Reasons and Intentions: An Introduction

Bruno Verbeek

Problems of Rationality

The central theme of this volume is the relation between intentions and reasons. There are various ways in which the relevant questions could be introduced. One particularly fruitful way is by considering a crucial problem in the area of practical rationality.

The dominant conception of practical rationality in the social sciences, especially within economics, as well as in philosophy, is that of instrumental rationality. Rationality, on this view, is concerned with the individual selection of actions that are most effective and efficient in realizing the preferences of the agent, given her beliefs. The principal theoretical expression of this conception is the so-called *theory of rational choice*.[1] Though the dominant position of this theory would suggest otherwise, it has been under constant attack. For a large part, the criticism comes from authors who reject the underlying conception of instrumental rationality. Rationality, so they claim, is concerned with the selection of ends as well as means. Instrumental rationality, and with it rational choice theory, at best partially expresses our concept of rationality. However, there is also internal criticism from authors who accept the notion of instrumental rationality, but deny that orthodox rational choice theory is the best expression of that notion. Such critics point to several anomalies and difficulties within the standard theory.

---

[1] Strictly speaking, there is no single theory of rational choice. Rational choice theory studies three, related fields. First, there is decision theory. Under this heading one finds theories dealing with individual choice, of which expected utility theory is the best known, though certainly not the only one. Secondly, there is the study of interdependent, strategic choice: game theory. Finally, there is the study of collective choice: social choice theory.

One of these difficulties is how to account within the theory for rational commitment to a course of action. In particular the question if, and if so how, future-oriented decisions or intentions provide reasons for action or commit the agent in some other way, is central to understanding the difficulties that the theory encounters. The problem of rational commitment is relevant not just for the justification of formal game and decision theory. As I shall come to explain, the phenomenon of commitment poses interesting and fundamental questions in the philosophy of action. Indeed, it is one of the aims of this volume to demonstrate how and why these two areas of investigation share similar problems.

However, rational commitment is not just important for these more abstract sub-disciplines of philosophy. It is significant for some well-known questions in practical philosophy and social sciences alike. I mention five of these questions. As I shall come to explain, these could all be addressed if we would have an adequate theory of rational commitment.

Promise-keeping

First, consider the traditional question in ethics as to why promises should be kept. Suppose you have a painting that I would like to own. You are willing to sell it to me for $100. Unfortunately, I don't have that much cash at hand. I could give it to you tomorrow after a visit to my bank. However, this is the last day you will be in the country. Tomorrow you leave for a far and isolated place and it is unlikely we will ever meet again. So I promise you that I will transfer the money to your bank account if you will let me have the painting now.

On the standard picture of instrumental rationality, I will not have any reason tomorrow to transfer the money to your bank account. Either, you will not have

accepted my offer, in which case there is no reason for me to deposit the money in your account. Or, you have accepted my offer and I already have the painting in my possession. Again, there will be no reason for me to deposit the money in your account. I already have the painting in my possession and there is nothing you can do about that anymore. (It is too costly and difficult for you to sue me or otherwise take action against me from your new residence.) It follows from this that it is irrational to honour promises of this sort.

However, is it really irrational? If you realize this, you will not accept my offer and this means that I have no way to acquire the painting (nor will you be in the position to receive the $100). Only if I can get you to trust that my promise is sincere will this be feasible. One way to achieve this is to commit to pay the money if I receive the painting. Such a commitment seems rational since it enables me to get the picture. The standard theories of instrumental rationality, especially rational choice theory, seem to recommend otherwise. On such views, I should not commit to deposit the money if I receive the painting. In real life on the other hand, we see otherwise reasonable people committed to honouring their promises. Which leads us to the following quandary: either we should be willing to accuse such people of systematic irrationality. Or, perhaps, we need to think carefully about the standard picture of rationality and investigate if there are ways to account in rational terms for promise-keeping.[2]

One more won't hurt

---

[2] I am not claiming that incorporating commitment is the *only* way to account for promise-keeping within an account of instrumental rationality. Alternative explanations appeal to such things as reputation, convention, biological evolution of cooperative traits, or a combination of these.

Suppose that you like wine a lot. Recently you inherited your Uncle Geoffrey's wine cellar, which contains several boxes of the finest Australian Shiraz. You look forward to lots of evenings with lots of nice wine, but then your physician strongly urges you not to drink any alcohol anymore. "Continue at this pace," he tells you, "and you will be dead within a year!" Shocked by his diagnosis, you resolutely decide not to drink ever again. That evening you sit by the fire pensively looking at a bottle of Penfolds Grange and the following thought occurs to you: 'One more won't hurt!' You can achieve all the salubrious effects of a life without alcohol whether you empty this bottle tonight or not. The marginal effect of the bottle on you overall future health is negligible, but the positive benefits from imbibing this wine are considerable (or so Uncle Geoffrey promised you). Rationality tells you that in this case the best thing to do is to empty the bottle and then never drink anymore. However, tomorrow evening the situation is remarkably similar. Perhaps you are slightly hung-over from the previous evening but again the marginal effects of the next bottle of wine on you health are tiny in comparison to the glorious pleasures of drinking the wine. Rational cost–benefit analysis, in other words, would counsel you to drink the wine tonight, tomorrow and as long as Uncle Geoffrey's estate has not been emptied, rather than leading a life of abstinence.

However, this seems clearly absurd. For no matter how delicious all those wines are, life is preferable to death, even if it means a life without drinking the best Shiraz ever made! Clearly, something has gone wrong. A rational person, so it seems, should resolutely commit herself to a life of abstinence. Again there is reason to investigate how we could incorporate such commitments in a theory of rational choice.

The law is an ass

Third, consider the problem of a judge applying the law. In general, it is desirable that judges do so. This way, the law, or rather, the application of the law will be predictable. Predictability is one of the key ingredients of the rule of law. Furthermore, law, thus applied, serves its main functions of coordinating and arbitrating between various (legal) agents. However, it is inevitable that there will be occasions where strict adherence to the law seems undesirable in that particular instance. Philip Howard relates of the remarkable experiences of the Missionaries of Charity, followers of Mother Teresa, who wanted to establish a homeless shelter in Manhattan.[3] The city offered them two fire-gutted buildings for the symbolic price of $1 each. For New York the proposed homeless shelter would be "a godsend". When the sisters presented their plans for reconstructing the abandoned buildings to the building commission, they were refused a permit on the ground that their plans lacked an elevator as is required by the New York building code. The Missionaries of Charity, however, are ascetics. Their rule explicitly forbids them to use elevators, dishwashers and other modern appliances. Since elevators would not be used in the building, the nuns did not want to incur the expense of installing them. The nuns were told that the law could not be waived in their case, even though adding an elevator did not make much sense. The nuns never appealed this decision, but we can predict how a judge would (and should?) have ruled. A judge committed to uphold the law would rule in this case against the nuns. This seems utterly irrational on the standard picture of instrumental rationality. There seems to be no reason for the judge in this case to stick to the letter of the law. Both the city, by their own admission, and the nuns would have been best served if a judge ruled in favour of the sisters. More general, why is it rational to follow a rule, for example the way a judge follows a law, when it is obvious

---

[3] Philip K. Howard, *The Death of Common Sense: How Law Is Suffocating America* (New York: Random House, 1994).

that doing so in the case under consideration is less than optimizing? "If the law is an ass, what does that make judges?" On the other hand, if judges can vacillate between ignoring and applying a law according to what seems desirable in each instance, some of the fundamental benefits of having laws in the first place cannot be achieved. Again, how could the decisions of a judge committed to apply the law in each case be justified by a theory of rational choice?

Coordination

Fourth, consider the problem of coordination. Suppose you and I have lost each other in a museum. Suppose, furthermore, that we have decided that if we lose each other we will go to the restaurant of the museum. It seems perfectly straightforward that we each should go to the restaurant: it is the uniquely reasonable place to go.

However, it is far from obvious in this case that the restaurant is the uniquely rational place to go. You and I simply want to find each other again. Any place will do. The reception area, the masterwork, the special exhibition are equally good places to meet. Why should we go to the restaurant – just because we agreed to do so? Perhaps one would argue that by agreeing to go to the restaurant, we have set a special light on the restaurant. Because of our agreement it somehow stands out from the other possible meeting points. It has become salient. This may very well be true psychologically, but this does not explain that I would be making a mistake if I were to go to the reception area instead of the restaurant. But how do we account for this intuition, when the standard theory so clearly rejects it?[4]

---

[4] See also Sanjeev Goyal and Maarten Janssen, "Can We Rationally Learn to Cooperate", *Theory and Decision* 40 (1996): 29–49.

Cooperation

Fifth, and finally, consider the problem of cooperation. The classical Prisoner's Dilemma is a case where it is clear that both parties would benefit if only they could commit to the cooperative course of action. David Gauthier has argued that rational agents should commit to cooperate conditionally. Conditional that is, on a similar commitment of the other. Yet traditional game theory teaches us that it is irrational to cooperate, for similar reasons as mentioned above under the heading of promise-keeping: either the other player cooperates, in which case you do best by defecting; or the other player defects, in which case again you do best by cooperating. The predictable outcome is that rational players will achieve sub-optimal outcomes. It is also clear that if rational players could commit to mutual cooperation, they could avoid sub-optimality. However, here, as in all cases mentioned above, rationality seems to have no room for such commitment.

Commitment and Autonomous Effects

Though widely diverging, these problems share some features. First, in each case it seems that orthodox rational choice theory is at odds with common sense. The reason for this is that in each case common sense tells us that an earlier commitment of the agents in question is a consideration in the assessment of the rationality of the action at t=2. First, in the case of promise-keeping, if the agent in promising to pay $100 for the painting somehow commits to actually paying tomorrow, the agent has reasons to execute his promise and, consequently, the other has reasons that trust the agent's promise. Promise-keeping, and with it, promise-accepting, seems perfectly reasonable. Second, in the case of Uncle Geoffrey's wine, if the decision not to drink any alcohol anymore after hearing

the doctor's advice is a commitment, it is irrational to drink one more. Third, for the same sort of reasons, judges, who are committed to uphold the law, would not be irrational if they stick to the law in cases where this seems irrational in the absence of such a commitment. Fourth, coordinating to meet in the restaurant seems rational and if we understand the earlier decision to go there as a commitment, we can understand how that earlier decision is a real reason to go there. Finally, the case of the Prisoner's Dilemma could also be resolved if the parties concerned could commit to cooperate.

However, such commitments are notoriously difficult to include in orthodox rational choice theory. It is almost axiomatic on this theory that a rational agent considers only the options ahead. That is to say, the rational agent on this theory only includes forward-looking considerations in her deliberation.[5] Commitments on the other hand are backward-looking considerations. So we immediately encounter the problem of how to characterize the commitments of promise-keepers, teetotallers, judges and coordinating and cooperating agents. Furthermore, suppose we find an adequate characterization of commitments as providing backward-looking reasons, how would such a characterization fit within an overall theory of rationality?

To complicate matters even further, in most of the cases described above the decisions of the agents involved have so-called autonomous effects.[6] Usually, decisions only have the effect of producing the action. For example, the decision to uphold one's promise has the effect of actually honouring one's promise at the time of execution. However, in several of the cases described above the earlier decision has additional

---

[5] McClennen identifies this as the main defect in standard rational choice theory. Edward F. McClennen, "Prisoner's Dilemma and Resolute Choice", in Richmond Campbell and Lanning Sowden (eds), *Paradoxes of Rationality and Cooperation: Prisoner's Dilemma and Newcomb's Problem* (Vancouver: University of British Columbia Press, 1985), pp. 94–104. See also his contribution to this volume.

[6] The term is that of Gregory Kavka, "The Toxin Puzzle", *Analysis* 43 (1983): 33–6.

effects that cannot be attached to the intended action or its effects. In the case of promise-keeping, my commitment to bring the money tomorrow if you will hand over the painting now has the effect that the other does actually hand over the picture (assuming he relies on your commitment). Thus, your receiving the painting is not the result of your paying the other, but it is the result of your commitment to pay the other. Similarly with judges who commit to uphold and apply the law. They achieve that citizens can continue to predict what the law demands, because they can predict to some extent how a judge would rule. The continuous reliance on judges to apply and uphold the law is not the result of past judicial decisions, but the result of the commitment of judges to that effect.[7] The case of coordination is a clear case of autonomous effects. When we lose each other in the museum, I go to the restaurant – not because you actually go, but because I believe you will go and the basis for this belief is your commitment that you would go to the restaurant if we lose each other. My going to the restaurant is the autonomous effect of your commitment to go there. Similarly in the case of cooperation in a Prisoner's Dilemma. The reason I would commit to cooperate with you is your commitment that you will cooperate with me provided I have a similar commitment.[8] In other words, my commitment to cooperate with you is the autonomous effect of your commitment – not of your cooperation. More generally, we can see that in many cases the intuitive rationality of the commitment is due to the autonomous effects of such a commitment. Such effects are notoriously difficult to incorporate in standard rational

---

[7] Earlier rulings in accordance with the law are not irrelevant for reliance on this view. Such rulings are not themselves reasons for reliance, instead they are indications of the commitment of the judge. It is this commitment which justifies the reliance, not the signals of the commitment.

[8] However, Holly Smith argues that such symmetrical commitments of the type "I will cooperate if you will…" will not commit the agents to actual cooperation. Holly Smith, "Deriving Morality from Rationality," in Peter Vallentyne (ed.), *Contractarianism and Rational Choice* (Cambridge: Cambridge University Press, 1991).

choice theory.[9] We will return to the problem autonomous effects have for the assessment of the rationality of commitment later. First, we need to discuss the nature of commitment itself.

External Commitment

So how should we characterize the nature of commitments that come with future-oriented decisions? One way in which defenders of the orthodoxy try to incorporate such commitments is through the notion of *external commitment*.[10] The idea is that agents committing to a course of action at t=1 take measures which reduce the number of future alternatives at t=2 in such a manner that the intended course of action at t=1 will be chosen at t=2. The paradigmatic example is that of Ulysses tying himself to the mast, thus making it impossible for him to jump overboard and swim toward the luring Sirens. This way, Ulysses made sure he would stay on his course past the Sirens on his way to Ithaca. Commitment is a form of binding oneself, of reducing one's freedom for choice, on this view.[11]

---

[9] For a discussion of some of these problems: Ken Binmore, *Playing Fair* (*Game Theory and the Social Contract*, Volume 1) (Cambridge, MA: MIT Press, 1994); Brian Skyrms, *Evolution of the Social Contract* (Cambridge: Cambridge University Press, 1996).

[10] Jon Elster, *Ulysses and the Sirens* (Cambridge: Cambridge University Press, 1979); R. H. Strotz, "Myopia and Inconsistency in Dynamic Utility Maximization", *Review of Economic Studies* 23 (1956): 165–80.

[11] In the literature (especially Elster, *Ulysses and the Sirens*.), one can find two additional types of external commitment. First, one could manipulate the costs and benefits of future options in such a way that the intended course of action at t=1 will also be rationally acceptable at t=2 on the orthodox, forward-looking view. For example, Ulysses could make a high-stakes side-bet with his companions, such that the lure of the Sirens is outweighed. Secondly, one could try to tamper with one's future decision-making capacities in such a way that one could be sure that the intended course of action at t=1 will be chosen at t=2. For example, Ulysses could choose to

Ulysses' intervention in his future options is a form of *external* commitment because Ulysses side-steps his own future decision-making. He ensures causally that the intended course of action will be followed. To be sure, decisions, including those which result in commitments to a future course of action, have a causal impact. What distinguishes external commitment from its counterpart is that the salient explanation of the subsequent actions does not refer to the earlier decision as a justifying reason for the action. Instead, such explanations mention the causal mechanism that is put into motion. In the example of Ulysses tying himself to the mast, the salient explanation as to why he does not give in to the lure of the Sirens is not the fact that he has decided not to. Rather, it is the fact that he is tied to the mast and that, therefore, steering his ship towards the Sirens is not an option. In other words, what makes external commitment external is that it is independent of the rational powers, capacities and decisions of the agent at the time of action.

Many would argue that this means that the evaluation of the action at t=2 in terms of its rationality is moot. After all, Ulysses could not do anything but continue his journey. Scott Shapiro (in this volume) demonstrates the flip-side of this argument. Rational decision-making is a two-staged process on his view. First, one takes stock of one's feasible options. Next, one selects from these options the best. If the feasible

---

undergo hypnotherapy so as to resist the lure of the Sirens. Alternatively, he could try to distort his own future judgments about what is reasonable in such a way that he will elect to continue the journey to Ithaca. I have left out these additional types. The latter is not easily identifiable as a form of commitment that is rational, whether on the orthodox view or alternatives. The former (i.e. making side-bets) seems too broad and includes phenomena that we would not readily recognize as a commitment. For example, on this view, buying a train ticket is a form of commitment because you just made the option of riding the train less costly. Even a simple deliberation about something will be a commitment, because you have just spent some mental energy, thus making the act of reconsidering less attractive (it is 'cheaper' to just stick to your decision).

options are reduced to one, it is also the rational option as this is the only one that can be selected.

If this is the only way in which agents commit, the suggestion is not very promising, for it is rarely the case that one can rule out future options simply by deciding or intending.[12] Furthermore, even where this is possible, it seems costly and therefore a second-best solution. In addition, in many cases, the only possible form of external commitment seems to be to authorize a third party to interfere in one's future actions. This leads to similar problems as the ones discussed above, for why would such a third party act in the desired ways?

Internal Commitment

These problems can be avoided if the existence of internal forms of commitment can be made plausible. The basic idea of internal commitment is that the agent at t=1 decides to pursue a course of action at t=2 and subsequently acts as intended, where the earlier decision plays a determining role in the justification of the action at t=2. More precise, if an agent at t=1 is internally committed to φ at t=2, then it is rational for her to φ at t=2 because of her earlier decision. This would involve a fundamental deviation from the standard theory of instrumental rational choice. For it is almost axiomatic on this theory that the rational agent considers all and only the options that lie ahead of her. Past decisions carry no independent weight in the deliberations of the rational agent according to the orthodox theory.

In developing a theory of internal commitment which can correct the orthodox view of rationality, several fundamental problems need to be solved. First, there is a

---

[12] But see the contribution of Shapiro in this volume.

question about the nature of internal commitments. So far it has been suggested that an agent committing internally to course of action, simply decides at t=1 to φ at t=2 and then φ-s at t=2. This suggests that commitments are object of choice.[13] However, we should be careful not to replicate the problems of external choice we signalled above. Exactly how does such a commitment function? Does it remove options from the feasible set? Or does it change the preference order of the agent at t=2? Moreover, is such a commitment something that can be entered through choice?

A first suggestion along these lines was made by David Gauthier some twenty years ago.[14] He argued in the context of the Prisoner's Dilemma that a rational agent should choose a disposition for 'constraint maximization'. The agent who adopts such a disposition has made the decision on standard utility maximizing grounds not apply utility maximizing reasoning in her future choices. What is important is to note that apparently an agent can commit by adopting a disposition. Gauthier was careful to stress that an agent who has disposed herself in this way does not commit herself in any way externally. Thus, the disposition of constraint maximization is not merely a psychological means to remove future options from the menu or a form of endogenous preference change. This still leaves open the question about the nature of the disposition that embodies the internal commitment. Furthermore, it is unclear whether such a disposition is indeed an object of choice. Can one decide how to decide?[15]

Edward F. McClennen has argued that the term 'disposition' is misleading. On his view, a rational agent is one who is capable of 'resolute choice', that is, a rational agent

---

[13] It turns out that this suggestion is quite problematic. See the contribution of Thomas Pink in this volume.

[14] David Gauthier, *Morals by Agreement* (Oxford: Clarendon Press, 1986).

[15] Velleman argues that this is question-begging. J. David Velleman, "Deciding How to Decide", in Garett Culity and Berys Gault (eds), *Ethics and Practical Reason* (Oxford: Oxford University Press, 1997).

takes his past decisions in some contexts as a (decisive) reason for action. In other words, an agent who commits simply decides in favour of some future course of action and subsequently (if she is rational) acts on that decision. [16]

Apart from the ramifications of this idea of formal rational choice theory, we now enter into a discussion as to how to account for resolute choice in a plausible theory of intention. What would intentions be like if they carry such a commitment with them? One suggestion, made among others by authors such as Korsgaard, Robins and Mintoff, is that in forming an intention to φ, the agent has given herself a reason to φ.[17] This would explain why a decision to φ at t=1 can rationalize φ-ing at t=2.

However, on the standard view, as developed by Anscombe and, in particular, Davidson, an intention consists of appropriate beliefs and desires.[18] If this view is correct as it stands, there seems to be no room for internal commitments. First of all, since desires can change, so can one's intention. If I have formed to intention to φ at t=2, nothing stops me from changing my mind whenever my desires change. If intentions are such ephemeral states they lack the stability and robustness of real commitments, for it is constitutive of the latter that they persist even when the agent's desires change. Secondly, many authors influenced by the standard Davidsonian view assume that an agent's

---

[16] It should come as no surprise that McClennen is highly critical of standard rational choice theory. In his contribution to this volume he rehearses some of his criticisms.

[17] Christine M. Korsgaard, *The Sources of Normativity* (Cambridge: Cambridge University Press, 1996).; Michael H. Robins, "Is It Rational to Carry out Strategic Intentions?", *Philosophia (Israel)* 25, no. 1–4 (1995): 191–221.; and Mintoff, in this volume. For a criticism, see John Broome, "Are Intentions Reasons?," in Arthur Ripstein and Christopher Morris (eds), *Practical Rationality and Preference: Essays for David Gauthier* (Cambridge: Cambridge University Press, 2001).

[18] G. E. M. Anscombe, *Intention*, second ed. (Oxford: Blackwell, 1963); Donald Davidson, "How Is Weakness of the Will Possible?", in Joel Feinberg (ed.), *Moral Concepts*, Oxford Readings in Philosophy (Oxford: Oxford University Press, 1970); Donald Davidson, "Agency", in Robert Binkley, Richard Bronaugh, and Ausonio Marras (eds), *Agent, Action, and Reason* (Toronto: University of Toronto Press, 1971).

reasons for action just are her desires combined with the relevant beliefs. (For example, my desire to quench my thirst and the belief that this is a glass of water and that water has thirst-quenching properties is a reason to drink the contents of this glass.) Since an intention is the result of all the relevant desires and beliefs, an intention is best understood as the result of weighing all the reasons for and against the action. However, it is not itself a reason. On the standard view, then, intentions are not the sort of mental states that provide reasons. Intentions do not 'bootstrap' reasons into existence – or so it seems.[19] Given these two problems, we can conclude that, against the background of a Davidsonian theory of intention, the only type of commitment that is available is external.

Therefore, those sympathetic to the idea that future-oriented intentions carry some sort of commitment face the task of formulating a theory of intention that allows for such commitments. Many take as their starting point the seminal work by Michael Bratman, whose 'planning theory' of intentions explicitly includes the future-oriented commitment in his notion of intentions.[20] However, the planning theory need not be the only way that allows for some measure of internal commitment as some of the contributions to this volume demonstrate.[21]

The Rationality of Internal Commitment

---

[19] The term 'bootstrapping' in this connection comes from Michael E. Bratman, *Intention, Plans, and Practical Reason* (Cambridge, MA: Harvard University Press, 1987). In his contribution to this volume, Bratman considers the question whether long-term policies for individual conduct can provide reasons for action. In spite of his resistance to 'bootstrapping' he answers positively.

[20] As can be seen in virtually all the contributions to this volume.

[21] In particular, the contributions by Pink and Den Hartogh.

In addition to these difficulties, there are problems with the rationality of such internal commitments. As we have seen, on the Davidsonian picture, forming an intention is the result of weighing the considerations for and against the action. Note, however, that in most of the cases I described above, these intentions have autonomous effects. Such effects, it seems, have no place on the Davidsonian theory of intention. We can illustrate these difficulties as follows. Suppose you find yourself in the predicament described by the promise-keeping case. It seems that there are reasons for intending to deposit money in the bank account of the owner of the painting, but there are no reasons for actually depositing the money. In other words, the reasons for the intention point into a different direction than the reasons for actually paying. There are at least two views possible on how these reasons relate to each other. First, one can adopt the view that the reasons for forming an intention depend completely on the reasons for the intended action. In other words, one should only intend to φ if and only if there are independent reasons to φ (independent, that is, from one's intention or decision to φ).[22] Call this the *primacy of action* view on the rationality of intentions. Secondly, one can adopt a more inclusive view and argue that all effects, whether autonomous or not, should be considered in the reasons for the intention. In that case, one subscribes to the *primacy of intention* view, the view that all and only the reasons for the intention determine the rationality of the intended action.[23]

---

[22] Note that primacy of action does not entail that one should intend to φ if and only if the reasons for φ-ing outweigh all other reasons. If that were true, one ought not to intend to pursue A rather than B or vice versa in cases where A and B are indifferent, i.e. cases where the reasons for A are as strong as the reasons for B, as in the case of Buridan's Ass. Surely, that would be an implausible view and primacy of action is not committed to it. See also Joe Mintoff's contribution to this volume.

[23] The distinction between primacy of action and primacy of intention is that of Robins, "Is It Rational to Carry out Strategic Intentions?"

On the Davidsonian picture, the primacy of action has to be the correct view. The desires and beliefs that constitute an intention to φ at t=2 are about φ-ing at t=2. Since these desires and beliefs are, according to Davidson, the reasons for φ-ing, it is clear that only reasons for φ-ing can justify the intention to φ. Note how this works out for the cases described above. In the case of promise-keeping the consideration that without sincerely intending to deposit the money tomorrow the preferred outcome will not be realized is irrelevant for the determining whether or not to intend to deposit the money. Similarly in the case of the Prisoner's Dilemma, there is no reason to intend to cooperate if the other will cooperate, since there is no reason to cooperate at the time of execution, even if this means foregoing cooperative outcomes with similarly committed agents.

Things get more complicated when we consider the case of the judge who intends to uphold the law in each and every case. First, he could ask himself whether it makes sense to intend to apply the law in each case. Since he is well aware that there could be cases like the Sisters of Charity's bid for a permit, it would not be reasonable to form such a categorical intention. What about the benefits of predictability and the rule of law that would be achieved by the judge's policy? At this point, it depends on our analysis about what produces these benefits. Suppose that these benefits achieved because citizens look at past decisions and evaluate whether the law was applied in those cases. Then the judge would need to consider for each case like that of the Sisters of Charity whether the benefits of upholding the law in this case, including the more remote effects of promoting the rule of law, outweigh those of ruling in favour of the Sisters. It could very well be that there would be cases where the total sum of considerations, including those more remote effects, favours not upholding the law. It would not be rational in such cases to intend to uphold the law by the thesis of primacy of action. Note, however, that it is hardly ever the case that one can rule out this possibility *ex ante*, at the beginning of her career as it were. A judge, even an ideal one, is never in a position that she can rule

out with certainty that such cases will occur. Therefore, a rational judge should never form the intention to uphold the law in all cases.

Suppose, however, we would not accept the idea that the benefits that relate to the rule of law and the predictability of arbitration is the effect of each individual ruling. Rather, it follows from the policy of upholding the law, regardless the merits of waiving its requirements in individual cases. The idea is that these remote benefits are the result of the firm advance commitment of the judge to uphold the law. If that is correct, these benefits are irrelevant for the formation of the intention to uphold the law by the primacy of action view.

In conclusion then, it seems that on the standard, Davidsonian view on intentions, many cases of internal commitment are simply irrational. This irrationality is the result of the implicit assumption of primacy of action in assessing the rationality of intentions. So if one wants to salvage the view that internal commitment is (at least sometimes) feasible and rational, we have to abandon this view on intentions and look for one that does not presuppose primacy of action but instead allows for the primacy of intentions. In other words, we need a theory of intentions that assigns a role to intentions that goes beyond a mere aggregation of the reasons for action at the time of execution.

The Contributions in this Volume

This discussion sets the stage for the contributions in this volume. All the contributions deal in one way or another with the two central questions that I have introduced in the context of commitment. First, the question about the nature of the commitments implicit in future-oriented intentions. Second, the question about the rationality of such intentions as well as the rationality of acting according to them. The contributions can be categorized into two groups. On the one hand, there are those who are sympathetic to

the idea that future-oriented intentions provide new reasons for action, though they vary in the way they try to express this idea theoretically (Gauthier, McClennen, Shapiro, Mintoff, Finkelstein and Bratman). On the other hand, there are those that reject the idea that intentions provide reasons for action (Pink, Den Hartogh and Van Hees and Roy). However, the latter all try to show how future-oriented intentions carry some sort of commitment, even though the commitment is not necessarily that of a newly created reason.

Arguably the first to formulate a theory of internal commitment that abandoned the primacy of action principle was David Gauthier, who is also represented in this volume. His theory of 'constraint maximization', laid out in *Morals by Agreement*, is the starting point for several of the contributions in this volume (most notably, those of Finkelstein and Mintoff). Since then, Gauthier's thinking has developed. Mostly, he has worked on the conditions under which it is rational to commit oneself to a cooperative course of action. For example, in *Morals by Agreement*, he argued that such general commitments are rational, if the expected pay-offs that become available through the having of such commitments exceeds those of its alternatives. However, in his 1994 paper, "Assure and Threaten", he qualified this blanket justification of commitments by arguing that one should not commit to a cooperative course of action if there is a positive chance such a commitment is not beneficial.[24]

In his contribution to this volume he qualifies his theory even further. In his contribution, Gauthier focuses on the relation between reasons for action and motivation. Unlike the theory developed in *Morals by Agreement*, he no longer holds that every rational agent necessarily has to observe moral requirements, if he can expect

---

[24] David Gauthier, "Assure and Threaten", *Ethics* 104, no. 4 (1994): 690–721.

others to do likewise.[25] He now allows for the possibility that a person can escape the rational hold of morality. Just as someone who has no place for friendship in his life has no reasons for friendly acts, a person who finds no place for morality in his life has no reason to be moral. However, according to Gauthier that does not mean that there are no reasons of friendship or morality for most of us who are convinced of the benefits of friendship and morality. Thus, Gauthier still believes that intentions to cooperate conditionally constitute genuine reasons for action. However, he no longer believes that a rational agent necessarily has reasons to intend this. If we apply this to the problem of rational commitment to a course of action, this means that Gauthier's new view is the following. If, at t=1, an agent decides to φ at t=2, the agent is committed to φ-ing at t=2. However, it is not the case that this entails that it is necessarily rational to decide to φ at t=2.

Ned McClennen is well known as the originator of the theory of 'resolute choice' according to which, under certain circumstances, it is rational to commit oneself internally to a course of action. In this contribution McClennen expresses his frustration with orthodox game and decision theory. He carefully argues that the standard model of decision-making commits us to an "autistic" view of coordination, both intra- and interpersonally. On this view, the choices of future selves as well as those of other persons are mere conditioning variables that need to be regarded as "states of nature". He distinguishes between "compatibilist" strategies – strategies that remain within the standard model – and "revisionist" strategies to deal with this problem. He concludes that resolute choice provides the correct compatibilist alternative to deal with the shortcomings of the orthodox view.

---

[25] Gauthier, *Morals by Agreement*.

Claire Finkelstein criticizes accounts (such as those of Gauthier in *Morals by Agreement*, McClennen, Bratman, Mintoff and Shapiro) that make use of an automatic or quasi-automatic intention-execution device, such as disposition, habit or habits of non-reconsideration. She argues that such accounts fail to "rationalize" the action they produce, so that they model rational irrationality, rather than fully rational intention execution. The further question, then, is whether "pragmatic" accounts of rationality that attempt to justify sub-optimal actions in furtherance of optimal plans can do without such devices. She argues that they can, and that the constrained maximization folks were wrong to think they needed such devices in the first place. Her solution proceeds from the idea that in deliberation we deliberate about complete "packages" consisting of the intentions plus its performance. It is the overall value to the agent of such complete packages that determines the rationality of each of its parts. The relation of this overall value to the value and disvalue of its constitutive parts is not a simple aggregation according to her. This explains why it can be rational to form an intention because of its autonomous effects and execute the intention even when this seems sub-optimal.

Joe Mintoff argues that intentions can "bootstrap" actions into rationality: that, under certain conditions, forming an intention makes an action rational which would not otherwise have been rational. Some argue that this is so only because intentions are reducible to, or supplemented by, combinations of preferences and beliefs. After critically evaluating these views, he argues that intentions can bootstrap actions into rationality because they (together with beliefs) in and of themselves provide reasons for further intentions and actions, and sketches a theory of deductive practical reason in which intentions play a central role. Building on previous work by Castaneda and Aune, as well as his own work on "minimally constrained maximization", he supplements his view with a detailed account how previously formed intentions for action supply reasons for acting *now* in the context of rational deliberation.

Michael Bratman's earlier work on intention is central to most of the contributions in this volume. Since then he has been working on the connection between his theory of intentions and the idea of agency. In his contribution to this volume he returns to the theme of temptation. In an earlier paper ("Toxin, Temptation, and the Stability of Intention") he argued for a "no-regret" principle to explain the rationality of withstanding temptations, like the case of Uncle Geoffrey's wine cellar.[26] There, he also argued that this approach to temptation (appropriately) gives different answers when applied to Kavka's much-discussed toxin case. In the present essay he returns to these issues to see how they are affected by a pair of ideas he has been developing since then. The first idea is that there is a kind of valuing that involves a policy about what considerations to give justifying weight to in one's deliberation. The second idea is that assessments of instrumental rationality – given their relativity to ends – will lean on assessments of "agential authority" of certain ends or the like.[27] In this way he arrives at a

---

[26] Michael E. Bratman, "Toxin, Temptation, and the Stability of Intention," in Jules L. Coleman and Christopher Morris (eds), *Rational Commitment and Social Justice: Essays for Gregoray Kavka* (Cambridge: Cambridge University Press, 1998).

[27] Bratman's analysis of agential authority is related to that of Frankfurt. Harry G. Frankfurt, "Freedom of the Will and the Concept of a Person", *Journal of Philosophy* 68 (1971): 5–20; "The Problem of Action", *American Philosophical Quarterly. AP* 15 (1978) : 157–62; *The Importance of What We Care About: Philosophical Essays* (Cambridge: Cambridge University Press, 1988). The basic idea is that there is a difference between what an agent happens to do, or finds herself doing, and full-blooded rational action. For example, consider an addict who really does not want to continue using drugs, but nevertheless every time finds himself taking drugs. Each time he is overpowered by his desire for drugs. This is something the addict does, quite deliberatively as well. Yet, at the same time it seems intuitively plausible to argue that the desire for drugs "does not speak for the agent". Frankfurt famously argued that in such and other cases, the person does not "identify" with her actions. He went on to argue that the relevant lack of identification is constituted by the absence of a special second-order desire, in this case, a desire to be moved by one's desire to take drugs. Bratman's take on this is that the functional role that Frankfurt and others have given to such second-order desires is played by future-oriented intentions.

more complex story about relations between valuing and the will, in general, and, more specifically, policies and temptation.

Scott Shapiro takes these arguments in a different direction. Whereas the previous authors have argued that we need to abandon standard choice theory, Shapiro argues in favour of an account of commitment that stays well within the orthodoxy. On his view, intentions do generate reasons. However, he does not believe that intentions, as a form of commitment, really differ from external modes of commitment. Instead, the resolve to pursue a certain course of action is in his opinion analogous to external commitment. The agent makes it psychologically very hard, if not impossible, to deviate from the intended course of action by making a decision. By making the alternatives unavailable, the remaining feasible course of action becomes rational "by default", so to say. In this way, an intention changes the reasons one has, in the sense that it – literally – removes competing reasons from the scene.

Tom Pink, in his contribution, distinguishes two related disputes about the nature of intentions. The first is a dispute about intention-rationality: how far is intention a state formed in response to its own desirability and so possibly for reasons unconnected with the desirability of subsequently acting as intended, and how far in response simply to the desirability of subsequently acting as intended? The second is a dispute about the relation of intention and action: what place does intention have in intentional action, and is intention-formation itself an intentional action? The paper explores the connection between these two disputes. Pink shows that the standard Davidsonian model of intentional action has a peculiar implication that does not match our common sense psychology of intention. Rational intentional action is voluntary action: it is the result of our will, our decision. Consider then the case of arriving at a decision to φ. Pink shows that on the Davidsonian model one is forced to deny that the decision to φ is in any way voluntary – only φ-ing is. The reason is that the Davidsonian model takes the decision to

φ to be the result of a pro-attitude (for example, the desire) to φ. The decision is the passive, causal result of this attitude. This explains on the one hand, as we have seen, why, on this model, internal commitment to a course of action is impossible. It also explains that attempts at bringing in voluntariness into an otherwise Davidsonian model of decision making inevitably pay the price of what Michael Robins calls "incoherence": the reasons for deciding to φ can differ from the reasons to φ.[28] One can remove the incoherence only by either ignoring (some) reasons for deciding to φ (primacy of action), or by ignoring (some) reasons for φ-ing (primacy of intention). Pink, however, finds both these responses unattractive and investigates an alternative model. The assumption of both responses is that the forming of an intention is a passive causal process, much like the forming of a desire. In contract, Pink argues that the forming of an intention should not be treated as the forming of a desire. Instead, it is best understood in terms of a practical reason-based understanding of intentional action: to act intentionally is to make a distinctively practical exercise of rationality.

Govert den Hartogh's contribution is perhaps the most outspoken of the one's that argue against primacy of intention. He argues that the presence of autonomous effects is irrelevant for assessing the rationality of one's decisions. He shows what the intuitive basis is for thinking that such effects matter, taking his cues from the contexts of coordination and deterrence. This does not mean, however, that Den Hartogh rules out the possibility of internal commitment altogether. He argues that we need to pay attention to the role that 'content independent reasons' play in our deliberations. Taking his inspiration from Bratman's earlier work, he shows that we often commit ourselves to a course of action because of the risk of last-minute mistakes, for example, the intention not to make hasty investment decision. If a stockbroker approaches me with the

---

[28] Robins, "Is It Rational to Carry out Strategic Intentions?"

proposal to invest all my life's savings in Acme Incorporated in the next five minutes before the market closes, this intention gives me a reason not to accept this proposal. I have such a reason, not because Acme Incorporated is a not a good investment, but simply because I cannot assess its investment value within five minutes. In other words, my reason for not accepting this proposal had nothing to do with the desirability of such investments, nor with the autonomous effects of the decision to accept (should there be any). Instead, it simply is part of a policy to deal with my epistemic and rational limitations. In this sense, and this sense alone, prior decision and intentions commit an agent to a course of action.

The contribution of Martin van Hees and Olivier Roy, finally, is quite different in tone and degree of technical analysis from the other contributions to this volume. It is an attempt to incorporate some insights in the role of intentions in rational deliberation in formal rational choice models. Van Hees and Roy first give an axiomatic treatment for intentions to realize state of affairs in relatively simple parametric choice situations. They subsequently investigate some strategic cases, in particular the problem of coordination referred to in the beginning of this introduction. They show that the introduction of intentions in formal choice models matter in the following ways. First, introducing intentions can rationalize "focal points". Secondly, they show that intentions can simplify choice problems in ways which traditional utility-based analysis cannot. Third, perhaps most importantly, Van Hees and Roy make a convincing case that the introduction of some of the insights of traditional analytic action theory into formal rational choice creates a richness in analysis which allows all kinds of new and interesting questions to be investigated.

Reasons and Intentions: Some Conclusions

In spite of their diversity in approach, in is possible to draw some tentative conclusions about the state of the debate from the contributions to this volume. First, all of them, including the more critical ones such as Den Hartogh's, reject the traditional, Davidsonian theory of intentions. They reject the idea that intentions are reducible to combinations of desires and beliefs. Instead, they opt for a theory that assigns a special role to intentions in our psychologies. They all stress the central place of future-oriented intentions as opposed to Davidson and Anscombe's emphasis on intentional action.

Second, all of the contributions accept that future-oriented intentions in some way and under some circumstances commit the agent to a course of action. There is disagreement about the way and the circumstances in which such commitments occur. However, it seems internal commitment cannot be laughed away as "magical thinking" as some authors did not so long ago.[29]

Third, all of the authors in this volume make room for the idea that such commitments can be rational to undertake. Furthermore, although they differ over the question as to what features of commitments provide reason to undertake them, there seems to be no disagreement about the rationality of executing them. All the authors assembled here do not question that failing to execute one's (rational) intentions counts as a failing – a rational failing. Such an agent is in some deep and fundamental way inconsistent.

From these points of agreement it is tempting to infer that the contributions in this volume offer what could be the building blocks of a new theory of reasons and intentions. And while I believe that this volume provides the reader with a state of the art

---

[29] For example, Elster, *Ulysses and the Sirens*, and "Sour Grapes – Utilitarianism and the Genesis of Wants", in Amartya Sen (ed.), *Utilitarianism and Beyond* (Cambridge: Cambridge University Press, 1982).

collection dealing with reasons and intentions, I merely hope that such a theory can be advanced by this work.