the contrary and even given the ability to diverge from that policy? If so, this would be a way in which, in certain temptation cases, on balance instrumental rationality may be anchored primarily in action policy rather than present evaluative ranking. Call this the *intention stability* strategy.

A second strategy notes that when we see certain attitudes as anchoring on balance judgments of instrumental rationality, we are seeing them as constituting a point of view that is in a strong sense the *agent's*. On balance instrumental rationality is rationality relative to the framework of considerations that is the *agent's own* framework. Your current evaluative rankings have priority for on balance judgments of instrumental

1. *See generally* MICHAEL E. BRATMAN, INTENTION, PLANS, AND PRACTICAL REASON (CSLI Publications 1999) (1987) (presenting the planning theory of intention).

rationality (when they do have such priority) because they have authority to articulate, in the face of conflict, where *you* currently stand.

Let's say that attitudes that establish where the person himself stands have "agential authority." And let's call this second approach the *agential authority* strategy.[2]

I turn first to the agential authority strategy.

IV. AGENTIAL AUTHORITY

The agential authority strategy sees on balance judgments of instrumental rationality as relativized to the *agent's* framework of relevant attitudes. So we need to ask which attitudes are not merely wiggles in the psychic stew, but rather help constitute the *agent's* relevant framework. In particular, we need to understand why valuings, or evaluative rankings, normally have agential authority, and whether this rationale may fail to apply in certain temptation cases.

I think there is an important connection between agential authority and the cross-temporal coordination and organization of one's practical thought and action: a main determinant of agential authority is the support of such cross-temporal organization of practical thought and action, in part by way of continuities and connections central to a broadly Lockean view of personal identity over time. We are agents who persist over time and whose agency is extended over time. Attitudes that play basic roles in supporting and constituting cross-temporal organizing structures central to our Lockean persistence over time have an important claim to agential authority.[3]

Given this approach, do valuings normally have agential authority? Well, what is valuing?

V. VALUING

As I see it, valuing is a pro attitude, one that bears a complex relation to value judgment. I can recognize a wide range of goods, judge that they are good, but still only incorporate some of them into my practical

2. For talk about where you stand, see generally HARRY G. FRANKFURT, *Identification and Wholeheartedness*, *in* THE IMPORTANCE OF WHAT WE CARE ABOUT 159 (1988) (understanding identification by appeal to wholehearted decision). For the terminology of agential authority, see generally Michael E. Bratman, *Two Problems About Human Agency*, *in* PROCEEDINGS OF THE ARISTOTELIAN SOCIETY 309 (A.W. Price ed., 2001) (distinguishing agential authority from subjective normative authority and sketching a model of both).

3. *See* Michael E. Bratman, *Reflection, Planning, and Temporally Extended Agency*, 109 PHIL. REV. 35, 35–61 (2000) (exploring inter-relations between strong reflectiveness, planning agency, temporally extended agency, and a Lockean approach to personal identity).

reasoning and action in ways that constitute valuing. Valuing, in a basic case, is a policy about one's motivationally effective practical deliberation: I value *X* when I have a policy of treating *X* as a justifying consideration in my motivationally effective practical reasoning.[4]

Now, for valuing to be an anchor for on balance judgments of instrumental rationality, it needs to help constitute the agent's framework of justifying reasons; so it needs to have agential authority. On the present approach, agential authority is largely a matter of role in Lockean cross-temporal organization of our practical thought and action. And I think we can see that valuing will normally support such cross-temporal organization. Valuing involves a policy of reasoning in certain ways over time, and of shaping one's actions over time in accord with that reasoning. So valuing will tend to support and to help constitute Lockean cross-temporal organization of practical thought and action. So there is a strong case for saying that valuings have agential authority, and so are anchors for on balance judgments of instrumental rationality.

But now we need to know whether this authority may be defeated in our case of temptation.

VI. VALUING AND AGENTIAL AUTHORITY

Faced with a second glass of wine, you value it this time more highly than refraining, though you continue to value an overall one-glass pattern more highly than an overall two-glass pattern. To explain in what your valuing two glasses just this once consists, however, we need to adjust our account of valuing. I have said that to value is to have a relevant policy about practical reasoning. However, your valuing of two glasses this one time is—unlike a general policy about practical reasoning—only an intention about present practical reasoning. It is a *singular commitment* to give relatively more justifying weight, in present motivationally effective practical reasoning, to a second glass.

As a singular commitment, the primary role of this singular valuing is to structure your present reasoning and action. In contrast, your general action policy of having only one glass of wine at dinner has the role of

4. For further complexities, see generally Michael E. Bratman, *Valuing and the Will*, 14 PHIL. PERSP. 249 (2000) (developing a model of valuing as a policy about weights in deliberation, within a Gricean strategy of "creature construction"); *see also generally* Michael E. Bratman, *Autonomy and Hierarchy*, *in* AUTONOMY 156 (Ellen Frankel Paul et al. eds., 2003) (analyzing the pressure on valuings to be higher-order attitudes).

organizing thought and action over time, in part by way of associated continuities and connections. So, given our approach to agential authority, there is a case for saying that this action policy, in contrast with your singular valuing, has the stronger claim to agential authority. So it may be on balance instrumentally rational for you to follow through with your one-glass action policy, despite your present directed evaluative ranking to the contrary.

VII. INTENTION STABILITY

Return now to the strategy of intention stability. We seek norms of intention stability that can, in certain cases, make (instrumental) sense of sticking with a prior action policy in the face of a conflicting evaluative ranking. And here I want to emphasize two ideas. First, given the roles our planning agency plays in coordination over time and socially, there are pragmatic pressures in the direction of stability of intention-like attitudes. Second, there is an important role here for a concern with future regret.[5]

Given that we are agents with limited cognitive resources, on many occasions we simply maintain our prior intentions as time goes by and even in the face of new information. However, sometimes we do stop and reconsider. When is it reasonable *not* to reconsider? Here I have proposed a two-tier pragmatic theory, one that seeks strategies of nonreconsideration that would promote the agent's ends. And because planning agents normally care about the cross-temporal integrity of their lives and normally identify with their anticipated ends in their planned for future planning agency, the ends relevant to stability will normally include such future ends. I conjecture that such a two-tier theory will support a defeasible default in favor of *non*reconsideration.

It is not clear, however, how to extend this approach to a temptation case in which you do, indeed, reconsider whether to stick with your one-glass policy. It is clear to you, once offered the second glass, that you value drinking it this time more highly than refraining. But you wonder whether you should nevertheless stick with your prior one-glass policy.

Now, the usefulness of stability of plan-type attitudes that lies behind the pragmatic account of reasonable nonreconsideration does, I think, also support some sort of defeasible, default presumption in favor of following through with one's prior intentions and policies even when one does reconsider whether to do so. Or, anyway, there is such a presumption so long as one does not distrust the earlier process of

5. *See generally* MICHAEL E. BRATMAN, *Toxin, Temptation, and the Stability of Intention*, *in* FACES OF INTENTION 58 (1999).

intention or policy formation. To grant this pragmatically grounded default in favor of the prior intention or policy is not, however, to see that intention or policy as providing a further—potentially bootstrapping—reason in deliberation. It is only to see it as establishing a burden of proof on a challenge to that intention or policy.[6]

The problem we now face, however, is that normally this default presumption will be overridden by a present evaluative ranking that ranks a specific alternative strictly higher than what one had intended. After all, your prior intentions and policies concerning action are themselves normally formed primarily on the basis of your evaluative rankings. To give an action-focused policy priority over such an evaluative ranking would normally be criticizable prior policy worship.[7]

How then are we to explain why it might be instrumentally rational for you to stick with your prior one-glass policy despite your present ranking in favor of a second glass? What we need is an explanation of why, in such a temptation case, this evaluative ranking may *fail* to override the default in favor of the prior policy of action.

VIII. ANTICIPATED FUTURE REGRET

It is here that I want to appeal to anticipated future regret. In the temptation case you know, let us suppose, that if you were to be guided by your evaluative ranking in favor of a second glass you would later regret it, and that if instead you were to stick with your one-glass policy you would later be glad that you did. A planning agent projects her agency into the future in a way that normally involves identifying with how she will see matters then. And this provides a ground for giving significance to anticipated future regret.

Because of concerns with policy worship, we needed to see the default in favor of a prior action policy as normally defeated by a present evaluative ranking to the contrary. But a planning agent's knowledge that she would regret acting on that evaluative ranking, and would be glad if she, instead, stuck with her prior policy, can *delegitimize* that

6. *Cf.* FREDERICK SCHAUER, PLAYING BY THE RULES 203–06 (Tony Honoré & Joseph Raz eds., 1991) (arguing that a rule can have a presumptive force that is not solely a matter of epistemic uncertainty, but is defeatible by strong reasons for not following it).

7. Analogous to what J.J.C. Smart calls "rule worship." *See* J.J.C. Smart, *Extreme and Restricted Utilitarianism*, *in* THEORIES OF ETHICS 171, 177 (Philippa Foot ed., 1967).

evaluative ranking. So it may sometimes be instrumentally rational for you to follow through with your one-glass action policy despite your present directed evaluative ranking to the contrary. And this argument avoids policy worship—for we continue to hold that the pragmatically grounded default presumption in favor of a prior policy of action is normally overridden by a present evaluative ranking to the contrary.

IX. Concluding Remarks

So we have two arguments for instrumentally rational willpower in the face of temporary, singular evaluative temptation. Both arguments see instrumental rationality as shaped by basic structures of our temporally extended planning agency.

This raises the question of how exactly these arguments interact with each other. It also raises the question of whether versions of these arguments extend to other domains—for example, to a *shared* policy of a group or institution. But these are matters for a different occasion.