TEMPORALLY EXTENDED RATIONALITY AND THE ETHICS OF BELIEF

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

In Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Emily L. Sherwin

December 2017

TEMPORALLY EXTENDED RATIONALITY AND THE ETHICS OF BELIEF

Emily L. Sherwin, Ph.D.

Cornell University 2017

Abstract

Actors may be called on to judge their reasons for action at two different points in time: once when they form an intention to act in the future and again at the time of action. At the time the actor forms her intention, her perspective is a general one, encompassing a range of possible circumstances that cannot be narrowed or fully specified in advance of action. At time of action, the actor's perspective is particularized, with more evidence available about reasons for action.

This difference in perspective presents a dilemma for rational agents. In many contexts, reliable advance planning has great value. It allows for intra-personal and interpersonal coordination and minimizes bias in favor of salient or emotionally charged facts. At the time of action, additional evidence about the context of the act clarifies, or appears to clarify, current reasons for action.

I describe this dilemma in two contexts: rule-following and promissory commitment. In each case there may be significant practical reasons for agents to stand by their original intentions, treating them as exclusionary reasons for action. Yet, if the agent revisits her intentions at the time of action, her reasons for action may support, or appear to support, a change in course.

I begin by examining theories of practical rationality that extend rationality

over time and thus permit agents to act on their initial intentions. Understood in this way, practical rationality may require agents to follow rules or honor promises without further consideration of reasons for action. I argue, however, that on a plausible understanding of epistemic rationality and epistemic responsibility to respond to evidence, acting without reflection may be epistemically irrational. If this is correct, then the dilemma of general and particular decision making persists, and affects important aspects of human life. We manage only by being imperfectly irrational.

BIOGRAPHICAL SKETCH

Emily Sherwin is Frank B. Ingersoll Professor of Law at Cornell Law School. She received her M.A. in Philosophy from Cornell in 2015, her J.D. from Boston University School of Law in 1981, and her B.A from Lake Forest College in 1977. She is the author of The Rule of Rules: Morality, Rules, and the Dilemmas of Law (Duke U. Press 2001)(with Larry Alexander); Demystifying Legal Reasoning (Cambridge U. Press 2008)(with Larry Alexander); Ames, Chafee, and Re on Remedies (Foundation Press 2012)(with Theodore Eisenberg), and many articles and book chapters on topics in law and legal philosophy.

ACKNOWLEDGMENTS

TABLE OF CONTENTS

Temporally Extended Rationality and the Ethics of Belief

Emily Sherwin

Chapter I: Overview

My aim in this project is to examine some practically and philosophically important instances in which epistemic rationality and practical rationality conflict. Much has been written about the conflicts that arise when an agent has practical reasons to adopt a belief that is not supported by, or is even contradicted by, the agent's evidence. It may be useful to believe the sky is green if adopting this belief will prevent an evil demon from demolishing the world.[1] It may be useful to believe in God, or in the efficacy of a cancer treatment, or in the innocence of a loved one, when evidence is lacking or cuts the other way.[2] Some have argued that the benefits that follow from beliefs of this kind can justify the belief; others have insisted that only evidence of truth can justify belief.

I have in mind a different type of conflict that arises when agents have practical reasons to place constraints on their own future actions. Agents who are neither omniscient nor omnipotent can control errors and conserve deliberative resources by adopting strategies that limit their own future all-things-considered deliberation about reasons for action. Theories of temporally extended practical rationality explain how strategies of this kind can be rational by incorporating prior intentions into the cognitive process that governs action. In doing so, they abandon or modify the traditional view of rationality as a synchronic match between actions and reasons for action. Thus, according to these theories, agents acting on prior rational intentions can be practically rational although their reasons for action, judged at the time of action, do not

support their acts. Theories of temporally extended practical rationality have been used to explain why it is rational for agents to follow general rules, and particularly rules of law, in at least some cases in which the agents' reasons for action appear to favor breaking the rules. They also have the potential to explain why agents are rational in honoring prior commitments in at least some cases in which the agent's reasons for action appear to favor abandoning the commitment.

To the extent these theories succeed, they solve very significant problems of rational action and morality. I will argue, however, that the most promising account of how agents can rationally constrain their own action and judgment over time succeeds only by endorsing epistemic irrationality. Specifically, the account on which I focus, developed by Michael Bratman, requires agents to disregard evidence that otherwise would require practically inconvenient adjustments to their current beliefs and avoid forming practically inconvenient beliefs based on that evidence. By plausible standards of epistemic responsibility, disregarding evidence and avoiding belief revision in this way is not epistemically rational. Thus, temporally extended practical rationality succeeds only at the price of epistemic rationality.[3]

This type of conflict between epistemic rationality and practical rationality is pervasive in daily life, but has not been much discussed in literature on either epistemology or rational choice. It differs from the problem of practically motivated belief not only because the agent is required to disregard evidence and avoid belief but also because the potentially inaccurate beliefs that result are closely entangled with the practical problem of what to do. When an agent chooses to believe in God or in a cancer treatment, the belief is the product of a practical choice but it is not a belief about the merits of the agent's choice. In the cases I describe the agent disregards evidence that otherwise would bear directly on the practical question the agent is trying to

resolve. I will also argue that in the contexts I describe, epistemic rationality and practical rationality interact in ways that make epistemic rationality prior to practical rationality. It follows that, although people often do follow rules and honor commitments, these practices cannot be explained as manifestations of rational agency.

A. Examples

Throughout this project, I will focus on two practices that exemplify the type of conflict between practical and epistemic rationality with which I am concerned: rule-following and promising. Each of these practices requires agents to form intentions at one point in time, then later to act on their intentions in particular cases. The temporal gap between intention and action, and the transition from general intention to a particular application of the intention, can work changes in the agent's reasons for action that in turn raise questions about the rationality of following through.

1. Rule-Following

The first example is drawn from legal theory, in which a persistent and important question is whether and how rational agents can follow rules.[4] A rule, as I use the term, is a prescription for action in set of circumstances described in general and reasonably determinate terms. The type of rule I have in mind is not simply a rule of thumb intended to guide the decisionmaker in close cases, but a rule intended to constrain the decision-maker's choice of action in all cases that fall within its terms. For example, an individual might adopt a personal rule in order to prevent procrastination: begin spring cleaning in the third week of April each year. A legal authority might adopt a rule prohibiting certain conduct: texting while driving a

vehicle is unlawful. Courts might adopt a set of enabling rules for private contracts: if certain conditions are met at the time a contract is made, the state through its courts will enforce the contract. In each case, the rule is designed to apply even when the rule-subject believes that her current reasons for action favor violating the rule.

Rules of this sort, designed to constrain choice, have a several potential benefits. If the rulemaker has special information or expertise, such as data on texting accidents or personal experience with the effects of procrastination, the rule may head off errors of judgment. Even if the rulemaker brings no special skills to bear, a firm rule can provide coordination. A fixed schedule can help an individuals coordinate her own actions over time. Rules of conduct create expectations about what others will do and so permit multiple actors to interact more confidently with each other. Transactional enabling rules allow actors to rely on the expected benefits of mutual agreements.

In each case, however, the rule is beneficial only if it is regularly followed. Rules designed for intrapersonal coordination break down quickly when treated as rules of thumb, because it will very often appear that some future occasion would be a more auspicious occasion for compliance with the rule. In the case of conduct rules and enabling rules, each violation or exception will indicate to observers that the rule is not always followed, and consequently will reduce the coordination value of the rule.

If all individuals subject to the rules were perfect reasoners, there would be no need for rules that constrain choice. In fact, however, human reasoners are not omniscient and are not capable of drawing complete and flawless inferences from evidence. Rule-subjects can take their own cognitive limitations into account in determining whether to act according to rules, but their assessments of these limitations will be similarly unreliable. If mistakes about whether to

follow rules were evenly distributed between mistaken compliance and mistaken violation, the effect on the coordination value of rules might not be significant. Individual rule-subjects, however, are not only likely to err but likely to error systematically against compliance with the rules. Cognitive studies suggest that salient or emotionally charged features of a decisional context have greater weight in deliberation than equally important background probabilities or long-term benefits such as maintaining the capacity of rules to provide coordination.[5] Rules of thumb, therefore, are not stable: the most important benefits associated with rules depend on the capacity of rules to constrain their subject's decision-making.

Thus, from the point of view of a rule-making authority or an individual who adopts a rule, good rules should be treated as authoritative at least within some range of cases. A rule is good rule, for this purpose, if the sum of outcomes over time will be better if everyone follows the rule than if everyone follows their own best judgment about reasons for action.[6] Not all rules are good rules in this sense, but if they are, then from the general, prospective viewpoint of the rulemaker, universal adherence to the rule is better than either unguided judgment or rule-guided judgment that treats the rule as a rule of thumb in cases of doubt.

The difficulty with rules is that even well-designed rules will require some outcomes that do not track the rule-subject's reasons for action at the time the rule comes into play. Because rules are announced in advance of conduct and cover generic types of cases, every rule will cover some particular cases that the rule-maker did not anticipate. Because rules must also be phrased in language that is determinate enough to enable subjects to comply, the terms of every rule will sweep in cases to which the rule ought not apply. Thus, a homeowner who has designated the third week of April for spring cleaning may conclude that some later date would be more auspicious. A driver may conclude on a particular occasion that she has a very good

reason to send a text. A judge may conclude that in the circumstances of a particular case, a technically valid contract ought not be enforced for reasons of hardship. In each case it will appear from the particularized point of view of the actor or the judge that following the rule will not produce the best outcome. This assessment is consistent with the initial assumption that the rule is, overall, a good rule, because all general and determinate rules carry the possibility of error in some cases. The individual reasoner simply believes that although it is best to follow the rule in most cases, it is best to ignore it in this case.

From the general point of view of the rule-making authority, however, this is not the right choice. The rule may be overbroad, but by definition, if the rule is a good rule reasoning errors will exceed errors of overbreadth. Moreover, reasoning errors will tend systematically to favor rule violation and, if observed, will reduce the coordination value of the rule over time. The result is a gap between the general point of view from which the rule is adopted and the particular point of view of those who are called on to follow the rule or apply the rule to others. Putting the problem in terms of rationality, it is practically rational to adopt the rule, but on a standard understanding of rationality as conformity to current reasons for action it may not be rational to follow the rule.

One possible solution to this dilemma is what Frederick Schauer has called "rule-sensitive particularism."[7] A rule-sensitive particularist considers not only the immediate consequences of following or not following the rule but also the effects that her own departure from the rule may have on the future capacity of the rule to reduce error and coordinate conduct. A rule-sensitive particularist would consider, for example, the likelihood that if she departs from the rule, others will observe this, conclude that the rule is not universally followed, lower their estimates of the coordination benefits of the rule, and consequently downgrade their own reasons

for following the rule. If this is a probably consequence of violating the rule, the rule-subject has an additional reason to comply, to preserve intact the benefits of the rule.

As Schauer points out, however, rule-sensitive particularism does not close the gap between the general reasons that favor adopting the rule and the rule-subject's current reasons for action.[8] Adding the value of a maintaining a reliable rule to the mix of current reasons for action does not ensure that the agent will calculate correctly. The effect of a single violation may seem negligible, and in any event, the rule-subject may undervalue remote consequences of her violation in comparison to the more salient immediate results of obeying the rule.

In response to the dilemma posed by general, determinate rules, Joseph Raz has proposed that rules should operate as exclusionary reasons for action.[9] Exclusionary reasons combine first-order reasons for action with second-order reasons not to act on a range of first-order reasons that fall within the ambit of the rule. Suppose, for example that a rule-making authority has enacted a rule "no texting while operating a vehicle." The authority enacted this rule after compiling statistics indicating that texting often results in serious accidents and that drivers regularly overestimate both the need to text and their own ability to compensate for the distraction by paying extra attention. According to Raz, this rule has the effect of excluding consideration of the probability of an accident, the importance of the driver's justification for texting, and the driver's capacity to concentrate on multiple tasks as reasons to text. Assuming the rule-maker was fairly thorough in considering possible reasons for and against texting at the wheel, the only reason for action open to a driver who wants to text is the first-order reason not to text provided by the rule.

From the point of view of the rulemaking authority, or of a rule-subject deciding whether to endorse this rule as a general rule for future cases, Raz's notion of exclusionary reasons

describes the way a rule should operate: assuming that rule-subjects err more often than they judge correctly in debatable cases, overall results will be best if all rule-subjects follow the rule in all cases. Yet it is difficult to explain how the enactment of a rule could have the exclusionary effect on rule-subjects that Raz describes at the time of application.

Anticipating an example that will recur later on, suppose that S is stuck in a slow-moving traffic jam when she hears on the radio that a lock-down is in place at her daughter's middle school. S would like to text her daughter at the school and her husband who works near the school. S understands the reasons for the no-texting rule and also understands that the rule is calculated to outperform individual judgment over time. But S also understands that the rule is general and overinclusive and she believes that in this case her first-order reasons for action favor breaking the rule. In the circumstances, there is no obvious mechanism by which, from S's point of view, the existence of the rule has an exclusionary effect on her current assessment of reasons for and against breaking the rule.

Thus, a rule-subject may judge that she has good reasons to adopt or endorse a rule that will govern her future actions, without exception. Nevertheless, when the time comes to apply the rule, the same rule-subject may judge that she has good reasons to break the rule. Both these judgments are practically rational, if rationality is understood as a response to current reasons for action at the time the judgment is made.

2. Promising and Other Forms of Interpersonal Commitment

Commitment plays a central role in moral and social life. Commitments of the kind addressed here arise most often through conventional mechanisms such as promising, which allow one person to assume an obligation to another voluntarily by communicating an intention

to be bound.[10] Assuming the mechanism is effective to bind the promisor, the result is to alter the rights and duties of the parties.

Promises and related commitments are similar to rules in that the promisor intends the promise to constrain her future choice of action. Thus, a promise might be understood as a self-imposed rule, to perform the promised act. Conversely, a self-imposed rule might be understood as a promise to oneself, requiring the promisor to act as the rule requires. The primary purpose of a promissory commitment, however, is not to reduce error but to create a new set of normative relations between parties. By promising, the promisor intends not only to constrain her own actions but also to confer a normative power on the promisee. If the promise operates as intended, the promisee can enforce the promise, or can choose instead to release the promisor from obligation.[11]

There is no consensus on the question whether promises and similar forms of commitment in fact alter normative relations between parties, or, if they do, on how exactly the alteration comes about. Hume took the position that there is no mental act capable of generating a moral obligation to act. Promises become binding, if at all, by virtue of social practices designed to advance social interests, which lead others to expect performance. The promise, in itself, does not create a new obligation that is independent of the consequences of non-performance.[12]

T.M. Scanlon has developed a theory that allows for promise-based obligation but does not depend on social practices and does not assume that promisors have normative power to impose new obligations on themselves.[13] Scanlon argues that when one person leads another to believe that she will take some future action, knowing that assurance of future action is important to the other, she becomes obligated to perform as expected. Because the promisee's interest is in

assurance that the promised act will occur, the duty that results is not just a duty to compensate for harm but a duty to perform. This type of promise-based obligation may be supported by conventions that encourage promisees to form expectations, but it also can arise from the interactions and understandings of the particular parties. The duty Scanlon describes, however, is not willed into existence by the promisor. Instead, the promisor acts in a way that brings into play a set of background normative obligations that one person owes to another when her actions generate expectations.

A different line of argument holds that autonomous individuals have the power to impose new, binding obligations on themselves through an act of will. Charles Fried, for example, has proposed developed a theory of contract law that links promissory obligations to "a morality of autonomy, respect for persons, and trust."[14] Seanna Shiffrin proposes that promising plays a key role in maintaining "morally decent" human relationships, which in turn are fundamental to human autonomy. The power to promise allows the parties to a relationship to make assurances, secure trust, settle differences, and right imbalances of power. Because we presumptively are capable of morally sound relationships, and because morally sound relationshipes depend on the power to obligate ourselves, we must possess this normative power.[15] David Owens argues that obligations and rights are important human goods, which we have an interest in controlling through mechanisms such as promising. This interest in control supports the existence of a normative power to stipulate that failure to perform a promised act will count as a wrong.[16] Similarly, Joseph Raz suggests that the ability to make binding promises provides promisors with "enhanced control" over their lives.[17]

The initial problem with theories of promissory obligation that rely on autonomy and other values that might be served by the power to impose new obligations on oneself is that the

potential benefits of having such a power do not establish that it exists.[18] A further difficulty is that, assuming the value of normative self-control supports the power to promise, it does not follow that the value of normative self-control supports an obligation to perform the promise. As Raz points out, keeping a promise may avoid a wrong, but it does not add to the promisor's self-control. A promisor who breaks a binding promise does not lose the power to promise and her later promises have the same moral significance as they would have had if she had kept the first promise. Accordingly, the value of self-control does not provide a full explanation for the bindingness of promises.[19]

Raz eventually locates a reason to perform promises in the value that promises provide to promisees. By making a promise, the promisor provides the promisee with assurance of performance. This idea initially seems compatible with Scanlon's view that the interaction between promisor and promisee generates expectations, which in turn bring harm-based moral obligations into play. Raz suggests, however, that in giving assurance of future performance, the promisee gives something of value to the promisee, consisting of control over the promisee's future actions.

Normally assurance of performance serves the promisee's interests, either directly or through the opportunity it provides to develop a new interest in performance of the promised act. Even if the promisee has no interest in performance, the promisor, by giving assurance of performance, gives the promisee the option either to enforce the promise or to release the promisor, and so relinquishes control of the question whether performance will serve the promisee's interests. The value to the promisee of having this assurance and control supports a reason for the promisor to perform the promise unless released.[20]

Raz's account has the virtue of tying ties the reasons for action created by a promise to performance of the promised act. Yet it does not fully explain how the promisor is able to give the promisee a valuable assurance of performance at the time of the promise. The promisor is obligated to perform at T_2 if the promisee gives valuable assurance to the promisee at T_1. The assurance is valuable to the promisee if the promisor will be obligated at T_2 to perform. The promisor will be obligated at T_2 to perform if the promisor gave valueable assurance of performance at T_1, and so forth. There is no obvious way to escape this loop of conditions.

Shiffrin offers two arguments for the bindingness of promises. The first of these explicitly endorces the idea of a transfer of control from promisor to promisee: in making a

promise, the promisor transfers to the promisee the right to decide whether the promisor should perform the promised act.[21] This explanation connects the power to promise to the duty to perform and also connects the duty to perform to the time of the promise, but it raises other difficulties. The nature of transferred right is not evident, and the rights the promisee receives may not perfectly correspond to the rights the promisor gives up.[22]

Shiffrin's second argument for the bindingness of promises at the time of performance is that promises, particularly within close relationships, solicit the trust of the promisee. By acting inconsistently with her solicitation of trust, the promisor commits a wrong against the promisee.[23] This argument, however, resembles Scanlon's argument, which is not an argument for normative power. Consequently, it fails to explain Shiffrin's initial assertion that the promisor binds herself by an act of will rather than by subjecting herself to pre-existing normative constraints against harm.

Assuming it is possible for an individual to provide herself with a new reason for action by intentionally undertaking an obligation, a further question is what weight such a reason carries. The reason generated by a promise is sometimes referred to as a content-independent reason for action, meaning that the reason reflects the promise itself, independent of what was promised and what consequences a breach of the promise may cause.[24] Raz, however, points out that a content-independent reason risks being a bare reason, meaning a reason that has no weight when balanced against contrary reasons for action. A bare reason not only lacks weight but also cannot generate secondary reasons for action such as promisee reliance or the negative effect of a breach on the general reliability of promises. Secondary reasons of this kind depend on the presupposition that the promise itself creates some initial reason to perform. If there is nothing that could, in principle, determine the weight of the initial promissory reason, then the reason is

not a reason at all and the secondary reasons that depend on it do not come into play.

Raz's answer to this problem is that the weight of the initial promissory reason for action generated by a promise is determined by the value of the assurance the promise provides, in the context of each particular promise.[25] The value of promissory assurance is not simply the value of the promised act. The act might occur in any event without help from the promise; or the promisee simply may not value the promised act. Yet promissory assurance continues to have value because it provides the promisee with an opportunity to develop an interest in performance, or simply because it gives the promisee the power to choose whether or not to hold the promisor to the oblication. According, Raz concludes that promissory reasons are not bare reasons, but reasons whose weight is ascertainable in principle, if not in practice.

Reviewing the possibilities: some reject the claim that promises generate obligations independent of the harm that may follow from a failure to perform. Others maintain that agents have normative power to impose obligations on themselves, regardless of prevailing social practice or the consequences of breach, by choosing to make a promise. Those who endorse the possibility of an independent self-imposed promissory obligation may disagree about how the obligation arises, whether it counts as a moral obligation, and what weight it carries when balanced against other considerations.

If no obligation arises from the act of promising alone, the problem of promising is much like the problem of adopting a rule. At T_1 the promisor intends to establish a rule for herself, that she will perform the promised act at T_2. Her intention is rational because following the rule will prevent harm to the promisee and help to avoid erosion of a social practice that supports reliable interaction, and because the promisor may not accurately judge at T_2 how much harm a breach of promise will cause. Yet at T_2, reasons not to perform the promised act may outweigh, or appear

to outweigh, reasons to perform. If rationality means responsiveness to current reasons for action at the time of action, then it may not be rational to follow the rule.

If promises generate content-independent obligations, then the picture appears to change. The obligation imposed by the promise becomes a reason for action, and consequently alters what otherwise would be the promisor's balance of reasons for or against performance at T_2. It now appears rational to perform the promised act at T_2, although in the absence of a promise it would not be rational to perform the same act.

The difficulty with this defense of the rationality of performing a promise lies in the nature of the promissory reason for action. If a promise generates a conclusive reason to perform or effects a transfer of the power of choice about performance from the promisor to the promisee, the constraint imposed on the promisor is complete. This interpretation of the effect of a promise, however, seems extreme: many promises are fairly trivial, and other obligations may intervene.

If the obligation to perform a promise is understood instead as an ordinary reason for action, to be weighed against other reasons, then an ideal reasoner with perfect information should perform if and only if the balance of reasons, including the promissory reason, favors performance. This characterization of the effect of a promise, however, poses several related difficulties. One is that, as already noted, the weight of a purely promissory reason for action is elusive.[26] Raz, for example, suggests that the weight of a promissory reason is limited and conditional, varies from promise to promise, and depends on the value of assurance to the promisee in context.[27] The second difficulty is that promisors, like rule-followers, are not ideal reasoners. They may misjudge the weight of promissory reasons, and in doing so they may overvalue concrete and immediate obstacles to performance in comparison to more the abstract

reason to perform generated by the promise.

As a result, the best way to protect the values underlying promissory obligation may be to treat the promise-based reason to perform the promised act as an exclusionary reason that not only favors performance in a balance of reasons but also blocks consideration of reasons that weigh against performance. Raz made this suggestion in his early writing on promising.[28] He first proposed a principle of voluntary obligation holding that a person who communicates her intention to undertake an obligation ought to perform that obligation. He then interpreted the "ought" in this principle as a reason for action that operates in two ways; it serves as a first-order reason to perform the promised act and also as a second-order exclusionary reason not to act on some range of reasons that contradict the first-order reason to perform. This proposal runs parallel to his solution to the promise of rules.[29]

An exclusionary reason of this type is exactly what is needed to preserve the values associated with self-imposed promissory obligation against subsequent changes of heart or miscalculations by promisors. Yet this approach faces significant obstacles. First, as just noted, the basis for an exclusionary rule is unclear. Assuming it is possible to ground an ordinary reason to perform in the value of assurance to promisees or the value of binding promises within relationships, these values do not necessarily support an exclusionary rule that blocks contrary reasons for action. At least, an additional argument is needed to justify this extension of the normative effect of a promise.

Second, and most important for my purposes, application of an exclusionary rule to block consideration of otherwise applicable reasons not to perform a promise may not be rational, if rationality is equated with responsiveness to current reasons for action. At T_1, the agent's reasons for action may support both making a promise and adopting an exclusionary rule that

will protect the unquantifiable values associated promising, such as autonomy, trust, or assurance. At T_2, however, circumstances may have changed so that the agent now believes that reasons for action, including the values served by promising, no longer support performance.

Thus, promises and other commitments pose a problem of practical rationality that is similar to the problem of rule following. In each case, the agent adopts a plan or assumes an obligation that calls for action at a future time. The plan or obligation is general in the sense that the exact circumstances in which it will apply cannot be known. Because the value of the plan or obligation depends on the agent's abiding by it in the future, it is rational for the agent to treat it as binding at the time of adoption. Yet it may not be rational to carry it out.

B. The Practical Problem

Practices such as rule-following and interpersonal commitment pose an internal problem for theories of practical rationality. Practical rationality, for purposes of this project, means instrumental rationality in action. An agent acts rationally if her actions are reasonably likely to advance her ends. I assume throughout that the agent's ends are rational: rational end-setting is a separate problem not considered here.

It is sometimes instrumentally rational for an agent to adopt a rule and form the intention to follow it in all future cases that fall within its terms. Similarly, it is sometimes rational for an agent to commit to perform a future act, intending to assume an obligation that will bind the agent within some defined range of future circumstances. In each case, the agent's intention is practically rational at the time she adopts it (T_1) because various benefits of rules and commitments depend on the agent's decision to place constraints on her own future decisionmaking. She must settle now on future action. Moreover, the rule or commitment the

agent accepts is not susceptible to case-by-case adjustment when reasons for action appear to call for a departure from the plan. A rule will not accomplish what the agent wants it to accomplish if she can always modify it to accommodate problems at the point of application.[30] A commitment will not have the normative force the agent desires it to have unless it will override contrary reasons for action in at least some of the circumstances in which the agent may be called on to honor it.

Although the element of constraint associated with rules and interpersonal commitments yields important practical benefits, constraint is at odds with a standard principle of rational choice. Following Edward McClennen, I will refer to this principle as the principle of separability. Separability holds that each time the agent makes a choice, her choice must be based on current reasons for action that reflect her current, forward-looking interests.[31] Only forward-looking considerations are relevant to the choice of what to do, because actions affect only the future. Thus, when the time comes for an agent to follow a rule or honor a commitment, Separability appears to require a fresh assessment of reasons for action. As a result, a rational agent cannot ultimately obtain the benefits of settlement and commitment because she cannot constrain herself in the manner she intends.

To solve this problem, a number of writers have proposed alternative theories of practical rationality that allow agents to plan ahead. In Chapter II below, I review competing theories in some detail. Each of them rejects the separability principle and all argue, in different ways, that it can be rational to act in accordance with prior intentions. For reasons I will explain at some length in Chapters II and III, I believe the most successful of these is Michael Bratman's theory of temporally extended agency.[32] Very briefly, Bratman argues that it is rational for an agent to act on a prior intention without further deliberation if, in doing so, the agent is guided by a set of

reasonable dispositions toward retention or reconsideration of intentions.[33] My later arguments about epistemic rationality will use Bratman's understanding of practical rationality as their starting point.

C. The Epistemic Problem

My central argument is that, although Bratman offers a plausible solution to the problems that rule-following and interpersonal commitment pose for practical rationality, his solution comes at the expense of epistemic rationality. In later Sections, I will explain the assumptions about epistemic rationality and epistemic responsibility on which this argument rests and develop the argument in more detail. At this point, I will preview the final argument by sketching the epistemic circumstances of an agent, S, who adopts a rule at time T_1 and is called on to follow the rule at time T_2.[34]

At time T_1, S's evidence indicates that S will do better over time by adopting rule R and following it in every case to which it applies than she would do by following her own unconstrained judgment. S's evidence indicates that R is overinclusive, so that in some unidentified set of future cases to which R applies, S would do better if she did not follow R; but it also indicates that S will not accurately identify all members of this set of cases. At T_2, S has the same evidence about the overall advantages of following R in all cases. She also has evidence indicating that in the particular case now before her, she will do better by not following R. She understands that following R is best in most cases, that her evidence about this case may be incomplete, and that her assessment of the evidence may be incorrect. But she also understands that by design, R is a general rule that does not prescribe the best result for every case it covers. If S has an epistemic responsibility to advert to her current evidence and form a

belief about her reasons for action, then on the assumptions I will make about epistemic rationality, epistemic rationality requires S to believe that she should not follow R in this case.

The epistemic circumstances of an agent who makes a promise or otherwise commits at T_1 to perform an action at T_2 are similar but not identical to the circumstances of an agent who adopts a rule. At T_1, S may have a variety of reasons to impose an obligation on herself to perform at T_2, including reciprocal benefits, a stronger relationship with the promisee, or the general ability to plan in concert with others. If in fact S has power to impose such an obligation on herself by willing it to be so, then the obligation will create a reason to perform at T_2. This new promissory reason to perform is likely to be salient in S's mind because it is linked to a sense of personal responsibility. Nevertheless, the difficulty of assigning weight to a purely promissory reason is likely diminish its role in S's practical reasoning at T_2. Therefore, as noted earlier, S has reason At T_1 not only to impose an obligation on herself but to resolve that the obligation will act as an exclusionary reason at T_2, blocking consideration of some range of reasons against performing the promised action.

At T_2, S will have the same evidence she had at T_1 about reasons to impose an exclusionary obligation on herself. She will also have evidence about promisee reliance and about the likelihood that a breach of her obligation will damage social conventions that rely on commitment. Yet, she may also have evidence indicating that the balance of reasons for and against honoring her commitment, taking account of reliance and harm to conventions. If S has an epistemic responsibility to advert to this evidence and form a belief about her reasons for action, then on my epistemic assumptions, epistemic rationality requires S to believe that she should honor her commitment in this case.

In settings of this kind, Bratman's theory of temporally extended practical rationality

holds that it is practically rational for S is to follow R, or to honor her commitment, without further deliberation. Specifically, if it was rational at T_1 for S to form an intention to follow to follow R or honor her commitment, and if S's reasonable dispositions toward prior intentions favor retaining the intention from T_1 to T_2, S should proceed to act on her intention. Suppose, for example, that S is generally and reasonably disposed to retain prior intentions unless the balance of reasons favors a change in plan by a factor of at least 2 on a scale of 10. S's evidence establishes reasons of strength 1 to change her plan. Practical rationality, as defined by Bratman, requires S to ignore her evidence and avoid the belief that she should not act on her prior intention. Yet, if S has an epistemic responsibility to advert to her evidence and form a belief about current reasons for action, this response may be epistemically irrational.

I will argue that in each of the cases I have described, plausible understandings of epistemic responsibility and practical rationality place epistemic rationality and practical rationality in conflict, in the manner just described. To do this, I must establish both that forming the belief that one ought not do an act affect the practical rationality of the act, and that agents have an epistemic responsibility to advert to current evidence that they ought not do an intended act. I attempt to do this in Chapter V below. If the argument succeeds, then practical rationality, in the temporally extended sense that makes effective rules and interpersonal commitments possible, requires S to be epistemically irrational.

I will also argue in Chapter V that, in the settings I describe, the conflict between practical rationality and epistemic rationality is not simply an impasse between two different ideals that can be noted and set aside. In situations of temporally extended agency, practical rationality and epistemic rationality are sufficiently entangled that epistemic rationality cannot be sacrificed without undermining key assumptions supporting practical rationality. Because my

arguments about conflict between epistemic rationality and practical rationality depend on the details of temporally extended practical rationality, I turn first, in Chapter II, to various theories of how agency can be extended over time.

REFERENCES

*References are to works listed in the bibliography (below).

[1] Foley (1987), p. 213.

[2] See, e.g., Chignell, pp. 7-9; James.

[3] Throughout this project, I use the term "epistemic rationality" to refer to a state of mind in which the agent's beliefs are justified and responsibly formed. Some who take a non-instrumental view of epistemic justification have refused to use the word "rational" to describe justified belief, to avoid the implication that beliefs are justified on instrumental grounds. See, e.g., Adler (2002), p. 11. Others have argued that epistemic rationality is a distinct form of rationality, not to be confused with instrumental rationality. See, e.g., Kelly (2003), pp. 612-613. Although I favor the non-instrumental view of epistemic rationality, I use the term in a broad sense that encompasses both instrumental and non-instrumental understandings of epistemic justification.

[4] See generally Alexander & Sherwin (2001), pp. 14-15; Raz (1986), pp. 49-50, 70-80 Schauer (1991), pp. 17-34, 135-66.

[5] See generally Judgment under Uncertainty: Heuristics and Biases (1982); Heuristics & Biases: The Psychology of Intuitive Judgment (2002).

[6] I assume that a good outcome is one that conforms to the rule-followers' reasons for action, but leave open the question of what should count as reasons for action. See Raz (1986), pp. 129-131 (discussing the "service conception" of authority).

[7] Schauer (1991a), pp. 94-100; Schauer (1991b), p. 676 & n. 66. Schauer describes rule-sensitive particularism this way: "Given that result *a* is indicated by rule *R*, you (the rule subject) shall reach result *a* unless there are reasons for not following rule *R* in this case that outweigh the sum of the reasons underlying *R* and the reasons for setting forth those underlying reasons in the form of a rule." Alternatively, the balance might be between reasons for not following R in this case and the effect of a violation in this case on the reasons underlying R and the reasons for setting them forth in the form of a rule.

[8] Schauer recognizes that, although an ideal approach to rule-following would be rule-sensitive, rule-sensitivity will in practice be ineffective to protect the value of rules because rule-subjects will err in making rule-sensitive judgments. Accordingly, he proposes a standard of "presumptive positivism," which adds a presumption in favor of rule-following to rule-sensitive particularism. Schauer (1991), at 691-94.

[9] Raz (1987), pp. 57-62; Raz (1979), pp. 16-19, 22-23, 30-33.

[10] See Raz (1977) and Raz (2012)(proposing an obligation-based account of promising, as opposed to an intention-based account, and discussing the normative functions of promises).

[11] See Raz (2012). In Hohfeldian terms, the promisee has a right to performance (or compensation for failure to perform) and a power to eliminate the obligation to perform. See Hohfeld (1913), p.

[12] Hume (1978), pp. 516-525. See also Rawls (1955)(describing the bindingness of promises as a rule imbedded in a beneficial practice and thus impervious to general utilitarian balancing).

[13] See Scanlon (1998), pp. 295-317.

[14] Fried (2012), p. 20; Fried (1981).

[15] Shiffrin (2008), pp. 497-498, 502-510, 517-519.

[16] Owens (2012), pp. 79-80, 95.

[17] Raz (2012), p. 61. For Raz, the obligation that results is normative in the sense that it creates a new reason for action, but it may or may not be a "moral" obligation. Whether a promissory obligation is a moral one depends on "the content and circumstances of [the] particular promise" rather than the fact that the promisor has made a binding promise. Raz (1977), p. 225.

[18] Raz acknowledges this difficulty, saying that "[i]t is impossible to have the power to promise, however good it may be to have it, unless [the fact] that one promised is a reason to do as one promised." Raz (2012), p. 69. "Yet," he goes on to say, "we do have it." Id.

[19] Id., pp. 67-69.

[20] Id., pp. 72-77.

[21] Shiffrin (2008), p. 517. Kant suggested a transfer theory of promising, stating that "". Kant (1991) § 20, p. 93. See Gold (2009)(proposing a transfer theory of contractual obligation limited to cases of fair mutual exchange); Owens (2012)(rejecting transfer theories).

[22] See Smith (2000), p. 120.

[23] Shiffrin (2008), pp. 518-519.

[24] H.L.A. Hart referred to reasons arising from promising and other forms of commitment as "content-independent" reasons for action. Hart (1982), pp. 254-55.

[25] Raz (2012), pp. 76-77.

[26] Some proponents of promissory reasons set aside the question of weight. Shiffrin, for example, states that it is wrong to break a promise, but does not say how wrong it is or consider whether and when performance of the promise might involve greater wrongs. Shiffrin (2008).

[27] Raz (2012).

[28] Raz (1977), pp. 221-223.

[29] See Raz (1987), pp. 57-62; Raz (1979), pp. 16-19, 22-23, 30-33.
[30] This is an argument I have made in previous work; I will not defend it at length here. See Alexander & Sherwin (2001), ch. 4.

[31] McClennen (1997), p. 229 (ultimately rejecting separability). McClennen describes the principle of separability thus:

> "*Separability*. The subplan you would prefer at a given node n_i within a given tree T (on the assumption you reach that node) must correspond to the plan that you would prefer at the initial node $n_1 = n_i$ in the modified tree $T(n_1 \rightarrow n_i)$."

See also Parfit (2001), p. 87 ("According to the standard version of the Self-Interest Theory . . . we should maximize at the level of our acts").

[32] Bratman (1987), p. 78.

[33] Id., pp. 60-101.

[34] In Section VI below, I will return to the agent's epistemic circumstances, adding details that conform to the epistemic assumptions outlined in Section V.

Chapter II: Some Theories of Temporally Extended Rational Agency

Nothing in an agent's synchronic reasons for action at the time of action can explain the special exclusionary force of authoritative rules and binding interpersonal commitments. Instead, the force of rules and commitments must come from the agent's past decision to adopt the rule or make the commitment. If these practices are to be justified as exercises of rational agency, standards of rationality must be extended over time in a way that makes this past decision relevant now.

In this chapter, I will examine four leading theories of temporally extended practical rationality, developed respectively by Edward McClennan, David Gauthier, Michael Bratman, and Scott Shapiro. Eache of these authors begins with an instrumental understanding of rationality: rationality requires agents to act in the way that will best advance their interests or desires. A plausible alternative equates practical rationality with responsiveness to a range of appropriate reasons that is broader than advancement of agent's interests.[1] For purposes of this project, however, I will assume a relatively narrow instrumental view.

A. McClennen and Resolute Choice

One theory of temporally extended rationality, notable for its directness, is the theory of resolute choice developed by Edward McClennen.[2] McClennen begins by describing two widely accepted forms of reasoning that, in his view, are not adequate to accomplish the instrumental objectives of a rational agent. The first is "myopic" reasoning, in which, at each point in time at which a choice is required, the agent chooses the action best calculated to advance the agent's preferences at that time. Myopic reasoning does not allow for effective planning: the agent may

prefer at T_1 to adopt a plan leading to action at T_2, but if the agent's preferences change between T_1 and T_2, the agent will not carry out the plan. As a result, long-term benefits may be lost.

Alternatively, the agent can engage in "sophisticated" reasoning. A sophisticated agent anticipates at T_1 what she is likely to prefer at T_2 and forms a plan that conforms to her probable future preferences. This method of reasoning permits the agent to plan, within the limits of what she is confident she will later wish to do. To some extent, the agent may also be able to manipulate her future preferences through precommitment strategies that create incentives for her future self to do what the plan requires. Sophisticated reasoning, however, does not enable the agent to secure the full benefits of long-term personal planning because there is no guarantee that her preferences at the time of action will in fact support completion of the plan.

McClennen notes that both myopic reasoning and sophisticated reasoning reflect the principles of consequentialism and separability. Consequentialism requires agents to choose the options best calculated to satisfy their preferences. Separability requires agents to determine, at each time when a choice is required, what option will best satisfy the preferences they hold at that time.[3] McClennen describes separability in terms of a decision tree: at each node on the tree, the agent must consult her current preferences for future outcomes. Past preferences or preferences or backward-looking preferences are inert because the future is the only period of time the agent's choice of action can affect.[4]

McClennan has no quarrel with consequentialism, but he rejects the principle of separability, which prevents agents from making effect plans by foreclosing consideration of past preferences. An agent may have a current preference for honoring a past intention, but unless that preference is based on the agent's current assessment that honoring her prior intention will have future good consequences, it has no role in present deliberation. Myopic reasoning is

strictly synchronic and therefore exemplifies the principle of separability. Sophisticated reasoning also follows the principle of separability, because permits the agent to carry out only those prior plans that reflect a correct assessment of the agent's current preferences at the time set for action. As a result, McClennen rejects both of the standard approaches to practical rationality.

McClennen's alternative proposal is for "resolute" choice, in which the agent "regiment[s] future choice to the originally adopted plan."[5] If at some node in the decision tree, the agent adopts a plan based on her current preferences for future outcomes, and if, when the time comes to carry out the plan, the agent's circumstances are compatible with the agent's initial expectations, then resolute choice requires the agent to act as planned. McClennen argues that resolute choice is instrumentally superior to either myopic choice or sophisticated choice because it allows for coordination between the agent's past and future selves. Myopic choice looks only to the present. Sophisticated choice allows for coordination of choices over time, but only through precommitment strategies that are costly and limited. In contrast, resolute choice Resolute choice accomplishes coordination without cost. Thus, as long as the outcome of resolute choice maximizes overall benefit to past and future selves, and apportions the benefit among the agent's selves in a fair manner, it represents a more perfect form of rationality.[6] Resolute choice means abandoning the separability principle, but McClennen argues that separability is not a valid principle but simply an unfounded assumption of traditional approaches to rationality.[7]

Although McClennen makes a good case for the practical benefits of resolute choice, there are gaps in the theory. Most important, McClennen does not fully explain the cognitive mechanism that enables an agent to honor a prior resolute choice in the face of current reasons

for a change in plan. It may be that prior choice changes the future self's preferences, or that it limits the choices that are now rationally "feasible," or that it creates a second-order preference with priority over first-order preferences. Without further explanation, however, each of these possibilities is metaphysically arcane.[8]

Another difficulty with McClennen's theory of resolute choice is that it does not account for the generality of many plans, including plans to follow rules or interpersonal commitments that call for action under a variety of unspecified future circumstances. McClennen indicates that an agent is not unconditionally bound to act on a resolute choice. Instead, a resolute choice is binding only when the facts at the time of action are as the agent expected them to be when she formed the plan.[9] There are at least two ways to understand this qualification. First, McClennen may have in mind that resolute choice means only that an agent who has chosen of plan of action based on her preferences at the time of choice cannot later change course based on a new preference about the exact set of circumstances she initially had in mind. If the facts the agent later faces are not facts that she visualized when she made the choice, she is not bound by her plan. If this interpretation is correct, the theory of resolute choice will not be effective in most cases involving rule-following or promising. An agent who forms a plan to follow a rule or honor a promise does not choose a particular response to a particular assumed set of facts; instead she chooses a type of response to a type of factual setting, the details of which situation that is not fully specified at the time of choice. As a result, she will always face new facts and her prior resolute choice will be excused.

Second, resolute choice may mean that an agent who has chosen a plan that calls for a type of action in a type of factual setting is bound by her former choice as long as the facts she now faces fall within the general type she initially had in mind. If so, resolute choice offers a

solution to the problems of rule-following and interpersonal commitment. Yet, if this is the correct interpretation of resolute choice, there is increased pressure for an explanation of why an agent who now believes that her prior self was wrong about the situation she now faces should not make an exception to her prior self's resolute choice. At the least, there is pressure for an explanation of how sure the present self must be that the prior self was wrong before abandoning the prior self's resolve. I will return to the problem of reconsideration in my discussion of other theories of temporally extended rationality and in later sections of the project.

B. Gauthier and Courses of Action

Another theory of temporally extended practical rationality, not far from McClennen's, is the theory developed by David Gauthier.[10] Gauthier's strategy is to redefine rationality in terms of courses of action or dispositions to act rather than discrete actions. According to Gauthier, a course of action, consisting of an intention to act in the future and a later act that conforms to the intention, will yield benefits the agent could not achieve through particularized decisions about what to do. For example, an agent who is able to follow a predetermined course of action over time can participate in mutually beneficial cooperative arrangements with others. In contrast, an agent who considers only the costs and benefits of particular acts at the time of action will be excluded from cooperative arrangements because she cannot plausibly assure partners that, having received a benefit, she will do her part in return. Therefore, adopting a course of action is often rational for the agent.[11]

Gauthier also claims that the practical benefits of a course of action give a rational agent a reason to perform the intended act, even if the act, viewed in isolation from the overall course of action, yields no benefit. To explain this conclusion, Gauthier begins with the premise that a

rational agent cannot adopt a course of action unless the agent expects when she adopts it that she will later follow through. If the agent expects at the outset that she will not follow through, the effect is a kind of inconsistency: the agent intends to do the act but also intends not to do the act. As a result, the agent cannot rationally form the initial intention to act and so cannot obtain the potential benefits of the course of action.[12] In other words, the agent must accept the course of action as a whole or not at all, and long-term practical rationality requires the agent to accept the whole. It follows, for Gauthier, that the agent's reason to adopt the cause of action is also a reason to do the act that completes the course of action.[13]

In an early version of his argument, Gauthier proposed that it is rational both to adopt and to carry out a course of action if, at the time the agent first adopts the course of action, the agent expects to do better overall by following it than she would do by not adopting it at all.[14] This formulation, however, can produce awkward results if the course of action does not meet expectations. Suppose, for example, that the cost to the agent of providing an agreed reciprocal benefit proves to be far greater than the benefit the agent has received, or that the agent's threat to retaliate against aggression fails to deter the aggression. In these cases, the intended course of action is no longer advantageous overall when the time comes to perform: the agent has no temporally extended reason to do the intended act and has current good reasons not to do the act. Yet, by Gauthier's early formula, the agent must carry through and perform.

In later writing, Gauthier revised his theory of rationality to encompass both the point at which the agent adopts the course of action and the point at which she completes the course of action. The agent must expect at the outset that the course of action will be advantageous, and also must continue to believe when she completes the course of action that the course of action as a whole, including the act that completes it, remains advantageous.[15] If the agent believes

when she is called on to complete the course of action that doing so will leave her worse off than she would be if she had never adopted the course of action, the agent can, and rationally should, reconsider the intention and decline to follow through. In these circumstances, following through is irrational even if the problem that caused the course of action to fail is something the agent anticipated as a possibility when she adopted the course of action.[16] Thus, if reciprocation proves to be more costly than expected, or if the agent's threat of retaliation fails to deter, the rational choice is to abandon the course of action; and once the agent abandons the course of action she has no further reason to perform the act.[17]

In this modified version of Gauthier's theory of temporally extended rationality, prior intentions exert considerably less constraint on the agent's future conduct than they did in the earlier version. The agent's initial conclusion about the *expected* benefits of adopting a course of action does not determine the rationality of completing the course of action at a later point. Instead, intentions are defeasible if the plan they set in motion proves to be disadvantageous as a whole. Yet, intentions continue to exert some force in later decisions to act: as long as the course of action as a whole remains beneficial, the agent can and should complete it even if the act of completion, viewed in isolation from the overall course of action, is costly. Direct maximization, act by act, is ruled out.

Initially at least, Gauthier's theory of temporally extended practical rationality appears to solve the problem of rules. A general rule prescribes a course of action: it instructs the rule-follower to act in a specified way in a class of future circumstances. According to Gauthier's theory of rationality, an agent considering whether to adopt the rule has two choices, either to accept the rule as a whole and follow it in all cases to which it applies or to reject the rule. If it appears, both at the time the agent first considers the rule and at the time the agent is called on to

follow the rule, that the course of action consisting of adopting and following the rule is more advantageous than never adopting the rule, the agent has reason to follow the rule. If, on the other hand, the agent determines when the time comes to follow the rule that following the rule is *not* more advantageous than never adopting the rule, the agent can and should reconsider and decline to follow the rule.[18] So - or so it seems - Gauthier's theory allows agents to captures the benefits of rules.

The difficulty is that Gauthier's theory does not easily accommodate the possibility of exclusionary rules. Because agents are likely to err in their judgments about reasons for action, and to err systematically in favor of violating rules, a rule is most effective if it not only prescribes conduct but also excludes exclude consideration of some range of contrary reasons for action. McClennen's theory of resolute choice excludes contrary reasons by treating the agent's initial assessment as conclusive unless circumstances have changed in some significant way. In contrast, Gauthier's course-of-action approach does not treat the agent's options as settled at the time of action. At least in the final version of the theory, the agent is expected to revisit the facts at the time of action to determine whether the course of action as a whole, including the act that completes it, is beneficial. The agent's prior decision to adopt a course of action prevents her from considering reasons for and against the act in isolation from the course of action, but these same considerations enter into her assessment of the overall costs and benefits of the cause of action. Ultimately, therefore, contrary reasons for action remain in play.

Thus, although Gauthier's theory of practical rationality allows agents to place certain constraints on their future actions by adopting a rule, it does not block consideration of reasons for action. Accordingly, it does not support the error-control function of rules. The problem may be even greater in the cases of promises and similar interpersonal commitments, which, at

least on some views, implicate values such as normative self-control and morally sound relations with others.[19] It is not clear how, or even whether, these values enter into the calculations Gauthier anticipates that agents will make at the time of action. Morever, exclusionary limits on deliberation, which appear to have no place in Gauthier's theory, are needed to prevent errors in assessing the weight of indeterminate values of this kind.

A more general difficulty with Gauthier's course-of-action approach, not limited to rules and interpersonal commitments, is that in both the earlier and the later versions of his theory, Gauthier achieves temporally extended rationality by giving up the standard and intuitively attractive association between the rationality of a particular action and the agent's current assessment of the consequences of the action.[20] In McClennen's terms, Gauthier achieves temporally extended rationality by giving up, or at least limiting, the principle of separability. In the early version of his theory, the agent's prior deliberative decision to adopt a course of action creates a reason to complete the course of action that is conclusive at the time of completion unless the agent's reasoning at the time of the prior deliberation was mistaken. In the later version of the theory, the agent's prior deliberative decision to adopt a course of action creates a conclusive reason to complete the course of action unless the agent was mistaken at the time of prior deliberation *or* the expected benefit of the course of action as a whole is disconfirmed at the time of action. In each case the reason for action generated by the agent's prior adoption of a course of action overrides the agent's current assessment of the costs and benefits of the particular act required to complete it. This allows mutually beneficial exchanges, and possibly authoritative rules, to go forward. Like McClennen, however, Gauthier fails to explain the cognitive process by which an agent can favor prior deliberation to current deliberation about reasons for action.

C. Bratman and Stable Intentions

In *Intention, Plans, and Practical Reason*, Michael Bratman develops a theory of temporally extended rationality for agents with limited deliberative resources that relies primarily on the continuing cognitive force of intentions over time. His objective is to show that an agent's prior intentions can constrain her actions without controlling or overriding her current assessments of reasons for action. In contrast to other proponents of temporally extended practical rationality, he begins with the view that "My intention today does not reach its ghostly hand over time and control my action tomorrow."[21]

My discussion of Bratman focuses on his early work, in which he developed an intention-based account of the cognitive process by which a rational agent can form and act on relatively stable plans. In later work, he extends his basic theory to problems of self-governance and shared agency.[22] Self-governance, for Bratman, refers to a form of autonomous agency in which the agent determines not only what she will do but what should count as justifying reasons in her deliberations about what to do, based on normative values she has accepted as her own. Shared agency comes into play when agents act together to achieve some end. Bratman's approach to each of these topics builds on, and preserves, the principles of temporally extended agency he first developed in *Intention, Plans, and Practical Reason*. Although Bratman's later work extends his early work in interesting ways, it is not directly relevant to the problems of rule-following and interpersonal commitment I address in this project. For the most part, therefore, I will leave aside later additions to the original theory.

Bratman begins with an instrumental understanding of practical rationality, in which actions are rational insofar as they promote satisfaction of agent's desires. He leaves open the

possibility that rationality may also place restrictions on admissible desires, but sets this problem aside.[23] He then distinguishes between the rationality of actions and the rationality of agents engaged in intentional acts. By standard accounts, the rationality of actions depends on the degree of correspondence between each of the actions the agents' reasons for action at the time action occurs. Agent rationality is a broader idea, that encompasses not only actions but also the reasoning process that leads to action and the "underling habits, dispositions, and patterns of thinking and reasoning that are manifested in" that process.[24] The question is how well the agent's processes of reasoning contribute through action to "long-term desire satisfaction."[25]

Bratman's norms of agent rationality are not designed to serve as internal guides to decisionmaking. Instead, they assess the agent's reasoning process from a mainly external perspective. The agent's beliefs and desires are taken as given. To avoid the problem of bootstrapping, however, her existing intentions are subject to scrutiny. The agent is rational only if she was asinitially reasonable in forming her intentions and also reasonable in retaining them over time.[26]

Although Bratmans' standards of rational agency are applied from an external perspective, and are strict in the sense that they do not require or imply fault on the agent's part, they require only a reasonable connection between the agent's decisionmaking process and instrumental success. In Bratman's words, "we do not . . . demand that such habits be *optimally* effective. . . As long as the expected impact exceeds an appropriate threshold, we can allow that there is room for some improvement and yet still judge the agent to be rational in the sense that she is not subject to criticism for the actions or intentions in which such habits issue."[27]

After laying this groundwork, Bratman turns to the problem of how agent rationality can be extended to capture the instrumental benefits of long-term planning. His theory of temporally

extended practical rationality rests on a particular understanding of the cognitive role of intentions. Intentions for Bratman, are not simply composites of beliefs and desires, but independent mental states. The feature that distinguishes intentions from beliefs and desires is an element of commitment to future action. Intentions, and the element of commitment they embody, are not themselves reasons for action, but they are part of the cognitive framework that motivates the actions of a rational agent. The result is a theory of practical rationality that locates rationality at the time of action but extends rationality over time by incorporating prior intentions into the process of current choice.

Bratman identifies two forms of commitment embedded in intentions: "volitional" commitment and "reason-centered" commitment.[28] Volitional commitment means that if an agent forms an intention and still retains it when the time comes to carry it out, the intention creates a "pro-attitude" that controls the agent's conduct without recalculation of reasons for action. Reasoning-centered commitment means that if an agent forms an intention and still retains it at the time of any future deliberation about actions or additional intentions, the prior intention is part of the framework for deliberation. Prior intentions affect the process of deliberation in two ways. They channel deliberation, for example by directing the agent's attention toward possible means for carrying them out; and they limit the options the agent is free to consider.[29] Both volitional and reason-centered commitment rest on a norm of consistency among intentions: an agent who has formed an intention in the past and has not abandoned it cannot rationally form conflicting intentions or consider conflicting intentional acts without first reconsidering and abandoning her prior, inconsistent intention.[30]

The key to Bratman's theory of temporally extended practical rationality is the idea that it can be rational for an agent to retain prior intentions over time and preserve the element of

commitment that accompanies them without revisiting the reasons that initially supported the intentions. To affect future action and deliberation intentions must, to some extent, be stable over time. Bratman assumes that intentions cannot and should not function as absolute constraints on future action or deliberation. If circumstances have changed significantly or if the agent simply realizes that her initial intention was a mistake, rationality requires the agent to reconsider.[31] In the run of ordinary cases, however, the agent can and should act on her prior intentions without further deliberation.

To give content to the idea of stable intentions, Bratman first distinguishes between ideal stability of intentions and reasonable stability of intentions.[32] This distinction tracks his initial assumption that standards of agent rationality are best understood as demanding only reasonable, rather than optimal, conduciveness to instrumental success. Under conditions of ideal stability, agents reconsider intentions if and only if the comparative benefits of a revised intention exceed the costs of renewed deliberation. This is not, however, an appropriate standard for non-ideal reasoners. Accordingly, Bratman develops a complex set of principles that link agent rationality to reasonable stability of intentions.[33] The general effect of these principles, which are described in more detail below, is to allow the agent to follow reasonable habits and dispositions pertaining to prior intentions, including a default rule that favors retention of prior intentions. Because the habits that guide the agent in retaining or reconsidering prior intentions need only be reasonable rather than ideal, the acts that follow will not always be practically rational acts, judged by the agent's current reasons for action. Under Bratman's understanding of agent rationality, however, the agent can be rational even if the act she performs does not conform to current reasons. Most of Bratman's discussion of temporally extended practical rationality is devoted to elaborating specific standards for retention of prior intentions, and the commitment

that accompanies them, over time. Because these standards are central to my discussion of practical rationality and epistemic rationality, I will set them out in detail. Bratman begins by noting that agents can respond to prior intentions in six possible ways, which are governed by differing norms.[34] The first response is "nonreflective nonreconsideration:" in this case, the agent simply retains and acts on an intention without further thought. Nonreflective nonreconsideration of an intention is not itself an intentional act; it is a simply a consequence of the agent's volitional and reason-centered commitments. In other words, nonreflective nonreconsideration follows from the agent's pro-attitude toward the intended act and the requirement of consistency that tells her not to entertain options that conflict with existing intentions when deliberating about further intentions.

The response opposite to nonreflective nonreconsideration is "nonreflective reconsideration," in which the agent spontaneously reopens the question whether to engage in an intended act, or implicitly reopens it simply by changing course. Nonreflective reconsideration is an intentional act, which may involve deliberation about current reasons for action. The agent does not, however, deliberate about whether to reconsider her prior intention. Instead, the agent simply abandons volitional and reason-centered commitments to her prior intentions and brings the full set of reasons for and against the intended action back into play.

The next two possible responses to prior intentions are deliberative. On reflection, the agent may decide not to reconsider a prior intention in light of on the costs associated with reassessment of reasons for action. Alternatively, if the costs seem worth incurring, the agent may decide on reflection to reconsider. Bratman notes, however, that deliberative decisions to reconsider prior intentions are rare: once the agent incurs the costs of deliberating about reconsideration, she might as well proceed to deliberate directly about the intention.

The last two responses to prior intentions occur when the agent either reconsiders an intention, or declines to reconsider the intention, based on a general policy the agent has adopted for managing classes of intentions. Bratman gives the example of a policy to reconsider insurance decisions once but only once per year. Policy-based reconsideration (or nonreconsideration) of intentions is deliberative, but deliberation takes place when the agent adopts the policy, rather than at the point of reconsideration.

Having identified the range of possible responses to intentions, Bratman develops norms of agent rationality for evaluating agents who engage in them. He begins with a general principle governing action on the basis of existing intentions, which he calls the "Intention-Action" principle. The Intention-Action principle holds that if it is rational for the agent to have an intention to do a certain act, and the agent acts accordingly, the agent is also rational in doing the act.[35] Bratman justifies this principle on the ground that, assuming that an agent's act are controlled by her current intentions, her rationality in doing the act and her rationality in holding the intention should be judged by the same standard.

Bratman then turns to principles governing the conditions under which is it rational for the agent to intend, at a point in time, to do an act. He ultimately settles on three principles, governing nonreflective retention of intentions, deliberative adoption of intentions, and application of policy-based intentions over time. All three principles are "historical" in the sense that they make practical rationality depend not only on circumstances at the time of action, but also on the combined processes by which the agent forms, retains, and applies intentions.[36] An agent is rational if she forms her intentions through an acceptable process of deliberation about reasons for action, reasonably retains them in the period intervening before action, and reasonably applies them to particular cases. The Intention-Action principle then makes it

rational for the agent to act on her intentions.

Bratman's first historical principle applies to what he calls the "basic case," in which the agent acts on a prior intention without further deliberation. The principle for this case holds that if (1) the agent rationally formed an intention to act, and if (2) it was rational for the agent not to reconsider the initial intention throughout the interim period between formation of the intention and action on the intention, then the agent is rational in holding the intention at the time of action.[37] Whether it was rational for the agent not to reconsider depends in turn on whether the agent's nonreconsideration was guided by generally reasonable habits and dispositions, reasonably applied to the particular intention in question.[38] Reasonable habits and dispositions pertaining to reconsideration of intentions typically include a disposition not to revisit an intention unless problems have arisen, such as changes in factual circumstances, desires, or related intentions. Beyond this, reasonableness turns on whether the "expected impact [of the habit or disposition] on the agent's long term interest . . . exceeds an appropriate threshold."[39] Bratman refers to this method of evaluating rationality as "a two-tier approach, an approach analogous in structure to certain versions of rule-utilitarianism."[40]

Bratman's second historical principle applies to the deliberative process by which the agent forms new intentions. The difficulty this principle addresses is that deliberation typically is not undertaken from scratch, but is constrained by other, pre-existing intentions of the agent. In deliberative cases, Bratman's historical principle holds that if (1) it is rational for the agent to hold the intentions that play a background role in the agent's deliberation, and if (2) the agent, deliberating about options that are consistent with the agent's background intentions, concludes that the act under consideration is at least as well supported by reasons for action as other admissible options, then it is rational for the agent to intend the act.[41] Clause (1) of this

principle, governing the agent's rationality in holding intentions that play a background role in current deliberation, refers back to the first historical principle, governing retention of prior intentions. Clause (2) requires the agent to assess new intentions consistently with whatever prior intentions she has rationally retained. The overall effect is that an intention that results from deliberation based either on irrationally formed background intentions or on background intentions the agent should have reconsidered is not rationally formed.[42]

Bratman also develops a historical principle of agent rationality for "policy-based" intentions. A policy-based intention is a general intention to act in a certain way in a recurrent type of situation, without new deliberation. In addition to being subject to reconsideration in the usual manner, policy-based intentions are defeasible in the sense that the agent may decide to "block" application of the policy on a particular occasion. Blocking a general intention, however, differs from reconsideration: an ordinary intention, once reconsidered, has no further effect on the agent's actions or deliberations, while a policy-based intention survives the particular instances in which it is blocked. In other words, the effect of blocking a policy-based intention is an exception to the intention rather than abandonment of the intention.[43]

Bratman's historical principle for the rationality of applying policy-based intentions is similar to the principle he applies to retention of ordinary, single-instance intentions. If (1) the agent rationally formed a general intention to do a type of act in a class of cases, and (2) it was rational of the agent not to reconsider this intention in the interim period before action, and (3) it was rational of the agent not to block the application of the intention to a particular case; then the agent is rational in holding and acting on the intention. Clauses (1) and (2) reiterate the historical principles for nondeliberative retention of intentions and deliberative formation of new intentions. Clause (3) establishes a separate standard, governing blocking or not blocking

policy-based intentions. This standard, like the standard for non-deliberative retention of intentions, refers to the agent's reasonable habits and dispositions. Rationality depends on whether the agent's blocking habits are likely to further the agent's interests over time, and the default position is not to block the application of the agent's established policy.[44] Although policy-based intentions do not play a major role in Bratman's early work, his standard for rational application of this type of intention is pertinent to the problems of rule-following and interpersonal commitment. In these contexts, the agent forms an intention to perform a type of act in a type of situation and later encounters particular circumstances that fall within the designated type.[45]

Bratman's historical principles of agent rationality have several advantages for his theory of agency and intention. First, they protect against the charge that the rationality of intentions can can be bootstrapped, by requiring that intentions be rationally formed at the outset and rationally retained to the point of action. Second, and most important for my purposes, Bratman's historical principles of agent rationality provide a means for extending practical rationality over time without completely abandoning the traditional focus on agents' current reasons for action. Bratman's principles allow an agent to bypass deliberation at the time of action and so to incorporate prior intentions into the overall cognitive state that determines what action she now will take. If the agent was rational in forming an intention, and if the agent's interim response to the intention is based on instrumentally acceptable habits and policies, then the agent is currently rational in acting on the intention even if the act might not be the best choice based on current reasons for action.[46] Rationality is temporally extended because rationality at the time of action includes not just rational calculation of reasons for action but also rational implementation of prior intentions. Intentions, however, do not control the agent's

present action from the past. Instead, they are continuing states of mind that play a cognitive role at the point of action.[47]

D. The Toxin Puzzle

Both Gauthier and Bratman have written on the subject of Gregory Kavka's toxin puzzle, in ways that shed light on their respective theories of temporally extended practical rationality.[48] In Kavka's puzzle, an eccentric billionaire offers an agent one million dollars if he can form an intention today to drink a vile of toxin tomorrow. The toxin causes misery for a day but has no permanent effects. If the agent forms the required intention, he will receive the money today whether or not he actually drinks the toxin tomorrow. Seen as a whole, the offer is advantageous: the agent will receive $1 million and will incur, at most, the cost of one bad day. Nevertheless, Kavka concluded that the agent cannot meet the condition. "[I]ntentions are . . . dispositions to act which are based on reasons to act."[49] Because the agent will have no reason to drink the toxin tomorrow when he faces the choice, he cannot today acquire the disposition to act. In making the statement just quoted, Kavka appeared to endorse the standard view of practical rationality as a measure of the responsiveness of an agent's actions to the agent's current reasons for action at the time he acts.

Gauthier purports to solve the toxin puzzle by treating the intention to drink toxin and the act of drinking as a single course of action, which the agent undertakes as a whole.[50] The agent cannot intend today to do an act tomorrow unless the agent expects to have reason tomorrow to do the act. According to Gauthier, however, this condition is met when the agent adopts a course of action. The agent expects today that when tomorrow comes, the intended course of action, including the final act, will appear superior to an alternative in which he never adopted the

course of action: drinking toxin "is part of the course of action with the best consequences."[51] The expected superiority of the course of action gives the agent a reason today to adopt the course of action in its entirety. Because the elements of the course of action are inseparable, the superiority of the course of action also gives the agent a reason tomorrow to do the act that completes it. Consequently, the agent can expect today that he will have reason tomorrow to complete the course of action by drinking the toxin, and this expectation enables him to form the necessary intention to drink.[52]

Although Gauthier's course-of-action theory offers an appealing solution to the toxin puzzle, his discussion of the puzzle suggests complications in the theory that may affect its application to more common problems such as rule-following or interpersonal commitment. Gauthier begins by explaining that to form an intention rationally, the agent must "expect to have adequate reason to execute the intention." The agent can expect to have adequate reason to execute the intention if he now expects that he will do better by forming and executing the intention than he would have done if he had not formed it.[53] Unless the agent's initial deliberation supports this conclusion, he cannot settle on a course of action and so cannot give himself a reasons to complete the course of action in the future.

Problems about the meaning of expectation emerge in the final section of Gauthier's essay on the toxin puzzle, in which he looks more closely at the nature of the agent's initial deliberation about a proposed course of action. One of Gauthier's reasons for adding this section is to generalize from the toxin puzzle to reciprocal benefit cases in which the agent expects to receive a benefit if she can provide sincere assurance to another party that she will confer a benefit in exchange. In this type of case, the benefit the other requests from the agent is not the agent's intention to perform an act but the act itself. The difficulty is that when the time comes

for the agent to perform the reciprocal act, the other will already have performed her part and the agent will have no immediate reason to reciprocate. Consequently, the only way the agent can provide sincere assurance of reciprocation at the outset is to form an intention to reciprocate, and, by Gauthier's reasoning, the agent cannot form this intention unless she expects that the course of action involving reciprocation is and will continue to be advantageous overall.[54] The agent's initial deliberation, therefore, must address both the costs and benefits of forming the intention to reciprocate and the costs and benefits of future reciprocation. Gauthier adds that "[d]eliberatively, the second question must be resolved prior to the first."[55]

So far, this is unsurprising. To illustrate his point, however, Gauthier poses a variant of the toxin puzzle in which the agent is offered $1 million to intend today to drink toxin tomorrow, with the added feature that the agent will be examined tomorrow before he drinks the toxin to determine whether he has a rare genetic condition that leads to permanent disability upon drinking toxin. In this case, forming the intention to drink the toxin and drinking the toxin is still the best course of action from the perspective of today, when the agent knows only that he has a very small chance of disability (one in one million). But it may not be the best course of action from the perspective of tomorrow, when he will know the outcome of the test. In Gauthier's words, the agent's expectation that he will do better by adopting the course of action "*may [tomorrow] have been falsified*."[56] Gauthier concludes that "realizing all this now, [the agent] cannot rationally form an intention to drink that would extend to the case in which [the agent] was found to have the adverse genetic configuration."

One way to read this statement is as a reiteration of the condition Gauthier endorses in later versions of his general theory of extended practical rationality, that an agent cannot rationally carry out an intended course of action if he learns before completing it that the course

of action as a whole is disadvantageous. If the course of action initially looked beneficial but proves with hindsight to be a mistake, the agent is released. Thus, in the revised toxin puzzle, the agent can form the required intention and win the money; but if, when the time comes to drink, he has learned that drinking the toxin is not worth the money, the intention is defeased, the course of action is terminated, and the agent need not drink.[57]

Gauthier's comments on the revised toxin puzzle, however, suggests a further limitation on temporally extended agency. His focus is not on the agent's deliberation at the time of action, but on agent's deliberation at the time he first forms an intention. Two conditions apply at this point: the agent must expect that the cause of action as a whole will be advantageous and he must not anticipate a situation in which he learns before acting that his expectation was wrong. Thus, in example, the agent realizes today that when the time to drink toxin arrives tomorrow he may know that he has an adverse genetic condition; consequently, he cannot form an intention today to drink tomorrow. The obstacle is not the one-in-a-million chance that the agent has the disabling genetic condition, because taking this small risk might be the agent's best course of action both today and tomorrow. The problem is that, when he forms (or would like to form) the intention to drink tomorrow, he is aware of a potential case in which he will know tomorrow that he has the condition. As Gauthier puts it, "I cannot form an intention rationally if I am aware that it would apply to circumstances in which I should do worse executing it than had I not adopted it."[58] As a result, the agent in the revised puzzle cannot form the intention, and cannot claim the million dollar prize.

This seems a fair resolution of the toxin problem. The original case is solved (the agent wins). If, however, the agent is already aware of a limitation that will prevent him from drinking toxin, it is difficult to say that he can form an intention now to do so. A similar limitation makes

sense in the real-life cases with which Gauthier is most concerned, such as undertakings for reciprocal benefit. Suppose, for example, that the agent is contemplating a course of action involving reciprocal benefits, in which the agent will receive something of value now in exchange for her undertaking to provide something of value in the future. If the agent currently has in mind a case in which she will not, and rationally should not, reciprocate, her assurance is insincere. It makes sense, then, that she should be disabled from making the offer.

Yet, a limitation of the type Gauthier proposes for the toxin case causes problems in the context of rule-following and interpersonal commitment. Rules are generalized instructions for conduct in a predefined range of cases; as such, they are not just likely but certain to produce the wrong result in some of the cases that fall within their terms. Thus, an agent who deliberates carefully about whether to adopt a rule will understand that in come specific cases that fall within the terms of the rule, the course of action will fail to yield a benefit. The same reasoning applies to promises and other interpersonal commitments whenever the commitment applies generally to a range of circumstances.

Thus, Gauthier's reasoning is in need of refinement to fit the context of generalized rules and commitments. If the relevant "course of action" is each application of the rule or each iteration of the commitment, the agent cannot form an intention to follow that course of action. As a result the agent is free to reject any rule or commitment at the point of application, even if following the rule or honoring the commitment in all cases would be beneficial. Alternatively, if the "course of action" is full compliance with the rule or commitment in all cases, then Gauthier's exception for bad outcomes the agent can foresee does not make sense in the settings I address.

Bratman's approach to the toxin problem differs from Gauthier's, both in reasoning and

in outcome. Bratman took a first pass at the toxin puzzle in *Intention, Plans, and Practical Reason*.[59] His concern at the time was that the toxin puzzle might pose a threat to his historical principles of agent rationality, in the following way. Bratman's action-intention principle holds that, at the time of action, if it is rational for the agent to intend (presently) to do an act, it is rational for the agent to do the act. His historical principle of non-deliberative rationality holds that it is rational for the agent to intend (presently) to do the act if it was rational for the agent to form an intention to act at some initial point, through deliberation, and also rational for the agent to retain the intention non-deliberatively up to the time of action, based on the agent's reasonable habits and dispositions.

Applying these principles to the toxin puzzle: the agent has reason today to form an intention today to drink the toxin tomorrow; and if conditions tomorrow are just as the agent expects them to be, it will be rational for the agent not to reconsider the intention tomorrow. It appears, therefore, that by Bratman's principles it is rational for the agent to drink the toxin tomorrow. For Bratman, however, this conclusion is false: it is not rational to drink the toxin tomorrow because from tomorrow's perspective, the agent is incurring a cost with no anticipated benefit. It seems to follow that one or both of Bratman's principles of agent rationality must be mistaken.

In *Intention, Plans, and Practical Reason,* Bratman escaped this conclusion by challenging the assumption that the agent is rational in forming the intention to drink and rational in retaining the intention tomorrow, at least as these conditions figure in the historical principle of non-deliberative rationality. His first point was that when the agent initially forms an intention to do a future act, she deliberates about the benefits of doing the act, not the benefits of adopting the intention. Therefore, benefits that follow from the intention independently of the

act are not reasons to form the intention. Benefits that follow from the intention may give the agent a reason to take some present action that will cause her to form an intention, such as submitting to hypnosis; but submitting to hypnosis may not count as a rational process of intention formation. Bratman's second point was that, even if causing oneself to acquire an intention through hypnosis is a rational process of intention formation, the historical principle of non-deliberative rationality is not applicable when the agent forms her intention in this way. As formulated by Bratman, this principle applies only when the agent formed a prior intention through *deliberation* - a condition that rules out hypnosis. Finally, Bratman argued that even if the historical principle were modified to encompass intentions formed without deliberation, reasonable habits governing reconsideration of intentions would require the agent to reconsider any intention formed by artificial means before acting on the intention. Thus, the historical principle is in applicable and the agent cannot rationally form an intention to drink toxin tomorrow.

Later, in the same volume in which Gauthier's toxin essay appeared, Bratman adjusted his view of the toxin puzzle.[60] In his later analysis, Bratman disavowed his earlier view that benefits resulting from an intention do not count in deliberation about whether to form the intention.[61] Instead, he isolates the toxin puzzle and the structurally similar case of reciprocal exchange as special cases, outside the scope of his full theory of temporally extended rationality. These cases differ from ordinary problems of advance planning in two ways. First, the benefits of holding the intention do not depend on doing the intended act, and second, all relevant facts are known in advance.[62] In these circumstances, stability of intentions is not at issue; therefore the agent does not need to rely on habits of reconsideration and Bratman's historical principle of non-deliberative rationality does not come into play.

Having isolated the toxin puzzle as a case not covered by historical principles of practical rationality, Bratman his new answer to the puzzle by defining the agent's "evaluative ranking." The agent's evaluative ranking is her ranking of possible actions at a particular point in time, given the agent's values, desires, and beliefs. Bratman then puts forward two general principles of deliberative rationality. The first is the "linking principle," which holds that at the time an agent forms a conditional intention (given circumstances C, do A), the agent must not have in mind that later, when circumstances C occur, it will not be rational for her to perform the intended act.[63] This principle is much like the initial premise of Gauthier's argument: a rational agent cannot form an intention if she expects from the outset not to follow through. The second principle is the "standard view" of rational action, which holds that the rationality of an action depends on the agent's preferences at the time of the act. This is the view that McClennen calls separability.

Next, Bratman invokes McClennen's distinction between sophisticated choice and resolute choice. Sophisticated choice, as Bratman describes it, combines the linking principle with the standard view of act rationality. In the toxin and reciprocal benefit cases, the result is that the agent, knowing today that her evaluative ranking tomorrow will favor not drinking toxin (or not rendering aid), cannot presently form the intention to drink toxin (or render aid). The agent's preferred course of action must give way to accommodate her expected evaluative ranking at the time of action.[64] So in each case, the agent loses.

Resolute choice accommodates the linking principle but abandons the standard view of instrumental rationality: rationality depends on the agent's evaluative ranking of options at the time she formed the intention. Thus, at the time of action, a prior plan will override the agent's current evaluative ranking. In the toxin and reciprocal benefit cases, the agent can form the

intention, but is then rationally bound to follow through. Bratman interprets Gauthier's theory of rationality as a weak version of resolute choice, subject to a second assessment of the intention at the time of action. Under either version of resolute choice, the agent succeeds in forming the needed intention.

Bratman, however, rejects resolute choice as inconsistent with the nature of agency. Deliberative agency has a temporal and causal location: it occurs in the present and looks forward. Accordingly, a rational agent ranks only what is under the agent's control, which is the agent's action from the present time onward.[65] Prior plans can have a degree of stability for rational agents, based on the agents' reasonable habits of non-deliberative non-reconsideration. But when an agent *deliberates* about action, her judgment cannot be based on her evaluative rankings at a previous time.[66]

Although Bratman treats the toxin puzzle and reciprocal benefit cases as anomalous, his discussion of them sheds light on his overall approach to the role of intentions in rational agency. Specifically, it clarifies that he does not mean to achieve a temporal extension of agency by giving up the standard model of deliberative rationality, in which deliberative agency is located at the point of action and is based solely on the agent's preferences at that time. McClennan and Gauthier are willing to abandon this model in pursuit of instrumental advantage. Bratman retains it, but adds to it a special form of agent rationality that permits agents to retain and act on prior intentions without further deliberation, based on generally advantageous habits and dispositions, in order to conserve deliberative resources. Past decisions exert no direct control over present actions. Yet, it may be rational for the agent to retain and act on intentions without reviewing their merits, based on reasonable dispositions toward prior intentions.

Thus, Bratman's version of temporally extended practical rationality is significantly

different from McClennen's or Gauthier's: rather than rely directly on prior deliberations as determinants of present action, he relies on the agent's reasonable disposition, at the time of action, to act on a prior intention without reflection. Consequently, the challenges McClennen and Gauthier face are significantly different from the challenges Bratman faces. McClennen and Gauthier owe an account of the degree of firmness built into prior intentions and the mechanism by which prior intentions modify or override current deliberation. Bratman owes an account of the degree of commitment the agent can rationally extend to prior intentions at the time of action.

I think Bratman is correct that agency is temporally located and that assessments of rationality should refer to the time of action. If, as McClennen and Gauthier suggest, a past decision to adopt a course of action, based on the agent's past preferences and expectations of future preferences, can directly affect the agent's reasons for action later, the location of agency is uncertain and agency itself becomes a changeable and mysterious concept. By comparison, Bratman's account yields a notion of practical rationality that may be more complex, but is conceptually stable.

E: Shapiro, Legal Plans, and Nonfeasible Options

Scott Shapiro has developed a theory of temporally extended practical rationality designed specifically to explain the authority of legal rules. Shapiro depicts law as a system of interconnected plans for ordering conduct in society.[67] Shapiro's theory of law is a positivist theory: in his view, the overall objective of law is to settle moral controversy; the authority of law over its subjects, however, depends not on its moral content but on the fact that legal actors have accepted a master plan setting conditions for development of particular rules. Thus, the authority of law rests ultimately on its subjects' intentions to follow legal rules.

Although Shapiro cites Bratman's plan-based conception of rational agency, his approach to rationality has more in common with McClennen's. Shapiro's theory, like the other theories discussed above, is instrumental. A rational agent may conclude that the best way to obtain the settlement and coordination benefits associated with a legal system is to endorse the system's master plan and commit to follow its rules. Shapiro explains that in deliberating about whether to make this commitment the agent is "strategically interacting with another agent - his later self."[68] Once made, the decision is binding on the agent's future selves, who no longer have the option of violating the system's rules.

This sounds very much like McClennen's resolute theory choice. Shapiro's theory, however, appears broader than McClennen's. Shapiro describes the effect of the agent's decision as "causal."[69] Once the agent accepts the system's rules, his deliberative reasons for action are exhausted and the only remaining reasons for action are the "implementation" reasons that require the agent future self to follow the rules.[70] Actions that do not conform to the agent's plan are simply not available, in Shapiro's words, they are not "feasible."[71] It follows that the agent can no longer choose not to comply in a particular case, because there is no choice left to be made.

McClennen's theory is more flexible in several ways than Shapiro's theory of nonfeasible options. First, McClennen avoids the notion of nonfeasibility.[72] He also places several conditions on binding effect of a resolute choice on the agent's future selves. First, the plan the agent fixes on at T_1 must result in a net benefit to all selves, distributed in some fair way to among those affected.[73] Second, circumstances at T_2 must in some sense be as the agent expected when she adopted the plan.[74]

Shapiro's theory does not include reservations of this kind, perhaps because it is designed

to enable rational to agents to adopt a system of legal rules. The point of a legal rule is to govern a class of future cases, the particulars of which cannot be specified in advance. The point of the master rule for a legal system is to authorize legal officials to enact and revise rules in response changing social conditions. Thus, in the context with which Shapiro is concerned, the agent will not anticipate all future applications of the rules, and may not even anticipate the terms of the rules. This means that unless Shapiro's theory is modified to allow for reconsideration under specified conditions, it places very significant restrictions on the agent's future selves. I will return to the problem of reconsideration in the following chapter.

F. Assumptions Going Forward

In the remaining sections of the project, I assume that most promising account of temporally extended practical rationality is Bratman's. Bratman's theory provides the best explanation of how prior intentions control action. It also takes seriously the intuitively attractive idea that assessments of an agent's rationality in performing an act should focus on the agent's intentions, desires, and beliefs at the time of action. Accordingly, it does not ambiguate agency by allowing past assessments of reasons for actions to directly control current assessments of reasons for action.

Chapter III: Some Special Problems of Temporally Extended Practical Rationality

The main objections I will make to theories of temporally extended practical rationality are that they depend on epistemic irrationality and that reliance on epistemic irrationality puts their own foundation into doubt. I pursue these objections in Chpaters IV and V. Before addressing the epistemic problems affecting temporally extended practical rationality, however, I will briefly discuss several other difficulties with theories of this type. One is the relationship among deliberation, intention, and action; another is the generality of intentions. These problems are connected in indirect ways to the epistemic problems I take up in the concluding sections.

A. Deliberation and Action

Standard descriptions of practical rationality assume that agents deliberate about current reasons for action, form an intention, and act on the intention, all in close succession. The agent's reasons for action reflect the expected value of the act as a means for realizing the agent's ends. The exact output of deliberation, however, is a matter of debate. Aristotle suggested that deliberation takes the form of a syllogism, the conclusion of which is the agent's act.[75] If finding food is necessary to satisfy hunger, and if S is hungry, then S finds food. The puzzle is how reasoning can conclude in action.

I will focus here on three different interpretations of what actually occurs when an agent deliberates and then acts, offered by Joseph Raz, John Broome, and Jonathan Dancy. Raz takes a restrictive view, arguing that the conclusion of reasoning is always and only a belief.[76] If S deliberates and concludes that she has a conclusive reason to find food, she will automatically come to believe that she should find food.[77] Deliberation, however, does not yield either an

intention to find food or the act of finding food: some further intervention of the agent's will is needed to bring about either of these responses.

Raz gives two explanations for his conclusion that deliberation cannot yield intention or action. The first of these relates to the role of will in rationality. Although S's conclusion that she should find food automatically yields a corresponding belief that she would find food, she may in fact never find food or form an intention to find food because she may not muster the will to do what she believes she should. Will, for Raz, is a rational power, but it is distinct from reasoning. S's reasoning comes to an end when she forms her belief and only then does will come into play. Under normal circumstances in which the agent's rational powers are functioning properly, will follows from belief; but if will fails, neither intention nor action will follow.

Raz considers and rejects two possible counterarguments to his claim that no reasoning occurs after the agent forms a belief about what to do. One is that the transition from belief to intention is the "practical" content of practical reasoning; in Raz's view, this argument fails because nothing that can properly be called reasoning occurs at this stage. A second counterargument is that when an agent forms an intention to act at a later time, her intention carries forward and prevents the agent from making inconsistent decisions in the interim before action; therefore her intention should be understood as a continuing embodiment of the agent's prior reasoning. In Raz's view, however, the fact that prior intentions affect interim reasoning does not imply that they embody prior reasoning; it means only that a prior intention may trigger additional and distinct deliberation about means for carrying out the intended act. Thus, Raz concludes that reasoning comes to an end when the agent forms a belief about what to do. The agent's intention, if any, is the product of this belief plus the agent's will, with no further

reasoning involved in the transition from belief to intention or from intention to action.

Raz's second explanation for his model of practical deliberation is that reasoning may lead the agent to conclude (and believe) that two or more actions are equally supported by reasons. The agent must then form an intention to engage in one of them and not the other. This choice cannot be the product of reasoning, because reasoning ranks the options as equivalent. Again, reasoning yields only a belief that the different options are permissible, and an additional element of will is need to generate the agent's ultimate intention or action.

Broome follows a different path, which leads to the intermediate position that deliberation can conclude in belief or intention, but not in action.[78] Action requires physical ability in addition to reasoning ability, and so cannot result from reasoning alone. Broome argues, however, that reasoning includes not only theoretical reasoning but also "intention reasoning." Intention reasoning is a form of instrumental practical reasoning in which an agent reasons from a general intention (to pursue some an end), and a belief about necessary means to the end, to a conclusion that consists of an intention to pursue the necessary means.[79]

Broome's defense of intention reasoning is grounded in a propositional approach to deliberation: reasoning is primarily a matter of inference from one proposition to another, rather than the application of rational powers to distill reasons from states of affairs. Beliefs and intentions are both proper outcomes for reasoning, because both have propositional content and both have connections to truth. Beliefs track true propositions, while intentions track (with some limitations) propositions the agent is "set to make true."[80] Thus, beliefs and intentions are sufficiently similar in propositional content that at least some least some inferences from one to the other are normatively required by the rules of reasoning.[81] The intention that results from this process of reasoning, however, is not itself normative because there is no guarantee that the

premises supporting it are justified.

Dancy takes a third view, that belief, intention, and action are all eligible outcomes of deliberation.[82] Dancy acknowledges the difficulties of maintaining that deliberation can conclude in action in any logical or semi-logical sense. In his view, however, the proper question is not whether action can be the conclusion of deliberation but whether the relation between actions and the states of affairs that inform practical deliberation is the same as the relation between beliefs and the states of affairs that inform theoretical deliberation. Having recast the problem in this way, Dancy argues that the relation in both cases is one of "favouring:" the states of affairs that inform deliberation and the considerations adduced in the course of deliberation may favor a variety of responses, including not just belief, but intention, action, and other possibilities as well.[83] Dancy adds that deliberation may lead directly to action if deliberative considerations favor action: there is no need for the interim step of belief.[84] Moreover, in any case in which relevant considerations favor action as well as belief that the agent should act, action is primary. That is, the considerations that favor the agent's belief that she should act favor it because they favor the action. Similarly, considerations that favor the agent's intention to act favor the intention because they favor the action.[85]

Theories of temporally extended practical rationality encounter difficulties under each of these models of practical deliberation. Beginning with McClennen's theory of resolute choice, McClennen claims that it is rational for an agent to deliberate at some initial point in time (T_1), form a resolute intention to act, and then carry through with action at a future time (T_2). McClennen's theory is compatible with both Raz's and Broome's views on deliberation concluding in intention. Applying Raz's model, deliberation at T_1 concludes in a belief that the agent ought to form a resolute intention to act, which can then be converted to an intention at T_1

by operation of the agent's will. Applying Broome's model, the agent reasons at T_1 from an intended end and a belief about means directly to an intention to act.

The difficulty, for McClennen, is explaining how the intention the agent forms at T_1 can yield an action at T_2. I will assume that there are no physical barriers to action. Even so, there is no guarantee that the agent will have either reasons to act or the will to act at T_2. The restriction McClennen adds to resolute choice, that a resolute choice requires action at T_2 only if circumstances at T_2 are substantially as the agent anticipated at T_1, does not solve the problem. Even when this condition is met, the agent may conclude at T_2 that acting on her prior intention is no longer advantageous, either because her preferences have changed or because she has already received the benefits she hoped to obtain by forming and communicating an intention to act. If so, she may prefer not to follow through. It follows that neither Raz's model of practical deliberation nor Broome's model can sustain resolute choice.

Similar problems arise in connection with Shapiro's theory of extended practical rationality. Shapiro holds that, subject to limited allowance for changes in plan (to be discussed in section B below), an intention formed at T_1 requires action at T_2 even when circumstances have changed between T_1 and T_2. This raises the possibility that the agent may conclude at T_2 that the intended act is no longer advantageous under current circumstances. If so, it is difficult to see how the agent can form a will to act at T_2. Again, neither Raz's model nor Broome's allows for completion of plans in the manner Shapiro describes.

Dancy's model of deliberation and action may at first seem more promising for McClennen's model of resolute choice and Shapiro's related model of nonfeasible options. For Dancy, action, like belief or intention, is a response to a set of considerations that stand in a favoring relation to the contemplated action. Because deliberation can lead directly to action,

problems about the role of will in transitions from belief to intention and intention to action do not arise.

The harder question is whether deliberation at T_1 can control future action, at T_2, in the manner McClennen and Shapiro describe. Dancy does not address temporally extended practical rationality and it is hard to predict how he would respond to the problems it poses. It is possible that considerations adduced in deliberation at T_1 can favor not just belief or other current responses, but also resolution at T_1 to act on a current intention at a future time, T_2. This seems unlikely: Dancy's model is grounded in a particularistic approach to ethics and the favoring relation he describes appears to be a relation between considerations adduced the time of deliberation (T_1, for McClennen) and an immediate response to those considerations. Assuming, however, that resolution is an eligible response to deliberation at T_1, there is nothing in either McClennen's theory or Shapiro's theory to prevent an agent who has resolved to act at T_2 from deliberating again at T_2. The only limitation is that the agent must act according to her prior plan. Yet, if the agent proceeds to deliberate further at T_2, and if the sum of considerations adduced at T_2 no longer favor the original plan, then Dancy's particularistic model of practical deliberation and action seems to require that the agent should do what her current deliberation favors. The agent's prior resolution is not itself a consideration favoring action at T_2; resolution is simply a response that is no longer favored by the agent's circumstances as a whole.

Gauthier's theory of intended courses of action poses a somewhat different puzzle under the various models of deliberation just described. In its final version, Gauthier's theory requires deliberation at two points in time: once at T_1, when the agent adopts a course of action, and once at T_2, when she completes the course of action. According to Raz's model of deliberation and action, deliberation at T_1 results in a belief that the agent ought to pursue a course of action. This

belief, combined with will, becomes an intention to pursue the course of action. On Broome's view, deliberation at T_1 results directly in an intention to pursue the course of action. On both views, deliberation at T_2 results in a belief that the agent ought either to complete the course of action, or not. If the agent concludes that she should complete the course of action, her current belief combines with will and physical capacity to yield an action. If not, there is no will and no action follows. So far, so good.

Difficulties arise because, to explain the constraint that intentions impose on agents, Gauthier places special conditions on the agent's deliberation at T_2. The reasons that inform deliberation at T_2 are not simply the agent's current reasons for action at T_2; she must compare her current reasons for action at T_2 with the combination of her current reasons for action at T_2 and any benefits she has already obtained by forming an intention at T_1. Thus, if S promised to provide a reciprocal benefit, formed the intention to do so, received a benefit from the promisee, and prefers the reciprocal exchange to her original position, she must, accordingly to Gauthier, will herself to act on the intention.

Although Gauthier's theory operates differently from McClennen's theory of resolute choice and Shapiro's theory of nonfeasiblity, Gauthier faces a similar problem in accounting for the will to act at T_2. At T_1, if the agent comes to believe that it will be to her advantage to adopt and later carry out the course of action, she normally will be able to will herself to form the necessary intention. At T_2, however, when she must carry out the intention, she may already have received the expected benefits of a course of action. If her original deliberation yielded only an intention, and if current reasons for action favor abandoning the intention at T_2, it is not clear how, at T_2, the agent can muster the will she needs to convert her intention into action.[86] Consequently, Gauthier's theory of temporally extended practical rationality fails to conform to

the models of deliberation to action proposed by Raz and Broome.

Dancy permits deliberation to belief or intention (or both) at T_1, and also permits deliberation directly to action at T_2, based on what response is most favored by considerations adduced at T_2. Under Dancy's particularist approach, however, the outcome of deliberation at T_2 is best understood as a response to the full range of actual, forward-looking considerations adduced at T_2. Consequently, it seems unlikely that Dancy's model of deliberation and action can accommodate that type of constrained comparison Gauthier envisions at T_2, in which deliberation is limited to a comparison between completing the intended course of action and never having undertaken the course of action.

Bratman's theory proceeds differently. According to Bratman, the agent deliberates at T_1 and forms an intention. If, as a result of the agent's dispositions toward prior intentions at T_2, the intention the agent formed at T_1 continues in force at T_2, the agent simply acts on that intention, without further deliberation and without forming a current belief about what to do. Action follows from the element of "volitional commitment" Bratman attributes to intentions. Thus, Bratman does not posit, as McClennen appears to, that prior intentions override the outcome of later deliberation. Nor does he posit, as Gauthier does, a second, limited deliberation at the time of action, assessing the cumulative benefits of the agent's intended course of action. Instead, the agent's initial choice is simply extended from T_1 to T_2, *without reflection*. According to Bratman's standards of agent rationality, this extension of the volitional component of intentions over time is rational if the agent is guided by reasonable dispositions toward prior intentions.

Translated to the terminology of Raz and Broome, Bratman's volitional commitment appears to be the agent's will to act, formed at the time of initial deliberation and somehow imbedded in her retained intention. I will assume that imbedding a will to act in a continuing

intention is metaphysically acceptable; and in fact, it seems plausible that conducting will is a primary function of intentions.[87] With this assumption in place, and setting aside problems of epistemic rationality, Bratman's theory can be reconciled with Raz's and Broome's models of deliberation and action. Deliberation at T_1 yields a belief about what to do in the future and, at least if combined with present will at T_1, a related intention. This intention, if not interrupted, contains an element of will and so yields an action at T_2. Bratman's standards of agent rationality then explain how it may be practically rational to leave the intention in place even though, objectively, the agent's current reasons support a different action. As long as the agent follows dispositions that are reasonable in the long run, the agent can proceed to act on the original plan.

Dancy's model of deliberation and action presents a more difficult challenge for Bratman. The agent's belief that she ought to make a plan is supported by relevant considerations at T_1, as is the agent's intention to act. The problem is that, as already described, Dancy's favoring relation appears to be a relation between considerations adduced in deliberation and a response at the conclusion of deliberation, which does not directly yield an action at T_2. Meanwhile, Bratman's conception of intentions rules out the argument that the agent's prior intention, formed at T_1, counts as a consideration favoring action at T_2: intentions, Bratman maintains, are not reasons for action but volitional commitments.[88]

A proponent of temporally extended practical rationality might argue that an intention carrying volitional commitment is an eligible response to considerations adduced at T_1, which, under Bratman's theory, remains in place T_2 if the agent is reasonably disposed to act on it unreflectively. Because no further deliberation occurs at T_2, no further considerations are adduced and no different response is substituted at T_2. Whether this argument works depends on

whether the set of possible responses to considerations adduced at T_1 is defined broadly enough to include a response that blocks further deliberation in a range of cases. I suspect that a narrower definition of eligible responses is more consistent with ethical particularism.[89] I will assume, however, that Bratman's theory of temporally extended practical rationality can be reconciled with Dancy's model of practical deliberation in the manner just described, and will focus instead on the epistemic arguments made in Section V.

B. Generality of Intentions and the Problem of Defeasibility

A complete theory of temporally extended practical rationality must explain the rationality of conforming to general intentions in particular cases. Temporal extension of agency depends on future-directed intentions. Any future-directed intention is general in the sense that, when the agent initially adopts the intention, neither the act nor the circumstances of its performance can be fully specified. Consider a fairly narrow intention to perform a single act on a single occasion: S intends to buy groceries tomorrow, from a certain store, according to a list. Even this intention can only be formulated in general terms. Substitutions S might make for unavailable items, the care S will take in making selections, and many other details needed to fully describe S's act, are either unpredictable or not worth the trouble of thinking through in advance. This is not a serious problem if, at T_2, the act and surrounding circumstances are in the range of S's expectations, although not precisely as S envisioned them at T_1. If circumstances are not roughly as expected, the rationality of acting on the intention is in doubt.

In the context of rule-following and some types of interpersonal commitment, intentions are general in the additional sense that the agent intends to perform an act in a class of future cases over a period of time. For example: S adopts an internal rule for herself, to write a letter to

her grandmother once a week. As formulated, S's rule is absolute - no exceptions - because S believes she will feel better about herself at the end of the year if she complies and because she also believes she may shirk if she does not fix a schedule for letter-writing. Setting aside questions about what constitutes a letter, the rule is general because it requires S to act every week regardless of variations in her physical and emotional circumstances. S must phrase formulate it in this general way, and adopt it as absolute, to obtain the internal coordination he seeks from the rule: an exception for weeks in which S's reasons for action favor skipping a letter would leave S with no resources against procrastination.[90] Nevertheless, assuming that S had good reason at T_1, and has good reason overall, to write every week, there will be weeks in which S's immediate reasons not to write will outweigh her immediate and overall reasons in favor of writing. The same reasoning applies if S promises her grandmother that she will write each week. To obtain the benefit of normative self-control, S must commit to write in at least some weeks in which his immediate and overall reasons for action favor skipping the letter.

The generality of future-directed intentions leads to the question when, if ever, a theory of temporally extended rationality should allow for rational abandonment of intentions or rational defection from intentions on particular occasions. McClennen, in his theory of resolute choice, appears not to permit abandonment of or exceptions to a resolute intention. This strict stance, however, may be attributable to a limited view of when intentions can be resolute. McClennen's main concern is not with resoluteness in unanticipated circumstances, but with resoluteness that overrides a change in the agent's preferences. Accordingly, he defines resolution in a way that significantly limits the effect of future-directed intentions: "if unfolding events, including any conditioning circumstances, are as you had expected them to be, you proceed to execute that plan."[91] This definition probably sweeps in circumstances that are within

the range of normal daily occurrence, even if not fully specified at T_1. Beyond this, however, any significant difference between expected outcomes and actual outcomes appears to release the agent from the intention. As result, McClennen's theory is not well-suited to sustain practices such as rule-following or ongoing interpersonal commitments, in which some actions that fall within the agent's general intention, as initially formulated, will not be favored by the agent's reasons for action at T_2.

Setting aside McClennen's carefully circumscribed notion of resolute choice, none of the authors I have discussed argues unequivocally that temporally extended practical rationality requires agents to act on prior intentions without exception, no matter how grave the consequences. Shapiro comes the closest to this position.[92] Shapiro's theory of rationality is designed specifically to support the authority of a system of legal rules, and thus assumes the generality of intentions. Those who accept the master plan of a legal system undertake to follow the system's rules in classes of cases that cannot be anticipated in all particulars. This open-ended obligation puts pressure on the theory to allow for reconsideration of intentions to follow the law. Shapiro recognizes the problem but his response is equivocal.

Shapiro's underlying theory of law supports a strict interpretation of the constraint that endorsement of a legal system places on agents. As a legal positivist, Shapiro believes that law claims practical authority over its subjects. In a well-known legal positivist account of the practical authority of law, Joseph Raz describes legal rules as exclusionary reasons for action, such that all reasons that contributed to the selection of the rule are subsumed in the rule and no longer open to deliberation by the rule-follower.[93] The explanation Shapiro gives for the rationality of rule-following is consistent with Raz's account: an agent who forms the intention to follow the rules of a legal system is disabled from consideration of deliberative reasons for

and against the actions required by the rules. Instead, the agent has only "implementation reasons for action," which require that she must follow the rule.[94] This suggests that an intention to follow a rule is not subject to reconsideration.

Yet, despite Shapiro's strong stand on the binding character of intentions, he resists the conclusion that intentions, including intentions to follow law, are absolutely binding at the point of application. He says instead that reconsideration of an agent's intention to follow legal rules is rational if supported by "good enough reason," or, at another point, "compelling reason," to reconsider.[95] Shapiro's standard for reconsidering intentions is imprecise, not only because he describes it in imprecise terms but also because he does not make clear whether reason to reconsider means reason to reconsider the agent's overall intention to follow the law or reason to depart from a specific rule on a specific occasion.[96]

Shapiro also equivocates about why agents can legitimately abandon their intentions to follow rules, or depart from their intentions in particular cases. One way to read his discussion of reconsideration of intentions is that rationality itself dictates that intentions must be defeasible to some degree. It is not rational for an agent at T_1 to impose a constraint on her future self that forecloses any and all deliberation about action. Instead, an agent who is rational at T_1 will insert some caveat into the intention, allowing for reconsideration when the agent has good or compelling reasons to reconsider.

Some of Shapiro's statements, however, suggest that, at least in a legal context, the agent's license to reconsider an intention to follow rules is not a matter of rationality but a contingent feature of the rules themselves. Thus, Shapiro states at one point that the authority of law

"should not be taken to mean that the law demands that its dictates be followed *come*

> *what may*. Laws, like all plans, are typically defeasible. When compelling reasons exist, the law will normally permit its subjects to reconsider its direction and engage in deliberation on the merits. The catch here is that the law claims the right to determine the conditions of its own defeasibility."[97]

This passage indicates that the source of the agent's ability to depart from legal rules is the rule itself. Legal rules may or may not permit agents to reconsider their intentions to follow the law in exceptional circumstances; when they do not, an agent's acceptance of the legal system precludes consideration of any and all deliberative reasons for failing to comply with its terms. This position aligns Shapiro's statements about reconsideration with Raz's theory of exclusionary reasons and with Shapiro's own explanation of the limits that intentions place on deliberation and choice. At the same time, it imposes a very strict constraint on any agent who forms an intention to act according to law.[98]

In contrast to Shapiro, Gauthier provides a determinate and relatively narrow formula defining the constraint imposed by prior intentions in the face of unanticipated circumstances. At T_1, the agent adopts a course of action if she expects that the course of action as a whole will be beneficial. At T_2, the agent compares the consequences of adopting and completing the course of action with the consequences of never adopting (and never completing) the course of action. If adopting and completing the course of action is preferable, the agent has a reason at T_2 to do the intended act; if not, the agent has no reason at T_2 to do the act and should abandon her intended course of action. Gauthier's formula, however, runs into difficulty when applied to general rules and standing commitments that call for application in a series of future cases.

Gauthier's broad objective in formulating a theory of temporally extended practical rationality is to explain the ability of self-interested agents to interact successfully within a

society, based on the advantages that follow from a disposition to cooperate over the course of multiple interactions.[99] His examples, however, typically involve discrete intentions that can be carried out by a single act. The agent forms an intention to complete the second half of a reciprocal exchange, or the agent forms an intention to retaliate if wronged.[100] In this type of situation, the comparison Gauthier proposes yields an answer: depending on the costs of performing the intended act, as they appear at T_2, the agent should either carry out the course of action or abandon it altogether. The agent can change her mind, but only if the overall plan now appears to be disadvantageous.

When an agent adopts a general course of action for all cases of a certain type, the simple comparison Gauthier proposes fails to cover all relevant possibilities. For any given case, the agent now has at least three options: never adopt the course of action, adopt the course of action and complete it in this case, or adopt the course of action and complete in all cases except this case. Judged instrumentally, the rational choice is to adopt the most advantageous of these three options. If the agent believes she can preserve the benefits of the general course of action without complying in the present case, the third option is best.

Suppose, for example, that S rationally forms an intention to follow rule R in all future cases, based on her belief that over time, she will avoid more errors by regular compliance with R than she would make if she decided case-by-case what to do. Sometime later, S faces a situation covered by R. S continues to believe that over time, compliance with R will be advantageous. She also believes that in this case, it would be better to violate R than to follow it, in part because this particular outcome of R looks wrong and in part because she believes that no one will notice her defection, so that defection will not undermine the benefits of the rule. Thus:

> (1) At T_1, S has belief B_1: S expects that following R regularly will be advantageous over the long run; therefore S forms the intention to follow R regularly.

(2) At T_2, S has beliefs B_2(a) and B_2(b):

> B_2(a): S believes that following R regularly has been and will continue to be advantageous in the long run.
>
> B_2(b) S believes that, for this case, a one-time defection from R is better than following R.

In these circumstances, it is not clear, even by Gauthier's reasoning, that S has a reason to follow R. S can obtain the benefits of her intended course of action by adopting the alternative course of action, "follow R in all cases but this one."

The availability of this third option, however, undermines the value of an authoritative general rule or an ongoing interpersonal commitment. The rule becomes a rule of thumb, to be followed when the agent judges that its outcome is correct. Similarly, the serial commitment imposes no meaningful constraint.[101] This result might be unobjectionable if agents never erred or had nothing to gain from normative self-regulation. Given the possibility of incomplete evidence and faulty judgment and the value of coordination and commitment, the loss can be significant.

Gauthier might answer that, for the purpose of the comparison he proposes, there is no third option: the relevant comparison for S is between adopting R and applying it to this case and never adopting R at all. Accordingly, all potential benefits of R over the long run weigh against the benefits of defecting from R in this case. Yet it is not clear why a rational agent should be limited to this comparison if it appears that she can defect now without losing significant long-term benefits of R.

Bratman's response to the problem of when agents should abandon intentions is more detailed. A quick restatement of his principles of agent rationality may be helpful here. Bratman

assumes that prior intentions, even when rationally formed, cannot be absolutely binding on a rational agent. He then draws a distinction between ideal stability of intentions, in which the agent reconsiders prior intentions if and only if reconsideration would lead to a change of mind, and reasonable stability of intentions, which requires only that the agent must follow reasonable dispositions toward retention of prior intentions.[102] Agents, therefore, are allowed a reasonable margin of error if their dispositions are instrumentally sound over the long run.

Bratman's historical principle of agent rationality incorporates reasonable stability of intentions as a key element of practical rationality. An agent who forms an intention after rational deliberation can later act on the intention without further thought, provided that the agent is guided by a disposition to retain intentions of this type, the disposition is reasonable, and the agent reasonably applies the disposition to the intention in question. The agent's disposition is reasonable if it generally works to the agent's advantage. The result is action without current deliberation, unless changes in circumstances exceed the limits of the agent's disposition to retain prior intentions.

As noted in Chapter II, Bratman also develops a supplemental principle of rationality designed specifically for intentions that generalize over classes of cases. Bratman refers to this type of intention as a "policy-based" intention.[103] According to Bratman, policy-based intentions are defeasible case by case. Specifically, the agent can block the intention on particular occasions if, in doing so, the agent is guided by reasonable dispositions toward blocking general intentions. Typically, reasonable dispositions toward blocking intentions make acting on a policy-based intention the default position, but allow for blocking in special circumstances. Again, the standard governing agent rationality is a standard of reasonableness and a disposition that generally works well is deemed to be reasonable.

Thus, in cases involving intentions to act in a series of like cases over time, Bratman's principles of agent rationality refer in two different ways to the agent's dispositions toward prior intentions. First, the agent must follow reasonable dispositions pertaining to retention of prior intentions. If the agent's reasonable dispositions support retaining the intention, and if the intention applies to a class of cases, the agent can still block application of the intention to a particular case within the general scope of the intention, if reasonable dispositions toward blocking, reasonably applied to the case, support an exception in this type of case.[104] If the agent is disposed not to block the intention, the agent proceeds to act. The upshot is that, rather than delineate a standard by which the agent determines at the point of action whether to retain or abandon, and apply or block, prior intentions, Bratman allows the agent's reasonable dispositions to govern the defeasibility of intentions.[105] If the agent's reasonable dispositions support retaining and acting on the intention, no deliberation occurs at the point of action, so the type of comparison Gauthier envisions does not take place.

Bratman's discussion of the stability of intentions raises three difficulties. First is uncertainty about when dispositions toward retention and non-blocking of intentions count as reasonable.[106] Reasonableness, and therefore the rationality of acting on prior intentions, depends for Bratman on what works well over time. What works well, however, may depend on counterfactual comparisons with alternatives not followed. The question of what works well over time is also affected by the epistemic problems discussed in Section V: I will return to the problem there.

The second difficulty is uncertainty about the process by which the agent determines when to reconsider or block an intention. In developing his historical principles for action on prior intentions in what he calls the "basic" case of single-instance intentions, Bratman assumes

a clean distinction between deliberative and non-deliberative responses to prior intentions: the agent can either proceed to act on the intention without deliberation or reconsider in full the reasons for the intention or its application to the particular case. If the agent proceeds to act, the complete absence of deliberation is important to the conclusion that the agent is rational, because it permits the agent's prior intention to carry forward and govern her action by force of the volitional commitment embedded in the intention. In practice, however, complete absence of deliberation may seldom occur: an agent who has formed a prior intention is more likely to engage in an abbreviated review of reasons before either acting on the intention or stopping to reconsider in full. Bratman's analysis does not easily accommodate this type of partial deliberation. Once the agent peeks at reasons for action and discovers that some have changed, it becomes more difficult to explain why rationality does not require a fully considered response.[107]

Third, the allowances Bratman makes for reconsideration of intentions and for blocking of general intentions in particular cases place significant limits on temporal extension of rationality. In particular, agents who consciously consider their reasons for action at the time of action, either because the reasons to abandon their prior intentions exceed the threshold fixed by their own dispositions toward intentions or because they simply happen to think about reasons for action, are not constrained. Accordingly, Bratman's theory does not easily accommodate rule-following in the strong sense described in Chapter II and may not allow for the type of normative self-control associated with interpersonal commitment. I will return to all these problems in Chapter V, after considering the epistemic side of the problem.

REFERENCES

*References are to works listed in the bibliography (below).

[1] E.g., Raz (2005).

[2] McClennen (1990), pp. 283-316; McClennen (1997).

[3] McClennen (1997), p. 229. See Chapter I, note 31, *supra*.

[4] McClennen (1990), pp. 12, 120-22; (1997), p. 229.

[5] McClennen (1997), p. 232.

[6] McClennen states these conditions most succinctly in McClennen and Shapiro (1998), p. 367. In comparison to McClennen, David Velleman suggests that a theory of practical rationality cannot be justified solely on grounds of instrumental benefit. Instead, the theory must be supported by independent reasons, such as the value of autonomy and volitional choice. See Velleman (2000); see also Broome (2001).

[7] Supporters of the separability principle view it as necessary to maintain consistency in choice, such that the process of decision at each node on the decisional tree should be the same whether the node is the first node or a subsequent node on the tree. McClennan dismisses consistency as an undefended axiom, which should give way to instrumental superiority. McClennen (1997), pp. 239-240.

[8] McClennen concedes that these options are not fully satisfactory. Id. at 238. In his 1998 collaboration with Scott Shapiro, he appears to distance himself from the explanations that rely on rational feasibility. McClennen and Shapiro (1998), p. 368. I return briefly to the idea of feasibility in my discussion of Shapiro in section II(E) *infra*.

[9] In McClennen's words: "[B]eing resolute in the context of changing information would typically be irrational. All that being resolute requires is that if, on the basis of your preference for outcomes, you adopt a given plan, and if unfolding events, including any conditioning circumstances, are as you had expected them to be, you proceed to execute that plan." McClennen (1997), p. 232.

[10] See Gauthier (1986), pp. 167-170; (1994); (1998b).

[11] See Velleman (2000).

[12] McClennen develops a similar principle, which he calls R-feasibility. McClennen (1990), pp. 211-213.

[13] See Gauthier (1998b), p. 46.

[14] Gauthier (1986), pp. 169-170.

[15] "[D]eliberating about what to do later, and so what to intend now, and deliberating about what to intend now, and so what to do later, both appeal to reasons, oriented not to the expected outcome of the action alone but rather to the expected outcome of the intention together with the action." Gauthier (1998b), p. 44.

[16] Gauthier makes this reasonably clear in his article *Intention and Deliberation,* using the example of a failed threat. "Suppose that I issue a threat and it fails. Then even though the circumstances I may find myself in are exactly those that I anticipated should the threat fail, and *even through I recognized that I might find myself in those circumstances, yet in making the threat, my expectation was that I should benefit thereby*, and I now know that expectation to have mistaken. Suppose on the other hand that I give an assurance and it succeeds - I agree to reciprocate and you in consequence help me. Then if the circumstances I find myself in are those that I anticipated should my assurance succeed, my expectation that I should benefit thereby has proved correct. And this difference, between the mistaken expectation associated with a failed threat, and the confirmed expectation associated with a successful assurance, is crucial to deciding the rationality of continuing one's course of action. It is not rational to continue a course of action if the expectation associated with adopting it has proved mistaken. . . ." Id., pp. 48-49 (emphasis added).

It is less clear how Gauthier would distinguish between a case in which the agent expects at the outset that the course of action will not be advantageous, and so cannot rationally form an intention to follow it, and one in which the agent expects overall advantages but also envisions particular cases in which the course of action will not be advantageous. This problem comes up in discussion of Gauthier and the toxin puzzle in section D below.

[17] I will return to this problem later in this section, when I discuss Gauthier's response to Gregory Kavka's toxin puzzle.

[18] More precisely, there are two conclusions open to the agent at this point. One is that following the rule in this case will leave the agent worse off than she would be if the rule did not govern this case. This conclusion matches the agent's conclusion in Gauthier's standard examples, which involve single-act plans such as a plan to perform the second part of a reciprocal promise. A possible conclusion is that following the rule in this and all other cases will leave the agent worse off then she would be if she had never adopted the rule as a rule for all cases. A conclusion of the first type indicates that the rule is not well-suited for the case; a conclusion of the second type indicates that the rule is not a good rule and should be abandoned. Gauthier does not distinguish between these two situations. I will return to this problem in the next chapter, where I raise some doubts about capacity of Gauthier's theory to deal with generalized courses of action such as rules. For now, I will set it aside.

[19] See Chapter I, supra, text at notes 14-25.

[20] See, e.g., Parfit (2001), p. 87 ("According to the standard version of the Self-Interest Theory . . . we should maximize at the level of our acts.").

[21] Bratman (1987), p. 5.

[22] See generally Bratman (2014); Bratman (2007).

[23] Bratman (1987), p. 52.

[24] Id., p. 51. See also Parfit (1984), p. 8 (drawing a simlar distinction between the rationality of acts and the rationality of motives); Dancy (1997) (questioning whether a theory of instrumental practical rationality can accommodate this type of distinction).

[25] Id., p. 53.

[26] See id., pp. 42-46. Suppose, for example, that a retired college professor is asked to teach a summer course at Ivy, where he earned his undergraduate degree. He enjoys teaching and would very much like to revisit Ivy; however, he suffers from a medical condition that makes travel risky and could affect his performance in class if not carefully controlled. Irrationally, he forms the intention to teach. He then begins a course of medication that will reduce his symptoms but may have adverse effects in the long run. Given his intention to teach, he should take the medication. Thus, from an internal, plan-constrained perspective, taking the medication is a rational act. From an external perspective, however, the professor is not rational in taking the medication because he was not rational in forming the intention that supports this act.

[27] Id., p. 53 (emphasis in original).

[28] Id., pp. 15-18.

[29] In Bratman's words, intentions act as a "filter of admissibility" on deliberative options. Id., p. 33.

[30] Id. Bratman also states that the agent's intentions must not be inconsistent with the agent's beliefs. In saying this, he appears to have in mind a particular set of beliefs, consisting of beliefs about what actions are feasible for the agent. He says, for example, "it should be possible for my entire plan to be successfully executed given that my beliefs are true." Id. at 31. In this statement, the beliefs referred to appear to be beliefs about what can be done. Thus, Bratman's remarks about consistency with beliefs do not imply that the agent's intentions must be consistent with the agent's beliefs what actions are best. I will return to the question of rational belief in Chapters IV and V below.

[31] This is clearest in Bratman (2012), pp. 78-79.

[32] Bratman (1987), pp. 72-73.

[33] Id., pp. 60-62.

[34] Id., pp. 60-62.

[35] In Bratman's words: "If it is rational of S to have a present-directed intention to A, and S successfully executes this intention and thereby intentionally A's, then it is rational of S to A." Id. at 55.

[36] Bratman begins with an "ahistorical" principle, which disregards the provenance of intentions

that predate the agent's act and assumes that all retained intentions were rationally formed. This principle is inadequate because it allows for bootstrapping of faulty intentions. Id. at 57-59.

[37] Id. at 80.

[38] Id. at 67.

[39] Id. at 72.

[40] Id. at 64.

[41] Id. at 85.

[42] In discussing both segments of his historical principle for deliberative intentions - the segment pertaining to nonreflective retention of prior intentions and the segment pertaining to deliberation on the basis of prior intentions - Bratman suggests that overall assessment of the agent's rationality may sometimes be mixed. For example, the agent may have been irrational in forming a prior intention, but rational in failing to reconsider before deliberating on the basis of that intention. In that case, the agent is not rational; nor is she entirely irrational. See id., pp. 85-86. I find a hybrid conclusion of this type unsatisfying, if rationality is a standard to which agents should aspire. Bratman, however, treats agent rationality as an external standard and, perhaps because of this, is willing to live with ambiguity about rationality.

[43] Id. at 89-91.

[44] Id. at 91.

[45] Most of the examples Bratman relies on in *Intention, Plans, and Practical Reason* involve single instance intentions in which the agent forms an intention at T_1 to do a specific act at T_2 and then completes the act at T_2. In later work, policy-based intentions become more prominent, particularly when Bratman takes up the problem of self-governance. Bratman proposes that self-governance, in the sense he has in mind, is best explained in terms of "self-governing policies, meaning intentions to give weight to certain reasons in practical deliberation. See Bratman (2007), pp. 239-240. Policies of this type ensure that the agent's thought and action will be guided by considerations that the agent judges to be normatively valuable, whether or not she could defend them in intersubjective terms. By adopting and implementing personal policies about reasons for action, the agent exercises "subjective normative authority" over her thought and action. Self-governing policies, like intentions generally, are subject to reconsideration but exert presumptive control over the agent's deliberation as long as they remain in place.

[46] See id., p. 78.

[47] Id., p. 5.

[48] Kavka (1983).

[49] Id., p. 35.

[50] Gauthier (1998).

[51] Id. at 48. At the time of forming an intention to drink, "I know that I should like [tomorrow] to change my mind. But if I am rational, and understand my situation, this knowledge is of no use to me. Either I suppose that I shall have no reason to drink the toxin, in which case insofar as I am rational I cannot have the mind to do so, or I suppose that I shall have reason to drink it, in which case I can have the mind to do so but no good reason to change it. Changing my mind is not part of a course of action I can embrace." Id.

[52] Gauthier adds that, although the agent can form an intention to drink toxin tomorrow on practical grounds (because he expects that drinking toxin is now and will continue to be part of the best course of action), the agent cannot form a belief that he will drink toxin tomorrow on practical grounds. In Gauthier's view, the agent can only believe what he believes to be *true*. Id, at 51-52. I will return to Gauthier's discussion of rational belief in sections IV and V.

[53] Id. at 48-49.

[54] Id. at 50.

[55] Id. at 57.

[56] Id. at 56 (author's emphasis).

[57] One might question whether this outcome would satisfy the terms of the eccentric billionaire's offer, either in the original puzzle or in Gauthier's variant. Consider another variant on the toxin puzzle, in which the agent is offered an equal chance of receiving $995 or $1 million if he forms the intention today to drink tomorrow, with the payoff to be determined tomorrow by a coin flip. In this version, unlike Gauthier's, there are no uncertainties about the effects of drinking the toxin: it will cause one day of illness and the agent places a negative value of $1000 on this day. Overall, intending-to-drink-and-then-drinking the toxin is the best course of action: the agent has a 50% chance of losing $5 and a 50% chance of gaining $999,000; but the agent can easily foresee the case in which he will lose $5. In these circumstances, allowing the agent to form an intention that is defeasible if the outcome is $5 looks like cheating: the intention is costless, but also pointless. In effect, the agent simply decides what to do when the time comes. I think the billionaire might object to paying for this intention.

[58] Id. at 57.

[59] Bratman (1987), pp. 101-06.

[60] Bratman (1998), pp. 59-83.

[61] Id., pp. 62-63 & n.14.

[62] Bratman seem to have in mind a stylized version of reciprocal benefit arrangements, in which the costs of reciprocating are known in advance.

[63] Bratman says:"If, on the basis of deliberation, an agent rationally settles at t_1 on an intention to A at t_2 if (given that) C, and if she expects that under C at t_2 she will have rational control of whether or not she A's, then she will not suppose that at t_1 she should, rationally, abandon her

intention in favor of an intention to perform and alternative to A." Id. at 62. The reference to rational control appears designed to exclude cases in which the agent artificially induces an intention.

[64] The standard view of instrumental rationality, as Bratman now interprets it, does not prevent the agent from considering what he calls "autonomous" benefits associated with the intention. Bratman continues to maintain that the relevant deliberation is deliberation about the act, not the intention; but benefits that follow from forming the intention can be reasons to decide in favor of the act. Id., n.13.

[65] "The point is not that a rational agent does not care about the past. The point concerns, rather, what is now under the control of the agent. What is now under her control are her alternatives from now on. So it seems she will want to rank those alternatives. . . Strong and moderate resolution , in seeking a strong role for planning in achieving the benefits of coordination over time and across agents, seem not to do justice to the basic fact taht as agents we are temporally and causally located." Id., p. 66.

[66] Although Bratman rejects resolute choice and generally endorses separability, he is dissatisfied with the outcome of separability and sophisticated choice in several specific cases. Each of these cases resembles the toxin puzzle in that the agent neither receives nor expects to receive new information between the time she forms an intention and the time she carries it out. In one such case, the agent has unstable preferences: she generally prefers one outcome but has a temporary preference for another outcome at the point of action. For example, she generally prefers a one-drink limit, but at this moment wants a second drink. In cases of this kind, Bratman wants to allow for "instrumentally rational willpower" that will enable the agent to override her temporary preference. The second problem case involves a Sorites sequence in which the agent prefers to advance incrementally along some continuum of action, but knows that a series of indistinguishable advances will eventually lead to an outcome she does not prefer. Here, Bratman wants to allow the agent to adopt a long term plan that incorporates a reasonable stopping point.

Bratman's solution to these special cases is to qualify separability and sophisticated choice with a further standard he refers to as "no regret." The no regret standard permits the agent to adopt and execute a course of action that overrides the agent's evaluative ranking at the time of execution, if at the time she adopts the course of action, she anticipates that, at a some later point in time, *after* the time when she executes the plan (1) she will be glad to have followed the course of action or (2) if she has not followed the course of action, she will regret that she departed from it. The relevant later time is, roughly, the natural conclusion of any larger plan of which the agent's preferred course of action is a part.

Bratman argues that when this no regret standard is met, the agent rationally can and should follow the course of action she initially intended to follow even if, at the time of execution, the agent takes account of the likelihood of future regret and still prefers to defect. For example, if, at the time of action, the agent ranks her desire for a second drink higher than she ranks avoiding the regret she expects to feel later, she should nevertheless follow her one-drink policy. Bratman's explanation, as I understand it, is that the likelihood of future regret is not a reason to choose one act over another (have another drink, or do not), but a reason to retain a prior intention, in this case the intention to have only one drink. The possibility of future regret can thus support adherence to a prior plan even if the act the plan currently requires is not

favored by the agent's current evaluative rankings.

With some effort, Bratman's no-regret standard can be reconciled at least partially with the standard view of deliberative rationality, which links rationality to the agent's evaluative rankings at the time of deliberation. The agent's ultimate choice is not based entirely on current evaluative rankings at the time of action; if it were she would not act as intended. Instead, the agent considers not only her current evaluative rankings but also her prior plan and the eventual superiority of following it rather than acting on current rankings. The comparison the agent makes, however, is a current comparison of current options, either to follow her plan or to follow her current rankings. For this reason, her past deliberation does not directly control her present choice as it would under a theory of Resolute choice. The agent may opt at the time of action to retain her prior intention as a way of controlling future regret, but her past evaluative ranking, at the time she adopted the intention, is no longer in play when she proceeds to act.

The toxin puzzle does not meet Bratman's no-regret standard because earning or not earning $1 million is a fixed part of the past when the time comes to drink the toxin. The agent may regret in the future that she missed the opportunity to gain the $1 million, but this fact does not play a part in her deliberation about whether to drink because it is not within her control at this point. Similarly, in reciprocal benefit cases, the possibility that the agent will regret the loss of a benefit does not play a part in her deliberation about whether to reciprocate, because, when the time comes to reciprocate, the agent will already have received (or not received), the other party's contribution: securing the arrangement is no longer within the agent's control. In contrast, in cases involving temporary preferences or Sorites sequences, it is still up to the agent to choose whether to retain her prior intention or indulge her current preference; therefore future regret is pertinent to the agent's present decision.

This pattern of outcomes confirms Bratman's understanding of agency as both temporally located and temporally extended. Rational agency is located at the time of action and is based on present or future evaluative rankings of outcomes over which the agent now has control. At the same time, rationality is defined broadly enough to allow for reasonable retention of prior intentions at the time of choice in cases that implicate concerns about effective planning.

[67] Shapiro (2011).

[68] Shapiro (1998a), p. 39.

[69] Id., p. 40.

[70] Id., pp. 39-40.

[71] Id. at 47; McClennen & Shapiro (1998b), p. 367. Shapiro is invoking a special notion of rational feasibility that he ties to reasons for action. An action is feasible, in Shapiro's sense, if the agent can perform it for a reason; if the agent has no reason to perform the action, it is not rationally feasible for him to perform it. Id. at 47-52. Cf. McClennen (1990), pp. 14-15 (discussing rational feasibility). In the rule-following case, Shapiro's conclusion that deliberative reasons are no longer in play after the agent has accepted the rule means that the agent has no reason for violating the rule, therefore the only feasible choice is to comply.

The scheme of the argument appears to be this:

(1) An action is not rationally feasible unless it is based on a reason.

(2) An agent's commitment to a plan of action exhausts the force of the deliberative

reasons on which the decision to commit was based and replaces them with an implementation reason to follow the plan.
(3) Therefore, the deliberative reasons on which the agent's decision to commit was based are no longer available to the agent as reasons for action.
(4) Therefore, defection from the plan cannot be based on the agent's deliberative reasons for or against committing to the plan.
(5) Assuming the agent has no new deliberative reasons for defecting from the plan, the only reasons available to the agent are implementation reasons.
(6) Therefore defection is not rationally feasible.

[72] McClennen & Shapiro (1998b), p. 366.

[73] Id., p. 367.

[74] McClennen (1997), p. 232; see note 8, *supra*.

[75] Some fairly clear statements and examples appear in Aristotle, DE MOTU ANIMALIUM 701^{a}9-25; Aristotle, NICOMACHEAN ETHICS, 1147^{a}25-30.

[76] Raz (2011), pp. 131-137. Raz draws this conclusion within a broader theory of Reason, reasons, reasoning, and rationality. Reason, for Raz, refers to the rational powers that enable agents to recognize reasons and are constitutive of personhood. One such power is reasoning; but belief, intention, emotion, and any other capacity that can be exercised irrationally are also components of Reason. Rationality is the proper functioning of all the rational powers, including, but not limited to, reasoning. In Raz's view, rationality is not normative. There is no reason to be rational; we simply are this way unless things are not working well. See id., pp. 83-101.

Raz also works from a primarily objective understanding of reasons: for Raz, epistemic limitations that prevent an agent from recognizing applicable reasons do not alter the agent's reasons unless the reasons are unknowable by the agent. Raz leaves the distinction between unrecognized and unknowable reasons largely undefined. See id., pp. 107-128.

Part of Raz's program is to argue that there is no significant difference between practical reasoning and theoretical reasoning, apart from the focus of practical reasoning on the agent's future acts. Comments on this view are outside the scope of my current project.

[77] Raz defines a conclusive reason as one that is undefeated by other reasons and defeats all other reasons. Id., p. 103.

[78] Broome (2002).

[79] Id., p. 86. Broome proceeds to then expand his model of intention reasoning to include reasoning from a general intention and a belief about the agent's best means for achieving that end to an intention to pursue those means. Id., pp. 107-109.

[80] Id., pp. 90-92. Broome points out that intentions do not encompass what the agent believes to be true already and do not track all available means of making something true.

[81] See id., pp. 93-95; see also Broome (2003).

[82] Dancy (forthcoming).

[83] Id., pp. 3-4. Dancy writes from a background of ethical particularism, in which the favoring relation plays an important role. See generally Dancy (2004), pp. 7-10, 73-78. One key feature of ethical particularism is "holism" about reasons: a reason that favors an action in one case may not favor, or may disfavor, the same action in another case. Accordingly, the agent's deliberative task is not to reason from general moral principlse but to recognize and respond to the "basic moral facts" presented in particular cases. Id. at 155-161.

[84] Dancy says "I can adduce considerations, deliberate, and act accordingly without needing to form an intermediate conclusion that this or that course of action is the one I have most reason to pursue. . . . [T]o respond to something as a reason is not and does not require, believing it to be a reason." Dancy (forthcoming), at 10. He adds that reasoning to a belief is not an implicit step in the process: "I would hope that not everything to which we are somehow rationally committed . . . is something we are already doing, even if only implicitly." Id. at 11.

[85] Dancy's argument for the priority of action over intention or belief appears to assume that the features of a situation that favor belief that the agent should act will also favor an intention to act and a corresponding action. My claim that temporally extended practical rationality and epistemic rationality conflict casts doubt on this assumption: if I am correct, reasons for action can diverge from reasons for belief about reasons for action.

[86] Depending on S's preference set, reliance by the other party may be a forward-looking consideration at T_2. But Gauthier does not assume that this is so; his theory is designed to apply even when S cares only about benefits received.

[87] This evidently is Bratman's assumption. See Bratman (1987), pp. 5-9, 18-23 (criticizing the "desire-belief" model of intentions). The assumption also seems well-founded: conceiving of intentions as conductors of will over time, rather than simply as fleeting composites of belief and desire, assigns them a more meaningful role in the process leading to action.

[88] One of the reasons Bratman gives for this conclusion is that intentions, while they remain in effect, are peremptory rather than simply persuasive. See Bratman (1987), p. 24.

[89] One reason for this is that ethical particularism is difficult to reconcile with practices such as rule-following or commitment in which an agent's future action is constrained by a past choice. As Dancy has described it, particularism in ethics rests on a strong form of holism about reasons: a reason that favors an action in one case may not favor, or may disfavor, the same action in another case; and this holds true of all reasons. Reasons, in other words, have no fixed polarity. Dancy (2004), pp. 7, 73-85. In comparison, the instrumental benefits of an authoritative rule depend on the constraint the rule imposes on the agent's actions and the exclusionary effect it has on her deliberation because the agent has adopted it as a rule. Similarly, the normative power often associated with promises depends on the constraint and exclusion that arise from the agent's commitment. The constraint and commitment needed to support the practical benefits of rules and commitments are content-independent, arising from the agent's intent to be bound rather than the merits of the rule or the nature of the commitment. It follows that the effect of the rule or commitment on the agent's choices must be unipolar: it must always favor following the

rule or honoring the commitment. Given the supposed content-independent force of rules and commitments, a contextual change in the polarity of the rule or the commitment makes no sense. There is nothing to indicate when or why such a change in polarity would occur.

For example: S offers to watch her friend's child while the friend completes a work assignment, and forms an intention to do so. Just after talking to her friend, S receives a call offering her an immediate job interview that cannot be postponed. Apart from the promise, the job interview might edge out S's desire to help out her friend. S's commitment to help her friend, however, is supposed to have a certain preemptive force in her deliberations, which will always favor honoring the commitment.

This is not to say that the commitment will prevail in all contexts. It may be outweighed by other considerations: if S badly wants the job, the correct choice is to go to the interview. But the commitment, as such, will always favor helping her friend rather than going to the interview.

Dancy might reply that a commitment has a "default" polarity, in favor of doing what S committed to do. Id., p. 112-113. Default polarity is consistent with holism about reasons, as long as polarity could sometimes change based on identifiable features of particular contexts. But if the effect of rules and commitments is content-independent and exclusionary, then a change in their polarity cannot be explained in terms of background facts. Consider the following two statements:

> (a) I am committed to doing x, so I normally will do x, unless stronger reasons outweigh my commitment.
>
> (b) I am committed to doing x, so my commitment to do x normally will be a reason for me to do x, but it may not be a reason for me to do x, or it may be a reason for me not to do x.

The first statement describes the possibility of defeasance: commitment creates a reason for action that in some percentage of instances will be overridden. This is how commitments are generally understood. But this statement does not assume that the polarity of reasons can change, and so does not imply holism about reasons. The second statement is a genuinely holistic statement, describing default polarity, which may change.

In the case of ordinary reasons for action, the move from a statement analogous to (a) to a statement analogous to (b) may be an improvement. Suppose, for example, that the question is whether to use salt in cooking. It is awkward to say that the effect of salt on the taste of food is a reason to use salt in cooking unless other reasons outweigh it; it makes more sense to say that the effect of salt on the taste of food is normally a reason to use salt in cooking, but not always, and sometimes a reason not to use salt. In context, if the normal polarity of the reason seems inapt, the agent should be able without much difficulty to formulate an explanation of why this is so. For this type of reason, holism makes sense.

In the cases of rule-following and commitment, which rely on a content-independent element of constraint, the holistic default reason (b) is unworkable. Changes in default polarity require some explanation. But, without reference to the content of the action S committed to take, there is no apparent source to which the agent could turn to explain a change in the polarity of her commitment. To the extent that intentions operate by imposing content-independent constraint on agents, as they appear to do in Bratman's theory, the same reasoning applies more generally to all intentions.

[90] For discussion of procrastination problems from an egotistic act consequentialist perspective, see Fumerton (1990), pp. 178-188.

[91] As noted in Section II, McClennen defines resoluteness to require that "if unfolding events, including any conditioning circumstances, are as you had expected them to be, you proceed to execute that plan." McClennen (1997), p. 232.

[92] See Shapiro (2011), pp. 118-129; Shapiro (1998).

[93] Raz (1979), pp. 16-19, 22-23, 30-33; Raz (1986), pp. 57-62.

[94] Shapiro (1998), pp. 47-52.

[95] Shapiro (2011), p. 124. Shapiro says: "[C]hoosing a plan does not set it in stone. Reconsideration is rational when, but only when, there is good enough reason for it." Several sentences later he adds that "it would defeat the purpose of having plans if [the agent] were to review their wisdom without an otherwise compelling reason to do so." Id.

[96] In Shapiro's context, reconsideration might mean reconsideration of the agent's acceptance of the master plan, or rule of recognition, of the legal system. If so, the sort of reason that is relevant to the agent's choice is a reason or set of reasons establishing that the system is not well-designed after all: following all of its rules in all cases they govern will not yield a better sum of results (taking into account coordination and related benefits) than the agent would obtain through unconstrained case-by-case judgement. If this type of reason is what Shapiro has in mind, the constraint that follows from accepting the master plan of a legal system is very strong. Alternatively, reconsideration might simply mean refusal to follow a particular rule in a particular case. If so, the sort of reason relevant to the agent's choice is a reason showing that the rule in question prescribes the wrong outcome for this case. If this is the standard for rational reconsideration, the constraint that legal rules impose on agents is much weaker: it is rational to break a rule whenever the agent believes there is a good or compelling reason to break it.

[97] Id. at 202. As an example, Shapiro refers to the necessity defense recognized in criminal law.

[98] Raz suggests one possible limit on rule-based constraint. As noted above, Raz argues that legal rules generate exclusionary reasons for action. He indicates, however, that the exclusionary reason provided by an authoritative rule excludes only those first-order rules on which the rulemaking authority relied in issuing the rule. In effect, reasons on which the authority relied are merged into the rule, but other reasons may still be in play. By analogy, Shapiro's "compelling reasons" could in principle be limited to those that the lawmaker did not consult in deciding to issue the rule. This interpretation of the constraint implicit in a rule, however, seems too narrow for Shapiro's purposes; and in any event how Raz's limitation on exclusionary rules applies in practice is something of a mystery. See Raz (1979), pp. 16-19, 22-23, 30-33; Raz,(1986), pp. 57-62; Raz (1985), p. 3 (giving the example of a rule based solely on economic considerations).

[99] See Gauthier (1986), pp. 1-20.

[100] E.g., Gauthier (1994).

[101] In both cases - rules and commitments - I am assuming that the agent will deliberate as a rule-sensitive particularist, taking into account any harm that violating the rule or dishonoring the

commitment will cause to the long-term advantages the rule or commitment provides. See Schauer (1991), pp. 94-100. This assumption does not ensure an ideal outcome, however, because the agent will not always judge correctly.

[102] Bratman (1987), pp. 72-73.

[103] Id., p. 87-92.

[104] There may be a rule-following problem imbedded in Bratman's explanation, because the various habits on which it relies themselves represent generalized courses of action that may begin with generalized intentions. This does not appear to generate a problem of regress although, as I will discuss later, there may be epistemic problems in applying Bratman's principles for retaining and not blocking intentions.

[105] In some cases, the agent may deliberate about whether to reconsider, taking into account, for example, interim steps she has made toward following her intention. In most cases, however, action is automatic if the agent is not disposed to reconsider and reconsideration is automatic if the agent is disposed to reconsider. See, e.g., id. p. 60.

[106] Bratman says that the reasonableness of a habits or dispositions depends on whether its "expected impact . . . on the agent's long-term interest . . . exceeds an appropriate threshold." Bratman (1987), p. 72. Thus, defeasibility appears to be a question of what works over time.
[107] See Schauer (1991), p. 677. Bratman appears to recognize this difficulty. He notes that "The historical principle for the basic case allows for a single, all-in assessment of the rationality of the agent for nondeliberatively intending to A. But sometimes relatively perfunctory reconsideration . . . is sufficient to have some impact on the assessment of the present rationality of the agent for his intention, but is insufficient to negative the link with the past. In such cases we may not be able to reach an all-in assessment of agent rationality." Bratman (1987), p. 95. A peek at reasons, however, is likely to be much more common than Bratman allows. If so, there are likely to be many cases in which his principles yield no answer to the question of agent rationality.

Chapter III: Some Special Problems of Temporally Extended Practical Rationality

The main objections I will make to theories of temporally extended practical rationality are that they depend on epistemic irrationality and that reliance on epistemic irrationality puts their own foundation into doubt. I pursue these objections in Chapters IV and V. Before addressing the epistemic problems affecting temporally extended practical rationality, however, I will briefly discuss several other difficulties with theories of this type. One is the relationship among deliberation, intention, and action; another is the generality of intentions. These problems are connected in indirect ways to the epistemic problems I take up in the concluding sections.

A. Deliberation and Action

Standard descriptions of practical rationality assume that agents deliberate about current reasons for action, form an intention, and act on the intention, all in close succession. The agent's reasons for action reflect the expected value of the act as a means for realizing the agent's ends. The exact output of deliberation, however, is a matter of debate. Aristotle suggested that deliberation takes the form of a syllogism, the conclusion of which is the agent's act.[1] If finding food is necessary to satisfy hunger, and if S is hungry, then S finds food. The puzzle is how reasoning can conclude in action.

I will focus here on three different interpretations of what actually occurs when an agent deliberates and then acts, offered by Joseph Raz, John Broome, and Jonathan Dancy. Raz takes a restrictive view, arguing that the conclusion of reasoning is always and only a belief.[2] If S deliberates and concludes that she has a conclusive reason to find food, she will automatically come to believe that she should find food.[3] Deliberation, however, does not yield either an

intention to find food or the act of finding food: some further intervention of the agent's will is needed to bring about either of these responses.

Raz gives two explanations for his conclusion that deliberation cannot yield intention or action. The first of these relates to the role of will in rationality. Although S's conclusion that she should find food automatically yields a corresponding belief that she would find food, she may in fact never find food or form an intention to find food because she may not muster the will to do what she believes she should. Will, for Raz, is a rational power, but it is distinct from reasoning. S's reasoning comes to an end when she forms her belief and only then does will come into play. Under normal circumstances in which the agent's rational powers are functioning properly, will follows from belief; but if will fails, neither intention nor action will follow.

Raz considers and rejects two possible counterarguments to his claim that no reasoning occurs after the agent forms a belief about what to do. One is that the transition from belief to intention is the "practical" content of practical reasoning; in Raz's view, this argument fails because nothing that can properly be called reasoning occurs at this stage. A second counterargument is that when an agent forms an intention to act at a later time, her intention carries forward and prevents the agent from making inconsistent decisions in the interim before action; therefore her intention should be understood as a continuing embodiment of the agent's prior reasoning. In Raz's view, however, the fact that prior intentions affect interim reasoning does not imply that they embody prior reasoning; it means only that a prior intention may trigger additional and distinct deliberation about means for carrying out the intended act. Thus, Raz concludes that reasoning comes to an end when the agent forms a belief about what to do. The agent's intention, if any, is the product of this belief plus the agent's will, with no further

reasoning involved in the transition from belief to intention or from intention to action.

Raz's second explanation for his model of practical deliberation is that reasoning may lead the agent to conclude (and believe) that two or more actions are equally supported by reasons. The agent must then form an intention to engage in one of them and not the other. This choice cannot be the product of reasoning, because reasoning ranks the options as equivalent. Again, reasoning yields only a belief that the different options are permissible, and an additional element of will is need to generate the agent's ultimate intention or action.

Broome follows a different path, which leads to the intermediate position that deliberation can conclude in belief or intention, but not in action.[4] Action requires physical ability in addition to reasoning ability, and so cannot result from reasoning alone. Broome argues, however, that reasoning includes not only theoretical reasoning but also "intention reasoning." Intention reasoning is a form of instrumental practical reasoning in which an agent reasons from a general intention (to pursue some an end), and a belief about necessary means to the end, to a conclusion that consists of an intention to pursue the necessary means.[5]

Broome's defense of intention reasoning is grounded in a propositional approach to deliberation: reasoning is primarily a matter of inference from one proposition to another, rather than the application of rational powers to distill reasons from states of affairs. Beliefs and intentions are both proper outcomes for reasoning, because both have propositional content and both have connections to truth. Beliefs track true propositions, while intentions track (with some limitations) propositions the agent is "set to make true."[6] Thus, beliefs and intentions are sufficiently similar in propositional content that at least some least some inferences from one to the other are normatively required by the rules of reasoning.[7] The intention that results from this process of reasoning, however, is not itself normative because there is no guarantee that the

premises supporting it are justified.

Dancy takes a third view, that belief, intention, and action are all eligible outcomes of deliberation.[8] Dancy acknowledges the difficulties of maintaining that deliberation can conclude in action in any logical or semi-logical sense. In his view, however, the proper question is not whether action can be the conclusion of deliberation but whether the relation between actions and the states of affairs that inform practical deliberation is the same as the relation between beliefs and the states of affairs that inform theoretical deliberation. Having recast the problem in this way, Dancy argues that the relation in both cases is one of "favouring:" the states of affairs that inform deliberation and the considerations adduced in the course of deliberation may favor a variety of responses, including not just belief, but intention, action, and other possibilities as well.[9] Dancy adds that deliberation may lead directly to action if deliberative considerations favor action: there is no need for the interim step of belief.[10] Moreover, in any case in which relevant considerations favor action as well as belief that the agent should act, action is primary. That is, the considerations that favor the agent's belief that she should act favor it because they favor the action. Similarly, considerations that favor the agent's intention to act favor the intention because they favor the action.[11]

Theories of temporally extended practical rationality encounter difficulties under each of these models of practical deliberation. Beginning with McClennen's theory of resolute choice, McClennen claims that it is rational for an agent to deliberate at some initial point in time (T_1), form a resolute intention to act, and then carry through with action at a future time (T_2). McClennen's theory is compatible with both Raz's and Broome's views on deliberation concluding in intention. Applying Raz's model, deliberation at T_1 concludes in a belief that the agent ought to form a resolute intention to act, which can then be converted to an intention at T_1

by operation of the agent's will. Applying Broome's model, the agent reasons at T_1 from an intended end and a belief about means directly to an intention to act.

The difficulty, for McClennen, is explaining how the intention the agent forms at T_1 can yield an action at T_2. I will assume that there are no physical barriers to action. Even so, there is no guarantee that the agent will have either reasons to act or the will to act at T_2. The restriction McClennen adds to resolute choice, that a resolute choice requires action at T_2 only if circumstances at T_2 are substantially as the agent anticipated at T_1, does not solve the problem. Even when this condition is met, the agent may conclude at T_2 that acting on her prior intention is no longer advantageous, either because her preferences have changed or because she has already received the benefits she hoped to obtain by forming and communicating an intention to act. If so, she may prefer not to follow through. It follows that neither Raz's model of practical deliberation nor Broome's model can sustain resolute choice.

Similar problems arise in connection with Shapiro's theory of extended practical rationality. Shapiro holds that, subject to limited allowance for changes in plan (to be discussed in section B below), an intention formed at T_1 requires action at T_2 even when circumstances have changed between T_1 and T_2. This raises the possibility that the agent may conclude at T_2 that the intended act is no longer advantageous under current circumstances. If so, it is difficult to see how the agent can form a will to act at T_2. Again, neither Raz's model nor Broome's allows for completion of plans in the manner Shapiro describes.

Dancy's model of deliberation and action may at first seem more promising for McClennen's model of resolute choice and Shapiro's related model of nonfeasible options. For Dancy, action, like belief or intention, is a response to a set of considerations that stand in a favoring relation to the contemplated action. Because deliberation can lead directly to action,

problems about the role of will in transitions from belief to intention and intention to action do not arise.

The harder question is whether deliberation at T_1 can control future action, at T_2, in the manner McClennen and Shapiro describe. Dancy does not address temporally extended practical rationality and it is hard to predict how he would respond to the problems it poses. It is possible that considerations adduced in deliberation at T_1 can favor not just belief or other current responses, but also resolution at T_1 to act on a current intention at a future time, T_2. This seems unlikely: Dancy's model is grounded in a particularistic approach to ethics and the favoring relation he describes appears to be a relation between considerations adduced the time of deliberation (T_1, for McClennen) and an immediate response to those considerations. Assuming, however, that resolution is an eligible response to deliberation at T_1, there is nothing in either McClennen's theory or Shapiro's theory to prevent an agent who has resolved to act at T_2 from deliberating again at T_2. The only limitation is that the agent must act according to her prior plan. Yet, if the agent proceeds to deliberate further at T_2, and if the sum of considerations adduced at T_2 no longer favor the original plan, then Dancy's particularistic model of practical deliberation and action seems to require that the agent should do what her current deliberation favors. The agent's prior resolution is not itself a consideration favoring action at T_2; resolution is simply a response that is no longer favored by the agent's circumstances as a whole.

Gauthier's theory of intended courses of action poses a somewhat different puzzle under the various models of deliberation just described. In its final version, Gauthier's theory requires deliberation at two points in time: once at T_1, when the agent adopts a course of action, and once at T_2, when she completes the course of action. According to Raz's model of deliberation and action, deliberation at T_1 results in a belief that the agent ought to pursue a course of action. This

belief, combined with will, becomes an intention to pursue the course of action. On Broome's view, deliberation at T_1 results directly in an intention to pursue the course of action. On both views, deliberation at T_2 results in a belief that the agent ought either to complete the course of action, or not. If the agent concludes that she should complete the course of action, her current belief combines with will and physical capacity to yield an action. If not, there is no will and no action follows. So far, so good.

Difficulties arise because, to explain the constraint that intentions impose on agents, Gauthier places special conditions on the agent's deliberation at T_2. The reasons that inform deliberation at T_2 are not simply the agent's current reasons for action at T_2; she must compare her current reasons for action at T_2 with the combination of her current reasons for action at T_2 and any benefits she has already obtained by forming an intention at T_1. Thus, if S promised to provide a reciprocal benefit, formed the intention to do so, received a benefit from the promisee, and prefers the reciprocal exchange to her original position, she must, accordingly to Gauthier, will herself to act on the intention.

Although Gauthier's theory operates differently from McClennen's theory of resolute choice and Shapiro's theory of nonfeasiblity, Gauthier faces a similar problem in accounting for the will to act at T_2. At T_1, if the agent comes to believe that it will be to her advantage to adopt and later carry out the course of action, she normally will be able to will herself to form the necessary intention. At T_2, however, when she must carry out the intention, she may already have received the expected benefits of a course of action. If her original deliberation yielded only an intention, and if current reasons for action favor abandoning the intention at T_2, it is not clear how, at T_2, the agent can muster the will she needs to convert her intention into action.[12] Consequently, Gauthier's theory of temporally extended practical rationality fails to conform to

the models of deliberation to action proposed by Raz and Broome.

Dancy permits deliberation to belief or intention (or both) at T_1, and also permits deliberation directly to action at T_2, based on what response is most favored by considerations adduced at T_2. Under Dancy's particularist approach, however, the outcome of deliberation at T_2 is best understood as a response to the full range of actual, forward-looking considerations adduced at T_2. Consequently, it seems unlikely that Dancy's model of deliberation and action can accommodate that type of constrained comparison Gauthier envisions at T_2, in which deliberation is limited to a comparison between completing the intended course of action and never having undertaken the course of action.

Bratman's theory proceeds differently. According to Bratman, the agent deliberates at T_1 and forms an intention. If, as a result of the agent's dispositions toward prior intentions at T_2, the intention the agent formed at T_1 continues in force at T_2, the agent simply acts on that intention, without further deliberation and without forming a current belief about what to do. Action follows from the element of "volitional commitment" Bratman attributes to intentions. Thus, Bratman does not posit, as McClennen appears to, that prior intentions override the outcome of later deliberation. Nor does he posit, as Gauthier does, a second, limited deliberation at the time of action, assessing the cumulative benefits of the agent's intended course of action. Instead, the agent's initial choice is simply extended from T_1 to T_2, *without reflection*. According to Bratman's standards of agent rationality, this extension of the volitional component of intentions over time is rational if the agent is guided by reasonable dispositions toward prior intentions.

Translated to the terminology of Raz and Broome, Bratman's volitional commitment appears to be the agent's will to act, formed at the time of initial deliberation and somehow imbedded in her retained intention. I will assume that imbedding a will to act in a continuing

intention is metaphysically acceptable; and in fact, it seems plausible that conducting will is a primary function of intentions.[13] With this assumption in place, and setting aside problems of epistemic rationality, Bratman's theory can be reconciled with Raz's and Broome's models of deliberation and action. Deliberation at T_1 yields a belief about what to do in the future and, at least if combined with present will at T_1, a related intention. This intention, if not interrupted, contains an element of will and so yields an action at T_2. Bratman's standards of agent rationality then explain how it may be practically rational to leave the intention in place even though, objectively, the agent's current reasons support a different action. As long as the agent follows dispositions that are reasonable in the long run, the agent can proceed to act on the original plan.

Dancy's model of deliberation and action presents a more difficult challenge for Bratman. The agent's belief that she ought to make a plan is supported by relevant considerations at T_1, as is the agent's intention to act. The problem is that, as already described, Dancy's favoring relation appears to be a relation between considerations adduced in deliberation and a response at the conclusion of deliberation, which does not directly yield an action at T_2. Meanwhile, Bratman's conception of intentions rules out the argument that the agent's prior intention, formed at T_1, counts as a consideration favoring action at T_2: intentions, Bratman maintains, are not reasons for action but volitional commitments.[14]

A proponent of temporally extended practical rationality might argue that an intention carrying volitional commitment is an eligible response to considerations adduced at T_1, which, under Bratman's theory, remains in place T_2 if the agent is reasonably disposed to act on it unreflectively. Because no further deliberation occurs at T_2, no further considerations are adduced and no different response is substituted at T_2. Whether this argument works depends on

whether the set of possible responses to considerations adduced at T_1 is defined broadly enough to include a response that blocks further deliberation in a range of cases. I suspect that a narrower definition of eligible responses is more consistent with ethical particularism.[15] I will assume, however, that Bratman's theory of temporally extended practical rationality can be reconciled with Dancy's model of practical deliberation in the manner just described, and will focus instead on the epistemic arguments made in Section V.

B. Generality of Intentions and the Problem of Defeasibility

A complete theory of temporally extended practical rationality must explain the rationality of conforming to general intentions in particular cases. Temporal extension of agency depends on future-directed intentions. Any future-directed intention is general in the sense that, when the agent initially adopts the intention, neither the act nor the circumstances of its performance can be fully specified. Consider a fairly narrow intention to perform a single act on a single occasion: S intends to buy groceries tomorrow, from a certain store, according to a list. Even this intention can only be formulated in general terms. Substitutions S might make for unavailable items, the care S will take in making selections, and many other details needed to fully describe S's act, are either unpredictable or not worth the trouble of thinking through in advance. This is not a serious problem if, at T_2, the act and surrounding circumstances are in the range of S's expectations, although not precisely as S envisioned them at T_1. If circumstances are not roughly as expected, the rationality of acting on the intention is in doubt.

In the context of rule-following and some types of interpersonal commitment, intentions are general in the additional sense that the agent intends to perform an act in a class of future cases over a period of time. For example: S adopts an internal rule for herself, to write a letter to

her grandmother once a week. As formulated, S's rule is absolute - no exceptions - because S believes she will feel better about herself at the end of the year if she complies and because she also believes she may shirk if she does not fix a schedule for letter-writing. Setting aside questions about what constitutes a letter, the rule is general because it requires S to act every week regardless of variations in her physical and emotional circumstances. S must phrase formulate it in this general way, and adopt it as absolute, to obtain the internal coordination he seeks from the rule: an exception for weeks in which S's reasons for action favor skipping a letter would leave S with no resources against procrastination.[16] Nevertheless, assuming that S had good reason at T_1, and has good reason overall, to write every week, there will be weeks in which S's immediate reasons not to write will outweigh her immediate and overall reasons in favor of writing. The same reasoning applies if S promises her grandmother that she will write each week. To obtain the benefit of normative self-control, S must commit to write in at least some weeks in which his immediate and overall reasons for action favor skipping the letter.

The generality of future-directed intentions leads to the question when, if ever, a theory of temporally extended rationality should allow for rational abandonment of intentions or rational defection from intentions on particular occasions. McClennen, in his theory of resolute choice, appears not to permit abandonment of or exceptions to a resolute intention. This strict stance, however, may be attributable to a limited view of when intentions can be resolute. McClennen's main concern is not with resoluteness in unanticipated circumstances, but with resoluteness that overrides a change in the agent's preferences. Accordingly, he defines resolution in a way that significantly limits the effect of future-directed intentions: "if unfolding events, including any conditioning circumstances, are as you had expected them to be, you proceed to execute that plan."[17] This definition probably sweeps in circumstances that are within

the range of normal daily occurrence, even if not fully specified at T_1. Beyond this, however, any significant difference between expected outcomes and actual outcomes appears to release the agent from the intention. As result, McClennen's theory is not well-suited to sustain practices such as rule-following or ongoing interpersonal commitments, in which some actions that fall within the agent's general intention, as initially formulated, will not be favored by the agent's reasons for action at T_2.

Setting aside McClennen's carefully circumscribed notion of resolute choice, none of the authors I have discussed argues unequivocally that temporally extended practical rationality requires agents to act on prior intentions without exception, no matter how grave the consequences. Shapiro comes the closest to this position.[18] Shapiro's theory of rationality is designed specifically to support the authority of a system of legal rules, and thus assumes the generality of intentions. Those who accept the master plan of a legal system undertake to follow the system's rules in classes of cases that cannot be anticipated in all particulars. This open-ended obligation puts pressure on the theory to allow for reconsideration of intentions to follow the law. Shapiro recognizes the problem but his response is equivocal.

Shapiro's underlying theory of law supports a strict interpretation of the constraint that endorsement of a legal system places on agents. As a legal positivist, Shapiro believes that law claims practical authority over its subjects. In a well-known legal positivist account of the practical authority of law, Joseph Raz describes legal rules as exclusionary reasons for action, such that all reasons that contributed to the selection of the rule are subsumed in the rule and no longer open to deliberation by the rule-follower.[19] The explanation Shapiro gives for the rationality of rule-following is consistent with Raz's account: an agent who forms the intention to follow the rules of a legal system is disabled from consideration of deliberative reasons for

and against the actions required by the rules. Instead, the agent has only "implementation reasons for action," which require that she must follow the rule.[20] This suggests that an intention to follow a rule is not subject to reconsideration.

Yet, despite Shapiro's strong stand on the binding character of intentions, he resists the conclusion that intentions, including intentions to follow law, are absolutely binding at the point of application. He says instead that reconsideration of an agent's intention to follow legal rules is rational if supported by "good enough reason," or, at another point, "compelling reason," to reconsider.[21] Shapiro's standard for reconsidering intentions is imprecise, not only because he describes it in imprecise terms but also because he does not make clear whether reason to reconsider means reason to reconsider the agent's overall intention to follow the law or reason to depart from a specific rule on a specific occasion.[22]

Shapiro also equivocates about why agents can legitimately abandon their intentions to follow rules, or depart from their intentions in particular cases. One way to read his discussion of reconsideration of intentions is that rationality itself dictates that intentions must be defeasible to some degree. It is not rational for an agent at T_1 to impose a constraint on her future self that forecloses any and all deliberation about action. Instead, an agent who is rational at T_1 will insert some caveat into the intention, allowing for reconsideration when the agent has good or compelling reasons to reconsider.

Some of Shapiro's statements, however, suggest that, at least in a legal context, the agent's license to reconsider an intention to follow rules is not a matter of rationality but a contingent feature of the rules themselves. Thus, Shapiro states at one point that the authority of law

"should not be taken to mean that the law demands that its dictates be followed *come*

> *what may*. Laws, like all plans, are typically defeasible. When compelling reasons exist, the law will normally permit its subjects to reconsider its direction and engage in deliberation on the merits. The catch here is that the law claims the right to determine the conditions of its own defeasibility."[23]

This passage indicates that the source of the agent's ability to depart from legal rules is the rule itself. Legal rules may or may not permit agents to reconsider their intentions to follow the law in exceptional circumstances; when they do not, an agent's acceptance of the legal system precludes consideration of any and all deliberative reasons for failing to comply with its terms. This position aligns Shapiro's statements about reconsideration with Raz's theory of exclusionary reasons and with Shapiro's own explanation of the limits that intentions place on deliberation and choice. At the same time, it imposes a very strict constraint on any agent who forms an intention to act according to law.[24]

In contrast to Shapiro, Gauthier provides a determinate and relatively narrow formula defining the constraint imposed by prior intentions in the face of unanticipated circumstances. At T_1, the agent adopts a course of action if she expects that the course of action as a whole will be beneficial. At T_2, the agent compares the consequences of adopting and completing the course of action with the consequences of never adopting (and never completing) the course of action. If adopting and completing the course of action is preferable, the agent has a reason at T_2 to do the intended act; if not, the agent has no reason at T_2 to do the act and should abandon her intended course of action. Gauthier's formula, however, runs into difficulty when applied to general rules and standing commitments that call for application in a series of future cases.

Gauthier's broad objective in formulating a theory of temporally extended practical rationality is to explain the ability of self-interested agents to interact successfully within a

society, based on the advantages that follow from a disposition to cooperate over the course of multiple interactions.[25] His examples, however, typically involve discrete intentions that can be carried out by a single act. The agent forms an intention to complete the second half of a reciprocal exchange, or the agent forms an intention to retaliate if wronged.[26] In this type of situation, the comparison Gauthier proposes yields an answer: depending on the costs of performing the intended act, as they appear at T_2, the agent should either carry out the course of action or abandon it altogether. The agent can change her mind, but only if the overall plan now appears to be disadvantageous.

When an agent adopts a general course of action for all cases of a certain type, the simple comparison Gauthier proposes fails to cover all relevant possibilities. For any given case, the agent now has at least three options: never adopt the course of action, adopt the course of action and complete it in this case, or adopt the course of action and complete in all cases except this case. Judged instrumentally, the rational choice is to adopt the most advantageous of these three options. If the agent believes she can preserve the benefits of the general course of action without complying in the present case, the third option is best.

Suppose, for example, that S rationally forms an intention to follow rule R in all future cases, based on her belief that over time, she will avoid more errors by regular compliance with R than she would make if she decided case-by-case what to do. Sometime later, S faces a situation covered by R. S continues to believe that over time, compliance with R will be advantageous. She also believes that in this case, it would be better to violate R than to follow it, in part because this particular outcome of R looks wrong and in part because she believes that no one will notice her defection, so that defection will not undermine the benefits of the rule. Thus:

> (1) At T_1, S has belief B_1: S expects that following R regularly will be advantageous over the long run; therefore S forms the intention to follow R regularly.

(2) At T_2, S has beliefs B_2(a) and B_2(b):

> B_2(a): S believes that following R regularly has been and will continue to be advantageous in the long run.
>
> B_2(b) S believes that, for this case, a one-time defection from R is better than following R.

In these circumstances, it is not clear, even by Gauthier's reasoning, that S has a reason to follow R. S can obtain the benefits of her intended course of action by adopting the alternative course of action, "follow R in all cases but this one."

The availability of this third option, however, undermines the value of an authoritative general rule or an ongoing interpersonal commitment. The rule becomes a rule of thumb, to be followed when the agent judges that its outcome is correct. Similarly, the serial commitment imposes no meaningful constraint.[27] This result might be unobjectionable if agents never erred or had nothing to gain from normative self-regulation. Given the possibility of incomplete evidence and faulty judgment and the value of coordination and commitment, the loss can be significant.

Gauthier might answer that, for the purpose of the comparison he proposes, there is no third option: the relevant comparison for S is between adopting R and applying it to this case and never adopting R at all. Accordingly, all potential benefits of R over the long run weigh against the benefits of defecting from R in this case. Yet it is not clear why a rational agent should be limited to this comparison if it appears that she can defect now without losing significant long-term benefits of R.

Bratman's response to the problem of when agents should abandon intentions is more detailed. A quick restatement of his principles of agent rationality may be helpful here. Bratman assumes that prior intentions, even when rationally formed, cannot be absolutely binding on a

rational agent. He then draws a distinction between ideal stability of intentions, in which the agent reconsiders prior intentions if and only if reconsideration would lead to a change of mind, and reasonable stability of intentions, which requires only that the agent must follow reasonable dispositions toward retention of prior intentions.[28] Agents, therefore, are allowed a reasonable margin of error if their dispositions are instrumentally sound over the long run.

Bratman's historical principle of agent rationality incorporates reasonable stability of intentions as a key element of practical rationality. An agent who forms an intention after rational deliberation can later act on the intention without further thought, provided that the agent is guided by a disposition to retain intentions of this type, the disposition is reasonable, and the agent reasonably applies the disposition to the intention in question. The agent's disposition is reasonable if it generally works to the agent's advantage. The result is action without current deliberation, unless changes in circumstances exceed the limits of the agent's disposition to retain prior intentions.

As noted in Chapter II, Bratman also develops a supplemental principle of rationality designed specifically for intentions that generalize over classes of cases. Bratman refers to this type of intention as a "policy-based" intention.[29] According to Bratman, policy-based intentions are defeasible case by case. Specifically, the agent can block the intention on particular occasions if, in doing so, the agent is guided by reasonable dispositions toward blocking general intentions. Typically, reasonable dispositions toward blocking intentions make acting on a policy-based intention the default position, but allow for blocking in special circumstances. Again, the standard governing agent rationality is a standard of reasonableness and a disposition that generally works well is deemed to be reasonable.

Thus, in cases involving intentions to act in a series of like cases over time, Bratman's

principles of agent rationality refer in two different ways to the agent's dispositions toward prior intentions. First, the agent must follow reasonable dispositions pertaining to retention of prior intentions. If the agent's reasonable dispositions support retaining the intention, and if the intention applies to a class of cases, the agent can still block application of the intention to a particular case within the general scope of the intention, if reasonable dispositions toward blocking, reasonably applied to the case, support an exception in this type of case.[30] If the agent is disposed not to block the intention, the agent proceeds to act. The upshot is that, rather than delineate a standard by which the agent determines at the point of action whether to retain or abandon, and apply or block, prior intentions, Bratman allows the agent's reasonable dispositions to govern the defeasibility of intentions.[31] If the agent's reasonable dispositions support retaining and acting on the intention, no deliberation occurs at the point of action, so the type of comparison Gauthier envisions does not take place.

Bratman's discussion of the stability of intentions raises three difficulties. First is uncertainty about when dispositions toward retention and non-blocking of intentions count as reasonable.[32] Reasonableness, and therefore the rationality of acting on prior intentions, depends for Bratman on what works well over time. What works well, however, may depend on counterfactual comparisons with alternatives not followed. The question of what works well over time is also affected by the epistemic problems discussed in Section V: I will return to the problem there.

The second difficulty is uncertainty about the process by which the agent determines when to reconsider or block an intention. In developing his historical principles for action on prior intentions in what he calls the "basic" case of single-instance intentions, Bratman assumes a clean distinction between deliberative and non-deliberative responses to prior intentions: the

agent can either proceed to act on the intention without deliberation or reconsider in full the reasons for the intention or its application to the particular case. If the agent proceeds to act, the complete absence of deliberation is important to the conclusion that the agent is rational, because it permits the agent's prior intention to carry forward and govern her action by force of the volitional commitment embedded in the intention. In practice, however, complete absence of deliberation may seldom occur: an agent who has formed a prior intention is more likely to engage in an abbreviated review of reasons before either acting on the intention or stopping to reconsider in full. Bratman's analysis does not easily accommodate this type of partial deliberation. Once the agent peeks at reasons for action and discovers that some have changed, it becomes more difficult to explain why rationality does not require a fully considered response.[33]

Third, the allowances Bratman makes for reconsideration of intentions and for blocking of general intentions in particular cases place significant limits on temporal extension of rationality. In particular, agents who consciously consider their reasons for action at the time of action, either because the reasons to abandon their prior intentions exceed the threshold fixed by their own dispositions toward intentions or because they simply happen to think about reasons for action, are not constrained. Accordingly, Bratman's theory does not easily accommodate rule-following in the strong sense described in Chapter II and may not allow for the type of normative self-control associated with interpersonal commitment. I will return to all these problems in Chapter V, after considering the epistemic side of the problem.

REFERENCES

*References are to works listed in the bibliography (below).

[1] Some fairly clear statements and examples appear in Aristotle, DE MOTU ANIMALIUM 701ª9-25; Aristotle, NICOMACHEAN ETHICS, 1147ª25-30.

[2] Raz (2011), pp. 131-137. Raz draws this conclusion within a broader theory of Reason, reasons, reasoning, and rationality. Reason, for Raz, refers to the rational powers that enable agents to recognize reasons and are constitutive of personhood. One such power is reasoning; but belief, intention, emotion, and any other capacity that can be exercised irrationally are also components of Reason. Rationality is the proper functioning of all the rational powers, including, but not limited to, reasoning. In Raz's view, rationality is not normative. There is no reason to be rational; we simply are this way unless things are not working well. See id., pp. 83-101.

Raz also works from a primarily objective understanding of reasons: for Raz, epistemic limitations that prevent an agent from recognizing applicable reasons do not alter the agent's reasons unless the reasons are unknowable by the agent. Raz leaves the distinction between unrecognized and unknowable reasons largely undefined. See id., pp. 107-128.

Part of Raz's program is to argue that there is no significant difference between practical reasoning and theoretical reasoning, apart from the focus of practical reasoning on the agent's future acts. Comments on this view are outside the scope of my current project.

[3] Raz defines a conclusive reason as one that is undefeated by other reasons and defeats all other reasons. Id., p. 103.

[4] Broome (2002).

[5] Id., p. 86. Broome proceeds to then expand his model of intention reasoning to include reasoning from a general intention and a belief about the agent's best means for achieving that end to an intention to pursue those means. Id., pp. 107-109.

[6] Id., pp. 90-92. Broome points out that intentions do not encompass what the agent believes to be true already and do not track all available means of making something true.

[7] See id., pp. 93-95; see also Broome (2003).

[8] Dancy (forthcoming).

[9] Id., pp. 3-4. Dancy writes from a background of ethical particularism, in which the favoring relation plays an important role. See generally Dancy (2004), pp. 7-10, 73-78. One key feature of ethical particularism is "holism" about reasons: a reason that favors an action in one case may not favor, or may disfavor, the same action in another case. Accordingly, the agent's deliberative task is not to reason from general moral principlse but to recognize and respond to the "basic moral facts" presented in particular cases. Id. at 155-161.

[10] Dancy says "I can adduce considerations, deliberate, and act accordingly without needing to form an intermediate conclusion that this or that course of action is the one I have most reason to pursue. . . . [T]o respond to something as a reason is not and does not require, believing it to be a reason." Dancy (forthcoming), at 10. He adds that reasoning to a belief is not an implicit step in the process: "I would hope that not everything to which we are somehow rationally committed . . . is something we are already doing, even if only implicitly." Id. at 11.

[11] Dancy's argument for the priority of action over intention or belief appears to assume that the features of a situation that favor belief that the agent should act will also favor an intention to act and a corresponding action. My claim that temporally extended practical rationality and epistemic rationality conflict casts doubt on this assumption: if I am correct, reasons for action can diverge from reasons for belief about reasons for action.

[12] Depending on S's preference set, reliance by the other party may be a forward-looking consideration at T_2. But Gauthier does not assume that this is so; his theory is designed to apply even when S cares only about benefits received.

[13] This evidently is Bratman's assumption. See Bratman (1987), pp. 5-9, 18-23 (criticizing the "desire-belief" model of intentions). The assumption also seems well-founded: conceiving of intentions as conductors of will over time, rather than simply as fleeting composites of belief and desire, assigns them a more meaningful role in the process leading to action.

[14] One of the reasons Bratman gives for this conclusion is that intentions, while they remain in effect, are peremptory rather than simply persuasive. See Bratman (1987), p. 24.

[15] One reason for this is that ethical particularism is difficult to reconcile with practices such as rule-following or commitment in which an agent's future action is constrained by a past choice. As Dancy has described it, particularism in ethics rests on a strong form of holism about reasons: a reason that favors an action in one case may not favor, or may disfavor, the same action in another case; and this holds true of all reasons. Reasons, in other words, have no fixed polarity. Dancy (2004), pp. 7, 73-85. In comparison, the instrumental benefits of an authoritative rule depend on the constraint the rule imposes on the agent's actions and the exclusionary effect it has on her deliberation because the agent has adopted it as a rule. Similarly, the normative power often associated with promises depends on the constraint and exclusion that arise from the agent's commitment. The constraint and commitment needed to support the practical benefits of rules and commitments are content-independent, arising from the agent's intent to be bound rather than the merits of the rule or the nature of the commitment. It follows that the effect of the rule or commitment on the agent's choices must be unipolar: it must always favor following the rule or honoring the commitment. Given the supposed content-independent force of rules and commitments, a contextual change in the polarity of the rule or the commitment makes no sense. There is nothing to indicate when or why such a change in polarity would occur.

For example: S offers to watch her friend's child while the friend completes a work assignment, and forms an intention to do so. Just after talking to her friend, S receives a call offering her an immediate job interview that cannot be postponed. Apart from the promise, the job interview might edge out S's desire to help out her friend. S's commitment to help her friend, however, is supposed to have a certain preemptive force in her deliberations, which will always favor honoring the commitment.

This is not to say that the commitment will prevail in all contexts. It may be outweighed by other considerations: if S badly wants the job, the correct choice is to go to the interview. But the commitment, as such, will always favor helping her friend rather than going to the interview.

Dancy might reply that a commitment has a "default" polarity, in favor of doing what S committed to do. Id., p. 112-113. Default polarity is consistent with holism about reasons, as long as polarity could sometimes change based on identifiable features of particular contexts. But if the effect of rules and commitments is content-independent and exclusionary, then a change in their polarity cannot be explained in terms of background facts. Consider the following two statements:

> (a) I am committed to doing x, so I normally will do x, unless stronger reasons outweigh my commitment.
>
> (b) I am committed to doing x, so my commitment to do x normally will be a reason for me to do x, but it may not be a reason for me to do x, or it may be a reason for me not to do x.

The first statement describes the possibility of defeasance: commitment creates a reason for action that in some percentage of instances will be overridden. This is how commitments are generally understood. But this statement does not assume that the polarity of reasons can change, and so does not imply holism about reasons. The second statement is a genuinely holistic statement, describing default polarity, which may change.

In the case of ordinary reasons for action, the move from a statement analogous to (a) to a statement analogous to (b) may be an improvement. Suppose, for example, that the question is whether to use salt in cooking. It is awkward to say that the effect of salt on the taste of food is a reason to use salt in cooking unless other reasons outweigh it; it makes more sense to say that the effect of salt on the taste of food is normally a reason to use salt in cooking, but not always, and sometimes a reason not to use salt. In context, if the normal polarity of the reason seems inapt, the agent should be able without much difficulty to formulate an explanation of why this is so. For this type of reason, holism makes sense.

In the cases of rule-following and commitment, which rely on a content-independent element of constraint, the holistic default reason (b) is unworkable. Changes in default polarity require some explanation. But, without reference to the content of the action S committed to take, there is no apparent source to which the agent could turn to explain a change in the polarity of her commitment. To the extent that intentions operate by imposing content-independent constraint on agents, as they appear to do in Bratman's theory, the same reasoning applies more generally to all intentions.

[16] For discussion of procrastination problems from an egotistic act consequentialist perspective, see Fumerton (1990), pp. 178-188.

[17] As noted in Section II, McClennen defines resoluteness to require that "if unfolding events, including any conditioning circumstances, are as you had expected them to be, you proceed to execute that plan." McClennen (1997), p. 232.

[18] See Shapiro (2011), pp. 118-129; Shapiro (1998).

[19] Raz (1979), pp. 16-19, 22-23, 30-33; Raz (1986), pp. 57-62.

[20] Shapiro (1998), pp. 47-52.

[21] Shapiro (2011), p. 124. Shapiro says: "[C]hoosing a plan does not set it in stone. Reconsideration is rational when, but only when, there is good enough reason for it." Several sentences later he adds that "it would defeat the purpose of having plans if [the agent] were to review their wisdom without an otherwise compelling reason to do so." Id.

[22] In Shapiro's context, reconsideration might mean reconsideration of the agent's acceptance of the master plan, or rule of recognition, of the legal system. If so, the sort of reason that is relevant to the agent's choice is a reason or set of reasons establishing that the system is not well-designed after all: following all of its rules in all cases they govern will not yield a better sum of results (taking into account coordination and related benefits) than the agent would obtain through unconstrained case-by-case judgement. If this type of reason is what Shapiro has in mind, the constraint that follows from accepting the master plan of a legal system is very strong. Alternatively, reconsideration might simply mean refusal to follow a particular rule in a particular case. If so, the sort of reason relevant to the agent's choice is a reason showing that the rule in question prescribes the wrong outcome for this case. If this is the standard for rational reconsideration, the constraint that legal rules impose on agents is much weaker: it is rational to break a rule whenever the agent believes there is a good or compelling reason to break it.

[23] Id. at 202. As an example, Shapiro refers to the necessity defense recognized in criminal law.

[24] Raz suggests one possible limit on rule-based constraint. As noted above, Raz argues that legal rules generate exclusionary reasons for action. He indicates, however, that the exclusionary reason provided by an authoritative rule excludes only those first-order rules on which the rulemaking authority relied in issuing the rule. In effect, reasons on which the authority relied are merged into the rule, but other reasons may still be in play. By analogy, Shapiro's "compelling reasons" could in principle be limited to those that the lawmaker did not consult in deciding to issue the rule. This interpretation of the constraint implicit in a rule, however, seems too narrow for Shapiro's purposes; and in any event how Raz's limitation on exclusionary rules applies in practice is something of a mystery. See Raz (1979), pp. 16-19, 22-23, 30-33; Raz,(1986), pp. 57-62; Raz (1985), p. 3 (giving the example of a rule based solely on economic considerations).

[25] See Gauthier (1986), pp. 1-20.

[26] E.g., Gauthier (1994).

[27] In both cases - rules and commitments - I am assuming that the agent will deliberate as a rule-sensitive particularist, taking into account any harm that violating the rule or dishonoring the commitment will cause to the long-term advantages the rule or commitment provides. See Schauer (1991), pp. 94-100. This assumption does not ensure an ideal outcome, however, because the agent will not always judge correctly.

[28] Bratman (1987), pp. 72-73.

[29] Id., p. 87-92.

[30] There may be a rule-following problem imbedded in Bratman's explanation, because the

various habits on which it relies themselves represent generalized courses of action that may begin with generalized intentions. This does not appear to generate a problem of regress although, as I will discuss later, there may be epistemic problems in applying Bratman's principles for retaining and not blocking intentions.

[31] In some cases, the agent may deliberate about whether to reconsider, taking into account, for example, interim steps she has made toward following her intention. In most cases, however, action is automatic if the agent is not disposed to reconsider and reconsideration is automatic if the agent is disposed to reconsider. See, e.g., id. p. 60.

[32] Bratman says that the reasonableness of a habits or dispositions depends on whether its "expected impact . . . on the agent's long-term interest . . . exceeds an appropriate threshold." Bratman (1987), p. 72. Thus, defeasibility appears to be a question of what works over time.

[33] See Schauer (1991), p. 677. Bratman appears to recognize this difficulty. He notes that "The historical principle for the basic case allows for a single, all-in assessment of the rationality of the agent for nondeliberatively intending to A. But sometimes relatively perfunctory reconsideration . . . is sufficient to have some impact on the assessment of the present rationality of the agent for his intention, but is insufficient to negative the link with the past. In such cases we may not be able to reach an all-in assessment of agent rationality." Bratman (1987), p. 95. A peek at reasons, however, is likely to be much more common than Bratman allows. If so, there are likely to be many cases in which his principles yield no answer to the question of agent rationality.

Chapter IV: Epistemic Rationality

In previous sections, I examined two significant problems of practical rationality and a number of proposed solutions to them. The first problem is that agents with limited time, limited information, and imperfect reasoning capacity can sometimes obtain better practical results by adopting and following general rules than by assessing reasons for action in each case. In some circumstances, however, reasons for action at the time of action may favor violating the rules. The second problem is that, assuming it is possible for agents to impose normative obligations on themselves by making promises or other commitments, interpersonal commitments can increase their capacity for normative self-control and improve the quality of their moral interactions with others. Yet, unless the agent assigns conclusive weight to the moral force of a prior promise, reasons for action at the time of action will sometimes favor abandoning the commitment to perform.

Theories of temporally extended practical rationality address these problems by permitting or requiring agents to act on their prior intentions. Michael Bratman's theory is particularly useful because it allows agents to act unreflectively on intentions, provided that in doing do, they follow reasonable general dispositions toward retention of prior intentions and application of prior intentions to particular cases. This aspect of the theory avoids the need to explain how a prior intention can create a current deliberative reason for action and preserves, at least in part, the standard understanding of practical rationality as a response to reasons in place at the time of deliberation. It also permits Bratman's theory to accommodate general intentions more easily than other candidate theories and allows for a better fit with plausible models of the manner in which practical deliberation leads to action. Although Bratman's theory is not without

difficulties, most references to practical rationality in this chapter assume that the practical rationality of agents is governed by Bratman's principles of temporally extended rationality.

A. Epistemological Assumptions

In the discussion that follows, I make a number of simplifying assumptions about the demands of epistemic rationality. For the most part, the assumptions I make are designed to place my discussions in the mainstream of epistemological thought and avoid overly strict epistemic demands that might ensure the success of my argument by definition. There are, however, plausible views of epistemic rationality that may undermine the conclusions I draw. I attempt to identify the problems these views might cause for me, but do not undertake a full refutation of contrary epistemological positions.

1. Epistemic Rationality as a Distinct Form of Rationality

My first assumption is that epistemic rationality is a special type of rationality, which pertains to an agent's beliefs[1] and requires a connection between grounds of belief, or processes of belief formation, and probable truth.[2] Standards of practical rationality assess the instrumental relationship between an agent's actions and her ends: actions should be effective to further ends. Standards of epistemic rationality assess the likely accuracy of the agent's beliefs as representations of the world: beliefs should in some way track truth.[3]

It follows that the substantive requirements of epistemic rationality and practical rationality will sometimes conflict. Acting on epistemically rational beliefs will not always maximize practical success. If optimism has positive effects on the body's ability to fight disease, then false confidence in the likelihood of a cure may help bring about the cure.

Similarly, if consistently following rules and honoring commitments leads to better practical and moral choices over time, then false confidence that current reasons for action support following the rule can further the agent's ends. To the extent that, in cases of this kind, practical rationality requires the agent to form the practically advantageous belief and epistemic rationality requires the agent to form the true belief, practical rationality may be inconsistent with epistemic rationality.

The form of conflict I will discuss is more subtle and arises when the most effective practical strategies require agents to limit their theoretical deliberation. The simple example of practically advantageous false beliefs, however, raises an important preliminary question, which is whether epistemic rationality is ultimately a form of instrumental rationality. I have assumed for the purpose of this project that practical rationality is an instrumental standard, and have set aside questions about the practical rationality of ends. In the area of epistemic rationality, there has been considerable debate about whether epistemic rationality is itself an instrumental standard. Some writers, notably Richard Foley, have argued that both the rationality of belief and the rationality of action ultimately are governed by instrumental standards. In Foley's view, epistemic rationality aims at satisfaction of a cognitive goal, which he defines as the goal of having true beliefs and avoiding false beliefs. An agent who has this goal ordinarily should believe propositions that, in the agent's own reflective judgment, are based on true premises and follow from truth-preserving arguments.[4] Foley adds, however, that epistemic rationality, so defined, is only one constituent of the rationality of belief: the agent may also have both practical goals and "long-term intellectual goals" that are best served by beliefs that do not meet the agent's epistemic standards. If so, a deliberately false belief can count as a rational belief, although it is not an epistemically rational belief. The rationality of belief, in turn, is only one

constituent of rationality, all things considered. Evaluation of an agent's rationality, all things considered, depends on the relative importance of the agent's various goals and the extent to which different means, including but not limited to accurate beliefs, are likely to advance those goals.[5] Thus, for Foley, purely epistemic rationality may conflict with rational belief or with overall rationality. The conflict, however, is not an impasse: it can be resolved by weighing one against the other according to their respective instrumental contributions to the agent's ends.

Foley gives an example in which a demon threatens to destroy the world unless an agent adopts a false belief; all things considered, it is rational for the agent to adopt the belief.[6] In Foley's view, it would be "epistemic chauvinism" to say that in these circumstances, rationality requires the agent to hold the belief most likely to be true.[7] Thus, for Foley, non-epistemic reasons, meaning reasons that are independent of probable truth, can in principle be reasons for belief.[8]

Other writers maintain that epistemic rationality is not instrumental but rests instead on the epistemically normative implications of evidence or the nature of belief. Thomas Kelly, for example, mounts a number of arguments against instrumental understandings of epistemic rationality, based primarily on common assumptions about evidence and belief.[9] We reflexively believe what our evidence supports and assume that our evidence provides a reason for belief. Moreover, we treat evidence-based reasons for belief as categorical for all agents who share the evidence, regardless of differences in their cognitive goals. In fact, not all agents wish to have accurate and comprehensive beliefs: particular agents may be indifferent or even averse to having true beliefs on particular subjects. Nevertheless, when confronted with evidence, they respond to it with belief in what the evidence supports. Thus, for Kelly, cognitive goals may affect an agent's reasons for seeking out evidence, but they do not define the agent's reasons for

belief. It follows both that evidence is normative in and of itself and that epistemic rationality is a distinct, non-instrumental form of rationality.[10] As further support for his non-instrumental approach to epistemic rationality, Kelly cites the "Incommensurability Thesis," which holds, contrary to Foley, that conflicts between epistemic rationality and practical rationality cannot be resolved. An instrumental conception of epistemic rationality undermines the Incommensurability Thesis because it suggests that instrumental considerations provide a basis for choice between epistemic rationality and practically rational beliefs.[11]

Jonathan Adler reached a similar conclusion about the non-instrumental character of epistemical rationality by a different route, arguing that evidence-based standards of epistemic rationality are intrinsic to the concept of belief.[12] Belief aims at knowledge; therefore an unqualified belief amounts to an internal assertion - an assertion to oneself - that the proposition believed is true.[13] It follows that what an agent should believe depends on what the agent, from a first-person perspective, can accept as true. Beliefs formed on instrumental grounds, without sufficient evidence of truth, do not satisfy this test and so are not rational beliefs.[14]

Nishi Shah makes a similar argument for a non-instrumental understanding of belief.[15] Shah relies on a feature of first-person doxastic deliberation he refers to as "transparency:" from an internal perspective, the question whether to believe a proposition *p* is indistinguishable from the question whether *p* is true. Transparency indicates that the standard of correctness implicit in the concept of belief is truth. This has normative implications for agents because any agent who possesses the concept of belief and deliberates about whether to believe *p* must accept the standard of correctness that belief entails. It follows that an agent considering what to believe must follow truth-generating reasons for belief.

I am most persuaded by the non-instrumentalist arguments in this debate. To believe *p* is

to believe that *p* is true, and this fact about belief is difficult, if not impossible, to suppress in the interest of practical goals. The normativity of the connection between truth and belief for agents is harder to pin down, but both Adler's argument from internal inconsistency and Shah's argument from deliberation under the concept of belief are plausible.

A non-instrumental approach to epistemic rationality is convenient to my project because it supports the distinctiveness of epistemic rationality. If epistemic rationality is non-instrumental, then it is natural to think that what epistemic rationality requires agents to believe can diverge from what practical rationality requires agents to believe or avoid believing, in aid of the agent's interests or goals. A non-instrumental approach to epistemic rationality also suggests that conflicts between epistemic rationality and practical rationality cannot be resolved through a balancing process. Both these assumptions simplify my argument. For reasons that will emerge below, however, they do not complete the argument that temporally extended practical rationality conflicts withe the requirements of epistemic rationality in the settings I examine.

Although a non-instrumental understanding of epistemic rationality advances my case, my arguments can accommodate instrumental views as well. All that is necessary is that, to the extent that epistemic rationality is instrumental, it is instrumental to a specifically epistemic goal such as having mainly true beliefs or seeking knowledge. This is enough to preserve the possibility of conflict between epistemic rationality and practical rationality. An agent who maintains a false belief simply because having the belief will serve her practical ends is not epistemically rational, even if epistemic rationality is an instrumental construct.

Thus, although Foley's instrumental view of the rationality of belief is not my preferred understanding of epistemic rationality, it does not undermine my arguments. Foley characterizes epistemic rationality as just one component of a broader standard governing

rationality of belief, but he nevertheless treats epistemic rationality as a distinct component of rationality, geared to the likelihood of truth. If one of an agent's goals is to have true beliefs and avoid false beliefs, then when the agent forms a belief, she has a reason, although not a conclusive reason, to believe what her evidence and reasoning indicates is likely to be true. Thus, even if standards for rational belief are instrumental in Foley's sense, a theory of practical rationality that demands a different set of beliefs may require agents to engage in epistemic irrationality.

2. Synchronic Rationality

Most theories of epistemic rationality treat standards of epistemic rationality as synchronic, at least in the sense that they assess agents' current justification for current beliefs.[16] What matters is not a commendable pattern of beliefs over time or an ongoing set of cognitive virtues that might lead to correct beliefs in the future, but the agent's grounds for holding a particular belief at a particular time. The purest form of synchronicity is found in "current time-slice" theories, which hold that epistemic rationality depends solely on the match between an agent's belief at a given time and the agent's evidence at that time.[17] The cognitive processes by which the agent formed the belief, or formed prior beliefs that now serve as evidence, are not relevant to the agent's epistemic rationality because epistemic rationality depends on whether current beliefs follow current evidence.

A strict synchronic approach to epistemic rationality conforms to the intuition that beliefs are temporally located and sensitive to changes in the agent's mental circumstances. It also helps to support my argument that theories of temporally extended (diachronic) practical rationality, which authorize agents to retain and act on prior intentions, will sometimes result in epistemic

irrationality. Regularly following a rule, for example, may be practically rational over time as a means to conserve cognitive resources, achieve internal and external coordination, and avoid errors in judgment; but it also may be inconsistent with the agent's rational beliefs about current reasons for action on particular occasions.

My argument, however, can be reconciled with theories of epistemic rationality that are diachronic in limited ways, including at least some forms of reliabilism. In his seminal discussion of reliabilism, Alvin Goldman argues that the justification for any belief depends on the cognitive process by which the belief is formed.[18] A belief-forming process is a function that maps inputs, such as perceptions or prior beliefs, onto output beliefs. Token outputs are justified if the type of process that generated them meets some standard of probability that it will yield true beliefs.[19]

In the course of his argument, Goldman rejects current time-slice theories of epistemic rationality and instead advocates what he calls a "historical" theory of justification. The historical component of the theory holds that, to the extent that inputs to the process of belief formation are prior beliefs, the inputs themselves must have been justified when formed. Goldman's theory of epistemic rationality is thus diachronic in one sense: it looks backward to assess the processes by which input beliefs were formed.

Nevertheless, Goldman's theory of process reliabilism is synchronic in important ways. The target of justification is a token current belief rather than a habit of belief formation or a pattern of belief over time: the role of the justifying process is only to ensure that the belief the agent *now* holds is likely *now* to be true. Moreover, in later refinements of his theory, Goldman allows for the possibility of defeat when the agent's current evidence indicates that her current belief is incorrect, however reliable the processes leading up to it may have been.[20] In

comparison, diachronic theories of practical rationality typically do not require that an agent's token actions must conform to the agent's current reasons for action; it is enough that acting on prior intentions is likely to advance the agent's interests over time.[21] Thus, an epistemic theory that is diachronic in Goldman's limited sense will not undermine my argument because practical and epistemic requirements continue to be at odds.

Taking up again the example of rule-following: adopting and following a rule may reduce the number and gravity of errors the agent makes over time in judging reasons for action. When the agent expects at T_1 that regular compliance with a given rule will provide this benefit, forming a plan to follow the rule is practically rational for the agent. Yet, it remains possible, even within a reliabilist theory of epistemic rationality, that following the rule in a particular case will be epistemically irrational. Epistemic rationality, from a reliabilist point of view, is a standard for assessment of the current truth of current token beliefs, based on the general reliability of the type of process by which they were formed but subject to defeat by current evidence. If by some epistemically rational process, the agent comes to believe at the time of action that she should not now follow the rule, or if her process-reliable belief that she should now follow the rule is defeated by her current evidence, then the agent's practically rational choice may differ from the agent's epistemically rational belief about what to choose.

In contrast, virtue theories of epistemic justification are unlikely to support my argument. Virtue theories may be motivated in part by the objective of securing mainly true token beliefs. Theories of this kind, however, tend to recognize a plurality of epistemic values and are more concerned with general, ongoing states of general cognitive virtue than the particular beliefs that follow from them. Accordingly, they are not synchronic theories in the sense that my argument requires. On the other hand, virtue theories do not really purport to be theories of rationality, so

there is no direct conflict between them and the arguments I make. Thus, although virtue theories do not support my argument, they do not appear to undermine it.

3. Justification for Belief[22]

The range of views on what justifies belief, or rationalizes belief, is very wide.[23] The most prominent general theories of justified belief are evidentialism and reliabilism. Evidentialists tie justification to the specific reasons supporting a target belief. Reliabilists tie justification to the general reliability of the type of process by which the agent formed the target belief.

Among evidentialists, one strategy is evidential internalism, which equates evidence with certain types of mental states the agent may have at the time of belief formation.[24] Questions then arise about what counts as an evidential mental state, with answers ranging from perceptual experience alone, through memory, occurrent beliefs, stored background beliefs that inform current beliefs, and inferences to the best explanation for any of these. Evidential internalists also may differ about the extent to which assessments of epistemic rationality should take account of the agent's subjective capabilities.[25]

Evidential internalism reflects an intuition that grounds of justification should be accessible to the agent, together with uncertainty about access to non-mental facts.[26] An evidential internalist approach is also consistent with conceptual approaches to epistemic rationality according to which belief, by its nature, demands that an agent who concludes that her evidence supports *p* must believe that *p*. Adler, for example, held that an agent's first-person judgment that her evidence does not support *p* is logically inconsistent with the same agent's belief that *p*. Similarly, Nishi Shah argues that an agent who deliberates about belief, and thus

invokes the concept of belief, must follow the intrinsic norms of belief. Norms imbedded in the concept of belief require the agent to adopt "a disposition to be moved by considerations that he regards as relevant to . . . truth."[27] An epistemically rational agent, therefore, must conform her beliefs to her internal evidence.

There are, however, a number of difficulties associated with evidential internalism. One of these is uncertainty about the capacity of mental states, however defined, to explain in a non-circular way the full range of beliefs that commonly are taken to be justified. Another difficulty is that strictly internal evidentialist theories ignore the importance in science and everyday life of the ability to share beliefs and explain their bases to others. Evidential internalism also invites skeptical arguments about worlds in which internal states are manipulated by scientists or demons.

These various criticisms have led to alternative proposals that rely on external elements to provide justification. One approach is to accept evidentialism but define evidence in ways that are not purely internal to the agent. Thomas Kelly, for example, endorses a form of "direct realism" in which evidence can include external facts.[28] Another form of evidential externalism is Timothy Williamson's knowledge-first approach to justification. Williamson accepts that beliefs are justified by evidence, then defines evidence as what the agent knows: "E=K."[29] Knowledge is an irreduceable mental state, which incorporates external conditions (as do all mental states, in Williamson's view). Because knowledge is irreducable, it can explain other epistemological concepts, such as justification for belief, without circularity. It also avoids the subjectivity of a purely internal theory of justification.[30] A possible drawback is that the agent will not always know precisely what evidence she has.

A different type of externalist approach is to reject evidentialism. Reliabilists, for

example, deny that evidence, as traditionally understood, is necessary to justify belief. Instead, they link justification to processes of belief formation that are reliable in the sense that they yield true beliefs with sufficient frequency,[31] or to reliable processes in combination with current evidence,[32] or to reliable cognitive functions.[33] Reliabilist theories imply an external standard for assessment of justification, insofar as reliability depends either on the actual probability that a type of cognitive process will lead to true beliefs or on other objective features of the agent's cognitive functioning. A reliabilist theory such as Goldman's, however, which holds that that beliefs formed by reliable processes are subject to defeat by current evidence, adopts a mixed perspective rather than a purely externalist perspective.

My objective in this project is to show that the best theories of practical rationality require agents to engage in epistemic irrationality by avoiding certain justified beliefs. In making this argument, I will adopt, at least initially, an evidential internalist understanding of justified belief and the epistemic rationality of agents in forming beliefs. On this view, the agent's evidence consists in her mental states, broadly defined to include the background beliefs that form her occurrent beliefs. The agent is epistemically rational if, judged from an external perspective, there is an acceptable fit between her beliefs and her evidence. This is a fairly standard position and a good starting point for discussion. Later, I will consider how my arguments would hold up under alternative approaches, such as process reliabilism or variations on evidentialism.

Turning to the problem of acceptable fit: some evidential internalist accounts of justified belief appear to require a perfect inferential match between the agent's evidence and her beliefs.[34] My focus is on the rationality of agents, and for this purpose a less exacting standard seems appropriate. Accordingly, I assume that the agent is epistemically rational in forming a

belief if the belief has reasonable support in her evidence. Reasonableness requires that the agent's reasoning from evidence to belief must be generally truth-conducive, but it does not require her to grasp all logical consequences of her evidence. Nor must the agent be free from common forms of cognitive bias that affect human reasoners, such as the tendency to give greater weight to emotionally salient evidence than to evidence of background probabilities.[35] This more moderate standard of fit between evidence and belief preserves the connection between epistemic rationality and probable truth without imposing a standard of fit between belief and evidence that no ordinary person can meet.[36]

The understanding of epistemic rationality just described is also compatible in a rough way with Bratman's description of practical rationality. For several reasons, my approach to epistemic rationality does not map perfectly onto Bratman's theory, or any other theory, of temporally extended practical rationality. Standards of practical rationality pertain to actions rather than cognitive functions. Moreover, Bratman's theory of practical rationality is designed to extend over time and not to be restricted to current reasons for action, while epistemic rationality is generally assumed to involve a synchronic relation between evidence and belief. Nevertheless, comparisons between epistemic rationality and practical rationality can at least be simplified if both are assumed to be forms of agent rationality and to allow for reasonable rather than ideal compliance by agents.

Given my assumptions about epistemic rationality, there is a least a gap, and possibly a significant conflict, between the requirements of epistemic rationality and the requirements of practical rationality in the contexts I describe. The fundamental practical problem Bratman's theory addresses is that agents must plan ahead: they will be more successful in advancing their interests over time if they are able to form intentions at one point in time that will constrain their

actions later. Constraint, for this purpose, means that the agent's prior intention to follow a rule, honor a commitment, or otherwise complete a plan, must govern the agent's conduct in at least some cases in which current reasons for action do not support carrying through on the intention. The practical need for constraint arises because the agent has limited decisional resources, is not omniscient, and will make epistemic mistakes, for example by favoring salient reasons over contrary background reasons such as coordination and normative self-governance.

The epistemic problem arises because, in some cases, the best choice at the time set for action is to violate the rule or dishonor the commitment. In other words, there are cases in which, if the agent reviewed her current evidence and formed the belief that current reasons for action favored abandoning her prior intention, her belief would be true. Yet, because the agent is not a perfect reasoner and is likely to be swayed by salient and immediate facts, her choices will be correct more often if she follows the rule in all cases then if she follows her own informed judgment. Thus, although it is best in some cases to act on her judgment, in the long run is best to act on her intentions in all cases. This in turn means that in some number of particular cases, the epistemically justified belief will be that the agent should not act on her intention and the practically rational choice will be to act on her intention. Accordingly, there is at least a potential for conflict between epistemic rationality and practical rationality. Whether this is a true conflict or one that can be dispelled by proper interpretation of the governing theories is a question I will pose and try to answer in Section V.

The foregoing preview of arguments rests on the epistemic assumptions outlined in this section. Specifically, it assumes an evidential internalist theory of justification for belief. Fit between the agent's beliefs and her evidence is judged by a semi-objective standard that allows for reasonable errors and omissions in reasoning. My arguments, however, retain at least some

significance under a range of alternative assumptions about how beliefs are justified, both internalist and externalist.

Suppose, for example, that the governing standard of epistemic rationality is a process-reliabilist standard. Process-reliabilism holds that token beliefs are justified if they result from a type of belief-formation process that is likely, by some measure of probability or propensity, to yield true beliefs.[37] This is an instrumental, mainly external standard of epistemic justification, which refers to a prior cognitive process rather than a relationship between beliefs and current evidence.

Process-reliabilism initially may look similar to the practical standard Bratman proposes for rational retention of intentions, which refers to reliable dispositions to retain prior intentions over time. The two standards, however, are quite different in their objectives and results. The epistemic reliability of a belief-formation process depends on its proven tendency to yield mostly true current beliefs, with truth judged at the time of belief formation. In contrast, the starting points for temporally extended practical rationality are that long-term instrumental success may diverge from current reasons for action and that when this occurs, long-term success should prevail. Thus, the practical reliability of a disposition to retain prior intentions depends on its proven tendency to yield mostly good practical outcomes for the agent over time, with outcomes judged in terms of their contribution to long-term success rather than their conformity to reasons for action at the time of action. As a result, there is no reason to think that truth-conducive processes of belief formation will exactly match practically reasonable dispositions toward retention or reconsideration of prior intentions. It follows that an agent could form a process-justified belief that her reasons for action favor violating a rule or dishonoring a commitment in circumstances in which her practically rational dispositions toward prior intentions favor acting

on a prior intention to follow the rule or honor the commitment. Consequently, there may still be a gap between the demands of epistemic rationality and the demands of practical rationality and a possibility that the two forms of rationality can conflict.

Similar reasoning shows how a conflict between epistemic rationality and practical rationality might arise under Williamson's knowledge-first theory of epistemic rationality . Williamson holds that beliefs are justified by the agent's current evidence and that the agent's evidence is what the agent knows. Beliefs, therefore, are justified by evidence that is not only currently in the agent's possession but also objectively true.[38] Again, there is no reason to think that an agent's beliefs about reasons for action, based on what the agent now knows, will match the practical choices of an agent following instrumentally successful dispositions toward retention of prior intentions. Because Williamson's theory is oriented toward true current beliefs rather than long-term practical success, epistemically rational beliefs about reasons for action do not necessarily correspond to practically rational actions.

More radically, it is difficult to imagine any synchronic theory of epistemic justification that would not result in a gap between temporally extended practical rationality, allowing agents to act unreflectively on to prior intentions, and epistemic rationality. The reason is that prior intentions are general in nature: they apply to types of future circumstances that are not yet fully specified when the agent forms the intention. Acting on them is practically rational if the agent is guided by a set of dispositions toward prior intentions that will produce better outcomes over time than case-by-case judgment. Epistemic rationality, understood synchronically, is particularized, matching beliefs about current reasons for action at a point in time to evidence about current reasons for action at that time.

Consider, for example, a strict and strictly objective theory of epistemic rationality

holding that an agent is epistemically rational only to the extent that her beliefs are true.[39] On this view, it is rational for an agent to believe that her reasons for action favor acting on a prior intention only if her current reasons do in fact favor acting on the intention. Under a theory of temporally extended practical rationality, however, concerns about reasoning error make it rational for an agent to act on a prior intention without further deliberation about current reasons for action if her instrumentally successful dispositions favor acting on the intention. As a result, there will be cases in which an agent rationally and correctly believes that she should abandon her intention to follow a rule or honor a commitment, but extended practical rationality recommends that the agent should carry out her intention. Again, there is a possibility of conflict between epistemic rationality and practical rationality.

4. Having and Responding to Evidence

Evidentialism holds that beliefs are justified by the *agent's* evidence. Accordingly, it becomes necessary to determine what evidence an agent has in her possession. On the evidential internalist view I have assumed, potential evidence consists of certain types of mental states, which may include perceptions, memories, current and stored beliefs, and inferences from all of these. Internality, however, may not be not enough to establish possession of evidence, because the agent may have forgotten, at least temporarily, much of her potential internal evidence at the time she forms a belief.

On this point, Richard Feldman considers a range of possible positions and settles on the view that evidence in the agent's possession is limited to occurrent mental states, meaning contemporaneous experiences or propositions the agent is "currently thinking of."[40] To this, Feldman adds a tentative allowance for certain types of background knowledge. Agents with

special learned expertise may have "feelings of certainty" about conclusions that draw on their expertise, and these feelings count as evidence for their conclusions even if the agent can no longer recall their source. More significantly, agents may have background beliefs that are unconscious and yet operative in the sense that they are "playing an active role in sustaining [the agent's] current state."[41] In Feldman's view, these too count as evidence. Beyond expertise and other active background beliefs, however, the agent's stored but non-occurrent memories and beliefs are not evidence in her possession.

Feldman reaches his fairly stark position on possession of evidence by elimination. His primary criterion for possession of evidence is accessibility to the agent, and it follows from this criterion that information that is forgotten or otherwise inaccessible at the time the agent forms her belief is not in the agent's possession. A variety of intermediate possibilities, such as matters the agent would mention if asked for evidence, or would think of if asked to think of evidence, or could easily call forth, are unsatisfactory because they are either dependent on behavioral characteristics of the agent, overly susceptible to prompting, or too vague to be useful. This leaves only occurrent mental states, subject to the possible extensions mentioned above. This is a quite narrow understanding of the agent's evidence. As Feldman recognizes, it may, depending on context, greatly restrict what agents reasonably can believe. A standard that limits evidence in the agent's possession to occurrent mental states and supporting background beliefs may overvalue superficial beliefs because the agent is not required to reflect in ways that might uncover submerged evidence.

For several reasons, however, Feldman's description of evidence in an agent's possession may be broader in practice than it initially appears. One possibility is that current thoughts may lead naturally to further reflection, bringing submerged evidence to the fore. Feldman's

inclusion of supporting background beliefs also tends to broaden the pool of evidence in the agent's possession. Therefore I will take Feldman's position as my starting point, subject to a generous allowance for background beliefs that play an active role in sustaining current mental states.

B. The Ethics of Belief

When an agent intends to follow a rule or honor a commitment but, at the time of action, her evidence suggests that she should not act according to her prior intention, the result is a gap between the agent's long-term reasons for action and her current grounds for belief about reasons for action. It remains for me to argue that gaps of this kind are not just reflections of the unsurprising fact that rationality takes a variety of forms, but instead represent genuine conflicts between the requirements of temporally extended practical rationality and the requirements of epistemic rationality. One point that bears on this argument is that criteria governing epistemic justification come into play only when an agent forms a belief. It follows that to establish a true conflict between practical rationality and epistemic rationality, I must first clarify the conditions under which epistemic rationality requires an agent to form a belief about reasons for action.

I have assumed an evidential internalist standard of justification for belief, holding that beliefs must match evidence and evidence consists of the agent's occurrent mental states and background beliefs that play an active role in supporting them. In applying this standard, the next question is what if any epistemic responsibility possession of evidence, or access to evidence, places on the agent to process her evidence and form a belief. An agent may often have practical reasons to obtain, advert to, and reflect on evidence, if acquiring more or better-supported beliefs will help her pursue her ends more effectively. For my present purpose,

however, practical reasons for belief are not important. The question is whether the agent has *epistemic* reasons to obtain and respond to evidence, or whether, alternatively, her epistemic responsibility is limited to ensuring that her beliefs match any evidence she has already adverted to in pursuit of a belief.

W.K. Clifford famously said that "it is wrong always, and everywhere, and for any one, to believe anything except on sufficient evidence."[42] Clifford told the story of a shipowner who had doubts about the seaworthiness of his ship. The shipowner suppressed these doubts and convinced himself that the ship was safe, based on its prior record and the benevolence of God. He then sent the ship on its way with passengers, who perished when the ship later sank. Clifford said that the shipowner:

> "*had no right to believe on such evidence as was before him*. He had acquired his belief not by honestly earning it in patient investigation, but by stifling his doubts. And although in the end he may have felt so sure about it that he could not think otherwise, yet inasmuch as he had knowingly and willingly worked himself into that frame of mind, he must be held responsible for it."[43]

There are several ways to read this passage. Clifford may mean only that the shipowner erred in forming a belief that did not match the evidence he considered. If so, the passage does not indicate an epistemic duty either to obtain evidence or to form a belief in response to existing evidence. All that is required is fit between evidence and beliefs actually formed.

Clifford's reference to "stifling doubts," however, suggests that the shipowner made an epistemic error when he ignored evidence that his ship was not seaworthy. The standard of epistemic responsibility underlying this suggestion can be formulated in several ways. The shipowner may have had a duty a to consider all evidence in his possession and then form a belief that fit that evidence. Alternatively, the shipowner may have had a less onerous duty, to consider evidence in his possession that cast doubt on a belief he had already formed, or to

consider evidence that cast doubt on a belief he was about to form. Each of these interpretations goes beyond fit between beliefs actually formed and evidence actually considered, by imposing a duty to advert to evidence and respond with a corresponding belief.

A more radical interpretation of Clifford is that shipowner failed to engage in "patient investigation" before forming a belief. On this view, fit between evidence and belief is not enough; nor is it enough for the shipowner to form a belief based on evidence in his possession. Instead, the shipowner must obtain "sufficient" evidence, or at least some evidence, before forming a belief.[44]

A further question in reading Clifford is the nature of the duty to obtain sufficient evidence, or to reflect on evidence and form a belief.[45] These may be epistemic duties. Alternatively, or in addition, they may be moral duties. Clifford stated, along these lines, that to believe on insufficient evidence is a "great wrong toward Man" because it fosters a general disregard of truth: "the credulous man is father to the liar and the cheat."[46]

More recent work suggests a variety of views on the question of epistemic responsibility to form beliefs in response to evidence. For example, as noted earlier, Thomas Kelly argues that evidence is categorically normative for agents: for all agents, having a particular set of evidence creates an epistemic reason to believe what that evidence supports regardless of the agent's epistemic goals.[47] Thus, in Kelly's view, "[i]f, despite my utter lack of interest in the question of whether Bertrand Russell was left-handed, I stumble upon strong evidence that he was, then I have strong epistemic reasons to believe that Bertrand Russell was left-handed."[48] The same conclusion follows if the agent does not want to form a belief, for example in response to a movie spoiler: "when someone inconsiderately blurts out the ending in my presence . . . it does not follow that I have no epistemic reasons to believe the propositions he asserts. Indeed, with

respect to the question of which epistemic reasons I possess, there is no difference between this case and a case in which I ask the individual to tell me the ending. . ."[49] These statements suggest that an agent who has evidence about current reasons to depart from a rule or break a promise may be required as a condition of epistemic rationality to process the evidence and form a belief.[50]

Kelly retreated somewhat from his initial position in a later exchange with Adam Leite about instrumental conceptions of epistemic rationality. Commenting on Kelly's earlier work, Leite presented a partial defense of the instrumental view. His primary argument was that although the meaning of evidence may be categorical for all agents, it does not follow that possession of the same evidence gives agents the same reasons for belief. Thus, the "support relation" between evidence and a target proposition may be non-instrumental, and yet particular agents may lack an epistemic reason to believe that proposition.[51] Leite gave the example of an agent who notices that many people leaving a certain building have dogs with them. As a result, the agent now has evidence that the next person to leave the building is likely to have a dog, but may not have reason to believe the proposition that the next person to leave is likely to have a dog.[52]

In response, Kelly maintained his position that evidence is normative: "there is no gap between possessing evidence that some proposition is true and possessing reasons to think that that proposition is true."[53] Yet he appeared to modify his initial position in response to Leite's companion-dog example. Kelly observed that in the situation Leite described, the agent never considers the target proposition. The implication of this observation is that evidence is normative only to the extent that the agent entertains a proposition on which the evidence bears. In contrast, an agent "who explicitly attends to some proposition p, at that moment recognizes

that she has overwhelming evidence that p is true, yet does not take up the belief that p," is not epistemically rational.[54]

So modified, Kelly's position on the normativity of evidence does not necessarily establish a conflict between epistemic rationality and temporally extended practical rationality. At least as described by Bratman, temporally extended practical rationality will sometimes call for unreflective action on prior intentions. Thus, an agent who is disposed to follow a prior intention without reconsidering her reasons for action will not, at the time of action, explicitly attend to any proposition about what she should do now. In these circumstances, the implications of Kelly's remarks about the normativity of evidence are ambiguous. Kelly's response to Leite's example may mean that evidence about reasons for action is not normative for an agent who does not reflect on propositions about current reasons for action. Yet it is also possible that Kelly would say that evidence bearing on reasons for action is normative for an agent who understands that she is about to act and understands that the evidence she has bears on some proposition about what action she should now take.

Exactly when Kelly would require an agent to form a belief is complicated by his apparent endorsement of "direct realism" about evidence, as opposed to the strict internalist view that evidence consists in occurrent mental states.[55] Agents are surrounded by evidence, but Kelly's response to Leite's companion-dog example suggests that they are not required to advert to all evidence in their vicinity or to believe all propositions that might be supported by evidence to which they advert. At the same time, Kelly's discussion suggests at least that an agent who has evidence for a proposition and considers that proposition must respond with a belief that conforms to the agent's evidence.

In contrast, Richard Feldman takes the view that epistemic rationality pertains only to the

agent's doxastic attitudes and does not imply a responsibility either to gather evidence or to respond to evidence.[56] Epistemic rationality, in other words, requires a proper fit between the agent's beliefs and the agent's evidence and nothing more. Questions about what evidence the agent should seek and whether the agent can ignore evidence or avoid belief are practical and moral questions that are not governed by epistemic standards.

Feldman's fully developed position, however, is not quite as absolute as his initial statements suggest. In the course of discussion, he indicates that the agent may be required to adjust her beliefs when negative evidence comes to her attention.[57] Specifically, Feldman gives the example of an agent who believes on limited evidence that ginkgo biloba will improve his memory. The agent then notices the following article title on the cover of a credible magazine: "Ginkgo Shown to Be Ineffective." At this point, the agent has evidence contrary to his belief and ought to give up, or at least reconsider, the belief. But an article entitled "Some New Information on Ginkgo" would not have the same effect. In other words, as a matter of epistemic rationality, it may not be open to the agent to ignore evidence that appears to defeat an existing belief. At the same time, that the agent has no duty to investigate ambiguous evidence or to look for evidence that is not now in his possession but might defeat his existing belief.[58]

Interestingly, Alvin Goldman, a process-reliabilist, makes a similar suggestion about defeat in an essay discussing possible evidential limitations on reliabilism.[59] Goldman provides an example in which an agent reads the weather forecast in the morning paper and rationally forms a belief that the weather will be sunny in the afternoon. That same afternoon he is caught in a rainstorm but declines to engage in any process of belief revision, ignores his current experiential evidence, and continues to believe it is sunny. Goldman says that although the agent's belief initially was justified under process-reliabilist standards, it is now defeated by the

agent's current evidence and can no longer be justified. Goldman's conclusion suggests not only that there are evidential constraints on reliabilism but also that agents have some epistemic responsibility to advert to evidence of which they are aware, at least when the evidence appears to defeat an existing belief.

Thus, discussions of epistemic responsibility suggest a range of possibilities. Epistemic rationality may require agents to seek evidence that will provide sufficient support for their beliefs. It may impose no duty of inquiry, but require agents to advert to evidence in their possession. It may require agents to advert to evidence in their possession that supports a proposition that has caught their attention, contradicts a belief they otherwise would form, or defeats an existing belief. All or any of these requirements may be limited to circumstances in which matters of practical or moral significance to the agent are at stake. At the far end of the continuum, epistemic rationality may require only that an agent's beliefs must match her evidence if she forms any beliefs at all.

A significant advantage of Feldman's minimal standard of epistemic responsibility is that it avoids placing an unmanageable burden on agents. Human reasoners navigate through great quantities of potential evidence, declining to advert to or draw inferences from most of it. Even if evidence is limited to occurrent mental states and background beliefs that illuminate occurrent mental states, we are in possession of many, maybe innumerable, items of evidence at any point in time, including unprocessed perceptions and barely discernable beliefs. Moreover, the evidence we notice might support a vast number of inferences. Not all of this can be epistemically normative for us, as Kelly recognizes: the responsibility would be too great. On the other hand, even Feldman is tempted to concede that there is an epistemic responsibility to process some of our evidence, particularly when the evidence suggests that the agent should

revise her existing beliefs. A bare responsibility to conform those beliefs that we happen to form to evidence we have that bears on them does not seem to capture everything that is needed to establish that we are believing rationally.

As will become clear in Chapter V, my argument that temporally extended practical rationality conflicts with epistemic rationality is considerably easier to sustain if there exists a general epistemic responsibility to form beliefs in response to available evidence or evidence in an agent's possession. Given a responsibility to respond to evidence by forming a belief, an agent who has formed an intention to act but encounters contrary evidence about reasons for action must reconsider her intention although it might be practically rational to retain the intention and act without further reflection. Yet, for the reasons given above, a general responsibility to respond to evidence with belief is too demanding. In Chapter V, I will look more closely at the practical and epistemic circumstances of a rule-follower or a promise-maker and will consider how various standards of epistemic responsibility might play out in these circumstances.

REFERENCES

*References are to works listed in the bibliography (below).

[1] I generally refer to beliefs in propositional terms, without meaning to assert that all beliefs are propositional.

[2] The truth-orientation of epistemic rationality is a starting point for most or all leading epistemic theories, including coherentism, foundationalist evidentialism, instrumentalism, reliabilism. See, e.g., Bonjour (1978), p. 5 ("the distinguishing characteristic of this particular species of justification is . . . its essential or internal relationship to the cognitive goal of truth"); Conee and Feldman (2004), pp. 83, 185, 252 (knowledge is the goal of belief; justified beliefs are beliefs based on evidence of their truth); Foley (2001), p. 217 (the "epistemic goal" is concerned with the accuracy and comprehensiveness of one's beliefs"); Goldman (1979), p. 95 (justification depends on reliability, "where . . . *reliability* consists in the tendency of a process to produce

beliefs that are true") (emphasis in original); Williamson (2000), p. 208 ("E=K supports the plausible equation of truth-directed justification with justification by evidence, and therefore with justification by knowledge").

[3] This description is patterned on a similar description offered by David Christensen. See Christensen (2000), p. 364.

[4] Foley (1987), pp. 6-10. In later writing, Foley redescribes the epistemic goal as the goal of now having accurate and comprehensive beliefs; in other words, the goal is not only that the beliefs one has be true and not false, but also that one believe all true propositions. See Foley (2001), p. 217. Presumably, the fact that this is impossible is not a problem becaue the need to pursue other ends will outweigh the goal of believing all truths.

[5] Foley (1987), pp. 11, 210-212.

[6] Id., p. 213. Foley's example is designed to show that evidentialism is not an acceptable approach to the overall rationality of belief. In Foley's view, evidentialism may be suited to the epistemic goals of having true beliefs and avoiding false beliefs but it overvalues these goals when applied as a general standard for rational belief.

[7] Id., pp. 213, 214.

[8] Although a belief formed for non-epistemic reasons can count as a rational (but not epistemically rational) belief, Foley suggests that this will rarely be the case. He notes first that it may not be possible for an agent to believe what the agent's current evidence suggests is false: belief is an automatic response that, according to common understanding, depends on the apparent truth of the proposition believed. Accordingly, to believe both that *p* and that the evidence does not support *p* is a "near-contradictory" state of mind. Id., pp. 215-216.

[9] Kelly (2003), pp. 621-630.

[10] Id., pp. 613, 625-626. Kelly articulates his position on the normativity of evidence emerges most clearly in an exchange with Adam Leite, discussed in more detail later in the Section. In his reply to Leite, Kelly modifies his suggestion that all evidence is normative in all cases, but stands by the position that epistemic rationality is not measured by instrumental standards. See Leite (2007); Kelly (2007); text accompanying notes 192-195, infra.

[11] Kelly (2003), pp. 618-620. Kelly refers to the Incommensurability Thesis as a "very natural reaction," although he does not adopt it explicitly. In support of the thesis Kelly cites, among others, Richard Feldman. Feldman argues that the various forms of rationality - epistemic and non-epistemic - are incommensurable, because there is no common value that can guide a choice among them when they conflict. Conee & Feldman (2004), pp. 193-94.

[12] See Adler (2002), pp. 1-14.

[13] "It is the constitutive claim of belief that its content is true." Id., p. 13.

[14] Adler's conceptual approach to the ethics of belief led him to endorse a strict evidentialist position, holding that the only acceptable reason for belief is evidence supporting the truth of the belief. If belief that *p* is equivalent to an internal assertion that *p* (and thus that *p* is true), it follows that a statement of the form "*p*, but I lack evidence for *p*" is contradictory from a first person perspective in the same way that Moore's paradoxical assertion "*p*, but I do not believe *p*" is heard internally as contradictory. In each case, both conjuncts might be true, but they are inconsistent from the point of view of the speaker (or believer). Id., 29-31 (referring to Moore (1942), pp. 500-503).

[15] Shah (2003), pp. 447-448, 469-470; Shah (2007).

[16] E.g., Conee & Feldman (2004), p. 233; Foley (1987), p. 8; Hedden (2014).

[17] Brian Heddon defends time-slice rationality on the grounds that it avoids the problem of personal identity over time and that it does not hold the agent accountable for prior mental states of which the agent is not currently aware. Hedden (2014).

[18] Goldman (1979).

[19] Goldman is deliberately vague about the standard of reliability for belief formation processes. See id., p. 95. Perfect reliability is not required; consequently some beliefs may be both justified and false. Goldman also indicates that justification will vary along a scale, and that the degree of process-reliability required will depend on the type of belief involved.

[20] Goldman's initial proposal to accommodate current defeat was to place an additional condition on justification, to the effect that the agent's belief is not justified if the agent had available, but did not use, an alternative reliable process that would have resulted in a contrary belief. Id., p. 102. More recently, he has expressed dissatisfaction with his initial test and proposed instead a hybrid reliabilist/evidentialist requirement holding that the agent's belief is not justified if the agent's current evidence, at the time of belief formation, defeats the agent's process-reliable belief. Goldman (2011a), pp. 275-76. This proposal pertains specifically to experiential evidence, but there is no obvious reason why it could not be generalized to other forms of evidence.

[21] Gauthier's theory is an exception insofar as it requires agents to reassess reasons for action at the time of action. Gauthier manages this, however, though a somewhat artificial argument that all benefits flowing from adoption of a course of action at T_1 count as reasons to complete the course of action at T_2. Gauthier (1998b), p. 46.

[22] In the following discussion of justification, I set aside problems connected to the "basing relation" between evidence and belief. See generally Kortz (2010). An agent may have good evidence for a belief, and may form a belief consistent with her evidence, and yet the agent's belief may not be based, causally or otherwise, on her evidence. In the example cases that guide my discussion of epistemic rationality, however, there typically is no serious question about the connection between the agent's evidence and her beliefs. Instead, the more difficult question is

the extent to which she must attend to evidence and form evidence-based beliefs. Accordingly, I will assume a proper basing relation between evidence and beliefs.

[23] As noted early on, I treat rational belief as synonymous with justified belief. I use the term 'rational' in order to emphasize the relationship between standards of practical rationality and epistemic standards.

[24] See Silins (2005), pp. 376-379 (defining evidential internalism as the view that if two agents are identical in their internal mental states, they have identical evidence). An example is the view expressed by Conee and Feldman. See Conee & Feldman (2004), pp. 57, 83, 178, 232; Conee & Feldman, (2011), pp. 302-03.

[25] Compare, e.g., Conee & Feldman ((2004), p. 99 (applying an objective standard of fit between evidence and belief) with Adler, pp. 9 (assessing justification from a first-person perspective that includes "full awareness of what one believes and one's reasons to believe it"); Foley (1987), pp. 9-10 (tying epistemic rationality to first person persuasive argument under conditions of careful reflection); and Kornblith (1983), p. 46 (arguing that "having justified beliefs is simply doing the best one can in light of the innate endowments one starts from, however reliable or unreliable it may be").

[26] See Kelly (2008), (2014) (outlining roles that evidence plays in epistemological theory and tensions among these roles).

[27] Shah (2003), p. 467.

[28] Kelly (2008), p. 950. This is the lawyer's view of evidence as comprising any relevant information about the world the agent encounters. See Austin (1962), p. 116 (proposing that any statement about the world can be a statement of one's evidence). As noted in the text, Kelly takes a non-instrumental view of epistemic rationality, but he does not rely on the internal relation between an agent's evidence and her beliefs to support this position. For Kelly, external evidence has normative epistemic implications for agents.

[29] Williamson (2000), p. 185. Williamson states that because agents cannot be sure that their evidence is true, they cannot always know that they know it. Therefore, to make their knowledge reliable, they must allow a for a margin of error such that there is space between the circumstances in which the agents knows that p is true (for example, she may know that she believes it is cold out), and those in which p is probably false. Within this buffer zone, p may be true but the agent does not know P is true. Nor can she know that she is within the buffer zone, because she cannot know the conjunction that (1) p is true and (2) p's truth is unknown. Id. pp. 15-20, 114-134.

[30] See id., pp. 15-17, 114-134.

[31] Goldman (1979).

[32] Goldman (2011).

[33] Plantinga (1993), pp. 19-20, 40 (justification depends on well-designed cognitive faculties, functioning properly and oriented to truth).

[34] Conee and Feldman indicate that, even if some inference from an agent's evidence are beyond normal human capacities, the beliefs to which those inferences lead are the beliefs justified by her evidence. See Conee & Feldman (2004), p. 87-89. One reason for this strict position is that Conee and Feldman are assessing beliefs rather than agents. Further, they revise their initial formula in an afterword and suggest that their original term "fit" may carry ambiguity. Id., pp. 101-02.

[35] See generally Judgment under Uncertainty: Heuristics and Biases (1982); Heuristics & Biases: The Psychology of Intuitive Judgment (2002).

[36] The position I describe falls somewhere between Foley's standard of careful reflection and Kornblith's standard requiring agent to do the best she can. Foley (1987), pp. 9-10; Kornblith (1983), p. 46.

[37] Goldman (1979).

[38] Williamson (2000), pp. 184-186.

[39] This does not appear to be Williamson's view. See id., p. 9; but cf. id. p. 11 (suggesting that belief might be governed by a standard of knowledge).

[40] Conee & Feldman (2004), ch. 9. Specifically: "S has p available as evidence at t iff S is currently thinking of p." Id., p. 232.

[41] Id., p. 239.

[42] Clifford (1886), p. 295.

[43] Id., p. 290 (emphasis in original).

[44] Richard Hall and Charles Johnson also posit a duty to seek evidence, although without the moral overtones suggested by Clifford. On their view, the duty to seek evidence follows from an instrumental understanding of epistemic rationality, aimed at truth, together with an assumption that current evidence is neutral on some propositions. When evidence is inconclusive, agents have an epistemic duty to avoid suspending judgment, by pursuing additional evidence that might establish the trust or falsity of the target proposition. Hall & Johnson (1998).

[45] See Van Imwagen (1996), p. 145.

[46] Clifford (1886), p. 294; see Wood (2002), p. 3.

[47] See Kelly (2003), Kelly (2007).

[48] Kelly (2003), p. 625.

[49] Id., p. 626.

[50] Adam Leite, Epistemic Instrumentalism and Reasons for Belief*:* A Reply to Tom Kelly's Epistemic Rationality as Instrumental Rationality: A Critique, in 75 Phil. & Phenomenological Research 456 (2007).

[51] Leite (2007), p. 457.

[52] Id.

[53] Kelly (2007), p. 468.

[54] Id.

[55] See Kelly (2008), p. 950.

[56] Conee & Feldman (2004), p. 178. Specifically, Feldman states the principle that "if S has any doxastic attitude at all toward p at t and S's evidence supports p, then S epistemically ought to have the attitude toward p supported by S's evidence at t." Feldman makes clear that the antecedent "if S has any doxastic attitude toward p" is intended to establish that S is under no obligation to form a doxastic attitude. Id., p. 179. He concedes that this view of epistemic rationality will not always maximize true belief, but suggests that it ordinarily will do so and this is enough. See id., pp. 184-186.

[57] Id. at 186-188.

[58] Id. at 187.

[59] Goldman (2011), p. 23. In his initial presentation of reliabilism, Goldman added a reservation to the effect that a belief that results from a reliable cognitive process is not justified if the agent had available an additional reliable process such that, if the agent had used both processes, the agent would have formed a different belief. Goldman (1979), p. 102. His later proposal reaches a similar result by a simpler route.

Chapter V: Practical Rationality and Epistemic Rationality

I have two objectives in this final chapter. The first is to show that temporally extended practical rationality requires agents to be epistemically irrational.[1] I focus on Michael Bratman because the theory of extended practical rationality he has developed offers both a plausible explanation of how it is possible for an agent to act on a prior intention and also what appears initially to be a way around possible conflict between practical rationality and epistemic rationality. Ultimately, however, the conflict between extended practical rationality and epistemic rationality is unavoidable because it reflects an irreconcilable difference between judging reasons for action from a general perspective and judging them from the perspective of a particular case.

As discussed in earlier chapters, Bratman's theory is not the only theory that extends practical rationality over time. Notably, David Gauthier, Edward McClennen, and Scott Shapiro all have proposed that practical rationality should not be judged synchronically, but instead should depend on the long-term instrumental consequences of the agent's action. Each argues that, to accomplish long-term rationality, agents must to some degree be bound by their prior intentions. Gauthier suggests that if the agent adopted a course of action expecting that it will be instrumentally beneficial, and continues to believe at the time of action that the overall course of action is beneficial, the course of action takes priority over current reasons not to perform the particular act that completes it.[2] McClennen proposes that a rational agent should treat her prior choices as "resolute," at least if doing so will yield a net benefit to all the agent's past and future selves, fairly distributed among them.[3] Shapiro argues that an agent who has formed an

intention to act has no rational option but to act as intended.[4]

Under each of these theories, the agent's beliefs about current reasons for action have no effect, or only a limited effect, on the rationality of her current actions. The agent is free to believe what she likes about current reasons for action, long as the actions she takes are instrumentally justified over time. Because practical rationality operates independently of the agent's beliefs, there is no conflict with epistemic rationality.

Gauthier, McClennen, and Shapiro are surely right that synchronic judgments about reasons for action are incomplete and do not lead to well-managed lives. Their proposals for temporal extension of practical rationality also provide an easy way to avoid the conflict I suggest: current beliefs about reasons for action do not determine the agent's choice of action. Nevertheless, I set these theories aside for several related reasons. First, as discussed in chapter III, none of these theories provides a cognitive explanation of how prior intentions can constrain an agent's actions when the agent has in fact formed a contrary belief about her current reasons for action. Thus, although they may suffice for evaluation of the instrumental rationality of particular acts, they do not provide an adequate account of agent rationality. Bratman, in comparison, provides an account of how diachronic choice might actually be possible for agents.

Second, it seems likely that, as Raz and others have argued, an intention cannot lead directly to action without the additional element of will to act.[5] An agent who believes that her current reasons for action, fully considered, do not support acting on her intention normally will lack the will to carry through. Bratman's account provides creative answers to both these problems, but in doing so it presents the possibility of conflict between aspects of practical rationality and epistemic rationality.

My second objective is to show why practical considerations cannot simply override

epistemic considerations. In the contexts I discuss, practical rationality and epistemic rationality are closely related because the beliefs at issue are beliefs about reasons for action. On a temporally extended understanding of practical rationality, agents should act on long-term reasons for action, judged in advance. Epistemic rationality requires that an agent's beliefs about reasons for action should be based current evidence, which may indicate that following a long-term plan is not now the right thing to do. It might seem that both practical rationality and epistemic rationality are aimed at reasons for action, maintaining the best long term course of action is more important than holding epistemically rational beliefs about what to do now. I will suggest, however, that in the contexts I describe, contexts, epistemic rationality and practical rationality are interdependent in ways that rule out a simple conclusion that practical objectives should override epistemic considerations.

A. Conflicting Standards of Rationality

1. Extended Practical Rationality: Bratman Revisited

I begin with a quick restatement of the major features of Michael Bratman's theory of temporally extended practical rationality. Bratman's standards of practical rationality are standards of agent rationality that evaluate the process by which the agent settles on particular actions. They are long-term instrumental standards, judging the extent to which the agent's decisionmaking processes will advance her ends over time. Bratman's objective is to develop a theory of rationality that can capture the various benefits of advance planning, including intrapersonal coordination of actions, coordination with other agents, and effective marshaling of deliberative resources.

To accomplish a temporal extension of practical rationality, Bratman relies on a special

understanding of the nature and function of intentions. Bratman views as independent cognitive states that are not simply composites of beliefs and desires. Conceived of in this way, intentions carry with them an element of volitional commitment that leads directly to action unless the agent reconsiders or blocks application of the intention. If the agent reconsiders the plan before acting on it, her prior intention dissolves and the volitional commitment that accompanied it is gone. If she does not reconsider or block her intention, she proceeds to act without further reflection on reasons for action.

The standards Bratman proposes to govern the practical rationality of an agent who acts unreflectively on prior intentions are designed to achieve reasonable, although not ideal, stability of intentions. The agent is rational if her intention is rational when formed and if she develops and follows a reasonable set of dispositions toward retaining or reconsidering prior intentions. The agent's dispositions toward prior intentions are reasonable if following them will produce better results over the long run than judging what to do at the time of action. In the sections that follow, I will explain why Bratman's theory might be thought to avoid conflict between practical rationality and epistemic rationality, and why in the end the conflict persists.

2. Epistemic Rationality Revisited

In Chapter IV, I outlined a number of assumptions about epistemic rationality that provide a background for my arguments here. Epistemic rationality, for my purposes, is a form of agent rationality, assessing the agent's justification in forming, maintaining, and reconciling her beliefs. Although my arguments can be adjusted to accommodate process-based standards of justification, I assume an evidential standard that makes justification depend on the fit between the agent's beliefs and the evidence in her possession. Assessments of epistemic rationality are

made from an external point of view, but justification is based on evidence in the agent's possession and the standard of justification is one of reasonable rather than ideal fit between evidence and belief. The agent is not required to draw comprehensive or flawless inferences or to avoid the types of cognitive bias that commonly affect human reasoners. Another important assumption is that epistemic rationality is synchronic: the question is whether the agent's current beliefs are supported by her current evidence.

As noted, questions about epistemic rationality in the context of rule-following and commitment pertain to beliefs about reasons for action. Thus, the function of epistemic rationality is to match beliefs about reasons for action to the world as it exists when she forms her beliefs. When the agent first adopts a rule to govern future action or makes a commitment to act, the relevant match is between the evidence the agent then has and the beliefs she forms about reasons to adopt the rule or make the commitment. If the agent reassesses her reasons at the time of action, then because standards of epistemic rationality are synchronic, the relevant match is between current reasons for action and current evidence. In contrast, under Bratman's theory of temporally extended practical rationality, reason for action are long-term reasons both at the time the agents forms an intention and at the time she acts, unless she reconsiders her intention.[6]

In the previous chapter, I left unresolved an important question about epistemic rationality, which is whether it imposes a responsibility on agents to respond to evidence in their possession. Maybe they should advert to their evidence, or draw inferences from evidence to which they do avert; maybe they have no responsibility other than to make reasonable efforts to match their beliefs to such evidence as they choose to consult. This is an important problem for my arguments. To show why, I will elaborate further on the examples of rule-following and interpersonal commitment. My primary focus will be on rule-following because the practical

motivation for rule-following is simpler and less controversial, making the potential for conflict easier to see. The problems raised by rule-following and interpersonal commitment, however, are closely related because both practices depend on the ability of agents to impose generalized constraints on their own future actions in particular cases.

3. Epistemic Circumstances of a Rule-Follower

An initial step in determining whether epistemic rationality poses a problem for temporally extended agency is to clarify the cognitive circumstances of an agent who forms an intention to follow a rule or honor a commitment. I begin with a case in which the agent, S, reflects about reasons for action both when she forms an intention to follow a course of action in the future and when she faces a particular case that falls within her intention. Later, I will consider the more complicated case in which S deliberates only when she forms an intention, and then later acts unreflectively on her initial intention.

At time T_1, S forms the belief at that she can advance her interests by adopting a rule, R. S's evidence for this conclusion includes a variety of beliefs about the benefits of R, which are themselves supported by experiential memory and well-founded background beliefs about how the world works. Specifically, S's evidence includes beliefs about the ways in which regularly-followed rules can provide coordination, both among S's actions over time and between S's actions and the actions of others. S's evidence also includes beliefs about S's own capacity for errors of judgment, beliefs about the savings in cognitive resources that are likely to result from following R, beliefs about the probability that R will prescribe the right outcome in most cases it governs, and beliefs about how other rule-followers are likely to respond if they observe a violation of R. All of these beliefs are occurrent mental states as S deliberates about adopting R

and therefore count as evidence in S's possession at the time of deliberation.[7]

Based on this evidence, S correctly infers, correctly, that R's prescriptions will be instrumentally superior over time to S's own synchronic all-things-considered judgments about reasons for action; that regularly following R may advance S's interests over time; and that a violation of R will have a negative effect on the general practice of following R and consequently on the coordination and other benefits that R provides for S and other members of society. S also infers, correctly, that because R is phrased in general terms, it is overinclusive: in some unidentified subset of future cases to which R applies, S would do better if she violated R. Yet, S's evidence at T_1, including evidence of her own capacity for error, indicates that in exercising her judgment case-by-case, S will not always determine accurately which cases belong to that subset. S concludes that overall, her reasons for action favor adopting R and regularly following R according to its terms, without further reflection in particular cases. S's conclusion is epistemically rational: it fits the evidence S now has. Based on her rational beliefs about reasons for action, S adopts the rule and forms the intention to follow it without exception.[8]

At T_2, S is called upon to follow R. At this time, if S reflects about what to do, she has at her disposal a memory of her belief at T_1 that she should follow R according to its terms, based on the evidence and inferences in play at T_1. S also has new evidence that bears on the question whether to follow the rule in the particular case now before her. Specifically, she has experiential evidence about her current situation; beliefs about the probable consequences of following R at T_2, supported by well-founded background beliefs about how the world works; and beliefs about how other rule-followers are likely to respond if she now violates R in a particular way. From this body of evidence, including the evidence first adduced at T_1, she can

draw an inference about her current reasons for action.

Notice that it is open to S to draw the inference that she should now violate R without revising any of the beliefs that led her to conclude at T_1 that she should adopt R and follow it regularly. S may continue to believe at T_2 that generally following R will advance her interests over time. She may continue to understand that her current evidence may be incomplete and that her current assessment of the evidence may be incorrect due to cognitive bias or other inferential mistakes. But she also will understand, as she did at T_1, that R, by its nature, is a general rule that does not prescribe the best outcome in each and every case it covers. Thus, if it appears to S that the sum of evidence in her possession supports the inference that she should now violate R, there is no inconsistency between her belief set at T_1 and a new belief that in her present circumstances, she should violate R.[9]

(a) **Example 1: Texting**. An ordinary example may help to clarify the problem. S lives in a state that has not yet enacted legal penalties for texting while operating a vehicle.[10] She belongs to a community group, Safety First, which was founded to promote the safety and well-being of local children. The members of Safety First recently voted to take a stand on texting by adopting a rule, RT. RT provides that members of Safety First "shall not read or send text messages while operating a vehicle, for any reason."

At the time the membership of Safety First voted on adoption of RT, S had heard reports from reliable sources about serious accidents attributable to texting, many of which had resulted in injuries to children. She had watched a television advertisement sponsored by a local car repair shop, showing video footage of a texting-related crash with the caption "We don't need your business this badly." Based on these reports, memories of her own behavior, and testimony by friends, she believed that in the absence of a rule, awareness of the risks of texting

while driving is not always enough to overcome self-interest and errors in judgment. She also believed that those who endorse a rule and commit to follow it regularly are likely to follow it. At the same time, S understood, based on reflection and experience, that general rules such as RT are likely to provide some undesirable results. Occasionally, given particulars such as traffic flow, speed, time of day, and a pressing need to communicate by text, and despite the risks involved and the possibility that violation will have a negative impact on the practice of following R, R will prohibit justifiable acts of texting.

S then weighed the advantages of adopting and following RT against the lost benefits of texting. Based on all her evidence at the time, she concluded that she would do better in the long run by adopting and following RT than she would by judging what to do in each case. Accordingly, she voted in favor of RT and formed the intention to follow RT consistently, treating it as an exclusionary reason for action.

So far, S's decision at T_1 appears to have been rational in both epistemic and practical terms. She was epistemically rational at T_1 because her beliefs fit her evidence; she was practically rational at T_1 because the action she intended to take (regularly following RT) appeared likely to advance her practical goals over time. The difficulty arises at T_2, when S must apply RT to a particular case. RT, like most rules, is phrased in general terms. The exact circumstances of its application cannot be known in advance; therefore S's body of evidence is certain to differ in some respects when the time comes to follow RT.

Continuing the example: S has a job that requires her to drive substantial distances during the day. She also has a daughter in middle school. At T_2 she is in slow, heavy traffic about one hour away from the school when she hears a radio bulletin reporting that a suspicious person has been detained outside her daughter's school and the school is currently on lock-down. S cannot

make it to school in less than one hour. S's daughter has a phone with her: she is required to keep the ringer off during school hours but is able to receive texts. S's husband, H, works near the school and could get there quickly. S knows that H is now in a meeting and that his practice in meetings is to turn off the ringer on his phone but to monitor incoming texts. She would like to send to texts to both family members, but she is currently unable to pull over and is several miles from any exit.

S's epistemic situation has now changed. S continues to hold all the beliefs that led her to form an intention at T_1 to follow RT, including beliefs about the dangers of texting, the value of RT, and her own fallibility. She understands that if she violates RT now she will risk harm to herself and others. She also understands that if she is somehow caught in a violation of RT, others may conclude that those who have adopted RT do not always follow it. If so, they may adjust downward their estimates of the value of RT and other self-imposed rules. This evidence favors the conclusion that S should follow RT.

On the other side, S continues to believe, as she did at T_1, that RT is overinclusive and that following it will sometimes be a mistake. Her new information suggests that her child may be in danger. Based on instinct, experience, and general background understanding, she believes that breaking RT now will enable her to provide some comfort to her child, to calm her own fears, and also to alert H and urge him to go to the scene.

Assume that at T_2 S adverts to her evidence, including both the evidence that supported her prior intention to follow RT and her new evidence suggesting that she should not follow RT, and concludes that the balance of reasons for action now favors violating RT. At this point, there are three possible interpretations of S's epistemic situation. The first is that S's conclusion is not reasonably justifiable on the basis of her evidence. Her reasoning may be seriously defective or

she may have limited her attention to her own immediate interests. Under a purely subjective standard of epistemic justification, S's conclusion might nevertheless count as epistemically rational, but on the standard of reasonable fit between evidence and belief that I have endorsed, S's belief is not epistemically justified. Consequently, this interpretation does not produce a conflict between epistemic rationality and practical rationality.

A second possibility is that S's conclusion is wrong but reasonable. S may have omitted to consider some element of evidence or drawn a mildly flawed inference as a result of cognitive biases that most reasoners share, such as the tendency to favor immediate and salient experiences over background beliefs about statistical risks or long-term probabilities. Otherwise, her beliefs fit her evidence. Under the standard of epistemic justification that I have adopted, S has an epistemically rational belief that her current reasons for action favor violating RT.

A third possibility is that S's inference about her current reasons for action is correct. On most standards of epistemic justification, S now has an epistemically rational belief that her current reasons for action favor violating RT. Her conclusion is also practically rational according to the traditional understanding of practical rationality that assesses rationality in terms of responsiveness to current reasons for action.

Under a theory of temporally extended practical rationality, however, both the second and third interpretations of S's epistemic situation might be thought to create a conflict between practical rationality and epistemic rationality. S has formed a practically rational intention to follow RT at T_1, on grounds that remain valid at a T_2. Yet S rationally believes that she should not now follow RT. In coming sections, I will consider whether these two cognitive states can be reconciled.

(b). Example 2: Library Books.[11]

In the texting example above, the primary benefit of rule RT was its capacity to curb reasoning errors by agents. A different example may help to illustrate the coordination benefits of general rules. Suppose that S^1 belongs to a private library maintained through membership dues.[12] Most members of the library find it personally useful to mark in books as they read. Most also prefer to read unmarked books, and most would prefer to to forgo marking in books if they could, in exchange, count on others not to mark in books. Without a rule, however, most expect that books will be marked in, and so continue to mark in books.

If all members of the library were both perfect reasoners and motivated to do what is morally right, they would have no need for a rule. No one would mark in a book unless her reason for doing so were strong enough to outweigh all the harm likely to result, including harm to other members' enjoyment of books and harm by example to the general, morally motivated practice of not marking in books. Each could rely on similar moral behavior by others and books would remain in good, if not excellent, shape.

Yet, even assuming that all members want to do what is morally right, the members will not get this right on their own because they are not perfect reasoners. One member cannot accurately assess the harm that her act of marking in a book will cause without knowing how many others will want to read the book, how much they care about marks, and how many will have good reasons, or think they have good reasons, to mark in a book. Nor is she likely to make a perfectly accurate assessment of her own need to make a mark. Thus, even under wildly optimistic moral conditions, spontaneous cooperation is unlikely to occur.

Suppose, then, that library officers propose a rule, RL, "no marking in library books." At the time the rule is proposed, S^1 has a variety of evidence in her possession. Her perceptual experience tells her that significant numbers of marks in a book distract from the pleasure of

reading and make the book's contents more difficult to follow. Prior experience and beliefs about human behavior tell her that people are likely to comply with a rule they have acquiesced in as long as others appear to be complying as well. S^1 understands that general rules are overinclusive: occasionally a reader may have a genuine flash of insight that can only be captured by marking in a book, and the world may be better off if this insight is recorded, even in a book. Yet, S^1 also understands that on average, she and other readers who believe their insights justify marking in a book will be wrong more often they are correct. Consequently, a strict rule, strictly followed, will provide the best set of outcomes over time. S^1 votes for RL and resolves to follow it regularly and treat it as an exclusionary reason for action.

At T_2, S^1 is in the library reading when she suddenly perceives a connection between a particular passage and a problem she has been trying for months to solve. The idea is complicated, and S^1's prior experience tells her that the best way to retain it is to mark the passage so that she can find it again and add a few words to the margin that will trigger her memory of the connected idea. S^1 is aware, of course, that this would constitute a violation of RL. S^1 continues to believe at T_2 that it is better on average to follow RL in all cases than to rely on her own case-by-case judgment. She understands that her judgment about current reasons for action may be faulty. She also understands that if she marks in the book, other library members will see her marks and downgrade their assessments of how widely the rule is followed. As a result, they will give less weight to the coordination value of RL in their own decisions about marking in books.

Despite her prior conclusions about the long-term benefits of following RL and her current concerns about harm to the value of RL if she does not follow it, S^1 may conclude that the best choice now is to mark in the book. Her idea is important, the risk of losing it is high,

and a few small marks will have a negligible effect on RL. S^1 may be wrong, but this is her honest and possibly reasonable assessment of current evidence. Her conclusion is consistent with the conclusion she reached at T_1, that it is best in the long run to follow RL in all cases, because her conclusion at T_1 was subject to the possibility that in some particular cases the best choice would be to break RL. Again, therefore, S^1 formed a practically rational intention at T_1 to follow RL and treat it as an exclusionary reason for action, but she rationally believes at T_2 that she should not follow RL.

4. The Epistemic Circumstances of a Promisor

The epistemic circumstances of a promisor are similar to those of a rule follower. Depending on how one conceives of promissory obligation, however, they may not be identical. I will assume for the purpose of discussion that promises give rise to promissory obligations that are independent of the consequences of breaking the promise and that serve as reasons of some weight to perform as promised.[13]

As in the rule-following examples, I begin with a case in which P, the promisor, engages in at least some degree of reflection about how to proceed both at T_1 when she makes the promise and at T_2 when she is called on to perform the promise. At T_1, P believes she has good reasons to make a promise. The benefits of a credible promise include the possibility of reciprocal benefits; the possibility of conferring benefits on, or at least not disappointing, the promisee; the possibility of reputational gains; and the satisfaction of self-direction. P also believes that her own future assessment of reasons for action at T_2 could be faulty: the value of self-direction is hard to quantify and the benefits of performance may seem remote in comparison to more immediate contrary concerns. Accordingly, P concludes at T_1 that she

should both make the promise and form an intention to treat it as an exclusionary reason to perform at T_2, without regard to contrary reasons for action. The exclusion need not be complete: there may be a threshold of contrary reasons beyond which P should set aside her intention and reconsider her promise. Within the excluded range, however, she should ignore contrary reasons and perform.[14]

At T_2, none of the evidence that supported P's initial conclusion has changed. Consequently, she continues to believe that, averaging over possible circumstances at T_2, the best overall plan is to treat her promissory obligation as exclusionary. Yet, P also believes, as she believed at T_1, that the exclusion is based, not on the intrinsic weight of the promissory obligation, but on the need to reinforce that weight with a broader exclusion that allows for reasoning errors in particular cases. P now has evidence suggesting that her current reasons for action favor breaking the promise. Taking into account the likelihood that her current reasoning is faulty, she may conclude that the exclusion should not apply and that she should not perform as promised.

Adding some details: P's neighbor N, whose husband recently died, is planning to move to a retirement home. P promises N that, on the weekend before the move, she will help N organize and pack her belongings. At the time P makes the promise, she has evidence that N will benefit significantly from P's assistance. She understands from her own observations and N's testimony that N is in reasonable health and has hired movers, but that she needs help and company sorting through personal items, arranging donations, and deciding what to discard. She believes, based on experience and her general background beliefs about how the world works, that making and then honoring commitments to help others has intrinsic value. Yet, she also understands that when the time comes to fulfill the promise, there may be other demands on her

time that she cannot now accurately anticipate. Because these other demands may be more salient than the background values associated with honoring promises, she may undervalue her promissory obligation and related benefits at T_2. Considering this possibility from the vantage point of T_1, P decides to treat her promise to N as an exclusionary reason to perform.

At T_2, the evidence that led P to make her promise to N and form the intention to treat the promise as an exclusionary reason for action remains in place. This evidence includes both the benefits that performance will confer on N, the intrinsic value of self-imposed promissory obligation, and the possibility that P will not correctly judge her reasons for action at T_2. Yet P also continues to understand, as she understood at T_1, that an exclusionary commitment will require her to perform the promise in some situations in which her current reasons for favor breaking it.

Suppose first that the day before P is due to help N, her pregnant daughter calls to say she has gone into labor 3 weeks early and needs P to come and help. P's daughter lives four hours away and there is no way both to fulfill her commitment to N and to help her child. In this situation, P's promise probably is not binding on her. Although P intended to impose an exclusionary limitation on her future choices, it is fair to interpret the exclusion as stopping short of cases in which significant moral obligations have unexpectedly intervened. As a result, P is free to take all reasons into account and ultimately to break the promise.

This strategy, however, will not work in all cases to reconcile promissory obligations with current reasons for action. P's self-imposed exclusionary reason for action is pointless if it does not block her from responding to some range of reasons for action, including reasons that, in a simple balance, might outweigh both the direct benefits of performing the promise and the intrinsic value of self-direction. The role of an exclusionary reason is to avoid errors by agents

who are not omniscient and cannot be expected to assess all the long-term and short-term consequences of an action with perfect accuracy. Thus, if, on the day before P is due to help N, her daughter calls to say she has a groupon discount for a spa visit this weekend and wants P to come with her, P's promissory obligation to N will preclude P from breaking the promise because she wants to go to the spa.

In the spa case, most agents are likely to recognize the priority of the promissory obligation, so exclusion may not be necessary. Suppose, however, that P's daughter calls in tears to say that she has a project due and her husband is traveling, so could P please come and help out with her three small children. This is a closer case, one in which P might plausibly conclude on reflection that her daughter's needs outweigh N's needs. At the same time, the exclusionary obligation P rationally undertook at T_1 may require P to fulfill her promise to N.

At least in the last variation on the facts, P's epistemic situation at T_2 raises the possibility of conflict between temporally extended practical rationality and epistemic rationality. P formed a practically rational intention to perform her promise to N and to treat her promissory obligation as exclusionary, on grounds that remain valid T_1. At T_2, however, if P reflects on her reasons for action, she may reach the epistemically rational conclusion that she should not now follow through.

5. Second Order Evidence

In the examples above, some of the agent's evidence might be characterized as second order evidence. In each case, the agent reasonably believes both at T_1 and at T_2 that she is subject to cognitive bias and errors of inferential reasoning. She believes that, if she judges that she should violate the rule or break the promise in particular cases, she will make more wrong

choices over time than she would if she always followed the rules. These beliefs are evidence about the agent's evidence or, at least, evidence about the agent's responses to evidence. Thus, they may indicated that the agent should discount or possibly disregard the beliefs she forms about reasons for action when faced with a particular case.

A partial answer is that the gent considers, or at least rationally should consider, the likelihood of flaws in her own reasoning as part of her first order evidence at T_2. She believed initially, and continues to believe, that she will make mistakes and that over time her mistakes will exceed the mistakes that result from the generality of the rule or commitment she intends to follow. Viewed as first order evidence, however, the evidence is not conclusive at T_2 because there will be some particular cases in which the general conclusion does not fold true and it would be better to follow her current judgment than to follow her intention.

Moving to a higher level does not change this result because the agent's evidence is mixed at this higher order as well. Evidence about likelihood of reasoning errors is now second order evidence, which bears on the reliability of the agent's first order conclusions about current reasons for action. Yet, evidence about the potential overbreadth of rules and commitments is also second order evidence, suggesting that the agent's judgment will sometimes be superior to unreflective compliance with her prior intention. Assessing the case at this higher level alone, it may appear, based on probability, that her evidence about agent reasoning error outweighs her evidence about the overbreadth of rules and commitments, because this is the statistically average result. But at T_2 when the agent must form a belief, she has arrived at a particular case. The agent now knows more first order facts, and the facts she knows affect the higher order conclusion. If the current case is a very strong case, as she may now conclude, then the statistical average no longer holds. Thus, the evidence is not higher-order in a sense that implies

preemption or exclusion of first-order evidence.

These observations about higher order evidence reflect another feature of the problem addressed in this project. An agent who forms an intention to follow a rule or perform a promise faces the question of reasons for action from two perspectives, one general and forward-looking and one particularize to a single action in circumstances that are better understood. The difference between these perspectives cannot be eliminated by according priority, or second order status, to practical concerns. I will say more about this toward the end of the chapter.

B. Is There a Conflict?

The principal question I address in this section is whether the divergence between the requirements of epistemic rationality and the requirements of temporally extended practical rationality is a true conflict or one that a well-crafted theory of practical rationality can explain away. There are two possible ways to argue that there is no inconsistency between practical rationality and epistemic rationality. The first path is to focus on practical rationality and argue that because practical rationality refers only to action, it is indifferent to what the agent believes about reasons for action at the time she acts. In other words, the agent can form a set of epistemically rational beliefs about current reasons for action and still follow the practically rational course of acting on her prior intention. As I will explain below, however, this argument is not available given Bratman's assumptions about reconsideration of prior intentions. It also fails to account for the element of will needed to convert intentions into actions.

The second path is to argue that, even if practical rationality requires the agent to respond to beliefs she has formed about current reasons for action, the agent may not form such a belief. This is a key assumption in Bratman's theory. If the agent is not required to form a belief about

current reasons, then a theory of long-term practical rationality can succeed in avoiding conflict with epistemic rationality by simply bypassing belief. This is a more promising argument for reconciling practical rationality and epistemic rationality. The question it raises is whether epistemic rationality includes a responsibility to advert to evidence about reasons for action before acting on a prior intention.[15]

Before proceeding, I should clarify that the problem I have in mind is not a problem of epistemic akrasia, but a problem of conflict between practical rationality and epistemic rationality. Epistemic akrasia comes into play when an agent's evidence supports one proposition but she believes another: the agent's beliefs conflict, not with her reasons for action, but with her reasons for belief. In an example taken from Daniel Greco, an agent who has good evidence indicating that air travel is relatively safe, is aware of this evidence, understands that it indicates that the plane he is thinking of boarding is unlikely to crash, and yet avoids flying because he believes the plane will crash, is in a state of epistemic akrasia. Epistemic akrasia may be attributable to conflicts between first order and second order evidence, or to conflicts between intuitive belief and reflective belief, or possibly to ambiguities inherent in some types of evidence.[16]

I suspect that in most cases of epistemic akrasia, something is wrong with the agent's reasoning: the agent either has failed to resolve a resolvable conflict among beliefs or has formed at least one of her conflicting beliefs without evidence. This is not the case in the situations I consider, in which the agent's theoretical reasoning is assumed to be reasonably sound and in some cases correct. The difficulty in my cases is an apparent clash between the requirements of practical rationality and the requirements of epistemic rationality. The only ways to resolve the inconsistency are to show that practical rationality and epistemic rationality can coexist or to

give priority to one set of requirements.

1. The Role of Beliefs in Extended Practical Rationality

I begin with the question what if any impact a new belief about reasons for action may have on the practical rationality of acting on a prior intention. If the practical rationality of acting on an intention is unaffected by the agent's beliefs about reasons for action when the time comes to act, then practical rationality and epistemic rationality can easily be reconciled. The agent can form an epistemically rational belief about her current reasons for action but proceed to act on a practically rational prior intention, although her prior intention does not track current reasons for action.

As noted earlier, some proponents of temporally extended practical rationality, notably Gauthier, McClennen, and Shapiro, appear to make this assumption.[17] Each suggests that a prior intention constrains current action, whatever the agent may currently believe about reasons for action. Because these theories fail in other ways to provide a convincing account of extended practical rationality of agents, I continue to focus on Michael Bratman's theory of temporally extended practical rationality.

Within Bratman's theory of extended practical rationality, the short answer to the argument that practical rationality is independent of belief is that agents are free to reconsider their prior intentions. When an agent comes to believe that she should not now act as she intended to act, she has implicitly reconsidered her either her intention or its current application and it no longer controls her action. Assuming that following the intention continues to be practically rational over the long term, this means that an epistemically rational belief about reasons for action can undermine extended practical rationality. The following sections provide

further explanation of, variations on, and examples of this point.

(a) Reconsideration of Intentions. At the center of Bratman's work on intentions is a set of standards that permit agents to form an intention at one point in time, then later act on the intention without further reflection and thus without considering current reasons for action. Intentions, for Bratman, are pro-attitudes that have the capacity, while they remain in place, to determine what action the agent takes. If the agent does not reconsider her intention, then when the time comes to act it will lead her to act as it prescribes.[18]

Bratman makes clear, however, that rational agents are not bound by their intentions: they can change their minds. To treat intentions as constraints on actions would, in his view, be irrational.[19] Thus, an agent's prior intentions control her actions not because they constrain her judgment but because she has not altered or rejected them. Bratman also indicates that whenever an agent enters into deliberation about what to do, this amounts to reconsideration of any prior intentions that would have governed her action.[20] By seriously considering another option, the agent implicitly abandons her intention and the volitional commitment it carries.

(b) Reasonable Stability of Intentions. Although the agent is free to reconsider her prior intentions, intentions will not perform their function of extending practical rationality over time unless they are reasonably stable. An agent who constantly reconsiders cannot plan effectively or obtain the benefits of coordination. [21]What makes an agent practically rational is that she possesses a reasonable set of dispositions toward prior intentions. The reasonableness of her dispositions depends in turn on their empirical tendency to serve her interests over the long run.[22] Typically, reasonable dispositions toward prior intentions will include a favorable disposition to act on prior intentions in a significant range of cases without reflecting on current reasons for action.[23] The agent can reconsider, but if she is practically rational, then up to some

threshold, her reasonable dispositions will guide her to act as intended.

(c) Policies. Bratman's early work on temporally extended practical rationality in *Intentions, Plans, and Practical Reasons* focused primarily on single-instance intentions: the agent forms an intention to do an act or type of act at a future time, then later carries out the intention without reflecting on her current reasons for action. Yet, Bratman also recognized the possibility of more general intentions, which he called "personal policies."[24] In later work, personal policies assumed a more prominent role as Bratman turned his attention to problems of self-governance over time.[25] The basic principles he set out early on, however, have remained in place.

Bratman characterizes personal policies as defeasible and develops an additional standard of extended rationality that permits agents to block their application to particular cases without abandoning the underlying general intention. The agent first forms a general intention ranging over all instances of a type of case. Then, upon recognizing that she is now or soon will be in circumstances governed by her personal policy, she forms a more specific "policy-based" intention to act in the token case. Unlike the initial formation of a general policy, formation of a specific intention to act on a personal policy is not deliberative. Instead, it follows unreflectively from the agent's dispositions toward adhering to personal policies in particular cases, which typically will include a disposition to follow the policy unless reasons for violating it reach some threshold of importance. If the agent's dispositions lead her to form a specific intention to follow her general personal policy, and if her dispositions are reasonable, then it is rational for her to act on her specific intention without reflecting on current reasons for action.

Alternatively, because personal policies are defeasible in particular cases, the agent may decide not to form a specific intention to act and instead to block application of her personal

policy to the case at hand. The process of blocking specific applications of general personal policies is similar to reconsideration of single-instance intentions. The rationality of a decision to block application of the policy depends on the agent's reasonable dispositions toward her personal policies, which normally include a disposition to apply the policy unless reasons not to apply it exceed some margin. If her current circumstances exceed the margin set by her reasonable dispositions, the agent should decline to act unreflectively and instead reflect on current reasons for action. Once the agent enters into the process of deliberation, she effectively has reconsidered the specific intention she otherwise would form, to apply her general personal policy to the current token case. She has not, however, reconsidered her general personal policy about what to do in cases of the same type. Thus, the only significant difference between the standards Bratman applies to personal policies and the standards he applies to single-instance intentions is that the notion of specific intentions allows for limited reconsideration, leaving the general personal policy otherwise in place.

(d) Texting. The problems of adopting and following rule and making or honoring commitments are versions of the problem of adopting and following general personal policies. The intention is general because it is designed to govern either a series of particular actions or an action that may be called for in various circumstances that are not yet specified. Continuing the texting example: at T_1, S voted for rule RT, which forbids texting while driving, and formed the intention to follow it regularly and treat it as an exclusionary reason for action. In Bratman's terms, she adopted a personal policy to follow the rule. In doing so, she was both epistemically rational and practically rational.

Assume first that S is disposed not to block the application of personal policies to specific cases unless the consequences of following the policy appear *significantly* worse than the

consequences of violating the policy.[26] This disposition toward personal policies is reasonable in the sense that, over time, it will enable S to realize her goals more effectively than she would if she always reconsidered her policies at the point of application to specific cases. At T_2, S has evidence consisting of the same beliefs that led her to adopt a policy to follow RT at T_1, including the benefits of RT, the potential harm to these benefits if she violates RT, and the imperfections of her own reasoning. She also has new evidence indicating that if she follows RT now, she will miss a chance to comfort and protect her child in what may be an emergency and in any event will worry a lot until she can acquire more information.

Assume further that if S assessed this body of evidence she would conclude that although RT continues to be a good rule, the consequences of following it in the current emergency are worse than the consequences of violating it. Assume also that she would view the case as close: the comparative advantage of violating R does not appear significant. S's assessment is epistemically reasonable, and may also be correct.[27]

In this case, standards of practical rationality tell S to follow her reasonable dispositions toward prior intentions, which in turn tell her not to block application of her policy to follow RT because her reasons to do so fail to meet the test of significance. Accordingly, she should form a specific intention to follow RT, then act on her specific intention. If, however, S adverts to her current evidence and forms a belief about current reasons for action, standards of epistemic rationality tell her to believe that her reasons for action slightly favor violating RT and sending a text. The margin may not be significant, but the evidence, as she reasonably evaluates it, lines up on side of violating RT.[28] Given Bratman's assumption that prior intentions lose their force when an agent reflects about contrary reasons for action, the outcome is that S will send a text.

This may not seem a serious blow to extended practical rationality. But if similar

situations arise over time, and if S tends as most people do to give more weight to immediate needs than to background concerns about coordination and error, and if each iteration of the example decreases the reliability of RT, then the long-term practical benefits of RT will diminish. Thus, even when the effect of new evidence is only to block the application of general intentions to particular cases, epistemically rational beliefs about reasons for action can may be at odds with long-term practical rationality.

It matters, here, that epistemic rationality is generally assumed to be synchronic. S was epistemically rational at T_1 when she formed the intention to follow RT, and her beliefs about the long-term benefits of following RT continue to be epistemically rational now. Yet, her new evidence supports additional beliefs about current reasons for action that, if she adverts to the evidence, will interrupt her practically rational long-term plan.

Thus, the first method of reconciling epistemic rationality and practical rationality is not available: although practical rationality is concerned with action and not with belief, it is not impervious to the agent's beliefs. Once the agent adverts to current evidence and forms a belief, that she should not now follow a prior intention, the prior intention no longer has the capacity to govern action directly. It follows that if the agent is required as a matter of epistemic rationality to advert to her evidence and form a belief about reasons for action, there is a conflict between epistemic rationality and practical rationality. I address this possibility below, in section V(B)(2).

(e) Practically Reasonable Dispositions and Exclusionary Rules. Joseph Raz has argued that authoritative rules should be understood as exclusionary reasons for action. As an exclusionary reason for action, a rule serves both as a first order reason to act as it prescribes and a second order reason to exclude from consideration a range of reasons for action whose

relevance and weight is settled by the rule.[29] Similarly, Raz has suggested that promises create reasons for action that block consideration of a range of reasons not to perform.[30] Each of the examples above incorporates this idea.

In my last discussion of the texting example, I suggested that one form of practically rational disposition toward blocking specific applications of a general personal policy would require that reasons to violate the rule should reach some threshold of weight. Another form of disposition toward blocking specific applications of policies might require the agent to treat the rule as an exclusionary reason for action. Within Bratman's theory, whether exclusionary dispositions are reasonable depends on the likelihood that they will help the agent achieve her ends.

Beginning with intentions to follow rules: a good rule is one that will, over time, yield a better set of outcomes than the agent would achieve if she reviewed her evidence and formed beliefs about reasons for action in each case. It follows that a disposition to treat good rules as exclusionary reasons for action meets the test of reasonableness. The set of outcomes produced by unreflective adherence to a rule might not be optimal, but it is preferable to the outcomes of case-by-case judgment. Accordingly, it is practically rational for agents to treat rules they judge initially to be good rules as exclusionary reasons.

Intentions to perform a promise are more complicated because the moral and other values associated with a promise are not quantifiable, making them hard to compare with the agent's judgment about whether to perform under a variety of possible circumstances. At the same time, it may be fair to assume that the agent's judgment over time will tend to under-assess the moral and other values of the promise in comparison to immediate and sometimes self-interested reasons to break it. If so, a disposition to treat the promise as an exclusionary reason for action is

likely to yield more value over possible future circumstances than regular exercise of judgment. In the promising context, this assessment is plausible. In the context of rule-following, however, it is unsettling think of blind adherence as reasonable.

(f) A Note on Belief and Acceptance. An early essay by Bratman on the nature of practical deliberation calls for one further comment on the argument that forming a belief about current reasons for action is equivalent to to reconsideration of a related intention. In the essay, Bratman propses that practical deliberation does not depend exclusively on beliefs about reasons for action.[31] Instead, it rests at least in part on a more tentative, manipulable, and context-dependent type of cognitive attitude he calls "acceptance." Although Bratman does not make a connection between the attitude of acceptance and non-reconsideration of intentions, the idea of acceptance complicates the relationship between current beliefs and prior intentions.

Bratman endorses a standard description of full-fledged beliefs: beliefs are evidence-based, truth-oriented, context-independent, and involuntary in response to evidence. Ideally, the set of an agent's beliefs also should be integrated in a coherent way. Beliefs, so described, are the "default background" for practical deliberation, but they do not always determine its outcome. Instead, practical deliberation may be guided by propositions the agent does not believe in the full sense of the term but accepts for the purpose of deliberation.

Acceptance refers to the agent's endorsement of a variety of assumptions that aid her in deliberating. Bratman does not propose that acceptance is an epistemically rational cognitive state. Instead it is a practically rational stance toward possible facts, controlled by the agent and shaped by the circumstances of her deliberation. For example, the agent may adjust the propositions on which her deliberation relies to reflect high or low stakes associated with a particular choice, her own temporary cognitive impairment at the time of deliberation, or other

specific features of a particular deliberative choice. These adjustments respond to context and are not aimed exclusively at truth.

The argument that reconsideration of intentions should turn on what the agent accepts rather than what she believes would go something like this. If an agent's initial intention rests on her acceptance of a set of provisional assumptions, then spontaneous reconsideration occurs only when the agent revises her assumptions. Therefore a new belief about reasons for action does not amount to reconsideration as long as the set of accepted premises remains intact. The intention remains in place and retains the capacity to govern the agent's actions.

The problem with this argument is that, assuming that deliberation based on accepted premises can support an intention to act, it does not follow that the intention should continue in place when the agent forms an evidence-based belief that contradicts those premises. As I understand Bratman's discussion of acceptance, he does not want to say that, for purposes of practical deliberation, an agent can accept propositions that are directly contrary to what the agent believes. Truth has great value in the context of instrumentally rational choice. Therefore, in a showdown between a belief and a proposition accepted for purposes of deliberation, the belief should win.

Bratman's examples of acceptance support this reading. Many deal with questions about how to deliberate under conditions of uncertainty. Others deal with settings such as group deliberation in which theoretical reasoning about reasons for action cannot proceed in the usual way. None involve direct contractions between acceptance and belief. The implication is that, although agents can rely on accepted propositions to fill empirical gaps in their practical deliberation, they cannot act on a prior intention when they currently believe they should not follow the intention.[32]

2. The Ethics of Belief

In the examples provided so far, the agent adverted to evidence currently in her

possession and formed the belief that she ought to abandon, or ought not now follow, her prior intention. My argument in the last section was that when this occurs, the agent can no longer act on intentions in the way Bratman's theory of practical rationality contemplates. Bratman, however, might still respond that, if all goes well and the agent's practically rational dispositions remain intact, the agent will not advert to evidence about reasons for action, will not reflect about whether to follow R, and will not form a belief.[33] And, Bratman might add, there is no epistemic reason why she should consult evidence and reflect about whether to follow a rule. Epistemically, she is free to do what practical rationality requires, which is simply to follow the rule based on her prior intention and her specific intention to act on it.

The line of argument just described, that temporally extended practical rationality bypasses epistemic difficulties, poses the questions about epistemic responsibility discussed at the end of the previous chapter. If epistemic rationality imposes a responsibility on agents to assess available evidence and form beliefs in response, then an agent is epistemically required to reconsider or block her intention when her evidence suggests that she currently has contrary reasons for action, even if long-term rationality might favor staying to the course. If there is no such epistemic responsibility, then the agent can act unreflectively on her prior intentions without offense to epistemic rationality, by simply declining to process her evidence. Duties of inquiry are not a significant issue in this context because the agent's new evidence normally consists of perceptions and beliefs about her present circumstances that bear on an act she is about to perform. The important questions are whether and to what extent epistemic rationality requires the agent to think about this evidence, draw inferences from it, and form a belief about her current reasons for action.

In the following subsections, I will consider three possible standards of epistemic

responsibility that require agents to process evidence in at least some circumstances. Each of these standards prevents agents from simply disregarding evidence that casts doubt on a prior intention. For reasons to be explained, I find the third of these standards best adapted to the problems of rule-following and interpersonal commitment.

(a) Strong Epistemic Responsibility. Starting with a strong version of epistemic responsibility to process evidence: assume that an epistemically rational agent must advert to all evidence in her possession, draw reasonable inferences from that evidence, and form beliefs that fit the evidence. Evidence in the agent's possession includes perceptions, beliefs, and memories the agent has in mind or that come to her mind and background beliefs that make sense of this evidence. S previously adopted a personal policy to follow rule RT's ban on texting while operating a vehicle. She now has new evidence about current reasons to send a text, such as the radio announcement she has just heard, the traffic around her, and various beliefs she holds about the current social climate and the competence of school officials. Strong epistemic responsibility requires S to advert to this evidence, to consider whether her reasons for action now favor violating RT, and if the answer is yes, to believe that she should now violate RT. For reasons set out in the last subsection, if S lives up to her epistemic responsibilities and concludes that she should send a text, then she will effectively have blocked the application of her intended policy.

The strong version of epistemic responsibility just described, however, is overly demanding. As noted in Chapter IV, even when evidence is limited to occurrent mental states, we acquire more evidence than we reasonably can process and translate into beliefs. It follows that not all evidence should be normative, or at least that not all evidence should require extended inferential reasoning of the sort required to form beliefs about reasons for action. In an example mentioned earlier, an agent standing outside a building sees a number of people leave

the building with dogs.[34] This observation might support a variety of propositions, for example that the building owner permits dogs on the premises, that a dog show is happening inside, or that the next person to leave is statistically likely to have a dog. Agents should not, however, have a categorical duty to treat these observations as evidence and draw conclusions about the propositions they might support; in most circumstances, they should be able to tune them out. Accordingly, I will set this version of epistemic responsibility aside and consider several alternative positions that rely on less burdensome standards of epistemic responsibility but also lead to conflict between epistemic rationality and practical rationality.

(b) Epistemic Responsibility in Practically Significant Circumstances. Another possibility is to recognize an epistemic responsibility to draw inferences from evidence and form beliefs but limit this responsibility to evidence that bears on practical questions of significant interest to the agent or, more narrowly, to evidence that bears on moral responsibilities of the agent. For example, Clifford's shipowner, knowing he was in the business of transporting passengers, and having possession of evidence suggesting that his ship was not seaworthy, might be required to advert to this evidence and form a belief about reasons for and against sailing as planned. He would not, however, be required to advert to evidence about how many of his sailors were wearing blue hats, which is practically and morally inert for him.

A standard of responsibility triggered by the practical and moral significance of the agent's evidence would frequently come into play in the contexts I examine. Intentions to honor interpersonal commitments have moral significance for agents. Intentions to follow rules have at least practical significance and sometimes moral significance for agents. In other situations, however, it will impose a less onerous burden on agents than the strong standard of responsibility first discussed, which requires them to process all evidence in their possession.

Initially at least, this second standard of epistemic responsibility appears to place epistemic rationality and temporally extended practical rationality in conflict in cases of rule-following and interpersonal commitment. If the agent has new evidence indicating that she should not take some practically or morally significant action, such as following a rule or honoring a promise, then the agent must advert to this evidence and form a belief about her current reasons for action. Once she considers the evidence, she can no longer act unreflectively in the way that Bratman's version of extended practical rationality requires. Assuming the evidence does not reach the threshold of weight at which the agent would no longer be disposed to act on her intention in any event, she must either be epistemically irrational or be practically irrational.

On closer examination, however, this standard of epistemic responsibility conflates epistemic and practical considerations. The standard is epistemic in the sense that it regulates the epistemic process of responding to evidence with belief and the agent's conclusions are guided by truth rather than practical advantage. Yet, the agent's responsibility to undertake the process of responding to evidence is practically motivated. Presumably, therefore, it could be overridden by contrary practical considerations such as the importance of making reliable plans in a business such as shipping.

Thus, the second standard of epistemic responsibility fails to establish an unresolvable conflict between epistemic rationality and practical rationality. If epistemic responsibility is based on practical concerns of the agent, then the apparent conflict between epistemic rationality and practical rationality reduces to a conflict among practical concerns. If the agent's long-term practical reasons to retain and act on prior intentions override her practical reasons to process evidence about current reasons for action, then the conflict dissolves and long-term advantage

prevails.

(c) Epistemic Responsibility to Respond to Belief-Defeating Evidence. A third, more promising account of agents' epistemic responsibility to process evidence and form beliefs would limit the agent's responsibility to cases in which the evidence tends to defeat a belief the agent currently holds. According to this formulation of epistemic responsibility, agents do not have a general epistemic responsibility to advert to evidence, draw inferences from evidence, or form beliefs. They must, however, process evidence in their possession that challenges their existing beliefs.

The motivation for this standard is epistemic rather than practical. The standard comes into play when the agent's evidence suggests that her current set of beliefs contains errors. Its effect is to restore epistemic order by correcting the erroneous beliefs. A defeat-based standard is also fairly modest, in comparison to the standard of strong epistemic responsibility described above, because it is triggered by existing beliefs. A further point in favor of this standard is that it appears compatible with both instrumental and non-instrumental understandings of epistemic rationality.[35]

Richard Feldman endorses a standard of epistemic responsibility much like this. Feldman states initially that epistemic rationality means only that whatever beliefs the agent forms must fit the agent's evidence, but then adds that agents may also be required to process evidence that threatens to defeat their current beliefs.[36] Alvin Goldman, in his revised version of process reliabilism, also suggests that agents should respond to evidence that defeats a reliably formed current belief.[37]

Feldman makes the point that challenge to prior beliefs should be fairly obvious. His example, described in Chapter IV, describes an agent who believes in the benefits of Gingko

Biloba, but then encounters a headline announcing that new research has shown Gingko to be ineffective. In most cases involving intentions to follow rules or honor commitments, this criterion will be met whenever the agents confronts a situation that brings her prior intention into play in unexpected particular circumstances.

At first glance, epistemic responsibility to advert to evidence that defeats existing beliefs seems well-suited to the contexts of rule-following and interpersonal commitment, in which the agent has formed an intention to act on the basis of a set of beliefs about reasons for action The argument for a conflict between epistemic rationality, defined to include responsibility to advert to belief-defeating evidence, and practical rationality, would be as follows. It is practically rational for an agent to form a general intention to take certain types of action in certain types of case, based on her beliefs about the advantages of following this intention regularly and the likelihood that she will err if she attempts to assess what to do case by case. When the agent confronts a situation of the intended type, she then forms a specific intention to implement her general intention in the case before her. If new evidence threatens to defeat the set of beliefs on which her general and specific intentions to act are based, then under the standard of epistemic responsibility now under consideration, the agent must advert to the evidence and form a new or modified set of beliefs in response. Doing so, however, will either eliminate or block her general intention and undermine its potential practical benefits.

(d) Texting, and a Problem. On close examination, the fit between the defeat-based standard, as just described, and problems of rule-following and commitment, is not perfect. Applying the standard to the texting example: S has a general intention to follow rule RT, which generates a specific intention to follow RT now. Practical rationality, extended over time, dictates that she should follow this intention as long as her practically dispositions toward rule-

following favor compliance with RT. Yet, if she encounters evidence suggesting that the set of beliefs that supports her general intention is incorrect as applied to this case, a defeat-based standard of epistemic responsibility requires her to advert to the evidence and determine whether to adjust her beliefs about reasons for action. If she does this and forms a new belief, she will effectively have reconsidered or blocked her intention to follow RT and the intention will no longer exert volitional control over her actions. She will send a text. By standards of practical responsibility, however, this outcome will, on average, be incorrect. Unless the evidence exceeds the threshold set by her practically rational dispositions, she should not send a text.

The difficulty in applying the defeat-based standard to this case is that S's evidence does not defeat the set of beliefs that supported her general intention and specific intentions to follow RT, in the sense of proving them untrue. The set of beliefs that supports S's general intention includes the belief that she will do better over time by consistently following her intention than by consistently following her judgment, the belief that some cases will arise in which she would do better by not acting on her intention, and the further belief that the second belief does not falsify the first belief. S's new evidence, indicating that she now faces a case in which she should not act on her general intention to follow RT, is consistent with all of these beliefs. Her original beliefs, in other words, are not in danger of defeat. Despite the new evidence, and even if her inferences from the evidence are correct, it continues to be true that she will do better over time by always ignoring her evidence and always acting on her general intention than she will by judging reasons for action in each case.[38]

(e) Epistemic Responsibility to Respond to Belief-Defeating Evidence, Reformulated. The defeat-based standard of epistemic responsibility, however, can be reformulated in a way that fits the problems of rule-following and interpersonal commitment but does not substantially

increase the epistemic demands placed on agents. In cases of this type, the agent has formed a general intention, or policy, to perform a type of action in a type of case. The agent's general intention is based on a belief that acting regularly on the general intention will have long-term practical benefits and a further belief that these benefits will more than compensate for the bad outcomes that result in some cases. Although the agent's new evidence does not contradict these existing beliefs, its effect is to reveal an epistemic flaw that results from their generality: unless revised, her beliefs do not properly account for her current situation. Another way to put this is that the agent's new evidence supports a better belief, that she generally should act on her intention but not in this case.

Modified to reflect this type of epistemic error, the third standard of epistemic responsibility would require the agent to respond to evidence showing that her existing general beliefs need revision to accommodate her present circumstances. From an epistemic point of view, consistently following RT, or treating a promise as exclusionary, is generally the right choice, but not now. A standard requiring belief-revision in cases of this type serves the same epistemic values as the standard of responsibility that both Feldman and Goldman appear to support, which requires agents to process evidence that may defeat existing beliefs. In both cases, responding to evidence and modifying prior beliefs corrects a problem that has come to light and restores the integrity of the agent's set of beliefs.

Of course, acting on the modified belief may not be practically rational over the long run, because the agent may be wrong. Her evidence may be incomplete and her inferences may be imperfect. This is why she formed a general intention to follow the rule, or to treat her promise as exclusionary, in the first place. Epistemically, however, what counts is that the agent's existing beliefs should be revised to match her reasonable assessment of current evidence

that bears on their justification.

Although this revised version of a defeat-based standard of responsibility has practical repercussions, from an epistemic standpoint its effects are limited. Agents are not required to advert to all evidence in their possession, or to draw all possible inferences from evidence, or to form all beliefs that match their evidence. They are not, for example, required to think through the implications of a random observation about people and dogs. As in the case of defeating evidence that Feldman describes, the agent is required only to draw ordinary inferences from readily accessible evidence, in order to determine how a prior belief holds up in a new case. Thus, the principal drawback of this approach is not the burden imposed on agents but the inconvenient consequences for temporally extended practical rationality.

The standard of epistemic responsibility I have described places epistemic rationality and practical rationality in conflict. Epistemic rationality requires the agent to advert to current evidence and form a belief about the agent's current reasons for action, while practical rationality is better served if the agent either fails to respond to current evidence or fails to form a new belief in response to evidence. As a result, the agent can only conform to standards of practical rationality by setting aside the demands of epistemic rationality.[39]

3. **Summary**

I have argued that, when S has rationally formed a general intention to follow a rule or honor a commitment in all cases of type C, temporally extended practical rationality and epistemic rationality place conflicting demands on S. The argument rests on two premises. The first is that, if S comes to believe that her current reasons for action in a type C case favor action contrary to her intention, her intention no longer controls her action. As a result, S cannot be

practically rational over the long term unless she avoids forming such a belief.

The second premise is that standards of epistemic rationality require S to advert to current evidence and form a belief about current reasons for action in circumstances that are typical of the problems I discuss. If S's evidence indicates that she should revise existing beliefs in response to new conditions, she has an epistemic responsibility to respond to the evidence. I explain and defend the first premise in section V(B)(1) and the second in section V(B)(2). Both premises are in keeping with a sensible understanding of what it means to be a rational agent. Rational agents should be able to change their minds and rationality of belief should include a responsibility to adjust general beliefs when evidence suggests that they are wrong as applied. In the contexts I describe, the implications of these two premises is that, because we necessarily view our choices from different perspectives at different times, components of our rationality are unavoidably at odds.

C. Priorities

I have argued that under a plausible understanding of epistemic responsibility, epistemic rationality requires agents acting on prior intentions to form beliefs about current reasons for action. In any event, agents will sometimes form such beliefs spontaneously as an epistemically rational response to evidence. I have also argued that under the most attractive theory of temporally extended practical rationality, practical rationality requires agents who act on prior intentions to avoid forming epistemically rational beliefs about current reasons for action. The next question is which type of rationality should prevail.

This may seem an odd question because in practice we, as agents, appear to have voted in favor of practical rationality. We follow rules in particular cases when we have evidence

supporting an exception to the rule; sometimes we follow them unreflectively and sometimes we follow them although we believe we have good reasons to break them. We treat commitments as exclusionary and disregard evidence of conflicting reasons for action.[40] This type of response to rules and commitments is not only common but also supported by powerful instrumental arguments. Practices that involve formation of plans, compliance with authoritative rules, and binding promises are essential to effective human agency and social interaction. It may seem obvious, therefore, that the long-term practical benefits of these practices should override whatever epistemic limitations theories of temporally extended practical rationality impose on agents.

Most others who have considered the relationship between practical rationality and epistemic rationality have concluded either that practical rationality ultimately should prevail over epistemic rationality or that practical rationality and epistemic rationality cannot usefully be set against each other for comparison. As noted in Section IV, Richard Foley maintains that conflicts between practical rationality and epistemic rationality can be resolved by weighing the agent's practical and epistemic goals and the likelihood that the actions under consideration will advance these goals. Consequently, epistemic rationality must sometimes give way to more pressing practical concerns.[41] In Foley's leading example, if an agent can save the world from a demon by adopting a false belief, it is rational, overall, for the agent to adopt the belief.[42] Of course, given the choice Foley describes, it surely is better to save the world. But for reasons I will explain below, Foley's example is distinguishable from the problems raised by rule-following and interpersonal commitment.

Those who take a non-instrumental view of epistemic rationality are more likely to say that epistemic rationality and practical rationality are incommensurable. Richard Feldman, for

example, states that there is no "generic ought" linking practical rationality and epistemic rationality. Consequently, there is no value that can guide a choice between these two aspects of rationality when they conflict.[43] Feldman may be correct that there is no metric for direct comparison between epistemic rationality and practical rationality. The examples of rule-following and interpersonal commitment, however, suggest that conflicts between epistemic rationality and practical rationality are daily occurrences. Thus, if rationality is to remain an ethically significant idea, then a choice must sometimes be made between the requirements of epistemic rationality and those of practical rationality.[44]

Andrew Reisner has proposed an intermediate view.[45] Reisner's approach to epistemic rationality is instrumental: epistemic rationality aims at the goal of true belief, which can be overridden by practical goals. Yet he does not simply weigh one type of rationality against the other. Instead he argues that in some normal range of circumstances, evidential reasons govern belief formation and practical reasons have no role. In exceptional circumstances, however, practical reasons for belief take over. Evidential reasons for belief are then "defeased" and no longer influence the agent's belief. Reisner admits, however, that he has no algorithm for determining, either as a general matter or in particular cases, where to locate the point at which evidence becomes inert and practical reasons for belief come into play. As a result, it is difficult to say whether and when a useful general rule, or a normatively important commitment, might tip the scales in favor of practical rationality.

In any event, temporally extended practical rationality, and particularly the examples of rule-following and interpersonal commitment, raise a special kind of problem. The examples of conflict cited by Foley and others are misleadingly simple because they posit cases in which practical rationality and epistemic rationality are unrelated apart from the stipulation that the

agent must give one up in order to maintain the other.[46] When an agent chooses to believe that the sky is green in order to save the world from a demon, the false belief plays no part in her decision whether to make the trade-off; it is simply the by-product of that decision. In other words, the agent's practical rationality and her epistemic irrationality operate independently of one another in these standard examples.

In the context of extended practical rationality through unreflective action on prior intentions, the problem is harder because epistemic rationality and practical rationality are interconnected. To be practically rational over the long term, the agent must avoid forming a belief about reasons for action that otherwise would dictate her decision about what to do. Avoiding this new belief can have practical consequences. If the agent's new belief would have been correct, the agent's action will be instrumentally inferior to the action she would have taken if she had evaluated and responded to her current reasons for action.[47]

Maybe this is not a serious problem. The practical harm resulting from avoidance of epistemically rational beliefs about reasons for action is limited in several ways. First, due to reasoning errors, the belief the agent failed to form may not have been correct. Second, if the belief is correct, the harm done may be more than balanced over time by the normally beneficial effects of following general intentions. In the case of rule-following, for example, the justification for adopting a rule is that, on average, compliance with the rule will head off errors of judgment and solve coordination problems. Over the long run, therefore, circumventing epistemically rational beliefs about reasons for action should result in less rather than more practical error. Similarly, in the case of a promise, the moral value that comes from successful normative self-control may, in the long run, overtake the negative practical consequences of disregarding evidence and avoiding epistemically rational beliefs about reasons for action.

Within Bratman's particular theory of temporally extended practical rationality, a third ameliorating factor is that any loss of practical benefits that follows from avoiding epistemically rational beliefs about reasons for action is limited to the range of error allowed by the agent's reasonable dispositions to retain general intentions and not to block their application in particular cases. The agent's dispositions count as reasonable only if they yield practical benefits over time. So it may seem that the danger that epistemic irrationality will destabilize practical rationality is both minimal and well-balanced by the long-run advantages of the agent's compliant dispositions.

The entanglement between epistemic rationality and practical rationality in theories of temporally extended practical rationality, however, may be greater than it first appears. The reason is that the intentions that facilitate temporal extensions of agency are not isolated from beliefs about reasons for action. Bratman's theory, and probably any fully articulated theory of diachronic rationality, involves webs of intentions that, together, permit rational agents to obtain the benefits of planning. In his full description of temporally extended agency, Bratman assigns two roles to prior intentions.[48] One is the volitional role, in which a prior intention motivates later action without further deliberation. The second is the "reasoning-centered" role, in which a prior intention constrains later deliberation: when an agent deliberates about forming a new intention, the options from which the agent can choose are limited to options consistent with the set of prior intentions she rationally retains at the time of deliberation. The standard of rationality for retaining intentions for deliberative purposes is the same as the standard for retaining intentions for the purpose of action: the agent must be guided by reasonable dispositions that will serve her well over time. If so, then the agent can and should conform her new intentions to her prior intentions without consulting evidence or forming beliefs that might

cast doubt on prior intentions.

Suppose, for example, that in the case of rule RT, the rule prohibiting texting while driving, S's geographic position one hour from her child's school is a result of her general intention to accept all work assignments without argument. Her decision to adopt this intention was constrained by her prior intention to do well in her present job, which in turn was constrained by her prior intention not to return to school and learn new skills until her daughter graduates from high school, which belonged to a chain of intentions leading back to a decision by S and H to live in a coastal area where the cost of living is high. At each stage, if S had adverted to her current evidence and formed a belief about reasons for action, she might have concluded that she should abandon her prior intention, consider a wider range of options than her prior intention allowed, and alter her course. At each stage, however, her reasonable disposition toward prior intentions indicated that she should retain, and therefore choose consistently with, her existing intentions. Accordingly, at each stage practical rationality required her to avoid the belief that she should not now be guided by her prior intentions.

The same is true of the intention S^1 formed, to follow RL prohibiting marks in books. The revelatory idea S^1 wants to record on her book probably can be traced through a long series of intentions. It is possible that S^1 could have improved on some of these, even if all fell within the margin allowed by S^1's reasonable disposition to retain prior intentions. Similarly, in the case of P's promise to help N, P's current circumstances, or P's initial promise to N, may have resulted from a series of imperfect non-deliberative choices about what plans to form and what commitments to make.

In each of these cases, temporally extended practical rationality allows the agent to plan and to make commitments. The agent's plans, however, as well as her current circumstances,

rest on what may be a long chain of intentions. At each link in the chain, practical rationality requires the agent to be epistemically irrational in ways that may undermine her practical rationality over time.

This aspect of temporally extended practical rationality puts substantial pressure on the notion of reasonable dispositions to retain prior intentions. The only limit on the extent of damage that epistemic irrationality may cause to practical success is the requirement that the agent's dispositions must prove advantageous over the long run. Yet, when the agent acts repeatedly on incomplete beliefs, long term advantage is difficult to verify. The agent may achieve workable outcomes, but this does not show that she could not have done better on at least some occasions by attending to all relevant evidence. Without a requirement of epistemic rationality at all stages of deliberation, there is no solid basis for the assumption that temporally extended practical rationality through action on prior intentions is in fact practically rational over time.[49]

C. Conclusion: Differing Perspectives

I have described a dilemma that does not seem escapable. Human reasoners are not omniscient and are not perfect reasoners. As a result, even if they are motivated to pursue a set of widely shared and morally sound ends, they cannot do this effectively unless they formulate plans, follow rules, and make commitments in a way that will control their future action. If instead they judge what to do in each case as it arises, they will make mistakes and will be unable to coordinate their actions internally or with the actions of others. Thus, it is practically rational over the long run for an agent to form intentions in advance and then act on them without reconsidering in every case. In other words, it is practically rational to avoid responding

in every case to evidence about current reasons for action. In some number of cases, however, the intended action will not in fact be the action best calculated to advance the agent's ends. An agent who fails to respond to evidence indicating that her prior intention will not advance her ends in current circumstances, and consequently that the set of beliefs that supports her prior intention should be adjusted to permit an exception, is not epistemically rational.

What makes this dilemma intractable is that agents who make plans, adopt rules, and undertake commitments confront their plans from two different perspectives. They form intentions from a general perspective, asking what actions will serve their ends over a range of possible circumstances that cannot be fully anticipated at the time of deliberation. They then act on their intentions from a particular perspective in which they have more and better evidence about how the intended action will work now. Given human error, it may continue to be true that always acting on the original intention will yield better results on average than always rethinking the question of what to do. Yet, the agent will sometimes be correct in her current judgment that she should not now act on her intention. From this perspective, acting against current judgment, or adopting an attitude toward evidence that suppresses current judgment, looks like a failure of rationality.

Bratman's approach to practical rationality allows agents to capture the practical benefits of planning, rule-following, and commitment. Agents can assess long-term reasons for action from the general, forward looking perspective and then act on them unreflectively at a later time. In this way, Bratman bypasses the particular perspective for the purpose of assessing agent rationality, through a process that is cognitively plausible.

Epistemic rationality, however, is synchronic, assessing the rationality of agents in forming and holding current beliefs about current evidence. If the agent forms no new beliefs at

the time of action, then the general perspective on action continues to govern. But if, as I have argued, the agent has an epistemic responsibility to process evidence that supports an adjustment in existing beliefs about reasons for action, epistemic rationality requires agents to assume the particular perspective.

Neither the general perspective not the particular perspective is mistaken. The general perspective can yield practical benefits, while the particular perspective has at least the potential to fix correctable errors in beliefs. In practice, we often give prior intentions priority over synchronically rational beliefs about reasons for action; ultimately, however, there is no reliable way to test the practical rationality of the dispositions that guide us to do this. As a result, temporally extended practical rationality, however necessary it may be to practical success, is an imperfect form of rationality in which epistemic rationality is permitted to lapse and practical rationality is left to rest on a questionable foundation.

REFERENCES

*References are to works listed in the bibliography (below).

[1] Bratman does not examine this possibility. David Gauthier, in his discussion of the toxin puzzle, takes passing notice of a potential for divergence, if not conflict, between practically rational intentions and epistemically rational beliefs. Gauthier (1998a). Gauthier argues that an agent who is offered a large sum of money if he can form an intention tonight to drink a vial of toxin tomorrow can form that intention and win the money if he reasonably expects tonight, and still believes tomorrow, that a course of action comprising formation of an intention tonight to drink toxin tomorrow and action tomorrow on the that intention will serve his interests. Intentions, in Gauthier's view, create reasons to act, as long as expectations hold steady.

Gauthier adds, however, that if the money is offered for belief rather than intention, the agent cannot believe tonight that he will drink the toxin tomorrow. The difference is that rational belief depends on the truth of the belief, not on the practical benefits of the belief. The agent's reason to drink tomorrow derives from his advantageous intention to drink, formed in the expectation that he will have reason to drink. But if the payoff is for the belief, not the intention, the agent has is no reason to form the intention. It follows that the agent has no reason to drink toxin tomorrow and therefore no reason to believe now that he will do so. Gauthier draws no general conclusions from this observation, except that beliefs and actions depend on different

types of reasons.

[2] Gauthier (1998b), pp. 44, 48-49.

[3] McClennen (1990), pp. 12, 120-22; McClennen (1997), p. 232; (McClennen & Shapiro (1998), p. 367.

[4] Shapiro (1998a), pp. 39-40, 47

[5] See Raz (2011), pp. 131-137; Broom (2002). Possibly McClennen could be read as addressing this problem. McClennen requires as a condition of extended practical rationality that acting on a prior intention must serve the collective interest of all the agent's selves over time and that benefits must be fairly distributed among selves. It might follow that if the agent, as now constituted, is viewed as embodying all her past and present selves, she could will action on behalf of the group. Will, however, is most plausibly understood as the will of the agent's current self alone.

[6] Upon reconsideration, reasons for action are again current reasons for action. See section B(1)(a), infra.

[7] S may or may not have knowledge of all these matters; those partial to Williamson's view of evidence can assume that she does without significantly changing the argument.

[8] Alternatively, she may form the intention to follow the rule unless she believes when the time comes to apply it that her reasons for violating the rule exceed some threshold that is greater than equipoise. In some cases, an exception of this kind is built into the rule itself; in other cases, the rule may be absolute but the agent's intention follow it is subject to an exception for unusual cases. Yet, to serve their instrumental purposes, rules must place some degree of constraint on the agent's future choice of action. In other words, the rule will not be effective unless a gap remains between the set of cases in which S must follow the rule and the set of cases in which S's current reasons for action at T_2 favor following the rule. See Alexander (1991).

[9] Assume that 'R' is a general rule; a 'Case' is any token governed by R; 'C' is the particular case now before S; and '>' means 'will better advance S's goals than.' S might reason as follows:

At T_1:
(1) S believes that '(∀x (Case (x) → S follows R in x) > (∃x (Case (x) → ~S follows R in x)'.
(2) S believes that '∃x (Case (x) & (~S follows R in x > S follows R in x)'.
(3) Therefore, S does NOT believe that ('∀x (Case(x) → (S follows R in x > ~S follows R in x))'.
At T_2:
(4) S believes (1)-(3).
(5) S believes that in C, '~S follows R > S follows R'.
(6) (5) does not contradict (4).
(7) (5) conforms to S's current evidence.
(8) Therefore, S is epistemically justified in believing that she should not follow R in C.

[10] This stipulation is designed to avoid questions about the moral authority of formally enacted

law in a just society.

[11] This is example is patterned on a similar example in Alexander and Sherwin (1991), pp. 59-60, 65-68.

[12] An example is the venerable Social Law Library in Boston. See http://socialaw.com/about.

[13] The nature of promissory obligation is discussed briefly in Chapter 1, Section A(2).

[14] This is the approach suggested by Raz and discussed in Chapter 1. See Raz (1977), pp. 221-223.

[15] It is possible that, even if epistemic rationality imposes no responsibility on agents to respond to their current evidence, an agent will simply do so, despite her prior intention to follow a rule or honor a promise. The agent may happen to notice new evidence and spontaneously form a practically inconvenient belief that she should not act on her intention. For reasons to be explained shortly, the effect of this new belief is to dispel the volitional commitment associated with her intention.

This is a problem for Bratman's theory of temporally extended practical rationality. Unconstrained, human agents miscalculate, they are short-sighted, and they lack internal resources to coordinate their actions with the actions of others or to manage their own procrastination. Thus, a rule or commitment that fails to curb assessment of current evidence in a significant way is not an effective tool for planning. At best, it will apply when all else is equal or when little is at stake. The agent can plan to watch the news if there is nothing better to do, or to get out of bed when the alarm clock rings.

Presumably, Bratman would answer that in the type of case just described, the agent is not practically rational because she has failed to develop and follow reasonable dispositions toward prior intentions. At least, the agent is not fully rational. Accordingly, there is no conflict between practical rationality and epistemic rationality because the conditions of practicaly rationality are not met. Bratman also suggests that in many common situations, agents are in fact rationally disposed to act on prior intentions without reflection, and thus without forming current beliefs about reasons for action. On the assumption that this suggestion is correct, I will examine the question of what epistemic responsibility an agent may have to advert to evidence and form beliefs even when it might be practically rational, in an extended sense, for her to act unreflectively.

[16] See, e.g., Greco (2014); Horowitz (2014).

[17] See notes 202-204, *supra*.

[18] Bratman (1987), pp. 15-18. Bratman also stipulates that the intentions of a rational agent must be internally consistent. From this it follows that an agent's prior intentions also function to limit the options the agent can consider in deliberation about further intentions. Id., pp. 9, 15-16. If I have formed a plan to take classes during an upcoming leave from my job, I must either avoid a plan to travel the world during my leave or reconsider the current plan.

[19] Id., p. 5.

[20] E.g., id., pp. 62-6, 71-72. Bratman suggests two and only two ways in which an agent might advert to the possibility of reconsideration without actually reconsidering her intention. First, she might take notice of additional reasons to act on the intention: in this case the intention is not reopened but merely reaffirmed. Second, the agent might consider the possibility of reconsideration, but decline to reconsider because reconsideration would be too costly. In each case, the grounds for the intention are never reopened. These exceptions suggest that in all other cases, reflection on evidence indicating that the agent should abandon her intention amounts to reconsideration of the intention.

[21] Id., p.16, 72-75. Early on, Bratman mentions "Buridan cases" in which intentions can solve the problem of what to do when reasons for action are in equipoise. Id., pp. 11-12. His overall theory, however, is designed to solve not only problems of equipoise but also problems of limited cognitive resources, internal coordination, and interpersonal coordination. These much broader ambitions suggest that in order to realize Bratman's goals, the force of intentions must extend beyond the case of equipoise to cases in which the agent's immediate reasons for action favor not acting on her intention but do not exceed some appropriate threshold.

[22] Id., p. 72.

[23] Id., p. 67.

[24] Id., pp. 56-57, 87-91. All intentions to act in the future are general in the sense that the circumstances in place when the agent fulfills the intention are not fully specified when the agent form the intention. Bratman's "personal policies" are also general in the sense that the agent intends to fulfill them repeatedly over time, under varying circumstances.

[25] Bratman's concern in this more recent work is when and how personal policies about what reasons to recognize and how much weight to give them count as authored by, and subjectively normative for, the agent. See Bratman (2007), pp. 6, 22-40. Bratman's answers to these questions are interesting and mostly persuasive, but do not bear directly on the problems I address here. I have no quarrel with the proposition that it is long-term practically rational, especially in matters of self-governance in Bratman's sense, to prefer continuity of plans over time-slice reasons for action. My problem is that, owing to the generality of the plan and the cognitive limitations of the agent, the long-term advantage of the plan will sometimes be overridden by the current time-slice advantage. When this occurs, adhering ot the pla may be long-term practically rational but may not be epistemically rational.

[26] Significantly" is, of course, a vague term. May be it connotes a 3 or more our of 10 on a scale of importance or badness. More likely, however, agents operate in vague terms, particularly when guided by dispositions rather than reflection.

[27] As noted earlier, one difficulty with Bratman's theory of retained intentions is that it is not clear how the agent can judge that a threshold such as significance is or is not met without reflecting on her current evidence. See text at note 131, *supra*.

[28] It might seem that if both practical rationality and epistemic rationality allow for reasonable rather than optimal compliance with standards of decisionmaking and belief formation, any conflict between them can be resolved by interpreting each to tolerate some reasonable level of

deviation in order to accommodate the other. Reasonable standards of epistemic rationality would allow agents to form beliefs that do not perfectly correspond to their evidence if this will lead to practical benefits. Reasonable standards of practical rationality would allow agents to adjust their dispositions toward prior intentions to achieve a closer fit between evidence and belief, at the expense of some practical gains.

The standards of reasonableness built into practical rationality and epistemic rationality, however, are specific to the type of rationality at issue. Thus, in the context of epistemic rationality, reasonableness means a reasonable fit between evidence and belief, given epistemic obstacles such as incomplete evidence and limitations on inferential reasoning. In the context of temporally extended practical rationality, reasonableness means a reasonable probability of long-run benefit, given limited time for deliberation and the need to coordinate by planning in advance for actions under circumstances that cannot currently be specified in full. These different forms of reasonableness are not commensurable in a way that allows for sensible trade-offs.

[29] Raz (1987), pp. 57-62; Raz (1979), pp. 16-19, 22-23, 30-33.

[30] Raz (1977), pp. 221-223. Although Raz does not elaborate on how to identify the range of excluded reasons, the range appears to be broad, corresponding to the political legitimacy of the rule-making authority and the moral value, and perhaps other values, associated with the promise.

[31] Bratman (1992).

[32] In his later work on self-governance, Bratman suggests a somewhat different type of deliberative assumption that come into play in the context of self-governing policies. Self-governing policies are second-order general intentions about the types of reasons the agent chooses to recognize as justifying reasons in her own deliberation about action. For this purpose, Bratman indicates that an agent can value a type of reason without being able to defend the reason's value in intersubjective terms. See Bratman (2007), pp. 209-210, 212.

[33] In fact, Bratman makes an argument very similar to this to show that his approach to practical rationality does not suffer from the classic dilemma of rule-utilitarianism. "[Rule-utilitarianism] sanctions utilitarian reasoning concerning rules but does not concerning particular acts. But given its commitment to the former it may seem unclear how it can block such reasoning in the latter case, the case in which we are assessing particular acts." Bratman sees no similar problem in his approach to practical rationality "because this is only an account of the rationality of an agent for (non)reconsideration [of intentions] that is *not* based on present deliberation. . . . In the sort of case the present account [addresses] there is no need to block direct consequential reasoning by the agent concerning his particular case of (non)reconsideration; for in the case in question there is no deliberation at all about whether to reconsider." Bratman (1987), pp. 209-210, 212.

[34] This example, from Adam Leite, is discussing in Chapter IV, at notes 190-94.

[35] Richard Feldman rejects instrumentalism, Alvin Goldman embraces it; as described below, both favor a standard that imposes responsibility on agents to process evidence that defeats current beliefs.

[36] Conee & Feldman (2004), pp. 186-188.

[37] Goldman (2011), p. 23.

[38] It does not help to recast the argument by claiming that S's new evidence defeats the beliefs that support her specific intention to follow RT now. According to Bratman, specific intentions to act on a general intention or policy are not formed deliberatively; instead, they arise spontaneously from the general intention without the intervention of any new reasoning or new beliefs. The agent's general intention generates a specific intention to act in each particular case, unless her current circumstances exceed the threshold fixed by her reasonable disposition to apply general intentions to specific cases. As a result, the agent's specific intention to act on her general intention is not supported by any beliefs other than the same prior beliefs that support her continuing general intention. As explained in the text, these beliefs are not defeated; therefore S has no responsibility to process her new evidence.

[39] Another possible source of epistemic irrationality in the cognitive process Bratman associates with temporally extended practical rationality is that agents must glimpse, or "peek" at, evidence in order to follow their reasonable dispositions toward prior intentions. See Schauer (1991), p. 677 (using the term "peek" to describe the momentary attention a presumptive rule-follower gives to current reasons for action). In some circumstances, the agent's habits and dispositions may not require any engagement with evidence. For example, the agent may simply conclude that not much is at stake and consideration of evidence is not worth the trouble. Or she may be disposed to follow a predetermined schedule without exception. See Bratman (1987), p. 88 (giving the example of regular insurance policy review). Often, however, the agent's reasonable dispositions toward prior intentions will involve a threshold of contrary reasons, as in the texting example. If S is disposed to follow RT unless there are significant reasons not to follow the rule, she must at least to glance at evidence about reasons to determine whether they are significant. Epistemic responsibility comes into play because once S has peeked at her evidence, she may have an epistemic responsibility to finish the thought by considering what the evidence recommends. I set this possibility aside because the burden it imposes on agents is not easily containable.

[40] I wait for the light to change at deserted intersections. I follow an exercise routine even when I have more important things to do. I keep (some) promises that the recipient probably does not care about and may even have forgotten. The practices of advance planning and interpersonal commitment in which these behaviors are imbedded are well-justified on practical grounds, but the practical advantages that support them may not always be present in particular cases. If I looked more closely at current evidence about reasons for action, I could see this easily; but I do not look more closely.

[41] See Foley (1987), pp. 212-225; see also Kelly (2003), pp. 618-620.

[42] Foley (1987), p. 213.

[43] See Conee & Feldman (2004), pp. 193-194.

[44] See Alexander (1998); Chang (1998).

[45] Reisner (2008).

[46] Foley's approach also rests on the debated assumption that epistemic rationality is instrumental. See Adler (2002); Kelly (2003).

[47] Foley makes a different point, that forming a false belief will not often be practically beneficial because the cognitive maneuvering needed to acquire a false but practically convenient belief is likely to affect a much wider range of beliefs in unintended ways. Foley (1987), pp. 222-225.

[48] Bratman (1987), pp. 15-18.

[49] I do not mean to claim that, by application of a conjunction principle holding that the likelihood of error in a set of conjoined cases is the product of the likelihoods of error in each case, below-threshold errors in a chain of beliefs lead to a massive error. See Clermont (2015). My argument is simply that it is impossible to verify the supposed long-term practical benefits of a disposition to act without full deliberation about current evidence and reasons for action.

BIBLIOGRAPHY

Adler, Jonathan E.
(2002) BELIEF'S OWN ETHICS (Cambridge, Massachusetts: M.I.T. Press).

Larry Alexander
(1991) *The Gap*, HARV. J. L. & PUBL. POL'Y 14:3, pp. 695-703.
(1998) *Banishing the Bogey of Incommensurability*, U. PENN. L. REV. 146: 1641-1649.
(2001) with Emily Sherwin, THE RULE OF RULES: MORALITY, RULES, AND THE DILEMMAS OF LAW (Durham: Duke University Press).

J. L. Austin.
(1962) Sense and Sensibilia (Oxford: Oxford University Press).

Lawrence BonJour.
(1978) *Can Empirical Knowledge Have a Foundation*?, AM. PHIL. Q. 15:1, pp. 1-15.

Bratman, Michael E.
(2014) SHARED AGENCY: A PLANNING THEORY OF ACTING TOGETHER (New York: Oxford University Press).
(2012) *Time, Rationality, and Self-governance*, PHIL. ISSUES 22, pp. 73-87.
(2009) *Intention Rationality*, PHIL. EXPLORATIONS 12:3, pp. 227-241.
(2009) *Intention, Belief, and Instrumental Rationality*, in REASONS FOR ACTION (David Sobel & Steven Wall, eds., Cambridge: Cambridge University Press), pp.13-36.
(2007) STRUCTURES OF AGENCY: ESSAYS (New York, Oxford University Press).
(1998) *Toxin, Temptation, and the Stability of Intention*, in RATIONAL COMMITMENT AND SOCIAL JUSTICE: ESSAYS FOR GREGORY KAVKA (Jules L. Coleman and Christopher W. Morris, eds.)(Cambridge: Cambridge University Press), pp. 59-83.
(1992) *Practical Reasoning and Acceptance in a Context*, MIND, 101:401, pp. 1-13.
(1987) INTENTIONS, PLANS, AND PRACTICAL REASON (Cambridge, Massachusetts: Harvard University Press).

Broome, John.
(2002) *Practical Reasoning*, in REASON AND NATURE: ESSAYS IN THE THEORY OF RATIONALITY (Jose Luis Bermudez & Alan Millar, eds.) (Oxford: Clarendon Press).
(2001) *Are Intentions Reasons*?, in PRACTICAL RATIONALITY AND PREFERENCE: ESSAYS FOR DAVID GAUTHIER (Christopher W. Morris & Arthur Ripstein, eds.) (Cambridge: Cambridge U. Press 2001), pp. 98-120.

Chang, Ruth.
(1998) *Comparison and the Justification of Choice*, U. PENN. L. REV. 146: 1569-1598.

Chignell, Andrew.

(2010) *The Ethics of Belief*, in STANFORD ENCYLOPEDIA OF PHILOSOPHY, http://plato.stanford.edu/entries/ethics-belief/

Christensen, David.
(1994) *Conservatism In Epistemology*, in NOUS 28:1, pp. 69-89.
(2000) *Diachronic Coherence versus Epistemic Impartiality*, in PHIL. REV. 109, pp. 349-371.

Clermont, Kevin M. (2015) *Conjunction of Evidence and Multivalent Logic, in* LAW AND THE NEW LOGICS (Lionel Smith ed., forthcoming 2016), *available at* http://ssrn.com/abstract=2472383

Clifford, W.K.
(1886) *The Ethics of Belief*, in CONTEMPORARY REVIEW 29, pp. 289-309.

Conee, Earl, and Richard Feldman.
(2004) EVIDENTIALISM: ESSAYS IN EPISTEMOLOGY (Oxford: Oxford University Press).
(2011) *Replies*, in EVIDENTIALISM AND ITS DISCONTENTS (Trent Dougherty, ed., Oxford: Oxford University Press) pp. 283-323.

Dancy, Jonathan.
(forthcoming 2014) *From Thought to Action*.
(2004) ETHICS WITHOUT PRINCIPLES (Oxford: Clarendon Press).
(1997) *Parfit and Indirectly Self-defeating Theories*, in READING PARFIT (Jonathan Dancy, ed.) (Oxford: Blackwell).

David, Marian.
(2001) *Truth as the Epistemic Goal*, in KNOWLEDGE, TRUTH, AND DUTY: ESSAYS ON EPISTEMIC JUSTIFICATION, RESPONSIBILITY, AND VIRTUE (Matthias Steup, ed.) (Oxford: Oxford University Press).

Dougherty, Trent.
(2011) EVIDENTIALISM AND ITS DISCONTENTS (Oxford: Oxford University Press).

Foley, Richard.
(1987) THE THEORY OF EPISTEMIC RATIONALITY (Cambridge: Harvard University Press).
(2001) *The Foundational Role of Epistemology in a General Theory of Rationality*, in VIRTUE EPISTEMOLOGY: ESSAYS ON EPISTEMIC VIRTUE AND RESPONSIBILITY (Abrol Fairweather and Linda Zagsebski, eds.)(New York: Oxford University Press).

Fried, Charles
(1981) CONTRACT AS PROMISE: A THEORY OF CONTRACTUAL OBLIGATION (Cambridge, Massachusetts: Harvard University Press).

(2012) *The Ambitions of* Contract as Promise, in PHILOSOPHICAL FOUNDATIONS OF CONTRACT LAW (Gregory Klass, George Letsas, & Prince Saprai, eds.)(Oxford: Oxford University Press), pp. 17-41.

Fumerton, Richard.
(1990) REASON AND MORALITY: A DEFENSE OF THE EGOCENTRIC PERSPECTIVE (Ithaca: Cornell University Press).

Gauthier, David.
(1986) MORALS BY AGREEMENT (Oxford: Clarendon Press).
(1994) *Assure and Threaten*, ETHICS 104:4, pp. 690-721.
(1997) *Rationality and the Rational Aim*, in READING PARFIT (Jonathan Dancy, ed.)(Oxford: Blackwell), pp. 24-41.
(1998a) *Rethinking the Toxin Puzzle*, in Jules L. Coleman and Christopher W. Morris, eds., RATIONAL COMMITMENT AND SOCIAL JUSTICE: ESSAYS FOR GREGORY KAVKA (Cambridge: Cambridge University Press), pp. 47–58.
(1998b) *Intention and Deliberation*, in MODELING RATIONALITY, MORALITY, AND EVOLUTION (P. Danielson, ed.)(Oxford: Oxford University Press), pp. 41-64.

Gold, Andrew S.
(2009) *A Property Theory of Contract*, 10 NW. U. L. REV. 1.

Goldman, Alvin.
(1979) *What is Justified Belief*?, in JUSTIFICATION AND KNOWLEDGE 89-104(George Pappas, ed.) (Dordrecht: D. Reidel Publishing Co. 1979).
(2011a) *Toward a Synthesis of Reliabilism and Evidentialism*, in EVIDENTIALISM AND ITS DISCONTENTS (Trent Doherty, ed.) (Oxford, Oxford University Press 2011), pp. 254-280.
(2011b) *Reliabilism*, STANFORD ENCYCLOPEDIA OF PHILOSOPHY http://plato.stanford.edu/archives/spr2011/entried/reliabilism.

Greco, Daniel.
(2014) *A Puzzle About Epistemic Akrasia*, in PHIL. STUD. 167, pp. 201-219.

Greco, John.
(2011) (with John Turri) *Virtue Epistemology*, in STANFORD ENCYCLOPEDIA OF EPISTEMOLOGY, http://plato.stanford.edu/archives/win2011/entries/epistemology-virtue/.

Hall, Richard J. & Charles R. Johnson.
(1998) *The Epistemic Duty to Seek More Evidence*, AM. PHIL. Q. 35:2, pp. 129-139.

Hart, H.L.A.
(1982) ESSAYS ON BENTHAM (Oxford: Oxford University Press).
Hedden, Brian.
(2014) *Time-Slice Rationality*, forthcoming in MIND.

HEURISTICS & BIASES: THE PSYCHOLOGY OF INTUITIVE JUDGMENT.
(2002) Thomas Gilovich, Dale Griffin, & Daniel Kahneman, eds. (Cambridge and New York: Cambridge University Press).

Horowitz, Sophie.
(2014) *Epistemic Akrasia*, forthcoming in NOUS.

Hume, David
(1978) A TREATISE OF HUMAN NATURE (2d ed., L.A. Selby and P.H. Nidditch, eds.) (Oxford: Oxford University Press).

Hurd, Heidi M.
(1999) MORAL COMBAT (Cambridge: Cambridge University Press).

James, William.
(1897) *The Will to Believe*, Ebrary Religion, Philosophy & Classics Subscription Collection, http://site.ebrary.com/lib/cornell/docDetail.action?docID=5000707

JUDGMENT UNDER UNCERTAINTY: HEURISTICS AND BIASES.
(1982) Daniel Kahneman, Paul Slovic, & Amos Tversky, eds.(Cambridge and New York: Cambridge University Press).

Kant, Immanuel
(1991) THE METAPHYSICS OF MORALS (Mary J. Gregor, trans.)(Cambridge: Cambridge U. Press).

Kavka, Gregory.
(1983) *The Toxin Puzzle*, in ANALYSIS 43:1, pp. 33-36.

Kelly, Thomas.
(2002) *The Rationality of Belief and Some Other Propositional Attitudes,* PHIL. STUD., 110:2, pp. 163-196.
(2003) *Epistemic Rationality as Instrumental Rationality: A Critique*, PHIL. & PHENOMENOLOGICAL RESEARCH: 66: 3 (2003), pp. 612-640.
(2004) *Sunk Costs, Rationality, and Acting for the Sake of the Past*, in NOUS 38:1, pp. 60-85.
(2007) *Evidence and Normativity: Reply to Leite*, in PHIL. & PHENOMENOLOGICAL RESEARCH 75:2, pp. 465-474.
(2008) *Evidence: Fundamental Concepts and the Phenomenal Conception*, PHILOSOPHY COMPASS 3:5, pp. 933-955.

(2014) *Evidence*, in THE STANFORD ENCYCLOPEDIA OF PHILOSOPHY http://plato.standord.edu/archives/fall2014/entries/evidence/.

Kornblith, Hilary
(1983) *Justified Belief and Epistemically Responsible Action*, in PHIL. REV. 92:1, pp. 33-48.

Kortz, Keith Allen
(2010) *The Epistemic Basing Relation*, in STANFORD ENCYCLOPEDIA OF EPISTEMOLOGY, http://plato.stanford.edu/archives/spring2010/entries/basing-epistemic/.

Leite, Adam.
(2007) *Epistemic Instrumentalism and Reasons for Belief: A Reply to Tom Kelly's* Epistemic Rationality as Instrumental Rationality: A Critique, in PHIL. & PHENOMENOLOGICAL RESEARCH 75:2, pp. 456-464.

McClennen, Edward F.
(2001) *The Strategy of Cooperation*, in PRACTICAL RATIONALITY AND PREFERENCE: ESSAYS FOR DAVID GAUTHIER (Christopher W. Morris & Arthur Ripstein, eds., Cambridge: Cambridge University Press), pp. 189-208.
(1998) with Scott J. Shapiro, *Rule-Guided Behavior*, in III NEW PALGRAVE DICTIONARY OF ECONOMICS AND THE LAW 363 (Peter Newman, ed.; New York: Stockton Press).
(1997) *Pragmatic Rationality and Rules*, in PHIL. & PUBL. AFFAIRS 26:3, pp. 210-258.
(1990) RATIONALITY AND DYNAMIC CHOICE (Cambridge: Cambridge University Press).

Mills, Eugene.
(1998) *The Unity of Justification*, PHIL. & PHENOMENOLOGICAL RESEARCH: 58:1, pp. 27-50.

Moore, G.E.
(1942) A Reply to My Critics, in The Philosophy of G.E. Moore (P.A Schlipp, ed.), pp. 500-503.

Owens, David
(2012) *Does a Promise Transfer a Right?*, in PHILOSOPHICAL FOUNDATIONS OF CONTRACT LAW (Gregory Klass, George Letsas, & Prince Saprai, eds.)(Oxford: Oxford University Press), pp. 78-95.

Pappas, George.

(2013) *Internalist v. Externalist Conceptions of Epistemic Justification*, in STANFORD ENCYCLOPEDIA OF EPISTEMOLOGY, http://plato.stanford.edu/archives/fall2013/entries/justep-intext/.

Parfit, Derek.
(1984) REASONS AND PERSONS (Oxford: Clarendon Press).
(2001) *Bombs and Coconuts, or Rational Irrationality*, in PRACTICAL RATIONALITY AND PREFERENCE: ESSAYS FOR DAVID GAUTHIER (Christopher W. Morris & Arthur Ripstein, eds., Cambridge: Cambridge University Press), pp. 81-97.

Plantinga, Alvin.
(1993) WARRANT AND PROPER FUNCTION (Oxford: Oxford University Press), reprinted in Oxford Scholarship Online (2003) DOI: 10.1093/0195078640.001.0001.

Plous, Scott.
(1993) THE PSYCHOLOGY OF JUDGMENT AND DECISION MAKING (Philadelphia: Temple University Press).

Rawls, John
(1955) *Two Concepts of Rules*, PHIL REV. 64:1, pp. 3-32.

Raz, Joseph.
(2014) *Is There a Reason to Keep a Promise*?,in PHILOSOPHICAL FOUNDATIONS OF CONTRACT LAW (Gregory Klass, George Letsas, & Prince Saprai, eds., Oxford: Clarendon Press (2014), pp. 58-77).
(2011) FROM NORMATIVITY TO RESPONSIBILTY (Oxford: Oxford University Press)
(1986) THE MORALITY OF FREEDOM (Oxford: Oxford University Press).
(1979) THE AUTHORITY OF LAW (Oxford: Oxford University Press).
(1977) *Promises and Obligation*, in LAW, MORALITY, AND SOCIETY (P.M.S Hacker & J. Raz, eds.)(Oxford: Clarendon Press), pp. 210-228.

Riesner, Andrew.
(2009) *The Possibility of Pragmatic Reasons for Belief and the Wrong Kind of Reasons Problem*, PHIL. STUD. 145:2, pp. 257-272.
(2008) *Weighing Pragmatic and Evidential Reasons for Belief*, PHIL. STUD. 138:1, pp. 17-27.

Scanlon, T.M.
(1998) WHAT WE OWE TO EACH OTHER (Cambridge, Massachusetts: Harvard University Press).

Schauer, Frederick. (1991a) PLAYING BY THE RULES: A PHILOSOPHICAL EXAMINATION OF RULE-BASED DECISION-MAKING IN LIFE AND LAW (Oxford: Clarendon Press).
(1991b) *Rules and the Rule of Law*, 14 HARV. J. L. & PUB. POL'Y 645.

Shah, Nishi.
(2006) *A New Argument for Evidentialism*, in PHIL. Q. 56:225, pp. 481-498.
(2003) *How Truth Governs Belief*, in PHIL. REV. 112:4, pp. 447-182.

Shapiro, Scott J.
(2011) LEGALITY (Cambridge, Massachusetts: Harvard University Press).
(1998a) *The Difference That Rules Make*, in ANALYZING LAW: NEW ESSAYS IN LEGAL THEORY 33 (Brian Bix, ed.; Oxford: Clarendon Press).
(1998b) with Edward F. McClennen, *Rule-Guided Behavior*, in III NEW PALGRAVE DICTIONARY OF ECONOMICS AND THE LAW 363 (Peter Newman, ed.; New York: Stockton Press).

Shiffrin, Seana Valentine
(2008) *Promising, Intimate Relationships, and Conventionalism*, PHIL. REV. 114:4, pp. 481-524.

Silins, Nicholas.
(2005) *Deception and Evidence*, PHIL. PERSPECTIVES 19, pp. 375-404.

Smith, Stephen A.
(2000) *Towards a Theory of Contract*, in OXFORD ESSAYS IN JURISPRUDENCE (J. Horder, ed.)(Oxford: Oxford University Press).

Steup, Matthias, ed.
(2001) KNOWLEDGE, TRUTH, AND DUTY: ESSAYS ON EPISTEMIC JUSTIFICATION, RESPONSIBILITY, AND VIRTUE (Oxford: Oxford University Press).

Velleman, J. David.
(2000) THE POSSIBILITY OF PRACTICAL REASON (Oxford: Clarendon Press).

Van Imwagen, Peter.
(1996) *It is Wrong, Everywhere, Always, and for Anyone, To Believe Anything Upon Insufficient Evidence*, in FAITH, FREEDOM, AND RATIONALITY, (J. Jordan & D. Howard-Snyder, eds., Lanham, Md.: Rowman & Littlefield), pp. 137-153.

Williamson, Timothy.
(2000) KNOWLEDGE AND ITS LIMITS (Oxford: Oxford University Press).

Wood, Allen W.

(2002) UNSETTLING OBLIGATIONS: ESSAYS ON REASON, REALITY, AND THE ETHICS OF BELIEF (Stanford: CSLI Publications).