7
Rational Self-Commitment

BRUNO VERBEEK*

1. Ulysses and the Sirens

After a long day at the office in Troy, Ulysses passes his favorite bar, the Sirens. From far away he can hear the laughing and singing of his friends. Back home, in the suburb of Ithaca, Penelope is waiting with dinner and he could really use a good night's sleep. Right now, all things considered, he would prefer to go home and leave beer and Sirens for another night. However, if he comes closer to the bar, his friends will notice him and call him over to have a beer or two with them. He also knows that if he accepts their invitation, he will never be home in time. Dinner will be cold, Penny will be cross and the good night's sleep will be too short. So Ulysses decides to go home and decides that if his friends call him over he will wave at them and tell them that he will see them some other night. Ulysses' decision seems sensible and nobody would be terribly surprised when Ulysses indeed passes the Sirens and declines the invitation of his friends. It is the rational thing to do given his decision.

This last observation—that it is rational for Ulysses to go home *given his decision*—is the one I want to discuss in this paper. Exactly why is Ulysses rational in this case? To understand why, we need to know more about the circumstances and the nature of his deliberations. Consider the following three scenarios. First, it could be the case that Ulysses prefers to be home with Penelope, both now ($t = 1$) and in the (near) future ($t = 2$). So that when his inebriated palls at $t = 2$ call out to him, he still prefers to be home. In this case, Ulysses' decision at $t = 1$ is rational as is his going home at $t = 2$. However, in this scenario it is unclear why Ulysses needed to

make a decision before he receives the invitation. He could simply stroll along and without any trouble decline his friends' invitation.

Second, it could be the case that Ulysses makes a prediction about himself. He could realize that, without any measures on his part, he will succumb to temptation when passing the bar. Therefore, he decides to go home and does go home. However, if this is correct, the role of Ulysses' prior decision again is unclear. Whereas in the former scenario the decision seem superfluous in accounting for the rationality of Ulysses' actions, in this scenario it is unlikely that the decision alone can account for Ulysses' actions, let alone their rationality. The standard way of understanding weakness of the will is as a lapse in rationality.[1] If Ulysses is tempted, he is not fully rational. If Ulysses is weak-willed, he will accept his friends' invitation against is better judgment. Therefore, at the time of action, he would not be rational. How could his decision at $t = 1$ to go home at $t = 2$, counteract the perverting influence of his friends' invitation on hi rational powers. It seems then that merely deciding to go home will not explain how a weak-willed Ulysses manages to go home.

A lot more can be said about weakness of the will and how to go about when one anticipated one's weakness. In particular, what sorts of external self-management one can use to deal with such lapses of rationality.[2] However, I will leave the suggestion that Ulysses predicts that he will be tempted when passing the bar here and move to the third scenario, which is the one that really interests me in this essay.

On this third scenario, Ulysses does prefer now (at $t = 1$) to go to Ithaca, but he anticipates that when his friends call out to him (at $t = 2$), he will change his mind about the attractions of Penelope, dinner and the benefits of a good night's sleep in comparison to a few hours in the company of his beer-drinking friends. In this scenario, Ulysses anticipates no (temporary) loss of his rational powers. Instead, he foresees that his preferences will change. On the orthodox theory of rational choice, giving in to the lure of the Sirens then would be rational: it would satisfy Ulysses' strongest preference at the time of action. Acting otherwise would involve a form of counter-preferential choice. How can Ulysses' decision and subsequent actions be rational in this scenario?

Jon Elster has argued that, in such cases, a rational agent has to treat her predicament like a case of anticipated weakness of the will.[3] In such situations, Elster argues, a rational agent should resort to commitment devices.

The fundamental reason for Elster's recommendation is his skepticism about the feasibility of counter-preferential choice. According to Elster, an agent prefers x to y if and only if she chooses x when offered the choice between x and y.[4] Given this conception of preference, genuine counter-preferential choice is not feasible. Therefore, Ulysses can only expect to go home if he uses a commitment device.

According to Elster, an agent commits herself if she acts at $t = 1$ in such a way that the set of her available alternatives at $t = 2$ is a proper subset of her alternatives at $t = 2$ if she did not act at $t = 1$, where it is the case that the act of binding is easier to perform and less costly than the preferred action at $t = 2$ without commitment.[5] Therefore, Ulysses should manipulate the situation in such a way that the undesired course of action is no longer available. For example, he could make it impossible to enter the bar by taking an alternative route home.[6] According to Elster, Ulysses needs to do something in addition to the decision that he will not enter the bar. In other words, on Elster's theory the only type of commitment that is feasible for Ulysses is a form of causal commitment. Ulysses needs to do something which subsequently exercises a causal—not a rational—influence on the available option. Rational agents are sometimes in a position that they have to protect themselves against their own future rationality through such methods.

Causal commitment then is a potentially costly way of avoiding future rationality. Making alternatives unavailable is expensive in terms of the freedom you might otherwise enjoy. To some this is sufficient reason to reject it.[7] However, the fundamental problem with Elster's suggestion is that it assumes that Ulysses' decision to continue on his way home does not carry any weight itself. His decision does not justify in any way his going home, while this seems to be exactly what is going on in this scenario. How is this possible? How can it be that simply by deciding to go home, Ulysses justifies his going home?

2. Decisions as causal commitments

The difficulty with interpreting Ulyssess' actions as the result of a causal commitment is that the role of his earlier decision is negligible in the justification of his subsequent actions. However, it is certainly possible to

build in this feature in an account of causal commitment. We can include the decision at $t = 1$ in the justification of his actions at $t = 2$ if Ulysses' earlier decision is itself a commitment device. Scott Shapiro has argued that if Ulysses genuinely decides to go home it must be the case that the earlier decision removes the option of entering the bar.[8] The decision to go home makes entering the bar no longer feasible. The idea is that once an issue has been decided there is no opportunity for choice any more: the issue is settled. So if Ulysses decides to go home, he can do only one thing and that is to go home. Trivially, this makes Ulysses' action rational. His changed preference for entering the Sirens is a mere wish, not a preference that can be satisfied.

Shapiro's theory has several advantages. First, it stays well within the traditional boundaries of rational choice theory. No appeal to counter-preferential choice is necessary. Second, it can make sense of the observation that Ulysses' going home is rational given his earlier decision. Finally, it can explain why Ulysses is successful in going home. He has—after all—no alternative.

In spite of these advantages, I am not convinced by Shapiro's theory at all. The most fundamental problem with it is that its answer to the question as to why Ulysses is rational in going home responds to a different question. The issue is not why Ulysses cannot enter the bar but why he should not enter the bar. This becomes clearer when we look at what would happen on Shapiro's theory if Ulysses were not rational when he passes the bar. If it is really true that because of his earlier (rational) decision entering the bar is not an available option any more, Ulysses would continue to go home even if the drunken singing of his friends in the bar is enough to drown out all sense in him. In other words, if Shapiro were right, we would have a wonderful panacea for weakness of the will. Simply decide, genuinely decide, to do the right thing and you will see that you always do the right thing.

The mistake Shapiro makes is that he does not treat Ulysses' case as a normative issue. Suppose that Ulysses, in spite of his earlier decision, spends the night boozing with his buddies at the Sirens. We would blame Ulysses in that case for his inconsistency. However, Shapiro's theory makes it impossible to find fault with Ulysses' actions in that case. The only way he could make room for the possibility of Ulysses entering the bar would be by arguing that Ulysses apparently did not really decide to go home after

all. However, in that case his entering the bar would be rational and his earlier decision would be irrational. In other words, on Shapiro's theory it just can never be the case that a person makes a genuine and rational decision and subsequently does not execute that decision.[9] This shows that Shapiro sees the relation between the decision at $t = 1$ and the actions at $t = 2$ as a causal relation and not at all as a normative one.[10]

From this brief discussion of Shapiro's proposal we can infer the following desiderata for any successful analysis of the case of Ulysses and the Sirens. First, it must be such that Ulysses' decision at $t = 1$ is rational. Second, it must be such that Ulysses' action at $t = 2$ is rational. Third, that (part of) the reason his action is rational is his earlier decision to go home. That is, the decision to go home is normatively—and not (just) causally—relevant for his going home.

3. Plans

We are rational agents—but only up to a degree. In our everyday lives we face constraints that limit the extent to which we can exercise rational choice. We have limited resources. Some limitations are internal in nature. We often do not have the acumen to oversee all the relevant information, nor can we determine which of the information is relevant. Furthermore, our reasoning capabilities are not perfect. We make mistakes and draw invalid conclusions. These limitations are further aggravated by external constraints. We are often, if not always, under time pressure to conclude our deliberations. As result of this pressure, we do not have the opportunity to gather, let alone inspect, all the available information. This pressure further compromises our ability to process the information correctly in our practical reasoning.

Given these constraints, we need a way to deal with these constraints on our deliberation. The solution is that we are planning agents. Rather than make our decisions on the spot each and every time we face a choice, we tend to structure our deliberations through plans. This is the fundamental insight of Michael Bratman.[11]

In order to fulfill this role in our deliberations, plans have certain characteristics. First, plans are typically *partial*. When I plan to go to a conference, I make some general decisions about when I will go. However,

I do not yet decide which sessions I will attend, or what I will wear or have for breakfast. Those decisions I leave up for later deliberation. In this way, plans relieve the burdens that time pressure puts upon us in an important way. By planning at t = 1 to do X at t = 2, I do not leave it up to the very last minute whether or not to do X. Instead, by settling on this plan now, when there is sufficient information and I am not under as much stress to come up with a decision, I can determine the more important things now, and leave trivial things (like what to wear or what to have for breakfast) for a later date.

Secondly, because of this partiality, plans are *hierarchical*. Plans concerning ends embed plans concerning the means. Once I have determined to attend the conference, I can settle on a plan about the means of transportation. Note that the decision for means of transportation only makes sense within the framework of this higher-order plan. The features of partiality and hierarchy enable us, planning creatures, to deal with the limitations of our rationality. Through planning we can coordinate between the opportunities we have for deliberation. Big complex decision problems can be tackled by planning the necessary moves one at a time. Further, the limitations necessitate that prior deliberation can shape later conduct. We simply do not have the time to make reasonable (let alone rational) decisions at the time of acting. Thus when we act upon a plan, the action is non-deliberative in the sense that it is not the immediate result of some practical deliberation. The decision-making process is in the past.

These two features of plans place demands on our plans. Plans need to be internally consistent. That is, it should be possible for my entire plan to be successfully executed given all my beliefs about my abilities and the world.[12] In addition, plans need to satisfy means–end coherence. Plans that satisfy these demands structure practical reasoning in the following ways. First, they create the context and problem for further deliberation and establish standards of relevance for options considered in deliberation. Once you have the plan to attend the conference, you face the problem of how to get there. Second, plans typically constrain further deliberation by placing "filters of admissibility" on it.[13] My decision to attend the conference rules out the option of attending a faculty meeting at the same time.

Two final remarks about plans conclude our discussion. First, it is not conceptually impossible to deviate from a plan (unlike Shapiro's theory). Rather, once I have settled on the plan to attend the conference, it is

no longer rational—but not impossible—for me to consider attending the faculty meeting at the same time. In other words, the "filters of admissibility" are normative not causal.[14] Second, plans can only function in these ways if they are inert. It is not impossible to reconsider one's plan, but in the absence of new information, plans resist reconsideration. They do so typically in a non-reflective way: once you have settled upon a plan, that is it; problem solved—no need to think about it any more.[15] Only when you encounter new information that was not available to you before, there may be grounds for reconsideration (provided there is time to deliberate). Therefore, plans come with certain commitments. Some of these commitments concern the sort of options I can reflect upon; others concern the circumstances in which I can reconsider my plan.

4. Rational commitment

Now that we have more or less fixed the notion of plans and the place plans have in our deliberative economy, we can explain the rationality both of Ulysses' decision and his actual carrying through of his decision. At $t = 1$, Ulysses considers his options. He could go home or he could go to the bar. He determines the reasons for either option and weighs them. The company of Penelope, warm dinner, and a good night's sleep weigh heavier than the joys of a beer or two with his pals at the Sirens. So in light of what he now knows and values about each option, going home is the better course of action. Therefore, Ulysses decides to go home and in doing so, settles on a plan, the execution of which involves declining the lure of the Sirens. At $t = 2$, his earlier plan to go home constrains his deliberation in that it makes entering the bar inadmissible. That is, his plan makes it irrational to enter the bar and rationalizes going home. Ulysses is committed, rationally committed, to go home.

What about Elster's worries about counter-preferential choice? These worries originate in a behaviorist notion of preferences that Amartya Sen has criticized, where preference and choice are synonymous.[16] Sen is correct in rejecting this view. The planning theory of commitment does not proceed from this view. On this theory, preferences are complex states that relate judgments of value with motivation, but not in the way the orthodox view suggests. There is no one-on-one relation between preferring A to B and

choosing A when confronted with this pair of options. Space does not permit me to elaborate here. For purposes of the argument it suffices to think of preferences as comparative judgments of value.[17]

At this point, there are two objections. First, it is far from clear that Ulysses' decision at $t = 1$ is in fact rational. Ulysses takes his preferences at $t = 1$ to determine the choice of plan. However, the plan is to be executed at $t = 2$, so why should Ulysses consider his preferences at $t = 1$ instead of those at $t = 2$ in settling on a plan?

We can distinguish between two views on the rationality of plans in general.[18] On the first view, the rationality of a plan is completely determined by the reasons for the intended action (apart from his decision). Thus, Ulysses' plan to go home is rational if and only if at $t = 2$ it is indeed rational to go home independent of his plan. This is the *primacy of action* view. This view is implicit in the suggestion that Ulysses' decision to go home is not rational. The alternative view—the view I endorse–is the *primacy of planning* view, which says that the rationality of a plan is not exclusively determined by the reasons for the intended action. On this view, the reasons for settling on a plan can include considerations other than those that obtain at the time of action. The truth of the latter view is obvious when we reflect on the pragmatic need for plans. Plans play such an important role in our deliberative psychology because we face time pressure and other constraints. In order to avoid last-minute mistakes; in order to cut down complex decision problems into manageable sub-problems; in order to coordinate intra- and interpersonally, we make plans. All of these reasons do not obtain at the time of action, but at the time of decision-making. In order to avoid the risk of last-minute mistakes at $t = 2$, we can settle on a plan at $t = 1$. Therefore, the rationale for a decision need not be confined to reasons for the intended action, but can also be found in other factors.

Still we need to establish that there are additional reasons to form the plan to go home that do not obtain at the time when Ulysses' drinking buddies request his company in the bar. One factor to consider is the regret Ulysses expects afterwards if he spends the evening with his friends ~~instead~~ with Penelope.[19] This, however, is a consideration that holds at $t = 2$ as well. It is not a reason for his decision at $t = 1$ that is not available at $t = 2$.

My suggestion is different. At $t = 1$, Ulysses prefers going home over entering the bar. Also at $t = 1$, Ulysses believes that at $t = 2$ he will prefer the Sirens to the domestic bliss that awaits him in Ithaca. Ulysses has this

belief at t = 1 and yet prefers to go home. Therefore, Ulysses considers his future preferences already at t = 1 and gives them some weight in his decision. So the question really is "why don't his future preferences have all the weight?" The reason is that even though Ulysses believes that he will value entering the bar over going home, he cannot be sure that his future preferences will reflect the value of both options correctly. The possibility of last-minute mistakes looms. Note that this is not the same as weakness of the will. Ulysses believes that without a decision now, he will act at t = 2 on his best judgment of the reasons that obtain. However, Ulysses is skeptical that his preferences at t = 2 reflect the reasons for either option as well as his preferences now. This is why he makes a decision now at t = 1 and why it is rational to do so. Furthermore, this explains why his decision to go home is rational. It is based, among other things, on his lack of trust in his future judgment. Note that Ulysses does not believe that he will necessarily be wrong at t = 2. It might be true that spending the evening at the Sirens is better for him than going home. As far as he can tell now at t = 1, this is not the case, but he might be mistaken. However, given the nature of the situation, the possibility of last minute mistakes, etc., there are reasons for settling on a course of action now, rather than leave it up to last minute deliberations. Therefore, there are reasons for deciding to go home at t = 1, that cannot be reduced to reasons for the intended action at t = 2.

This argument for the rationality of Ulysses decision at t = 1, explains some other things as well. Consider the case of Mark, a typical 11-year-old boy. He firmly believes that girls are totally yucky, but at the same time he observes how boys not much older than he suddenly lose their cool and go completely gaga over girls. Being somewhat precocious, he expects that in due time he too will be drawn to these yucky creatures. So we have a similar shift in his preferences as that of Ulysses. At t = 1, Mark prefers to avoid his female classmates, whereas he actively will seek their company (at least some of them) at t = 2. Mark believes that this shift will not be the result of a collapse in his rational abilities (such as they are) but simply reflect a change in values. Should Mark decide to lead a chaste life and spurn any interaction with these alien creatures? The difference between Ulysses and Mark is that unlike Ulysses, Mark has no reason to be skeptical about his future judgment in the same way as Ulysses is. Ulysses distrusts his future judgment on the merits of the Sirens versus Penelope. Mark

has no reason to distrust his judgment in that way. He can anticipate a genuine change in what is good for him (what is good for an 11-year-old boy is not always good for a 16-year-old boy). Therefore, Mark should not make any decision at $t = 1$ and simply "go with the flow." That is, rational commitment of the type I have been discussing is a means to deal with the constraints agents face in their deliberations. If Mark is not expecting such constraints (at least not relevant ones), there is no need to commit. On the other hand, if he does expect that his judgment might be clouded, he should commit. And, finally, if he has reason to expect that he will not be able to act on his best reasons or his prior decision, he should take causal measures to deal with that situation. Thus, this rationale for settling on a plan allows for a nice threefold distinction about the rationality of such commitments. We can distinguish situations where there is a need for rational commitment, situations where there is a need for causal commitment, and situations where there is no need for commitment at all.

This concludes the discussion of the first objection against the idea that Ulysses is rational when he decides to go home to Ithaca at $t = 1$. There is a second objection. Suppose that Ulysses is rational in deciding at $t = 1$ to go home. What prevents him from reconsidering at $t = 2$? After all, no plan is sacred. We can, and often do, change our plans. Why should Ulysses not change his plan?

To answer this worry, we need to return to the pragmatic rationale for forming a plan in the first place. As I argued above, the reason we form plans and make future-oriented decisions are tied up with the limited nature of our rationality. We need plans to reduce the possibility of last-minute mistakes and to reduce complex decisions to several smaller, simpler decisions. Ulysses' decision at $t = 1$ to go home and not enter the Sirens at $t = 2$ is such a plan. Plans can only have this function if they are inert. They should not come up for reconsideration arbitrarily. Plans have a default stability. Once you have decided, you have no reason to revise the plan, unless you discover a reason to revise it. That is to say, Ulysses is committed to his decision, unless there is new information between $t = 1$ and $t = 2$. However, everything that was relevant to his decision at $t = 1$ remains the same, the beliefs about his own abilities for rational decision-making, the beliefs about the nature of his options, the beliefs about his shifting preferences, and so on. In other words, there is no reason for Ulysses to

revise his plan. Since there is no reason for reconsideration, Ulysses is not free to reconsider.

Furthermore, suppose that we change the story a bit. Suppose that his shift in preferences is not expected at $t = 1$ and that as he strolls along the bar, Ulysses discovers that his friends are there. It still would not follow that Ulysses necessarily has good reasons to start deliberating about his earlier decision and determine whether he should reconsider. Because at $t = 2$, the pressure is on: he needs to make a decision now and the chance that he would make a mistake is considerable. That fact in itself provides a reason to stick with the earlier decision, even though there is new information.[20]

Therefore, if we interpret Ulysses' decision at $t = 1$ as a plan to go home and not enter the bar at $t = 2$, we have a genuine case of rational commitment. Ulysses decision at $t = 1$ is rational and so is his execution of his decision at $t = 2$. The reason why going home at $t = 2$, despite his shift in preferences, is rational is his earlier decision. Finally, the influence of his earlier decision on his actions at $t = 2$ is not a form of causal commitment. It is not the case the entering the bar at $t = 2$ is no longer an option as result of his decision at $t = 1$. Rather, entering the bar is no longer rational. It seems, therefore, that our account fulfills all desiderata we formulated at the end of section 2.

5. Rational commitment or bootstrapping?

In sections 1 and 2, I argued that rational commitment is different from causal commitment. However, at this point it might be argued that this is not really the case. I have argued that by forming a plan at $t = 1$, Ulysses has a reason at $t = 2$ to go home. Is that not just saying that by forming a plan at $t = 1$, the option of going home has become relatively more attractive than staying at the Sirens? It seems that Ulysses, through his decision at $t = 1$, has started a causal process that resulted in making the option of going home more attractive. If that is the case, Ulysses' plan at $t = 1$ looks like a form of causal commitment after all.

Strictly speaking, however, this is not the case. By deciding to go home, Ulysses does not change anything about either home or the Sirens. He does not alter the relative desirability of these options. Ulysses, in forming his plan, puts himself in a state in which it is no longer rational to act on

his strongest preference at $t = 2$. This state, the state of having-decided-to-go-home provides Ulysses with a reason to ignore his current preference (that is, his judgment at $t = 2$ about the relative values of home versus the Sirens), since he has reason to doubt that judgment.[21] Obviously, this has a causal effect on his subsequent actions. However, this causal effect is not what commits Ulysses; his decision at $t = 1$, on the other hand, does.

Be that as it may, the objection shows us an obvious further objection to the analysis of rational commitment. If the analysis were correct, agents can create their own reasons for action. Simply by deciding to go home at $t = 2$, Ulysses has created sufficient reasons to go home. That seems wrong. We either have reasons to do something or we don't—we cannot create these reasons. Of course, we can and often do create reasons by changing things in the world. For example, if I decide to go to a conference and I buy a non-refundable ticket, then, arguably, I have changed the balance of reasons for going to the conference. However, that is not the case with Ulysses. If my suggestion is correct then right after deciding so at $t = 1$ Ulysses has a decisive reason to go home at $t = 1 + \varepsilon$ before anything else has changed in the world. The only thing that is different is that Ulysses has decided. How could that difference create a genuine reason? Arguing that it does, constitutes, so the argument goes, an unacceptable form of bootstrapping.[22]

Bootstrapping seems to spell problems for the feasibility of rational commitment. If decisions do not create reasons for a course of action, decisions do not rationalize the future action. Therefore, if bootstrapping is impossible, rational commitment is not feasible and we have to revert to a version of the causal theory of commitment. So it seems we face something of a dilemma: either we insist that Ulysses' decision justifies his future choice, in which case we have to accept that we can bootstrap our own reason into existence, or we avoid bootstrapping but then we have to give up on the feasibility of rational commitment altogether.

Let me state at this point the conclusion that I will reach after the next two sections. I accept the dilemma. However, I do not believe that it is vicious. Intentions bootstrap reasons for action into existence that did not exist prior to the decision. In order to reach this conclusion, I will discuss two recent attempts to circumvent the dilemma. Both John Broome and Govert den Hartogh have argued, on different grounds, that the dilemma is false. Both authors try to account for the intuition that Ulysses decision justifies his

future conduct, without accepting bootstrapping. In the remainder of this paper I will discuss their attempts at avoiding the dilemma. I will argue that they are unsuccessful. The dilemma is unavoidable, but it is unclear if accepting bootstrapping is as objectionable as the critics make it seem.

6. Broome on reasons and requirements

In several publications, John Broome has argued that the exclusive focus on reasons obscures an important distinction in the normative landscape.[23] Two types of normative considerations, so he argues, govern rational deliberation. On the one hand, there are considerations that can best be labeled "pro tanto reasons," or "reasons" for short.[24] On the other hand there are "normative requirements," or "requirements" for short.[25] What distinguishes reasons from requirements is that the former have "weight" whereas the latter are "strict." Let me explain. Reasons have weight. A reason to F is a consideration in favor of F-ing that could result in the conclusion to F, other things being equal. For example, the sunny weather is a reason to go to the beach. If there were no other reasons to the contrary (such as the deadline for this paper), I would go. In determining what to do, reasons enter deliberation like weights on a scale: they could tip the balance one way or the other. That is, reasons "add" to your decision. Reasons continue to play this role, even when they are outweighed by alternative considerations. As Broome puts it, reasons are "slack" (they do not necessarily determine the outcome of rational deliberation) and "absolute" (they reside in fact and continue to exist even when outweighed).[26]

Requirements, on the other hand, have no weight. The requirement to F does not admit of degrees: either you are required to F or not. For example, you are required to choose the necessary means to your ends. This requirement cannot be outweighed by other considerations. If you have the end q and p is the necessary means to q, you ought to p. It does not mean that you have a reason *to* p. In fact, it could be the case that there are good reasons against p. Let q be the end of achieving world domination and let us suppose that the necessary means for that is killing 5,000 people; it does not follow that you have a reason to commit mass murder. However, that does not remove the requirement. In this case you ought to give up

the end to achieve world domination. In general there are two ways to deal with this particular requirement: choose the necessary means or give up the end. Requirements, therefore, are "strict" (they cannot be outweighed) and as a result not absolute since they do not remain "on the scene" even when defeated.

This highlights a further distinction that Broome makes in this context, that of scope. Both reasons and requirements can vary in their scope. For example, there is a reason to go to sleep in time this evening. It may not be a particularly strong reason for some, but that is not relevant right now. What is relevant is that it has a very narrow scope: it concerns just one action. Contrast this with the requirement that you ought to *p* if your end is *q* and *p* is the necessary means to *q*. This requirement has a wider scope, for it holds between *q* and *p*. It is "relational."

With these distinctions in place, we can now discuss Broome's solution to the problem of bootstrapping. According to Broome, decisions do not generate reasons. However, once Ulysses has decided to go home, he is required to see to it that he goes home. Thus, Ulysses makes a rational mistake if he were to enter the bar, even though the balance of reasons at $t = 2$ favors going to the Sirens. It is not the case that by deciding to go home, he has given himself a reason to go home. Instead, having decided to go home, he is required to go home or to reconsider his original decision. Note that it is not the case that Ulysses has created this requirement. It was there all along. By deciding to go home at $t = 2$, it became applicable to Ulysses. According to Broome, Ulysses' decision does not bootstrap a reason into existence that was not there before; however, Ulysses would be irrational if he did not go home (or repudiate his decision). Reasons and requirements play a different role in the deliberative economy of rationality. If we appreciate these differences, so Broome claims, we will see that the bootstrapping dilemma is false.

This is an elegant and plausible proposal. However, I am not convinced. There are two problems with Broome's solution. First, though we avoid bootstrapping, we return to one of the problems that I disposed of above. There is nothing in the concept of requirements that looks like the necessary inertia of plans. Perhaps we could solve this by arguing that decisions create requirements as well as some (weak) reasons for non-reconsideration. However, this would get us back to the bootstrapping that Broome wanted to avoid altogether.

There is a second, more serious problem with the proposal. Sometimes, decisions create strong reasons for a course of action. Consider the following example.[27] Like most mothers, Mom loves her two children, Peter and Jane. She would like to give them both a treat. Unfortunately, she can give only one of them a treat. Since both children are equally deserving, needy and desirous, she is indifferent which of the two should get it. Therefore, she is indifferent between the outcome in which Peter receives the treat (P), and the outcome in which Jane receives the treat (J). However, she prefers to flip a fair coin and let it decide who gets the treat. Note that this is fair: giving the treat to one of her children using this device is better (since fair) than giving it to one of them straightaway. Note further that Mom has two possible plans at her disposal that would do the trick. She could give Peter the treat if ''heads'' comes up and give Jane the treat when ''tails'' comes up. Alternatively, she could decide to give Jane the treat if ''heads'' comes up and Peter if ''tails'' comes up. We can put this scenario in a schematic representation (see figure 7.1). (A square node represents a point in the tree where a choice has to be made, whereas a circular node represents a coin flip.)

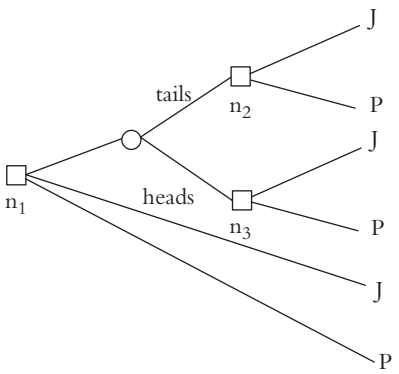

Figure 7.1. Decision tree

Suppose that Mom decides for the first plan. She decides she will give the treat to Jane if ''tails'' and to Peter if ''heads.'' Suppose ''tails'' comes up. She is now in n_3. Again she faces the choice between giving the treat to Peter or Jane. They are equally deserving, needy and keen on the treat, so it seems that the reasons in Broome's sense are equally strong and that it is just Mom's earlier decision that requires her to give the treat to Jane.

However, that is incorrect. Having made the decision to give the treat to Jane if "tails" makes it the case that it is simply fair to give the treat to Jane. Mom's decision thus has created a reason of fairness to give the treat to Jane. In other words, in this situation Mom's earlier decision has created a strong reason. She has bootstrapped this reason into existence simply in virtue of her decision.

In conversation about this example, Broome argued that what makes it the case that it is only fair that Jane gets the treat is that Peter would have received it if "heads" had come up. In other words, it is not Mom's *decision* that creates the reason, but the fact—a counterfactual fact so to speak—that Peter would have received it had the world been otherwise. So it is not the decision that creates the reason but the world (i.e., the world that would have been the case if "heads" had come up).

At the time I did not know how to reply, but now I do. It is true that Jane's receiving the treat if "tails" is fair because Peter would receive it had things turned out differently. However, we should ask ourselves *why* Peter would receive the treat if "heads" had come up. The answer is, obviously, Mom's earlier decision, her plan, in n_1 to give the treat to Peter if "heads" comes up. In other words, Mom's decision is what made it the case that it is only fair that Jane gets the treat when "tails" comes up. So this really is a case where decisions bootstrap reasons into existence. Therefore, the distinction between requirements and reasons does not help us to avoid the complaint of bootstrapping in every case of rational commitment to a future course of action.

This is not to say that the distinction is not a useful one. On the contrary; it elucidates some very important features of rational deliberation. However, I doubt that it avoids the dilemma that the bootstrapping objection seems to pose to the idea of rational commitment.

7. Den Hartogh on the authority of intentions

A different way of avoiding the bootstrapping objection has been proposed by Govert den Hartogh.[28] Den Hartogh makes an interesting analogy between rational self-commitment and the exercise of authority. If Ulysses' earlier decision at $t = 1$ creates reasons to act in a certain way at $t = 2$, it is as if his earlier self executes authority, rational authority, over his later self. Just

as the decisions of an authority carry weight in the subsequent deliberations of those subjected to the authority, so it seems that the decision of the self at $t = 1$ carry weight in the subsequent deliberations of those subjected to the authority of the self at $t = 1$, and this can only be the later self of $t = 2$. Den Hartogh thinks that the similarities are so striking that he proposes to adopt the analysis of interpersonal authority for the analysis of intrapersonal authority.

According to Den Hartogh, rational authority—as opposed to all other (that is, irrational) forms of authority—comes in two types only. First, there is coordinative authority. This is the sort of authority that helps to solve coordination problems. Imagine a busy road with cars coming from both ends. Suppose that there is no rule as to what side of the road cars should pass. Obviously, if others will keep to the right, you have a reason to keep to the right as well. However, the same holds for the left. What is more, nobody has any independent reasons to stay on one side of the road. So there is a coordination problem. Suppose that a person, say, a traffic warden, makes a clear signal to the effect that all cars should stay on the right-hand side of the road. Now all car drivers have a reason to stay on the right-hand side of the road. The warden's signal functions as a focal point for the relevant expectations of all drivers. Prior to the signal, we had no special reason to expect that others would drive on any side of the road, but now we do. I believe that you have seen the signal and I believe that you believe that I believe that you have seen the signal. What is more, I believe that you believe that I have seen the signal, etc., etc. Our relevant beliefs refer to each other. Because we now have these nested interdependent mutual expectations, we have a reason to drive on the right-hand side of the road. In short, the signal of the traffic warden to stay on the right-hand side of the street has created a new reason to do so that did not exist before.[29] He has coordinative authority.

Second, there is epistemic authority.[30] This is the sort of authority where the subject takes the command of the authority as a reason because the authority is better placed than she is to oversee and weigh the relevant reasons. One turns to the authority for his or her judgment because the judgment of the authority is more likely than one's own to be correct. For example, a medical doctor has authority (and I defer to his authority) when he orders me to take antibiotics. I surrender my judgment in this case. Why would it be rational to do so? Obviously, because the doctor is better

placed than I am to oversee the merits of the case. He is in a better position than I am to see what is in my interest and what is not. And as long as the doctor's orders are not obviously off (e.g., when he would prescribe huge quantities of cocaine for a common cold), it is perfectly rational to accept his decisions and act accordingly.[31]

Den Hartogh argues that future-directed decisions only have epistemic authority. Thus Ulysses' decision to go home creates a reason for his actions at $t = 2$, but these reasons depend crucially on the question whether Ulysses at $t = 1$ is better situated to determine the merits of a visit to the Sirens compared to the journey home. This characterization of the sort of reasons that decisions give is very helpful. It explains, first, why prior decisions have inertia. Just like a rational agent does not continue to question the doctor's orders, provided she trusts his judgment, a rational agent does not reconsider her earlier decision. Unless, of course, there is new information available that gives reason to doubt the earlier judgment. Furthermore, such reasons can be ignored if the agent has no grounds for assuming she is likely to make a better decision than the earlier self. So even if Ulysses' preferences change as result of some new, unanticipated information, he can still be rationally committed to go home. Finally, the proposal makes clear when it is rational to reconsider: this is when the earlier decision loses its authority. This will be the case, first, when the reasons for referring to the authority are no longer valid. That will be the case when one is relatively sure that one will not commit any last-minute mistakes. Second, when one has reason to doubt the superior epistemic position of the authority (relative to one's own), the prescriptions of the authority lose all their force. This could happen when one receives new information about the alternatives, which would have altered the initial decision.

Den Hartogh's position implicitly makes a distinction between vicious and benign bootstrapping. It avoids the sort of unwarranted bootstrapping that would occur if the self at $t = 1$ would have coordinative authority as well as epistemic authority. So it is not the case that Ulysses creates entirely new reasons; however, by deciding to go home, he specifies a reason that is there, which is his reason to avoid last-minute mistakes. In other words, rather than accepting bootstrapping entirely, or trying to avoid it completely, Den Hartogh accepts a very limited sort of bootstrapping. The sort of reason that Ulysses' earlier decision creates is a function of his finite, constrained rational abilities. If the danger of last-minute mistakes

is minimal, if there is more than sufficient time to oversee the merits of a visit to the Sirens relative to the comforts of home, Ulysses need not, perhaps even should not, make a decision about what to do. I agree with Den Hartogh in this case. This is exactly my defense for the rationality of Ulysses' decision at $t = 1$.[32] The reasons to decide at $t = 1$, rather than postponing his decision until $t = 2$, are determined by the time pressure and the danger of last-minute mistakes.

However, I am not convinced by Den Hartogh's claim that the reasons that decisions create are only of this epistemic type. Two obvious problems for this view are, first, Buridan's Ass cases (where the agent is absolutely indifferent between the two alternatives), as well as cases of incomparability. In both cases, the reasons for either alternative do not settle the issue. In such cases, Den Hartogh argues, we do not bootstrap any reasons into existence by deciding for one of the available alternatives. That would be a case where decisions have more than epistemic authority. Rather, in such cases the prior decision simply creates the causal "umpf" necessary to prevent indecision. In those cases, decisions just have causal impact on the agent, but no rational weight. The argument, in a nutshell, is the following. Suppose you face two indifferent options A and B. Suppose at $t = 1$ you decide for A, but at $t = 2$ you take B instead. Den Hartogh denies that in such cases you have made a rational mistake.[33]

I have doubts about this particular argument similar to my reservations about Shapiro's theory. However, there is a more fundamental objection to the entire theory. There is a counter-example to Den Hartogh's claim that future-oriented decisions only have epistemic authority. The example of Mom and her two children from the previous section is a case where decisions do more than relieve the burdens of decision-making at the last minute. On Den Hartogh's view, Mom would not have created any reason to give the treat to Jane. Mom's decision to flip a coin is merely a device to create the push that the whole machinery needed. However, this seems to miss something. Mom's decision to flip the coin is not just a way to break the deadlock between the reasons for Jane and Peter. Rather, flipping the coin and letting its outcome determine who gets the treat is fair—and that is a decisive reason in favor of coin flipping. Thus having flipped the coin, there really is a reason and not just a causal push to give the treat to Jane. So this seems to be a case where a decision has more than epistemic authority.

8. Conclusion

I conclude that neither Broome's nor Den Hartogh's attempt at avoiding bootstrapping works in all cases. Sometimes we can and do create our own reasons, just like the rational commitment model I suggested predicts. Therefore, I seriously doubt whether bootstrapping is a problem for the analysis of rational commitment. The unavoidability of bootstrapping at least suggests that it is not a problem in the first place. As planning agents we do it all of the time. The dilemma I introduced above is not vicious. We can safely take the first horn and accept that rational agents sometimes create reasons for their actions simply in virtue of their earlier decisions.

However, this does not settle the matter in the specific case of Ulysses. First, both Broome and Den Hartogh could try to argue that in the central case of Ulysses and the Sirens, there is no objectionable bootstrapping. Broome then still has to account for the inertia of intentions. Den Hartogh's theory of the epistemic authority of intentions provides such an account in terms of the circumstances of choice at $t = 2$ (time pressure and last-minute mistakes). So maybe all of Ulysses' reasons to ignore his preference for the Sirens and continue his course home are reducible to his inferior epistemic position at $t = 2$. Though I tend to agree with this analysis of Ulysses, I doubt that it holds for all cases of rational self-commitment. Regardless of this, though, I conclude that rational self-commitment is not only feasible, it is also advisable for limited rational agents like us. Without it, the Penelopes of this world would be very lonely indeed.

Notes

* This paper is a revised version of "The feasibility of rational self-commitment" which was presented at the workshop on Rationality and Commitment at the University of St Gallen, 13–15 May 2004. I want to thank the participants of the workshop, John Broome, Govert den Hartogh and Luc Bovens, for their helpful comments.

1. For example, Davidson (1970).
2. Elster (1979) is still the *locus classicus*.
3. Elster (1979).
4. Elster (1979: 6 ff.).

5. Elster (1979). This is not a good definition of commitment, even in Elster's sense. The central feature of a commitment according to Elster is that certain options at $t = 2$ are made impossible. It is irrelevant whether the "act of binding" is easier or less costly than the most preferred action at $t = 2$, for the act to be efficacious in restricting the options at $t = 2$ to a proper subset of those at $t = 1$. Secondly, Elster's statement seems to rule out that actions which affect the options at $t = 2$ in such a way that these partly overlap with those at $t = 1$ as a form of commitment. For example, if Ulysses attends a lecture in the evening, rather than walking home, he has excluded the option of the Sirens from his options at $t = 2$ but he also has new, other options that he would not have without this action. Now imagine that Ulysses attends the lecture in order to avoid his friends hailing him from the Sirens. Should this not count as a form of commitment?
6. Elster discusses two additional techniques of "self-management" that could be relevant for Ulysses-like cases. First, Ulysses could manipulate the situation in such a way that his future preferences will continue to favor going home to Penelope. He could, for example, deposit a large sum of money with a friend and tell this friend that he is free to keep it, should he enter the Sirens. Second, Ulysses could attempt to tinker with his rational decision-making powers at the time of action, in such a way that he will not be completely rational and, as a result, go home. For example, Ulysses could undergo hypnotherapy or take special drugs, which make him unreceptive to the lure of the Sirens. (Note that this is the mirror image of the scenario of weakness of the will discussed above.) Elster (1979: chapter 2).
7. For example, McClennen gives such pragmatic arguments to abandon "sophisticated choice" and argue for the rationality of "resolute choice" (McClennen 1990).
8. Shapiro (forthcoming). Similar suggestions have been made by Isaac Levi in. (Levi 1994)
9. This assumes, of course, that there is no external interference.
10. There are other problems with this approach as well. In Verbeek (2002a) I argue that this position assumes an incoherent notion of feasibility which implies (if correct) that at the time of choice one has only one option. That is, there is no choice in the first place which makes the requirements of rational choice empty.

11. Bratman (1987).
12. There are some complications here concerning self-prediction. If I can predict that I would not do X at $t = 2$, would a plan to do X be feasible? Could it be rational to plan to X under such circumstances? In Verbeek (2002a) I argue that there are good reasons to resist certain types of self-prediction in the determination of what plans are feasible. Furthermore, there can be situations where (part of) my plan cannot be successfully executed given my beliefs about my abilities and the world yet it is rational to adopt such a plan. The standard example is that of a conditional plan (I will do X if C obtains), where ~~you are~~ certain that the condition C will not obtain, for example, the plan to leave one's spouse if he or she is unfaithful.
13. To a large extend, the debate about rational self-commitment is a debate about the question of how these "filters" are to be characterized.
14. Obviously plans also have a causal impact on our deliberation. Such impact does not rationalize the future choice.
15. This is not the only type of non-reconsideration: it is possible to resist reconsidering plans in a reflective way. However, the typical way plans function is such that you only start worrying about your plan if you have new reasons to doubt its rationality. See also Bratman (1987: esp. 64–72).
16. Sen (1977).
17. I am fully aware of the many questions this view might raise. For example, what to say to the type of objection that David Lewis has raised in. Lewis (1988, 1996)? In addition, there is the following puzzle. If preferences are belief-like states about the comparative value of options, then it looks like Ulysses is in the following predicament. He now believes (p) going home to be better than entering the Sirens, but he also believes that he will come to believe that $\sim p$, and this seems paradoxical. See also Van Fraassen (1984). I have to leave all these problems for a future occasion.
18. For a similar distinction, see Robins (1995).
19. Bratman (1998) takes up this suggestion in the context of the so-called "toxin puzzle."
20. ~~Joseph~~ Raz makes a similar point in his discussion of legal authority ~~(Raz 1978)~~.

21. Remember that I suggested that the best way to think of preferences is as comparative judgment with motivational effect.
22. See Bratman (1987: 24–7, 86–7).
23. Broome (1999, 2001a, 2001b, 2002).
24. I follow the name the Broome gives this type of considerations in Broome (2004).
25. This term comes from Broome (2001a).
26. Broome (2001a).
27. The example is due to Diamond (1967). I discuss this version of it at length in Verbeek (2001).
28. Den Hartogh (2004, forthcoming).
29. This is a widely shared analysis that can be traced to Lewis (1969), Schelling (1960), Sugden (1986). I discuss their ideas in Verbeek (2002b). Den Hartogh (2002) gives a detailed discussion of this and other types of authority.
30. Raz (1985).
31. According to Den Hartogh, what makes these two phenomena forms of authority is that both issue so-called content independent reasons. That is, it does not matter *what* the authority commands, only that he or she commands it. However, both coordinative and epistemic authority provides content-independent reasons only within a certain range. It is rational to accept an authority (i.e., the other person has rational authority over the subject) when his decisions are sensitive to the reasons at hand, including the reason why you refer to him. For example, a coordinative authority who commands the car drivers to pick a side of the road at random has no authority, since his command obviously does not do anything to improve the coordination problem. However, whether he commands us all to drive on the right side or the left side of the road makes no difference to the reasons he issues for driving on the right side or the left side respectively. Both are equally authoritative commands.
32. See section 4.
33. Note that Broome's proposal that decisions fall under a requirement to execute them or reconsider them can explain why such an agent would be inconsistent and, therefore, make a mistake.

References

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

Bratman, Micheal E. 1998. "Toxin, Temptation, and the Stability of Intention." In Christopher W. Morris and Jules Coleman (eds.), *Rational Commitment and Social Justice*. Cambridge: Cambridge University Press, pp. 59–83.

Broome, John. 1999. "Normative Requirements." *Ratio* 12/4: 398–419.

Broome, John. 2001a. "Are Intentions Reasons?" In Arthur Ripstein and Christopher Morris (eds.), *Practical Rationality and Preference: Essays for David Gauthier.* Cambridge: Cambridge University Press, pp. 98–120.

Broome, John. 2001b. "Normative Practical Reasoning I." *Aristotelian Society* 75: 175–93.

Broome, John. 2002. "Practical Reasoning." In Jose Luis Bermudez (ed.), *Reason and Nature: Essays in the Theory of Rationality*. Oxford: Oxford University Press, pp. 85–111.

Broome, John. 2004. "Reasons." In R. Jay Wallace, Philip Pettit, Samuel Scheffler and Michael Smith (eds.), *Reason and Value: Themes from the Moral Philosophy of Joseph Raz*. Oxford: Oxford University Press, pp. 28–55.

Davidson, Donald. 1970. "How is Weakness of the Will Possible?" In Joel Feinberg (ed.), *Moral Concepts*. Oxford: Oxford University Press, pp. 93–113.

Den Hartogh, Govert Arie. 2002. *Mutual Expectations: A Conventionalist Theory of Law*. Law and Philosophy Library. Dordrecht: Kluwer Academic Publishers.

Den Hartogh, Govert Arie. 2004. "The Authority of Intentions." *Ethics* 115/1: 6–34.

Den Hartogh, Govert Arie. Forthcoming. "Intending for Autonomous Reasons." In Bruno Verbeek (ed.), *Reasons and Intentions*.

Diamond, Peter. 1967. "Cardinal Welfare, Individualistic Ethics, and Interpersonal Comparisons: Comment." *Journal of Political Economy* 75: 765–6.

Elster, Jon. 1979. *Ulysses and the Sirens*. Cambridge: Cambridge University Press.

Elster, Jon. 1982. "Sour Grapes—Utilitarianism and the Genesis of Wants." In Bernard Williams and Amartya Sen (eds.), *Utilitarianism and Beyond*. Cambridge: Cambridge University Press, pp. 219–38.

Levi, Isaac. 1994. "Rationality and Commitment." In Carol C. Gould (ed.), *Artifacts, Representations and Social Practice*. Dordrecht: Kluwer, pp. 257–75.

Lewis, David. 1969. *Convention: A Philosophical Study*. Cambridge, MA: Harvard University Press.

Lewis, David. 1988. "Desire as Belief." *Mind* 97(387): 323–32.

Lewis, David. 1996. "Desire as Belief II." *Mind* 105 (418): 303–13.

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

Raz, Joseph (ed.) 1978. *Practical Reasoning*. Oxford, New York: Oxford University Press.

Raz, Joseph. 1985. "Authority and Justification." *Philosophy and Public Affairs* 14: 3–29.

Robins, Michael H. 1995. "Is it Rational to Carry Out Strategic Intentions?" *Philosophia (Israel)* 25/1–4: 191–221.

Schelling, Thomas. 1960. *The Strategy of Conflict*. Cambridge, MA: Harvard University Press.

Sen, Amartya K. 1977. "Rational Fools: A Critique of the Behavioral Foundations of Economic Theory." *Philosophy and Public Affairs* 6: 317–44.

Shapiro, Scott. Forthcoming. "The Difference that Rules Make." In Bruno Verbeek (ed.), *Reasons and Intentions*.

Sugden, Robert. 1986. *The Economics of Rights, Co-operation and Welfare*. Oxford: Blackwell.

Van Fraassen, Bas. 1984. "Belief and the Will." *Journal of Philosophy* 81/5: 235–86.

Verbeek, Bruno. 2001. "Consequentialism, Rationality, and the Relevant Description of Outcomes." *Economics and Philosophy* 17/2: 181–205.

Verbeek, Bruno. 2002a. "Feasible Intentions." Unpublished manuscript.

Verbeek, Bruno. 2002b. *Instrumental Rationality and Moral Philosophy: An Essay on the Virtues of Cooperation*. Dordrecht: Kluwer Academic Publishers.

Queries in Chapter 7

Q1. please check closing quotes is missing here.