RATIONALITY AND SUCCESS

BY PRESTON GREENE

A dissertation submitted to the

Graduate School—New Brunswick

Rutgers, The State University of New Jersey

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

Graduate Program in Philosophy

Written under the direction of

Andy Egan

and approved by

_____

_____

_____

_____

New Brunswick, New Jersey

October, 2013

ABSTRACT OF THE DISSERTATION

Rationality and Success

by Preston Greene

Dissertation Director: Andy Egan

Standard theories of rational decision making and rational preference embrace the idea that there is something special about the present. Standard decision theory, for example, demands that agents privilege the perspective of the present (i.e., the time of decision) in evaluating what to do. When forming preferences, most philosophers believe that a similar focus on the present is justified, at least in the sense that rationality requires or permits future experiences to be given more weight than past ones. In this dissertation, I examine such theories in light of the expected success of the agents who follow them. In Chapters 2 and 3, I show that this bias toward the present is a *liability*: it tends to make agents less successful than they might otherwise be. I also show how these problems can be avoided: In the case of rational decision making, we must privilege *the beginning* rather than the present (what I call "inceptive maximization"). In the case of rational preferences, we must be *completely temporally neutral*.

In chapters 4 and 5 I introduce a larger framework in which to interpret these results. My core thesis is that practical rationality is a form of *conditional reliability*. Practically rational decisions, preferences, intentions, or other relevant factors reliably produce whatever we take to be of value, *conditional on an agent's beliefs.* This focus on *value-conduciveness*

is thus the analog of the focus on *truth-conduciveness* in reliabilist theories of epistemic norms. Like reliabilism in epistemology, I show that practical reliabilism is supported by a methodologically naturalistic approach to normativity. In this way and others, I argue that epistemic and practical reliabilism interconnect to create an overarching theory of normativity.

Acknowledgements

I owe much to my committee members—Ruth Chang, Holly Smith, and Brian Weatherson—for their indispensable suggestions on drafts. Christopher Meacham and Larry Temkin also provided very helpful comments and support. I further benefited from conversation with Nick Beckstead, John Broome, Tim Campbell, Heather Demarest, Alvin Goldman, Jim Joyce, Branden Fitelson, Ben Levinstein, Duncan MacIntosh, Peter Railton, Mary Salvaggio, and Robbie Williams, as well as audiences at Rutgers, the Unversity of Iowa, and York University.

Without doubt my greatest debt is to Andy Egan, who I was lucky enough to have as my dissertation director. Throughout my last three years at Rutgers, Andy was overwhelmingly generous with his time, and he was relentlessly encouraging yet incisive in his comments.

Table of Contents

List of Figures

Chapter 1
Introduction

Most of us are "present biased" in the decisions we make and the preferences we form. This attitude is so natural that one may be unaccustomed to specifying the assumptions that underlie it. If we were to attempt such a specification, it might look something like the following:

> The present and near-present are important, the past is not. When making decisions, the present is all that matters: In deciding what to do one should privilege how things stand in the present—how things stood in the past doesn't really matter. When we form preferences, the present is just as important: present pleasurable experiences are best, and future pleasurable experiences are better the nearer they are to the present. Painful future experiences, on the other hand, are better the more distant they are from the present. Past experiences don't matter at all—they're over and done with.

The influence of this attitude is not limited to common practice, it also inspires philosophical theories. According to standard decision theory, rational agents are *present-directed*: they choose options that are best in the present (i.e., at the time of decision). Regarding preferences, many philosophers believe that *future bias*—preferring future pleasurable or past painful experiences to past pleasurable or future painful ones—is at least permissible if not rationally required. In each instance present bias can be thought to generate corresponding principles of rationality.

In the first two chapters of this dissertation, I argue that present bias is a *liability*: it makes agents less successful than they might otherwise be. The adoption of principles

generated by present bias thus produces theories of rationality that are a liability. In each instance I show that by rejecting present bias we can create theories of rationality that avoid this result. In the case of rational decision making, I develop a new theory—*inceptive maximization*—to replace the standard present-directed maximizing account of rationality. In the case of rational preferences, I argue that rationality requires *complete temporal neutrality*—the rejection of both near and future bias.

In Chapters 3 and 4 I develop a broader framework in which to interpret these conclusions. In formulating theories of rationality, my overarching thesis is that practical rationality is a form of *conditional reliability*. I argue that theories of practical rationality should emphasize the *value-conduciveness* of decisions, intentions, preferences, or other practically relevant factors, *conditional on an agent's beliefs.* This results in theories of rational decision, intention, and preference that are sometimes similar to orthodoxy, but (most notably in the rejection of principles generated by present bias) also sometimes radically different. This is to be expected, as standard approaches to building theories of practical rationality adopt a radically different methodology. They begin by attempting to determine axioms—or constraints on potential axioms—from which full theories can be inferred, and only subsequently consider the practical upshot of such theories. Present bias, when considered in isolation from its practical upshot, is extremely intuitive. For this reason non-reliabilist approaches treat principles generated by present bias as inviolable fixed points—constraints on reasonable axioms—and thus any practical problems that present bias may create are thought to be unavoidable. As a result standard approaches to theories of practical rationality lack a connection to the goals of practical reasoning: they are committed to the idea that we should expect rational agents to often be less successful than they might be if they were irrational.

Theories of rationality, as I understand them, attempt to provide answers to the normative questions that remain after the goals have been fixed. In line with this, I include little discussion of theories of value—understood as proposals for fixing the goals. My project is in fact compatible with many ways of understanding value, though certain aspects may

favor a consequentialist approach. This dissertation is therefore not so unlike any of the swelling number of self-help books written for popular audiences, which promise guaranteed techniques for accomplishing one's goals, whatever those goals may be. (This might appear to some to be a liability for a philosophical dissertation—let me reassure you by noting that it is written at a level of generality that promises less practical use than most self-help books.)

It may be true, however, that once we account for differences in belief accuracy, agents who follow the proposals offered here can expect to be more successful at accomplishing their goals (whether that involves an egoistic pursuit of personal wealth, or a utilitarian pursuit of global happiness) than they would otherwise be. This allows for an account of practical rationality that is undeniably real, rather than one that is the product of making refinements to our existing evaluative attitudes. We are, for example, able to make predictions about future observations using the concept of rationality that I offer: accounting for differences in belief accuracy, we can predict that rational agents will tend be more successful than irrational ones. We are also able to explain our observations using this concept of rationality: again accounting for differences in belief accuracy, we can often cite an individual's irrationality in explaining why she ended up worse off. This account of rationality is therefore unique in having an existence obviously independent of our judgments, and in that way obliterates non-cognitivist worries and stakes a claim to objectivity.

1.1 Chapter Summaries

The rest of the dissertation consists of four chapters. Each chapter is presented as a paper that can be read in isolation from the others.

Chapter 2: Non-Classical Theories of Rational Decision Making

Straightforward maximizers of utility or value are "present-directed" in always choosing the act with the greatest expected utility or expected value at the time of decision. Several

theorists have pointed out that straightforward maximization leads to seemingly avoidable disasters for "transparent" agents, and several theorists use this fact to motivate what I call "non-classical theories of rational decision making." In an effort to avoid the problems created by transparency, non-classical theories of rational decision making work within the maximizing framework but reject present-directedness. Two prominent examples are David Gauthier (1986)'s "constrained maximization" and Edward McClennen (1990)'s "resolute choice."

The motivation for non-classical decision theories from worries over transparency, however, is difficult to grasp because each theorist defines transparency differently, and perhaps more importantly, they have all defined it incorrectly. I argue that transparency must be understood as *susceptibility to reliable prediction*, and that the cases used to motivate the idea that straightforward maximization leads to disaster are *open Newcomb problems*—they are identical to a version of Newcomb's problem in which the million dollar box has been opened. Once this is made clear, I show that non-classical decision theories should endorse *inceptive maximization*, which differs from straightforward maximization by providing recommendations based on the plans that would maximize expected value *at the beginning*, rather than the acts that maximize expected value in the present. This finally completes the project started by Gauthier, and can be used as a template for a formal decision-theoretic description.

Chapter 3: Time-Biased Preferences

Most of us display a *bias toward the near*: we prefer pleasurable experiences to be in our near future and painful experiences to be in our distant future. We also display a *bias toward the future*: we prefer pleasurable experiences to be in our present or future and painful experiences to be in our past. Among economists, it is often thought that both of these biases are permissible as long as they do not produce preference patterns that are *dynamically inconsistent.* Among philosophers who discuss the issue, however, there is general agreement that near bias is a rational defect (if there are any non-structural rational

requirements on preferences). However, almost no one finds future bias objectionable. On the contrary, some philosophers take a theory to be refuted if it conflicts with the rationality of future bias. I argue that this position is untenable: the same considerations that are used against near bias also apply to future bias. First, as with near bias, future bias leads to imprudent planning. Second, the most plausible error theory for near bias also applies to future bias. I conclude that those who reject near bias should go a step further and endorse complete temporal neutrality.

Chapter 4: Practical Reliabilism

The at-the-beginning evaluation of plans required by inceptive maximization and the temporal neutrality that results from rejecting future bias have the benefit of making rationality a positive capacity rather than a liability, but they might still strike us as bizarre. Is it really rational for agents to consider how things stood in the past when making decisions? Are past pleasures really just as preferable as future pleasures? I argue that these considerations reveal a tension that is generated by two competing approaches to theory building. Straightforward maximization and future bias result from an *axiomatic* approach to theories of rationality, which takes *a priori* intuitive constraints on axioms to be the building blocks for normative theories. As an alternative, the *reliabilist* approach does not place any constraints on what a theory must look like other than that it reliably produce the best results.

I understand practical reliabilism in a general way: as the claim that we should focus on the *value-conduciveness* of practically relevant factors. The evaluation of practical norms, however, is conditional in nature. Instead of evaluating what an agent believes (as in epistemology), we evaluate what an agent does *given* what she believes. I propose a procedure for determining value-conduciveness appropriate for conditional reliability, and show how this procedure makes practical reliabilism immune to the problem of demon worlds.

Chapter 5: Normative Naturalism

Given that there are reasons to prefer both the reliabilist and axiomatic approaches, it might appear that we are left at an impasse. However, metanormative commitments may have a strong influence on which approach one takes to be correct. I show that the reliabilist approach to practical norms is more compatible with a *methodologically naturalistic* approach to normativity. Specifically, it allows us to *explain* and *predict* the world around us using the concept of rationality, while the axiomatic approach does not. When reliabilist theories diverge from axiomatic ones—as in open Newcomb problems—those following the axiomatic theories tend to experience negative "feedback" in the form harmful or otherwise undesirable features of their experiences. Those following the reliabilist theories, on the other hand, tend to experience positive feedback. This creates pressure in favor of reliabilist theories and allows us to make the following prediction: given enough environmental variation, and enough time, there will be a progression toward the adoption of reliabilist norms. By endorsing the reliabilist metatheory we are able to explain this as a progression toward greater *rationality*.

The reliabilist approach to practical norms has the further benefit of creating a direct interface with epistemic normativity. Since practical rationality produces principles that are only conditionally reliable, practically rational agents will tend to be successful to the degree that their belief-forming processes are *unconditionally* reliable. In this way practical and epistemic normativity are interrelated and mutual supportive.

Chapter 2
Non-Classical Theories of Rational Decision Making

Standard decision theory, as it applies both to theories of self-interested rationality and to consequentialist morality, requires agents to be *straightforward maximizers*. Straightforward maximizers are "maximizers" in that they accept a *maximizing conception of rationality*, according to which the best options are those that maximize expected utility or expected value. They maximize "straightforwardly" in that they are *present-directed*: they choose options that are best in the present (i.e., at the time of decision). Straightforward maximization is a common assumption for both intuitive and technical treatments of decision making; it strikes many as a paradigm of good sense, and it is a shared feature of causal and evidential decision theory.

In this chapter I focus on theories that accept straightforward maximization's commitment to a maximizing conception of rationality, but reject present-directedness. Since present-directedness enjoys wide appeal, I use the term "non-classical" to refer to these theories. I argue that previous attempts at formulating satisfactory non-classical decision theories have failed, and I then propose my own version.

A prominent historical motivation for non-classical decision theories starts with dissatisfaction over the practical upshot of straightforward maximization. The thought, in general form, is this: straightforward maximization makes perfect sense in most situations, but when agents are "transparent" straightforward maximization becomes a *liability*. Transparent straightforward maximizers are less likely to accomplish their goals (whatever those goals may be) than they would be by following some other theory.[1] Some take this to show

[1] Cf. McClennen, 1990, 118: "This is a brief for rationality as a positive capacity, not a liability—as it must be on the standard account."

the need for creating non-classical decision theories that work effectively for both situations of transparency and non-transparency. Given this, in order to properly understand the motivation for non-classical decision theories, we must first understand what transparency is, and how exactly it creates problems for straightforward maximizers.

2.1 Defining Transparency

Straightforward maximization plus a transparency condition is supposed to lead to problematic cases. In this section I reveal the three most prominent accounts of the transparency condition. We will see that all three suffer fatal problems and thus make the motivation for non-classical decision theories seem unclear. I then introduce a way to understand transparency that solves these problems, and I show how this understanding creates a clear motivation for non-classical decision theories.

2.1.1 Beliefs

In *Reasons and Persons*, Derek Parfit introduces the following case:

> *Parfit's Hitchhiker*
>
> Suppose that I am driving at midnight through some desert. My car breaks down. You are a stranger and the only other driver near. I manage to stop you, and I offer you a great reward if you rescue me. I cannot reward you now, but I promise to do so when we reach my home. Suppose next that I am transparent, unable to deceive others. I cannot lie convincingly. Either a blush, or my tone of voice, always gives me away. (1984, 7)

A straightforward maximizer, Parfit contends, would be left stranded in the desert. Firstly, if Parfit is a straightforward maximizer he would not be willing to pay the reward (let us assume, arbitrarily, that it is $1000): when he arrives home he will have already received all the help he needs, and so paying $1000 will not be the best option for him. Secondly,

because he is transparent he cannot convincingly lie, and so the stranger will have no reason to save him. Here it seems that a straightforward maximizer fails where an agent following some other theory would succeed.

As is common with these sorts of vignettes, there are a large number of factors that need to be specified in order to reach the desired conclusion. We must assume, for example, that Parfit only cares about getting home and paying as little money doing so as possible, and so does not care about keeping promises or rewarding those who help him. We must also assume that he is not able to "bind" himself to paying the reward; that is, he does not have the ability to irrevocably commit himself to paying before the stranger takes him home. Relatedly, we must assume that he does not have the ability to change what the situation will be when he reaches his home; nothing he does now can change the fact that he will have a simple choice between keeping or losing $1000. This also requires that we suppose that the stranger does not have the ability to impose a cost on Parfit that is greater than or equal to $1000 should he refuse to pay.

We can simplify our thinking about this case by focusing on three key elements. Let a *decision problem* be defined by the acts available, the possible outcomes of those acts and their utilities, and the probabilities of those outcomes. In Parfit's Hitchhiker, there are two acts available to Parfit, he can pay $1000 or he can keep it. Next, we can suppose that keeping his money has much greater utility for Parfit than losing it. Finally, we suppose it is nearly certain that Parfit will keep his money should he choose not to pay the reward and also nearly certain he will lose his money should he choose to pay it. In this decision problem, straightforward maximization does indeed recommend that Parfit not pay the reward, since it has greater expected utility at the time of decision.

"Binding acts" should not be given a special status compared to other types of acts. If Parfit is capable of irrevocably committing himself to paying prior to your deciding whether you will rescue him—perhaps through the use of some drug or external device—then he faces a different decision problem than that specified above. In this new problem, there are again two acts available to Parfit: he can choose to bind himself to paying the reward

or not. Whether this is recommended by straightforward maximization will again depend on how we specify the relevant outcomes, utilities, and probabilities. If set up correctly, straightforward maximization will indeed recommend that Parfit choose to bind himself. However, it is important to realize that this is a different decision problem from the one Parfit intends to reveal in his example, and so not relevant to the current discussion. A similar point applies to decision problems in which Parfit has the ability to change the probabilities or utilities of the outcomes.

Let us now turn to how Parfit is understanding the transparency condition. We might take lying to involve, among other things, asserting what one believes to be false.[2] Parfit is therefore imagining that a straightforward maximizer would *believe* that he would not pay the reward, and that this belief would be revealed to you when he fails to lie. If this is so, then Parfit seems to understand "transparency" in terms of beliefs: to be transparent is to have beliefs that are detectable by others. Therefore, Parfit's point is that straightforward maximization can lead to problems for agents with detectable beliefs.

2.1.2 Dispositions

David Gauthier understands transparency differently—as a condition in which others are aware of one's dispositions. In *Morals by Agreement,* he writes, "Since our argument is to be applied to ideally rational persons, we may simply add another idealizing assumption, and take our persons to be transparent. Each is directly aware of the disposition of his fellows." (1986, 173–4). With that understanding of transparency, he discusses cases like the following:

> *Hume's Harvesters*
>
> My crops will be ready for harvesting next week, yours a fortnight hence. Each of us will do better if we harvest together than if we harvest alone. You will

[2]Cf. Williams, 2002, 96: "I take a lie to be an assertion, the content of which the speaker believes to be false, which is made with the intention to deceive the hearer with respect to that content."

> help me next week if you expect that in return I shall help you in a fortnight. (1993, 692)[3]

Gauthier's conclusion is that if he were a transparent straightforward maximizer then he would be left to harvest alone. He writes, "Consider my decision about helping you. I have gained what I wanted—your assistance. Absent other not directly relevant factors, helping you is now a pure cost to me" (1993, 692). If this is correct, then a straightforward maximizer is not disposed to fulfill his side of the bargain, as doing so does not maximize expected utility. And, secondly, since the potential bargaining partner is aware of this disposition, no bargain would be made. So here, as before, a straightforward maximizer fails where others would succeed.

While Parfit and Gauthier understand transparency differently, in other respects their cases appear to be the same. In order to show their similarities, let us again arbitrarily assign monetary value to the relevant outcomes. Let us say that for Gauthier the cost of helping with your crops next week is equivalent to $1000. Let us also say that the benefit of receiving your help with his crops is very large—about as large as being rescued from a possible desert demise (this makes the utilities of the possible outcomes of the two cases identical). Let us arbitrarily take both the crop-helping and the desert-rescuing to be equivalent to $1,000,000. Parfit and Gauthier's decision problems are now identical. In each instance there are two acts available: the agent can choose to pay $1000 or keep it, and it is nearly certain that the outcome will be much better if he chooses not to pay it. Straightforward maximization therefore recommends not paying. Nevertheless, a transparent straightforward maximizer is likely to never be given this choice, and instead suffer a terrible loss, equivalent to about $1,000,000. Parfit takes transparency to amount to belief

[3] Adapted from David Hume 1888, 520–1. Hume writes, "Your corn is ripe today; mine will be so tomorrow. Tis profitable for us both, that I shou'd labour with you to-day, and that you shou'd aid me tomorrow. I have no kindness for you, and know you have as little for me. I will not, therefore, take any pains on your account; and should I labour with you on my account, in expectation of a return, I know I shou'd be disappointed, and that I shou'd in vain depend upon your gratitude. Here then I leave you to labour alone: You treat me in the same manner. The seasons change; and both of us lose our harvests for want of mutual confidence and security."

detectability, and Gauthier takes it to be disposition detectability.

2.1.3 Intentions

John Broome describes the following case:

> *Broome's Shepherd*
>
> You are leading your flock of sheep down from the mountain. In the last narrow defile before reaching the safety of the plain, you meet a wolf. If the wolf lunges into the flock, trying to grab a sheep, it will probably end up catching only a scrawny one. However, half the flock will die of fright or be lost over a cliff. Both you and the wolf know this, and you both know there is another course of action that would suit both of you better. It would be better for both if you handed over the juiciest of this year's lambs to the wolf, and in return the wolf allowed the rest of the flock to pass unmolested. Is there some way this desirable outcome could be achieved? (2001, 101)

Broome then adds the transparency condition, writing: "Assume you are transparent to the extent that the wolf can accurately predict your intentions" (101).[4] With this condition in place we are presented with the familiar conclusion: transparent straightforward maximizers will receive a worse outcome in situations like this than other agents—agents who can sincerely intend to hand over the juiciest sheep *after* the wolf has allowed the flock to pass (when they are no longer in any danger). Broome, however, understands transparency in terms of intention-detectability, rather than belief- or disposition-detectability. Since the shepherd's decision to hand over the sheep will occur only after the flock is safe, choosing to keep his juiciest sheep maximizes the shepherd's expected utility. However, since the wolf is aware that the shepherd intends not to hand over the sheep, he chases half the flock over a cliff.

[4]Some theorists take intentions to amount to a special sort of disposition. On this account, Broome's understanding of transparency is close to Gauthier's. Kavka (1983, 35), for example, takes intentions to be dispositions "that are based on *reasons to act*."

Upon reflecting on its structure, we can see that Broome's case is the same as Parfit's and Gauthier's. In order to make this clear, let us again arbitrarily designate the utilities of the various outcomes in monetary units. Let us take the value of the juiciest sheep to be $1000, and the value of half the flock to be $1,000,000. As in Parfit's Hitchhiker and Hume's Harvesters, the shepherd therefore has a choice between paying $1000 or keeping his money. Straightforward maximization recommends not paying. However, if the shepherd is a transparent straightforward maximizer, then he is likely to suffer a terrible loss and end up $1,000,000 poorer.[5]

2.1.4 Reliable Prediction

We have reviewed three prominent accounts of how "transparency" causes problems for straightforward maximizers. As we've seen, Parfit, Gauthier, and Broome are discussing the same case, with the exception of their understanding of the transparency condition. Parfit takes transparency to amount to belief-detectability, Gauthier takes it to be disposition-detectability, and Broome takes it to be intention-detectability. Should we conclude that there are three separate phenomena that can cause problems for straightforward maximizers, all of which operate under the name "transparency"? In what follows, I will show that all three understandings of transparency are unsatisfactory and must be replaced by a single account that understands it in terms of susceptibility to *reliable prediction*.

Consider Parfit's case first. You are willing to rescue Parfit in exchange for a reward, which we were imagining is $1000. Since Parfit is taking transparency to amount to belief-detectability, he imagines that you would save him upon learning that he believes that he would pay. But this is not true. Information about what Parfit believes is only relevant to you in so far as it indicates that Parfit is likely to pay. You might, after all, think it likely that Parfit will change his mind. If you think that Parfit will change his mind, then the fact that he now believes that he would pay is irrelevant.

---

[5]Kavka is another example of a theorist who understands transparency in terms of intentions. See the discussion of "special deterrent situations" in his 1987, 16.

The point works in the converse situation as well. If you detect that Parfit believes that he won't pay (as in the original example), this is again only relevant in terms of what it indicates about the likelihood of Parfit actually paying. If you think that it is likely that Parfit will change his mind and pay you $1000 after all, then you might rescue him in spite of what he currently believes.

The key premise being revealed here is that Parfit's problem, as a straightforward maximizer, begins and ends with his susceptibility to reliable prediction. If you are able to predict what someone in Parfit's situation would do, then *you are able to selectively rescue only those who would pay*. This allows for a connection between what Parfit would do and his chances of being rescued. If, however, you are not able to predict what Parfit would do, then there is no connection between what Parfit would do and his chances of being rescued. Without this connection, straightforward maximizers are no less likely to be rescued than anyone else.

Given this, we can see that belief-detectability is not what matters most fundamentally; rather, it is a *tool* that you might use to *predict* what Parfit is going to do. What is important to the case is that Parfit be susceptible to reliable prediction, and having his beliefs be revealed to you is merely a way to accomplish that. Therefore, we should understand transparency as susceptibility to reliable prediction.

The same is true of Gauthier's case. Gauthier imagines that your decision to help him with his crops would depend on learning the status of his current disposition to return your help a fortnight hence. But it is not enough to learn that Gauthier is currently disposed to return your help if you also believe that this disposition will change. Therefore, you will only be willing to help Gauthier if you *predict* that he would help you in return, and learning of his disposition may be a *tool* for doing that. Again, in order for Gauthier (as a straightforward maximizer) to be less likely to receive help than someone else, there needs to be a connection between what he would do and his chances of receiving help. This is the case only if Gauthier is susceptible to reliable prediction.

Finally, Broome's understanding of transparency is inadequate for similar reasons.

Since intentions can change, the wolf will only be interested to know that the shepherd presently intends to hand over a juicy sheep if he takes this to indicate that the shepherd is likely to hand over a juicy sheep in the future. What the wolf cares about is what the shepherd's intentions will be *in the future*: will he or won't he hand over the sheep? Again, in order to create a problem specifically for straightforward maximizers the wolf needs to be able to predict this reliably.

To sum up: the crucial element in the cases that make trouble for straightforward maximizers is not the availability of their beliefs, dispositions, or intentions, but the ability of others to make reliable predictions about what they are going to do. It is consequently a mistake to identify transparency with anything other than reliable prediction. This is an easy mistake to make, since knowing someone's beliefs, dispositions, or intentions is usually a useful tool for making a reliable prediction about what they will do. But in each instance, the tool has been mistaken for the finished product: these cases require that a straightforward maximizer's future actions be reliably predicted, and having access to their beliefs, dispositions, or intentions is neither necessary nor sufficient for that. We can, of course, call whatever condition we like "transparency," but in so far as we aim to show that straightforward maximization plus a transparency condition leads to disaster, transparency needs to be understood in terms of reliable prediction.

Even with this improved understanding of transparency in hand, we still need to specify exactly what a reliable prediction amounts to. What is entailed, for example, by the fact that you are a reliable predictor of what Parfit is going to do? The simplest way to understand reliable prediction is in terms of *conditional chances.* To say that a predictor is reliable is exactly to say that there is a high chance of his predicting ϕ in situations where ϕ will in fact occur. A predictor's reliability will usually be restricted to a domain, and in this instance we can take you to be reliable regarding predictions of Parfit's future willingness to pay. If this is right, then Parfit is representing the situation as being one in which there is a very high chance of rescue conditional on being in a world in which he would pay the reward, and a very low chance of rescue conditional on being in a world in which he

would not pay the reward. For artificial clarity, we can specify that Parfit takes there to be a 90% chance of rescue in a payment world, and a 10% chance of rescue in a non-payment world. Similarly, in Gauthier's example, we can specify that Gauthier takes there to be a 90% chance of receiving help with his crops in a world in which he would reciprocate, and a 10% chance of receiving help in a world in which he would not reciprocate.

While this is just one of the ways we could cash out what it means to be a reliable predictor, something in this neighborhood is required to make sense of how transparency causes problems for straightforward maximizers, since there must be some connection between what they would do and the outcomes they receive. The conditional chance model provides a particularly illuminating and clean version of it, but alternatives can be pursued if one is skeptical of the notion of conditional chance.

Once we understand transparency in terms of reliable prediction, and reliable prediction in terms of conditional chance, it is clear how these cases are similar. In a world in which the agent would perform a certain non-maximizing act that results in the loss of $1000 (i.e., give away money, volunteer to reciprocate harvesting, or hand over a juicy sheep), the agent has a 90% chance of receiving a large benefit worth about $1,000,000 (i.e., rescue, help harvesting, or safe passage). In a world in which the agent would not perform the non-maximizing act, the agent has only a 10% chance of receiving the benefit. All three examples are therefore different versions of the same case. The case is depicted by Figure 2.1, where T stands for "the agent chooses to keep $1000" and M stands for "the agent receives $1,000,000."

Parfit, Gauthier, and Broome have thus independently converged on the same case in trying to reveal a potential problem for straightforward maximization. In the next section we will see that this case, which is the key to understanding the motivation for non-classical decision theories, is a *Newcomb problem*.

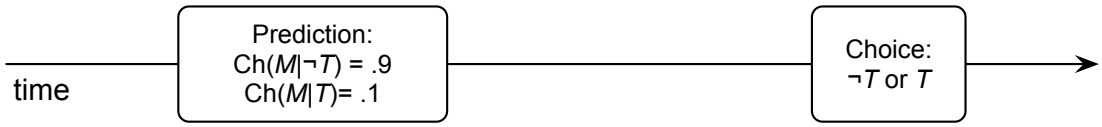

Figure 2.1: Parfit's Hitchhiker, Hume's Harvesters, and Broome's Shepherd

2.2 Open Newcomb Problems

Here is an example of a Newcomb problem that James Joyce provides:

> Suppose there is a brilliant (and very rich) psychologist who knows you so well that he can predict your choices with a high degree of accuracy. One Monday as you are on the way to the bank he stops you, holds out a thousand dollar bill, and says: "You may take this if you like, but I must warn you that there is a catch. This past Friday I made a prediction about what your decision would be. I deposited $1,000,000 into your bank account on that day if I thought you would refuse my offer, but I deposited nothing if I thought you would accept. (1999, 146-7)[6]

In Robert Nozick's classic formulation of the problem the bank account is an opaque box in front of the agent and the money is sitting in a transparent box rather than the predictor's hand.[7] Joyce's otherwise equivalent formulation does a good job of revealing the connection to the cases presented by Parfit, Gauthier, and Broome. As before, an agent is likely to receive $1,000,000 if he would make a non-maximizing choice costing $1000. In fact, this case can also be represented by Figure 2.1.

Joyce's case differs from that discussed by Parfit, Gauthier, and Broome in one respect: you do not know what has been predicted when it comes time to decide. Recall that at

[6]Adapted from a version presented by Sobel (1985, 198-9 n. 6). Sobel points out that the classic formulation of Newcomb's problem can lead to confusions that make evidential decision theory seem appealing.

[7]See Nozick 1993, 41.

the time of his decision Parfit would know that his payment has been predicted, since this is revealed by the fact that he has been rescued. The same is true in Gauthier's and Broome's cases. At the time of decision, you know whether Gauthier has helped you, and the shepherd knows whether the wolf has allowed the flock to pass. However, it is easy to make a slight modification to the standard Newcomb problem to make it exactly like these cases: consider a variant in which upon receiving the psychologist's proposal you quickly access your bank balance on your smartphone, and then afterwards you are free to make your decision.

Making a similar adjustment to the box presentation of the problem is just as easy. In fact, in their seminal paper on the topic, Gibbard and Harper (1981, 181) do exactly that. They write: "Consider a variant on Newcomb's story: the subject of the experiment is to take the contents of the opaque box first and learn what it is; he then may choose either to take the thousand dollars in the second box or not to take it."[8] The case now has the same structure as the cases with which we started: an agent learns what has been predicted and then must decide whether to accept or refuse the smaller reward. In honor of Gibbard and Harper, we should call cases like this "open" Newcomb problems (as if the million dollar box has been opened) to differentiate them from the standard version (in which the million dollar box remains closed). We can then see that the cases discussed by Parfit, Gauthier, and Broome are all open Newcomb problems. Once we see that in studying strategic interaction puzzles like Parfit's Hitchhiker, Hume's Harvesters, and Broome's Shepherd we have really been studying Newcomb problems all along, we gain a better understanding of the mechanism by which problems are created for straightforward maximizers. There is really only one problem for straightforward maximizers—the Newcomb problem—and it presents itself in two ways. There are *open* Newcomb problems, in which the agent knows what has been predicted before he must make his choice, and there are *closed* Newcomb problems, in which the agent does not know what has been predicted before he must make

[8]Gibbard and Harper introduce this case as part of an attack on the "why ain'cha rich?" objection to causal decision theory. They take it to show that "why ain'cha rich?" considerations also affect evidential decision theory, since in this variant evidential decision theorists also end up poor.

his choice. But this difference is mostly irrelevant when it comes to thinking about whether to accept the $1000.[9] The real problem for straightforward maximizers is that they suffer massive losses before they have any choices to make.[10]

2.2.1 Newcomb Problems are Problems for Everyone

When thinking about the cases introduced by Parfit, Gauthier, and Broome, it is easy to get the impression that they are only relevant to theories of self-interested rationality. The cases seem to involve, after all, strategic interaction between egoists.[11] However, the significance of Newcomb problems is actually much broader. To create problems for act utilitarians, for example, we need only imagine that the psychologist gives you the option of having $1000 donated to an effective charity, with the information that if he predicted that you will turn down the offer, then he will have already donated $1,000,000 to the same charity. According to an impartial notion of utility that weighs the interests of others equally, it would be much better if $1,000,000—rather than $1000—were donated, but a utilitarian straightforwardly maximizing expected utility can expect only $1000 to be donated.

The same result can be achieved in any of the open Newcomb problems discussed above. For example, in Parfit's Hitchhiker we can imagine that it would be best, from an impartial perspective, for Parfit to be rescued from the desert but then not pay the reward. Parfit may, after all, have more benevolent plans for the money than you do. We can also imagine that the worst outcome—again, impartially—is Parfit being left in the desert to die. But if Parfit is a transparent straightforwardly-maximizing utilitarian we should expect the worst outcome.

---

[9] Unless one is an evidential decision theorist. See Section 5 for more details.

[10] In the words of David Lewis (1981b, 377), "When we made our choices, there were no millions to be had."

[11] Parfit (1984, 24–8)'s discussion of the hitchhiker case, for example, is presented as part of his argument that *self-interest* theories are "self-defeating."

The point generalizes to other theories of the good as well, such as objective list theories. All that is required to create problematic cases is that transparent agents straightforwardly maximize whatever conception of value they favor. We should conclude, therefore, that Newcomb problems are not isolated to self-interest theories. The only theories that are immune to them are ones that somehow make sure that regardless of how the situation is set up it could never be best to keep the $1000. This is difficult to accomplish within a maximizing framework without adopting implausible constraints.

The importance of these cases is therefore assured for both egoists and moral consequentialists. Even non-consequentialists who differentiate between morality and prudence will find thinking about these cases worthwhile in so far as they concern prudence, and unless they absolutely prohibit keeping the $1000 in any situation, they will find them relevant to thinking about morality as well.

2.3 Only Newcomb Problems Are a Problem

I suggest that the prediction structure characteristic of Newcomb problems is a necessary condition for producing problematic cases for straightforward maximization. While there are difficult details to be worked out at the level of formal decision theory, straightforward maximization—understood as a general formula for thinking about rational decisions—is an airtight theory for producing decisions that tend to bring about the best consequences in light of an agent's present perspective. Therefore, in order to create situations in which the recommendations of straightforward maximization tend to produce worse consequences than than those of rival theories, we need to tie those consequences to what a straightforward maximizer will choose in the future, or would choose counterfactually, rather than what he actually chooses in the present. Predictions are indispensable in accomplishing this: the only tool for tying consequences to future and counterfactual decisions is reliable prediction.

In order to illustrate this idea, consider a closely related puzzle: Kavka's toxin puzzle.

Here is Kavka's original formulation of the puzzle:

> You have just been approached by an eccentric billionaire who has offered you the following deal. He places before you a vial of toxin that, if you drink it, will make you painfully ill for a day, but will not threaten your life or have any lasting effects.... The billionaire will pay you one million dollars tomorrow morning if, at midnight tonight, you *intend* to drink the toxin tomorrow afternoon. He emphasizes that you need not drink the toxin to receive the money; in fact, the money will already be in your bank account hours before the time for drinking it arrives.... All you have to do is sign the agreement and then intend at midnight tonight to drink the stuff tomorrow afternoon. You are perfectly free to change your mind after receiving the money and not drink the toxin. (1983, 33–4)

The toxin puzzle is similar to a Newcomb problem, but the two are importantly different when it comes to the conditions under which an agent is rewarded. As Kavka emphasizes, in the toxin puzzle an agent is rewarded if he possesses the *intention* to drink the toxin at midnight. Newcomb problems are different. In a "newcombized" version of the toxin puzzle, an agent is rewarded if it is *predicted* at midnight that they will drink the toxin. This difference is crucial.

In the non-newcombized version of the puzzle, consider what straightforward maximization recommends prior to midnight. Prior to midnight, if given a choice between possessing the intention to drink the toxin (or doing something that will result in the agent possessing the intention) or not possessing it, straightforward maximization recommends that the agent choose to possess it (assuming, of course, that there isn't a cost associated with possessing the intention that exceeds $1,000,000). Now consider what straightforward maximization recommends after midnight. If given a choice between discarding the intention to drink the toxin or keeping it, straightforward maximization recommends discarding it. These are the correct recommendations, I suggest, and more importantly, they are the recommendations of the decision theories that tend to produce the best consequences in this

case. Therefore, Kavka's toxin puzzle does not present any special problem for straightforward maximizers; they are just as likely as anyone else, based on the decisions they make, to come away with the $1,000,000.

In the newcombized version of the toxin puzzle, however, straightforward maximizers *are* less likely to come away with the $1,000,000 because of the choices they make. In the newcombized version, an agent is rewarded if there is a high chance that he will drink the toxin, as revealed by a reliable prediction about whether he would choose to drink it. Consider what straightforward maximization recommends regarding drinking the toxin—it recommends not drinking it. This makes an unfavorable prediction likely, and thereby causes straightforward maximization to produce worse outcomes than theories of decision making that recommend drinking the toxin.[12]

The toxin puzzle is an important case for thinking about the nature of intention and its associated norms. Thinking about the toxin puzzle can be useful in attempting to answer important questions; e.g., can one rationally intend to do what one thinks one will have no reason to do? However, the toxin puzzle is not relevant to thinking about the relative success or failure of straightforward maximizers, and this is because it lacks the prediction structure characteristic of Newcomb problems.[13]

It may, however, be possible to create problematic cases that do not involve reliable prediction by invoking infinities. In fact, Arntzenius et al. (2004) have accomplished this by creating cases that involve infinite decisions and infinite utilities. But reliable prediction is required to create finite cases that are problematic for straightforward maximizers, and Newcomb problems provide us with the most clear and uncontroversial instances of this

---

[12]That the two cases are importantly different becomes even clearer when we consider the differing effects of "binding" oneself to drinking the toxin prior to midnight in the newcombized and non-newcombized versions. In the newcombized version, by ensuring that one will drink one thereby ensures the reward. However, this is not so in the non-newcombized version. Kavka actually makes the mistake of thinking that binding oneself to drinking will ensure the reward in his case, but as McClennen (1990, 227) points out, "Precommitment at midnight ensures that you will drink the toxin tomorrow; but it does not ensure that you intend to drink the toxin."

[13]I offer a more detailed proposal for judging the relative success and failure of decision theories in Chapter 4.

phenomenon. This should be unsurprising when we reflect on what straightforward maximization recommends and what Newcomb problems entail. Straightforward maximization tells agents to always choose the best available act, but Newcomb problems are situations in which agents are punished if it is predicted that they will choose the best available act. It therefore makes sense that Newcomb problems are the Achilles heel of straightforward maximization.

2.4 Realism of Newcomb Problems

In discussions of closed Newcomb problems, the predictor is often thought of as a supernatural entity or technologically advanced scientist. This has the unfortunate effect of making the case seem very artificial. Tying together open and closed Newcomb problems therefore has the further benefit of showing the realism of Newcomb problems, and may be seen as an important part of the motivation for non-classical decision theories.

In order to highlight the realism of Newcomb problems, let us once again compare them to toxin puzzles. Toxin puzzles are imaginary. Firstly, as we have seen, other agents tend to care about what an agent *will do*, and not solely about what an agent intends to do. Secondly, there are no obvious ways to detect for the presence of intentions at the level of reliability presupposed by the toxin puzzle. Imagined toxin puzzles therefore reward agents for possessing a potentially impossible-to-detect property that no one would actually want to reward in the first place.

Newcomb problems, on the other hand, abound. Closed Newcomb problems may be very rare, but open Newcomb problems are not. This becomes clear when we notice how little reliability is required of the predictor in many cases. Consider that in the standard formulation of Newcomb's problem, in which $1,000,000 is at stake, the predictor need only be more than .5005 reliable for those who accept the $1000 to tend to end up with less money than those who do not.[14] As the open Newcomb problems discussed above

[14]The reliability actually needs to be a bit higher if we take the diminishing marginal utility of money into

reveal, whenever we bargain, threaten, or otherwise interact with other agents, rewards and penalties are often tied to predictions regarding our future or potential behavior. Often the rewards and penalties are immense: examples might include making tenure, getting married, being granted parole, or getting a security clearance, all of which seem to involve predictions regarding our future behavior, and in each instance it is reasonable to think that the reliability of others in predicting our future behavior is at least slightly above chance.

A particularly common yet often overlooked type of Newcomb problem involves a single agent acting as both predictor and predictee. If you are anything like me, you face this sort of Newcomb problem all the time. Imagine that you would like to go running, but you are having trouble motivating yourself to actually do it. You may find that an effective motivational tool is to make a "deal" with yourself to do less running than you normally do. Perhaps, instead of running your normal four laps around the park, today you will only run only three. Here's the catch: this motivational tool might be effective only if you predict that you will actually run only three laps. You might instead predict that you will break the deal and run four laps. You might anticipate that after having run three laps you are liable to judge that continuing on for the fourth lap would be, at that point in time, the best thing to do all things considered. Upon realizing that you will likely break the deal and so a typical amount of running awaits you, the motivational tool is ruined, and you end up with no exercise at all.

In this example, the larger reward (in the place of $1,000,000) is getting three laps of exercise done today instead of zero. The smaller reward (in the place of $1000) is the extra lap. Following the structure of an open Newcomb problem, you receive the larger reward only if the predictor (in this case, you) predicts that after you attain the larger reward you will forego the smaller reward. Therefore, if you are a reliable predictor of yourself, then things will tend to turn out better if you would in fact forego the smaller reward. Given that the difference between three laps and zero may be *much* more important to your health than

---

account. Lewis (1981a, 9–10) points out that the reliability of the predictor needs to be greater than .68 if we use Bernoulli's logarithmic scale for the utility of money [u($$x$)=ln(1+$x$)].

the difference between three laps and four, the required reliability of the prediction need not be very high. It is reasonable to think that those unskilled in self-deception do indeed possess this amount of reliability.

There is nothing special about exercise in generating these sorts of open Newcomb problems. Examples like this may be present wherever agents attempt to motivate themselves to accomplish unappealing tasks. However, their ubiquity does force us to take note of the fact that situations similar to that just described are likely to be repeated throughout one's life. Given this, one might think that even a straightforward maximizer has reason to forego the smaller reward, since by doing so she can give herself evidence that she will make similar choices in the future, and thereby influence her future predictions. The mechanism at work here, however, is suspect. In the interpersonal case, making non-maximizing choices in order to influence future predictions may make sense, since it gives other people evidence about the way you make decisions. But this effect is lost when the agent is aware that her present decision aims to influence her future predictions. It is easy to imagine agents with the following attitude: "In the future I'm going to choose the best option available, regardless of whatever I've attempted to fool myself into thinking I'll choose in the past."[15] This seems reasonable, and so, again, it seems that a straightforward maximizer's only hope is that they are skilled at self-deception.

2.5 Causal vs. Evidential Expected Utility

Debates over the rational action in closed Newcomb problems have historically been understood as a disagreement between causal and evidential decision theory. However, the way in which causal and evidential decision theory differ is mostly irrelevant to the project of constructing alternatives to straightforward maximization. This is because both causal and evidential decision theory endorse straightforward maximization. We can differentiate

[15] Arntzenius et al. (2004, 19) entertain a similar argument in discussing their "Satan's Apple" case.

between i) a theory's understanding of expected utility, and ii) whether the theory recommends that expected utility be straightforwardly maximized. Both causal and evidential decision theory recommend that expected utility be straightforwardly maximized; they differ in how they understand expected utility.

2.5.1 Two Forms of Supposition

At the heart of the divergence between a causal and evidential understanding of expected utility is a difference in two types of *supposition*. Consider two ways in which one might suppose that Shakespeare didn't write *Hamlet*. First, "If Shakespeare didn't write *Hamlet*, someone else did." Second, "If Shakespeare didn't write *Hamlet*, someone else would have."[16] In the first instance an *indicative* supposition is used, in which one supposes that the condition is actual: one might imagine *receiving the news* that Shakespeare actually didn't write *Hamlet*, and then consider whether given this news the play has, in fact, been written (yes). The second instance features a *subjunctive* supposition, in which we suppose *counterfactually*. One imagines, counterfactually, that Shakespeare had failed to write *Hamlet*, and then reflects on whether the play would have been written by someone else (probably not). We might call the first form of supposing "A-supposition" ("supposition-as-actual"), and the second form "C-supposition" ("supposition-as-counterfactual").[17]

Given this set up, we can distinguish between evidential and causal decision theory based on the type of supposing they endorse for calculating expected utilities. Evidential decision theory takes A-supposition to be the proper way to evaluate the expected utility of acts. According to evidential decision theory, in order to calculate the expected utility of an act one considers the probabilities of potential outcomes *on the A-supposition* that the act is performed. One way to build A-supposing into a technical model is by using *conditional credences*. In order to evaluate the expected utility of an act, we take the product of an agent's credences in possibilities conditional on the act being performed and the utility of

[16]The example is from Jonathan Bennett, 1988, 523–4.

[17]Following Daniel Elstein and Robert Williams (manuscript).

those possibilities. Taking the "w_i"s to refer to elements of the space of possibilities and "A" to the act in question, the equation for expected utility becomes:

$$\mathrm{EU}(A) = \sum_i \mathrm{cr}(\mathrm{w}_i|A)\mathrm{u}(w_i)$$

Thus, calculating the *evidential expected utility* of an act is a matter of calculating the probability of possibilities given that one receives the news that A is performed. Because of this, we might think of evidential expected utility as representing the "news value" of an act.

Causal decision theory, on the other hand, takes C-supposing to be the proper way to evaluate the expected utility of acts. There are several ways to understand C-supposing, but for the current purposes I will restrict the discussion to two. According to the *counterfactual* approach, C-supposing might in the first instance be understood in terms of subjunctive conditionals. Thus, when one evaluates the expected utility of an act, one asks what possibilities would be likely were the act performed. Therefore, we might calculate the expected utility of acts by substituting the probabilities of subjunctive conditionals—$\mathrm{P}(A \mathbin{\square\!\!\rightarrow} W_i)$—for the conditional credences that are used to calculate evidential expected utility.

Using subjunctive conditionals to model C-supposing that A focuses our attention on how things stand in the closest A-world. What if one is skeptical that we can identify the probabilities of a world being closest to the actual world? Joyce (1999, 172–6) (following Lewis (1981a) and Sobel (1994)) suggests that C-supposing is best understood in a different way. Subjunctively supposing, he thinks, is best thought of as the process of accepting a proposition and making the minimum changes in other opinions to accommodate it. This process, Joyce argues, is importantly different from judging the probability of $A \mathbin{\square\!\!\rightarrow} W_i$. In order to better capture this process, we need to use *imaging*, rather than simple subjunctive conditionals. Imaging is an epistemic operation in which an agent rules out worlds in which A is false, and then redistributes the probability of these worlds to worlds in which A is true, giving greater additional weight (redistributed from the not-A worlds) to worlds that are "more similar to" the actual world. This process is thus intended to respect an

agent's judgments about the relevant overall similarity of worlds without requiring a closest possible world to be identified. Joyce represents the expected utility formulation using the notation "P^A":

$$EU(A) = \sum_i P^A(w_i)u(w_i)$$

Where P^A represents the "image" of P under *A*.

Another way to understand causal expected utility is in terms of *conditional chances.* In this formulation, we use propositions about the chance of possibilities conditional on the act being performed, and then sum the probabilities of these chance hypotheses:

$$EU(A) = \sum_i \sum_x xP[ch(w_i|A)=x]u(w_i)$$

One benefit of this formulation is that it allows for a direct interface between expected utility and the understanding of reliable prediction presented above, since they are both understood in terms of conditional chance. However, defenders of both the conditional chance and counterfactual approaches agree that not much likely rides on which way causal expected utility is understood. Harper and Skyrms (1988) state: "It can be argued...that the various forms of causal decision theory are equivalent—that an adequate version of any one of these approaches will be interdefinable with adequate versions of the others." Lewis (1981a, 5) concurs: "We causal decision theorists share one common idea, and differ mainly on matters of emphasis and formulation." The recommendations of causal decision theory are likely to be the same or very similar, regardless of how causal expected utility is formulated.

2.5.2 Why Causal Expected Utility is Superior

Let me highlight two major advantages of understanding expected utility causally. One way to see that something has gone wrong with evidential notions of expected utility involves comparing open and closed Newcomb problems. Consider Joyce's closed Newcomb problem presented above. In this situation, the evidential expected utility of refusing the $1000

is greater than that of accepting it, because an agent's credence in having the $1,000,000 in his account conditional on refusing it is greater than his credence in having the $1,000,000 conditional on accepting it. However, if we make the case an open Newcomb problem, the impact of accepting or refusing the $1000 on his conditional credences changes. If the agent already knows whether the $1,000,000 has been deposited, then his decision to accept or refuse the $1000 has no effect on his credences. Therefore, in the open version, accepting the $1000 has greater expected utility than refusing it.

This switch in the expected utility of refusing the $1000 is curious. Let *"$ "* refer to the situation in which $1,000,000 has been deposited in the agent's bank account, and "*~$ "* to the situation in which nothing has been deposited. In an open Newcomb problem, the agent knows which of *$* and *~$* is the case. In either situation, evidential decision theory recommends accepting the $1000. Now, when we move to the closed version, the agent no longer knows whether *$* is the case or *~$* is the case, but he still knows that *either* $ or ~$ is the case. Here is the kicker: he further knows i) if he knew whether *$* or *~$*, evidential decision theory would recommend accepting the $1000, and ii) nothing he does now will raise or lower the chance of either possibility. So, if an agent knows that $ and ~$ partition the space of possibilities, and that the chance of either possibility is the same regardless of what he now does, and that in both situations evidential decision theory recommends accepting the $1000, how can it be that refusing the $1000 has greater expected utility? It cannot be, at least without breaking what seems to be an extremely firm and intuitively supported principle of dominance. This gives us the following requirement:

> **Dominance Requirement:** any satisfactory decision theory must give the same recommendations in the open and closed versions of a Newcomb problem.

This successfully debunks evidential decision theory, and so if we are restricting our attention to only these two theories, as has historically been the case, then causal decision theory is the winner. However, as we shall see, by adopting causal expected utility but *rejecting*

straightforward maximization, a theory of rational decision making can also satisfy this requirement.

Another popular argument against evidential understandings of expected utility appeals to so called "medical Newcomb problems." One example of such a case, which we can call "The Smoking Lesion," asks us to imagine that it is discovered that smoking does not, in fact, cause lung cancer. The reason for the correlation between the two, the story continues, is that a common brain lesion (the development of which is governed by random chance) causes one to be at a high risk for lung cancer, and also produces a predisposition to smoke. In fact, of those with the lesion, 90% smoke, whereas only 10% without the lesion smoke. Given this information, should one be deterred from smoking? The standard answer is that if one wants to smoke, one should. In this case, one's decision to smoke cannot possibly influence whether one does or does not have the lesion (and hence whether or not one is at higher risk for cancer), and since one cannot affect the chance of cancer, one is free to smoke without any added penalty.

Evidential decision theory, however, seems to disagree with this reasoning. This is because an agent's credence in having the lesion might increase or decrease depending on whether he decides to smoke, and so the evidential expected utility of not smoking may be much greater than that of smoking. If so, evidential decision theory would emphatically recommend that the agent choose not to smoke.

Cases like this are really troublesome for evidential decision theory because it seems to recommend refusing a reward with no associated costs, and unlike with standard Newcomb problems, there is no connection between an agent's choice—not even in the form of past prediction of that choice—and the development of cancer. For this reason, medical Newcomb problems should not be thought of as Newcomb problems at all; they lack the prediction structure that we have seen is crucial for making trouble for straightforward maximizers. Whether an agent receives the larger reward (freedom from cancer) does not depend on a prediction of the agent's behavior; rather, it solely depends on the completely

random processes that govern the development of the lesion. In a true Newcomb problem, as we are understanding them, the development of the lesion would need to be tied to an agent's future behavior by a reliable prediction—being more likely if the agent would choose to smoke and less likely if the agent would choose not to smoke. This difference, at the very least, forces us to put medical Newcomb problems in a category different from the ones that feature reliable predictions.

Since there is no connection between an agent's actual, future, or potential choices and the development of cancer in The Smoking Lesion, it should be clear to both supporters and critics of straightforward maximization that evidential decision theory gives the wrong recommendations. This gives us the following requirement:

> **Medical Newcomb Requirement:** any satisfactory decision theory must recommend the nondominated option in a medical Newcomb problem.

While this may exclude evidential decision theory, we shall see that a theory can satisfy this requirement by, again, accepting a causal notion of expected utility but rejecting straightforward maximization.

2.6 Solving the Problem

We have seen that open and closed Newcomb problems are ubiquitous cases in which straightforward maximizers tend to be less successful because of the choices they make. Is this a genuine problem for straightforward maximization? And, if so, what is the solution?

2.6.1 The Binding Response

A popular strategy for defending causal decision theory in light of (closed) Newcomb problems is to accept that casual decision theorists will generally end up much worse off than those who refuse the $1000, but then deny that this speaks poorly of their rationality. Gibbard and Harper (1981, 181) take the moral of Newcomb problems to be the following:

"If someone is very good at predicting behavior and rewards predicted irrationality richly, then irrationality will be richly rewarded." In the same spirit Lewis (1981b, 377) writes, "They have their millions and we have our thousands, and they think this goes to show the error of our ways.... The reason why we are not rich is that the riches were reserved for the irrational." Generally speaking, the proposal is this: in Newcomb problems the predictor is punishing for predicted rationality and rewarding for predicted irrationality. We therefore should expect rational agents to be punished in these cases.

Arntzenius et al. (2004) have subsequently attempted to bolster this defense by pointing out the benefits of being able to self-bind. Causal decision theorists, they show, will avoid punishment if they are able to foresee potential difficulties and bind themselves to more favorable courses of action. In Parfit's Hitchhiker, as we noted above, it would be wise for Parfit to bind himself to paying prior to being rescued. This makes sense, because by doing so he thereby causes himself to have a better chance of rescue. Arntzenius et al. write: "The lesson is that under certain circumstances, the following ability can be incredibly helpful: the ability to have one's present choices causally influence one's future choices" (268). That much is surely correct. However, they then draw a much stronger conclusion: "Rational individuals who lack the capacity to bind themselves are liable to be punished, not for their irrationality, but for their inability to self-bind." They add: "Certain situations exploit rational agents who are unable to self-bind" (269).

We therefore see a shift from thinking that causal decision theorists are punished for their rationality, to thinking that they are punished for their inability to self-bind. In reality, neither of these proposals is an adequate way of understanding what causal decision theorists are "punished for." Let us first consider the suggestion that they are being punished for their irrationality. Consider the following: in a Newcomb problem, agents who accept the \$1000 *either rationally or irrationally* are punished, while agents who refuse the \$1000 *either rationally or irrationally* are rewarded. How might a causal decision theorist rationally refuse the \$1000? A causal decision theorist who expects huge psychological penalties for refusing to pay the rescuer in Parfit's Hitchhiker might rationally pay the money, and he

would be rewarded. Conversely, if it were likely that this same causal decision theorist would *irrationally* choose to keep the money, he would probably be punished. This is because what is being predicted is an agent's *choice*, and not the *reasoning process* by which he comes to make that choice. The predictor does not care how an agent comes to decide to pay $1000, he only cares that the agent will pay it. For all we know, the predictor might believe that paying the money is rational.

Therefore, agents are punished if it is predicted that they will accept the $1000, whether they do so rationally or irrationally. If this is true, then the idea—stressed by Gibbard and Harper, and later by Lewis—that agents are punished for their irrationality is false.

What of the idea that agents are being punished for their inability to self-bind? This is also false. We noted above that decision situations are partly individuated by the acts available, and we can distinguish between situations in which agents are able to self-bind and situations in which agents are unable to self-bind. It is true that in "binding cases" causal decision theorists can expect to do just as well as anyone else. However, this fails to show that causal decision theorists are being punished for their inability to self-bind. It simply does not follow from the fact that self-binding agents avoid punishment that other agents are being punished for their inability to self-bind.

Allow me to emphasize this point by drawing an analogy. Imagine what might happen if a causal decision theorist were given another kind of ability: the ability to charm others into acting as they desire. Given this ability, Newcomb problems like Parfit's Hitchhiker are similarly unproblematic for causal decision theorists, since they can persuade the predictor to rescue them regardless of what is predicted. Here too, we can see that a causal decision theorist can expect to do just as well as anyone else in "charm cases." Using Arntzenius et al.'s argument we can therefore conclude that causal decision theorists are being punished for their lack of charm. Indeed, using this same argument we can ostensibly show that causal decision theorists are being punished for many things. Instead, of course, we should resist the idea that the predictors in Newcomb problems punish causal decision theorists for lacking special abilities. Sometimes binding acts, or charming acts, are available to agents,

but sometimes they are not. Newcomb problems are instances of the latter.[18]

I suggest that we abandon attempts to write off Newcomb problems as punishing rationality or the lack of special abilities. It should be clear that in Newcomb problems agents are being punished for their *future decisions*. This idea fits the cases discussed above perfectly: what the predictor cares about is whether $1000 will be refused; he does not care whether an agent is able to self-bind, or whether the agent will be rational. He only cares about what the agent will do. We may—by way of independent argument—believe that what the agent will do is irrational, but it is not the case that the agent is being *punished for* being irrational. The agent is being punished for the future decisions he is likely to make.

We are therefore left with two possibilities: we can continue to defend the rationality of straightforward maximization but accept that rationality is a liability (even when one has perfect information), or we can reject straightforward maximization and amend our understanding of rationality so that it isn't a liability.[19] The appropriate response will partly depend on whether the latter is possible. Let us therefore turn to non-classical theories of rational decision making, which reject straightforward maximization.

2.6.2 Gauthier's Proposal

A peculiar feature of Newcomb problems is the way that the desirability of an act changes over time without any corresponding change in an agent's telic desires or information about the act's likely consequences. Prior to prediction, agents in Newcomb problems have most reason to hope that they will choose to forego the $1000 (as this will make a favorable prediction likely), but after the prediction they have most reason to hope that they will choose to accept the $1000 (as this will make them $1000 richer with no corresponding cost). However, straightforward maximizers take this peculiar feature to be irrelevant. Since they advocate making decisions only in light of their desirability *in the present*, they only take

[18]Meacham (2010, Section 4) presents a similar argument against Arntzenius et al.'s claim.

[19]In Chapters 4 and 5 I discuss the reasons why we might choose one of these options over the other.

the post-prediction situation to be relevant to an agent's choice.

Consider an agent who makes decisions in a different way. Imagine someone who is disposed to consider not simply which acts maximize expected utility in the present, but also which acts it would have maximized expected utility to *plan on choosing* in the past. In thinking about this idea, a lot is owed to the work of David Gauthier. In much of *Morals by Agreement*, his discussion exclusively concerns strategic interaction between bargaining partners, but he does provide a generalization of his basic idea that covers decision-making more generally. He proposes the following:

> *Gauthier's Proposal:* if at some time t_0 it maximizes expected utility to follow a plan that involves A-ing at some subsequent time t_1, then the agent should A at t_1. (1986; 1988/89)[20]

In Newcomb problems, there is a divergence between the best act in the present and the best plan in the past. Consider again Parfit's Hitchhiker. At any time before your prediction, Parfit can see that at that point it would maximize his expected utility to follow a payment plan. Therefore, if we set t_0 to a time prior to your prediction, then Gauthier's proposal would recommend paying the \$1000. However, the location of t_0 is crucial. For illustration, consider Figure 2.2. Payment is recommended if t_0 is before your prediction, but it is not recommended if t_0 occurs after your prediction. As we noted above, in Parfit's Hitchhiker the best plan after your prediction involves not paying. Therefore, if Gauthier's proposal is to give a different recommendation than straightforward maximization in Newcomb problems, then it is crucial that the point at which plans are evaluated is not located in this post-prediction region. But as it stands now the proposal leaves the point of evaluation for the utility of plans unspecified. We therefore turn to suggestions for specifying this point.

---

[20]I have taken a few liberties with Gauthier's proposal in light of the discussion above. In *Morals by Agreement* Gauthier formulates his theory in terms of an agent's dispositions to act. He claims that if it is rational to adopt a decision making disposition then it is rational to subsequently act in accordance with that disposition. Formulating the theory in terms of dispositions raises problems beyond the ones discussed here. See, e.g., Holly Smith, 1991.

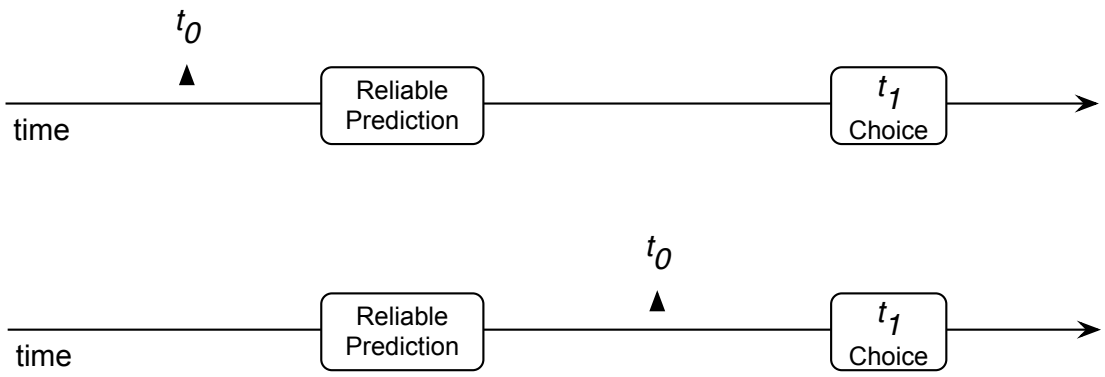

Figure 2.2: Evaluation Points for Plans

2.6.3 Actual Commitment Theories

One suggestion for specifying the location of t_0 is to tie evaluation of the utility of plans to actual commitments or resolutions made by the agent. This is McClennen's proposal in *Rationality and Dynamic Choice*. When one of McClennen's "resolute choosers" determines that a plan p is utility maximizing, he resolves to follow p. When subsequently faced with a decision, the agent ends up "intentionally choosing to act on that resolve" by choosing the act specified by p (1990, 157). I suggest that for our purposes it is best to discard references to the "resolve" of McClennen's agents, and simply focus on the way in which his proposal might amount to a promising alternative to straightforward maximization. In this spirit, we can imagine that a "resolution" amounts to an agent dropping a temporal anchor when he recognizes a utility-maximizing plan. In the future, when the agent has a decision to make, the utility-maximizing status of the action at the time of the anchor dictates what the agent will do. So, in Parfit's example, a resolute chooser who is still temporally located before your prediction may see that a plan of paying has greater expected utility than a plan of not paying. He then resolves to pay, thus dropping the anchor. When the time to choose whether to pay arrives, resolute choice recommends that the agent pay due to the utility-maximizing status of paying at the time of the anchor.

Resolute choosers act just like straightforward maximizers in most situations, but differ in sticking to a utility-maximizing plan when they recognize it. This proposal makes absolutely clear the point at which plans are to be evaluated, and it does sometimes avoid the troubles that Newcomb problems give straightforward maximizers. However, McClennen's resolute choice gives the same recommendations as straightforward maximization in situations where agents are not forewarned of upcoming predictions. This is the case in the classic Newcomb problem, where it is commonly assumed that the agent does not know of the prediction until he is offered the $1000. The agent therefore has no opportunity to make a resolution, and without a prior resolution McClennen's theory recommends that the agent accept the $1000. Such agents are therefore expected to receive bad outcomes in these cases. This is part of a general pattern: decision rules that require an actual precommitment will sometimes produce worse outcomes than ones that do not.[21] Any theory that requires that an agent actually "resolve," "intend," or otherwise "commit" to a plan will fail if an agent is not forewarned of impending predictions.[22]

2.6.4 Hypothetical Commitment Theories

In order to solve this problem, a theory needs to focus on *hypothetical* commitments, rather than actual ones.[23] An example of such a theory is given by Christopher Meacham (2010) In "Binding and Its Consequences," Meacham's main goal is to respond to Arntzenius et al.'s suggestion that casual decision theorists are sometimes punished for being unable to

[21] See Chrisoula Andreou, 2008, 415–22. Andreou calls open Newcomb problems in which the agent is not forewarned of the prediction "Newxin Puzzles."

[22] In "Are Intentions Reasons?" Broome takes himself to be attacking the intention version. He sums it up thusly: "One argument for the claim that intentions are reasons is pragmatic. It has two steps. The first is to demonstrate that, if indeed intentions are reasons, our lives will go better than they would if intentions were not reasons. The second is to argue from there to the conclusion that intentions are reasons" (100). However, as the above argument shows, this theory fails on simply pragmatic grounds

[23] In a footnote, McClennen writes: "I also want to leave open the possibility that even if the agent did not as a matter of fact resolve at some point before n_i to choose in a certain fashion at n_i, still one can consider as relevant to the question of what is to be chosen at n_i what one would have resolved to do at some antecedent point if one had (counterfactually) considered the matter" (1990, 285 fn. 10). However, he does not develop this idea further.

self-bind, but he does formulate, without explicitly endorsing, a non-classical decision theory. He accomplishes this by adjusting the mechanisms of Bayesian decision theory. According to Meacham's "cohesive decision theory," agents should make decisions according to the "comprehensive strategy" (a function that maps every decision problem to one of its available acts) they would choose for themselves from the perspective of their "initial credence function." Meacham imagines that such agents would maximize "cohesive expected utility," which is defined thus:

$$\text{CoEU}(CS) = \sum_i \text{ic}(w_i{:}CS)\text{u}(w_i)$$

Where "ic" is the agent's initial credences and "*CS*" is a comprehensive strategy for selecting acts given decision problems. Meacham uses ":" as a neutral connective between strategies and worlds, leaving it open whether this should be understood causally or evidentially, thus allowing for versions of his theory that resemble either causal or evidential decision theory. Meacham's formulation of expected utility thus builds on the standard formulation [i.e., $\text{EU}(A) = \sum_i \text{cr}(w_i{:}A)\text{u}(w_i)$] by giving control to the agent's initial credences rather than her current credences, and by evaluating plans ("comprehensive strategies") instead of acts.[24] Meacham's theory therefore allows us to conceptualize the cohesive decision theorist as following the plans that would maximize expected utility from the perspective of one's initial credences.

The interpretation of the "initial credence function" is crucial for evaluating the theory. As Meacham notes, we might take it to be an agent's "ur-priors," or alternatively as the initial credences of an ideal subject in the agent's situation (69 fn. 33). Consider first the ur-prior interpretation. Giving control to the ur-priors is like stripping away all of an agent's evidence, and in this way it is similar to asking the agent to figure out what plans would be endorsed from behind a "veil of ignorance." Unlike with Rawlsian theory, however,

[24] As Meacham notes, a more accurate specification of the theory would rely on what an agent currently thinks her initial credences might have been, rather what her initial credences actually were. The leads to the theory being formulated thus:

$$\text{CoEU(CS)} = \sum_i \text{cr}(\text{ic}_i)\sum_j \text{ic}_i(w_j{:}CS)\text{u}(w_j)$$

Where "cr(ic)" is the agent's current credences regarding her initial credences.

the proposal is not that agents select rules for social interaction from behind the veil of ignorance; rather, the proposal is that they select plans to follow for every possible decision situation they might encounter.[25]

Unfortunately, in this form Meacham's theory fails to provide a decision theory that adequately handles the challenges imposed by Newcomb problems. Newcomb problems create temporal problems for straightforward maximizers: in Newcomb problems straightforward maximizers are punished for the future decisions they are likely to make, but once the time comes to make these decisions the punishment has already occurred. In order to avoid the problems created by Newcomb problems, therefore, we need to adjust the *temporal* perspective from which straightforward maximizers make decisions. They must consider not only which acts maximize expected utility in the present, but also which acts it would have maximized expected utility to plan on choosing in the past. Giving control to the ur-priors, however, adjusts the *epistemic* perspective from which plans are evaluated. On this interpretation, cohesive decision theorists do not consider what plans would maximize expected utility in the past, but rather what plans would maximize expected utility were they to lack certain information. This leads the theory to fail.

For example, consider a version of Parfit's Hitchhiker in which your prediction takes place on July 1st, 2013. Let A refer to the set of days prior to July 1st, 2013, and B to the set of days after July 1st, 2013. Let us now consider which comprehensive strategy would

[25]If this were a fruitful way of spelling out Gauthier's proposal it would come as some surprise, since one supposed difference between Gauthier and Rawls is that Gauthier does not require bargainers in the original position to be ignorant. But if the foregoing is correct, then Gauthier's proposal in its most powerful form should indeed require agents to think about what plans would have greatest expected utility from behind a veil of ignorance. Unlike with Rawlsian theory, however, the proposal is not that agents select rules for social interaction from behind the veil of ignorance; rather, the proposal is that they consider decision rules for themselves. Thus, the contractarian is able to retain an excellent answer to the "compliance problem." The compliance problem asks why it is that rational agents should be compelled to comply with the rules they would have agreed to from the original position. On the Rawlsian approach, it is not clear why an agent is compelled to abide by the rules that benefit the disadvantaged, given that she now knows that she is one of the advantaged for which such rules do not benefit. Some believe that this question cannot be adequately answered. For the contractarian, however, the reliabilist metatheory that I develop in Chapter 4 provides a ready answer to the compliance problem: those that follow the decision rules it would be rational to select from behind the veil of ignorance are the ones maximally suited to do well in life, and are therefore rational. See Susan Dimock, 2003, 395–414 for discussion of the standard contractarian reply to the compliance problem on which this answer builds.

maximize expected utility for Parfit from the ur-priors. As we saw above, in a Newcomb problem an agent's temporal location determines whether it is better to plan on paying the $1000. In our example, if Parfit is located in *A*, then the best strategy to adopt would include handing over the $1000. This is, after all, the course of action an agent would "bind" himself to. If, on the other hand, an agent is located in *B*, then the best strategy would include keeping the $1000. It would not make sense, while located in *B*, for Parfit to bind himself to handing over the $1000. Whether cohesive decision theory would recommend handing over the $1000 will therefore depend on the probabilities assigned to *A*- and *B*-location from the ur-priors. It is difficult to think about the probabilities an agent would assign to his possible locations in time if he had no evidence whatsoever. However, given that the best strategy in Newcomb problems crucially depends on these probabilities, we must find some way of determining them if the ur-prior interpretation of cohesive decision theory is to produce a recommended course of action in Newcomb problems.

Problems ensue even if we find a principled way to assign these sorts of locational possibilities an ur-prior probability. Perhaps we can, for example, take the probability that the agent is located in *A* to be given by the sum of the probabilities of all the days located in *A*. This will depend on the total number of possible days an agent could be located, and the probabilities assigned to each day. If the ur-prior probability of being *A*-located falls below a certain threshold (e.g., .001 if the predictor has near-perfect reliability), then cohesive decision theory will recommend that Parfit not pay the $1000. We can make the probability of *A*-location this low by simply moving the day of prediction to near the beginning of the total possible days that an agent could be located. Therefore, according to cohesive decision theory whether Parfit should hand over the $1000 depends on what day the prediction takes place in relation to the total possible days that he could be located. Most of the time the theory will recommend that Parfit hand over the $1000, but if the prediction occurs near the beginning of the period of time that Parfit could be located, then the theory will recommend that he keep it.

Furthermore, the fact that this threshold is so low in this example is simply a product of the large difference between the potential reward ($1,000,000) and the required payment ($1000). If the gap between these two is made narrower—perhaps by making the required payment larger—then cohesive decision theory will recommend not paying at higher probabilities for *A*-location. Even if the probability of being *A*-located is relatively high—say, .5—cohesive decision theory will recommend that Parfit not pay if the required payment exceeds $500,000 (again, imagining that the predictor has near-perfect reliability for convenience).

We therefore see that on the ur-prior interpretation, cohesive decision theory is too sensitive to both the ur-prior probability of being located in the pre-prediction region of Newcomb problems, and the amount of money that an agent has to give up. It is too sensitive because i) it seems to make arbitrary distinctions when providing recommendations, and ii) agents following the theory sometimes suffer the same problems as straightforward maximizers (when the ur-prior probability of *A*-location is low or the required payment is high).

2.6.5 Inceptive Maximization

The way to fix these problems, I believe, is to adapt Meacham's evaluation-from-initial-conditions idea to the temporal framework within which Gauthier and McClennen are working. To do this, we must adopt an explicitly temporal reference point when specifying the perspective from which an agent should evaluate the expected utility of plans. Rather than have agents evaluate plans from the perspective of an initial credence function, I suggest that they should evaluate plans from the perspective of an initial temporal location. In that spirit, we can adjust Gauthier's proposal accordingly:

> *Inceptive Maximization*: If at *the beginning* it maximizes expected utility to follow a plan that involves *A*-ing at some subsequent time t_1, then the agent should *A* at t_1.

As with cohesive decision theory and initial credences, how we determine an initial temporal location ("the beginning") will be important.

2.6.5.1 How to Interpret "the Beginning"

One option is to take "the beginning" to mean the beginning of the agent's life. The crucial question, then, becomes this: if an agent were to consider, at the beginning of his life, all the potential decision problems that he might face, which plans would maximize expected utility (using his present utility function) at that time? Agents following the theory would then choose options in accordance with those plans.

To generalize about these proposals: the crucial difference between straightforward and inceptive maximization is the point at which plans are evaluated. Straightforward maximizers always take the relevant point of evaluation to be the present, while inceptive maximization takes it to be the beginning. It is this feature that makes inceptive maximizers immune to Newcomb problems. Since an inceptive maximizer's actual location in time is irrelevant to her choice of what to do, her ability to reap the rewards of favorable predictions is not blocked by the temporal structure of the situations she faces.

It is important to note that inceptive maximization allows for agents to change what they plan on doing in light of learning new information about their situation. Contrary to actual commitment proposals like that offered by McClennen, such plans can take a conditional form and need only be hypothetical. The agent might reason as follows: "Say, at the beginning of my life, I had imagined that I would face the decision situation I now face. What plan would maximize expected utility then?" In Parfit's Hitchhiker, whether it maximizes expected utility for Parfit to plan to hand over the $1000 from the beginning partly depends on the reliability of the predictor. In a situation in which your reliability is at least slightly above chance,[26] then it would maximize expected utility for Parfit to plan on paying you from the beginning. However, in a situation in which you are not reliable in

[26] As argued above, the chance of a correct prediction need only be about .5005.

predicting Parfit's future actions, it would not maximize expected utility for Parfit to plan to pay, since without reliability in predicting there is no connection between your prediction and what Parfit would choose. Therefore, even after being saved, if Parfit learns that your prediction was not reliable, then he can adjust his assessment of what would maximize expected utility from the beginning. This proposal thus allows for agents to learn about their situation and adjust which option they choose accordingly—just as straightforward maximization does—rather than locking them into irreversible commitments.

If we understand an agent's "initial credence function" in the second way mentioned above—as the initial credences of an ideal subject in the agent's situation—then the theory may produce similar recommendations to inceptive maximization's. This will depend on whether such an ideal agent would be certain of his temporal location. If such an agent would be certain of his temporal location, then cohesive decision theory may indeed produce the same recommendations. However, the crucial element is that the agent consider a change in temporal perspective, rather than merely an epistemic one.

Perhaps a more promising way to interpret "the beginning," though, is as the beginning not of an agent's life, but of everything. This interpretation is supported by the thought that an agent's actual coming into existence is an arbitrary point at which to make distinctions regarding the rationality of his future behavior. For example, if our understanding of the beginning is restricted to an agent's lifetime, it will make a large difference whether a prediction occurs a few moments after one's life starts or a few moments before. This will make the difference between the agent being *A*- or *B*-located. It will therefore make the difference between whether inceptive maximization and the beginning-of-an-ideal-agent's life interpretation of cohesive decision theory recommend that the agent pay the $1000.

A further consideration is that understanding "the beginning" without restrictions allows for agents to receive good outcomes in Newcomb problems that feature predictions that occur before they are alive. Whether this is important will depend on whether one thinks improving on straightforward maximization in these sorts of cases is desired (again, however, it seems arbitrary to hold that it is desired when predictions occur shortly after an

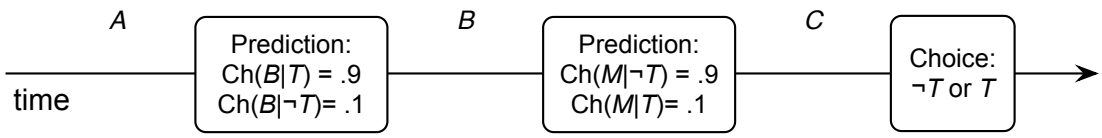

Figure 2.3: A Nested Newcomb Problem

agent comes into existence but not shortly before). This raises some difficult issues regarding the type of practical success that non-classical decision theories should aim to produce, which I discuss more in Chapters 4 and 5.

2.6.5.2 Retrograde Maximization?

It may be suggested that we need not specify an exact temporal point, as inceptive maximization does, but instead adopt *retrograde maximization*: if *at some prior time* it maximized expected utility to follow a plan that involves A-ing at some then subsequent time t_1, then the agent should A at t_1. This theory would handle the cases discussed so far, but it fails in *nested* Newcomb problems. Imagine, for instance, that we add another prediction point to Joyce's Newcomb problem. Just as before there is a rich psychologist who awards you $1,000,000 if he predicts that you will turn down $1000 at a future point. However, prior to this, a more kind-hearted psychologist has also made a prediction regarding whether you will turn down the $1000, but she has decided to award you $1,000,000,000 if she predicts that you will *accept* the $1000. Let us imagine that both psychologists are 90% reliable for these sorts of predictions. The case is then depicted by Figure 2.3, where, as before, T stands for "the agent chooses to keep $1000 " and M stands for "the agent receives $1,000,000," and adding B: "the agent receives $1,000,000,000." While you are located in C, planning on accepting the $1000 maximizes your expected utility, but while you are located in B, planning on refusing the $1000 maximizes your expected utility. However, while in A it again maximizes expected utility to accept the $1000. Therefore, there is a

prior time at which it maximized expected utility to follow a plan that involved refusing the \$1000 (*B*), and also a prior time in which it maximized expected utility to follow a plan that involved accepting the \$1000 (*A*). Retrograde maximization thus fails to produce a coherent recommendation. However, in this case refusing the \$1000 is plainly irrational: those who refuse it tend to end up many millions poorer, and neither causal nor evidential decision theory recommends doing so.

The nesting problem shows the need to evaluate plans from as far back as possible. Only by doing so are we able to account for the gains and losses created by all reliable predictions of one's future decisions. Therefore, if an agent is to be successful in Newcomb problems, then he must consider the plans that would maximize expected utility from a point prior to any predictions of his future behavior that may be relevant to his current decision. Inceptive maximization accomplishes this by placing the point of evaluation at the beginning.

2.6.5.3 The Medical Newcomb Requirement

Both causal decision theory and inceptive maximization are able to satisfy the dominance requirement. Causal decision theory recommends accepting the \$1000 in both open and closed Newcomb problems, while inceptive maximization recommends refusing it in both. What about the medical Newcomb requirement? By adopting a causal understanding of expected utility, inceptive maximization satisfies the medical Newcomb requirement as well. At any time prior to the point at which the lesion may form, no-smoking plans do not maximize causal expected utility. Imagine, for instance, that an agent is temporally located prior to the point at which the lesion might form. At this point it would not make sense for the agent to bind herself to not smoking, since doing so has no effect on the chance of the lesion forming. This is because random chance—rather than a reliable prediction—determines whether the lesion forms. This factor explains why refusing to smoke is irrational on any reasonable theory, including inceptive maximization. It is different with Newcomb problems that do involve predictions. Prior to a prediction, an agent is wise to bind herself to certain plans, since by doing so she can indeed influence what is predicted.

Nozick once challenged two-boxers to explain why the dominated choice in a medical Newcomb problem like The Smoking Lesion is obviously irrational, while the dominated choice in standard Newcomb problems is not.[27] The idea that a rational agent should inceptively maximize causal expected utility answers this challenge in a bold way: refusing to smoke is irrational, but refusing the $1000 is not.

[27] See Nozick, 1969, 135.

Chapter 3
Time-Biased Preferences

In the previous chapter we examined some of the practical problems associated with a straightforwardly maximizing conception of rationality. We saw that the problems are generated in particular by straightforward maximization's commitment to *present-directedness*—the idea that an agent should choose the options that are best in light of their present perspective. According to the alternative—inceptive maximization—agents should not privilege the perspective of the present in deciding what to do. It is this feature of inceptive maximization, I argued, that allows inceptive maximizers to avoid the problems faced by straightforward maximizers.

In this chapter, I argue for a similar conclusion regarding rational preferences. Most of us display a *bias toward the near*: we prefer pleasurable experiences to be in our present or near future and painful experiences to be in our distant future. We also display a *bias toward the future*: we prefer pleasurable experiences to be in our present or future and painful experiences to be in our past. Among philosophers who discuss the issue, there is general agreement that near bias is a rational defect (assuming there are any non-structural rational requirements on preferences). But almost no one finds future bias objectionable. On the contrary, some philosophers take a theory to be refuted if it conflicts with the rationality of future bias. In this paper, I show that there is an argument against future bias parallel to the usual argument against near bias. I conclude that those who reject near bias should reject future bias as well, and thus endorse complete temporal neutrality.

3.1 Three Views on the Rationality of Time-Biased Preferences

We can distinguish between two prominent accounts of the rationality of time-biased preferences. First, there is the economic view, which holds that both near and future bias are permissible. Second, there is the philosophical view, which holds that future bias is permissible but near bias is impermissible. A third option is that rationality requires complete temporal neutrality: both near and future bias are impermissible. In this section I introduce each of these views in turn.

3.1.1 The Economic View

Standard theories of rational preference in decision theory and economics are primarily concerned with preference coherence. While not all candidate constraints are uncontroversial, it is thought that consistency requires, for example, that one's preferences be transitively ordered—if one prefers A to B, and B to C, then one also prefers A to C—and that they be complete, in the sense that any two outcomes are comparable. Otherwise, such theories take a laissez-faire attitude to nonstructural features (the content) of preferences. They therefore allow agents to be *biased toward the near*. All else equal, agents who are biased toward the near prefer pleasurable experiences to be in the near future rather than the distant future, and painful experiences to be in the distant future rather than the near future. Such agents may also be biased toward the near when all else is not equal, sometimes choosing less pleasurable but nearer experiences to more pleasurable but more distant ones, and more painful but more distant experiences to less painful but nearer ones.

In order to capture these phenomena, appeal is made to near-biased agents' engagement in *temporal discounting*. The preferences of pure temporal discounters are affected by a discount rate applied to future experiences, in which the discount factor increases as the experience becomes more temporally distant.[1] It is common to distinguish exponential

[1] As John Broome (1994) argues, there is an important difference between "pure" discounting, in which well-being is discounted, and the discounting of commodities used by economists in cost-benefit analysis. The discussion here focuses exclusively on the rationality of pure discounting.

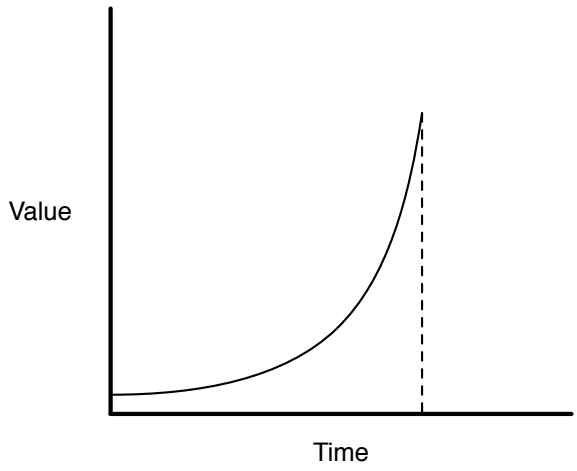

Figure 3.1: Exponential Discount Curve

discounters from hyperbolic ones. Let V_t be the value of an experience when you will receive it, i be the interval of time between now and when you will receive it, and D the discount rate. For exponential discounters the calculation of the present value of an experience is done using the following formula:

$$\text{Exponential Discounting: } V_{now} = V_t/(1 + D)^i$$

This discount function is called "exponential" because it calculates value by an exponential function of the discount rate. This produces discount curves with pleasantly uniform characteristics—as seen in Figure 3.1, where the dotted line represents V_{t} and the solid line represents V_{now}.

Experiments, as well as intuition, suggest that most people are not exponential discounters, but rather *hyperbolic discounters.*[2] Hyperbolic discounters are more "present-biased" than exponential discounters—for hyperbolic discounters the value of a reward increases more significantly as it gets close to the present. A hyperbolic discounter might, for example, prefer a $100 check that can be cashed immediately to a $200 check that cannot be cashed for three years, while at the same time preferring a $100 check in six years to a $200

[2]See, e.g., Ainslie, 2001.

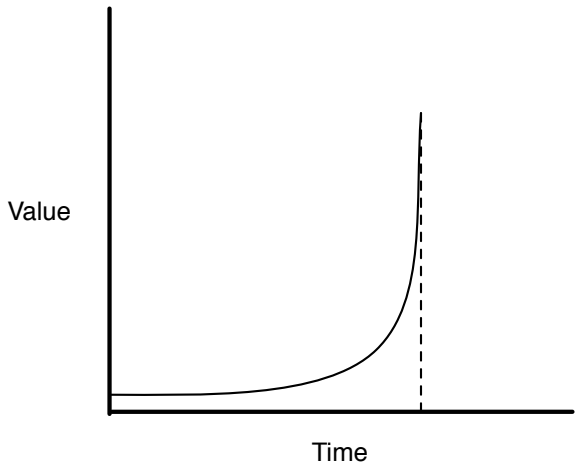

Figure 3.2: Hyperbolic Discount Curve

check in nine years.[3] Since in each instance the delay is three years, hyperbolic discounters must be affected by whether or not the monetary reward is close to the present, rather than (as with exponential discounters) solely by the amount of delay. Here is an example of a hyperbolic formula:

$$\text{Hyperbolic Discounting: } V_{now} = V_t(1/(1 + Di))$$

This equation produces discount curves that are more bowed; the perceived value of an experience increases more dramatically when it gets close to the present. This is represented by Figure 3.2.

An interesting feature of hyperbolic discounting is its *dynamic inconsistency.* In certain situations the discount curves will "cross," making a hyperbolic discounter's preferences over the same outcomes change over time. This is suggested by the example above. A hyperbolic discounter may originally have a preference for the $200 check in nine years over the $100 check in six years. However, as time passes, and the $100 reward gets closer to the present, a hyperbolic discounter may come to prefer the $100 check over the $200 check. Exponential discounters, on the other hand, are not present biased, and this

[3]The example is from Ainslie, 2001, 33.

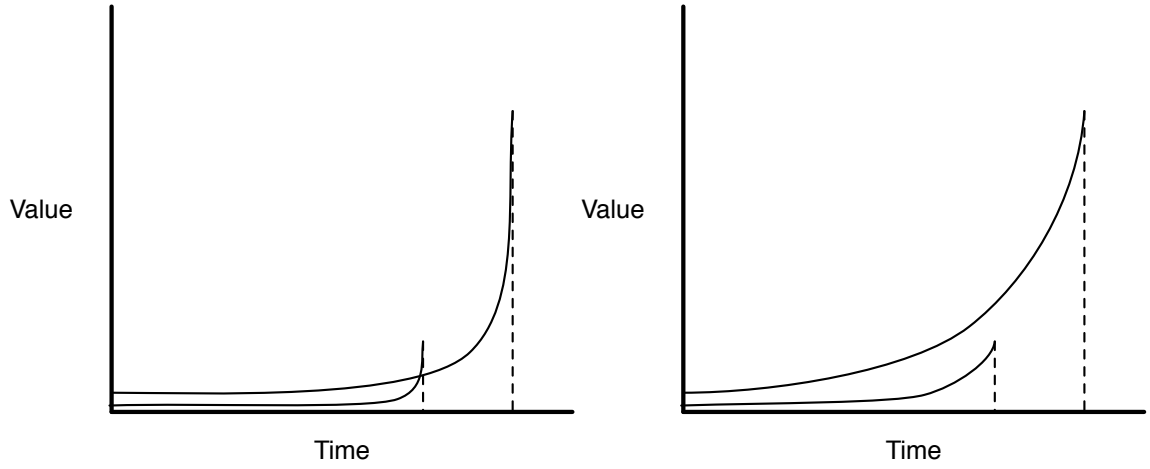

Figure 3.3: Hyperbolic Inconsistency (left) vs. Exponential Consistency (right)

makes them dynamically consistent: their preference for one reward over another does not depend on the distance between themselves and the closer reward. Their preference for one check over the other will therefore stay constant as time passes. Figure 3.3 illustrates these features of hyperbolic and exponential discounting.

In line with their emphasis on consistency, those endorsing a standard economic approach to rational preference often accept that dynamically-inconsistent discounting is irrational. We thus arrive at the *economic view* of time-biased preferences, according to which any dynamically consistent form of temporal discounting is permissible.

3.1.2 The Philosophical View

Among philosophers who have considered the issue, many conclude that near bias, even in its dynamically consistent forms, is unjustifiable. For example, Sidgwick (1884, 380), in his discussion of the various formulations of his impartiality principle, urges: "The mere difference of priority and posteriority in time is not a reasonable ground for having more regard to the consciousness of one moment than to that of another. The form in which it practically presents itself to most men is 'that a smaller present good is not to be preferred

to a greater future good' (allowing for difference in certainty)." Rawls reiterates the point in *A Theory of Justice*: "Rationality requires an impartial concern for all parts of our life. The mere difference of location in time, of something's being earlier or later, is not a rational ground for having more or less regard for it." In line with Sidgwick's caveat, he continues:

> Of course, a present or near future advantage may be counted more heavily on account of its greater certainty or probability, and we should take into consideration how our situation and capacity for particular enjoyments will change. But none of these things justifies our preferring a lesser present to a greater future good simply because of its nearer temporal position. (1971, 293–4)

Distant experiences are typically less certain than near ones, and this gives us a reason to prefer near pleasures and distant pains over distant pleasures and near pains. However, there is no reason to prefer certain experiences based solely on their temporal position.

It is therefore part of what we might call the *philosophical view* that agents should not be biased toward the near.[4] There exists an enduring challenge for those that attempt to defend the philosophical view solely on the basis that, intuitively, near-biased agents seem to make arbitrary distinctions. The challenge is to explain why the arguments in favor of temporal neutrality do not also apply equally in favor of agent neutrality. Thomas Nagel (1970) famously relies on the apparent parity of the considerations in favor of temporal neutrality and those in favor of agent neutrality in order to argue against the rationality of agent bias. Derek Parfit (1984, Part II) stops short of endorsing Nagel's conclusion, but he does go to great lengths to show the difficulty in finding a coherent justification for temporal neutrality that does not also apply to agent neutrality.

The challenge does not necessarily assume that we are not required to be agent neutral. Rather, the worry starts with the thought that agent bias is not based on obviously arbitrary distinctions, and the rejection of it is considered to be a weighty philosophical hypothesis. Given this, presumably the considerations in favor of temporal neutrality are supposed to

[4]Other examples of philosophers endorsing this view include Lewis (1946, 493), Nagel (1970), Elster (1986), and Broome (1991).

be more obvious than those in favor of agent neutrality, but how is this so? According to Nagel, the considerations used to support the arbitrariness of time bias apply equally to agent bias.

The best answer to this question, I believe, invokes what David Brink (2010, 360–6) calls "compensation." In line with Sidgwick, it is common to feel that temporal neutrality is characteristic of prudence; specifically, in the farsighted trade of smaller present goods for greater future goods when the probabilities are right. Prudence is something that children need to learn—as famously demonstrated by Walter Mischel's experiment, in which many children fail to resist eating one marshmallow now to gain a second one a few minutes later. On the economic view, implausibly, these children are simply agents who discount the future at steeper rates than the rest of us. Less extreme future discounting may explain why many adults fail to save enough for retirement, persist in smoking, and neglect to get minor dental problems corrected before they become root canals.

When one makes prudent sacrifices one is usually compensated for the loss later. Even though an agent neutralist like Nagel may claim that our concern for our own welfare over that of others is arbitrary, he cannot claim that the distinction between one individual and another is similarly arbitrary. The separateness of persons does seem to suggest that the considerations in favor of temporal neutrality are indeed more obvious than those in favor of agent neutrality. Prudent agents who sacrifice smaller near rewards for larger distant ones are compensated for their sacrifice when their investment pays off, and these investments are worthwhile if the chances of receiving compensation are appropriate. However, agents who sacrifice their own smaller rewards to provide larger rewards for others do not (in many instances) have any chance of compensation. Such trade-offs are therefore not investments but rather pure costs.

In order to provide some rationale for thinking that the argument in favor of temporal neutrality is more obvious than that in favor of agent neutrality, we must therefore buttress intuitive considerations about the arbitrariness of near bias with facts about the practical upshot of temporal neutrality *for the individuals making the sacrifices.* This is how it should

be. The allure of prudence, I suggest, is not that prudent people steer clear of making arbitrary distinctions. Rather, it is that prudent people make choices that result in their leading better lives (featuring *inter alia,* cushy retirements, better health, and extra marshmallows). This is the difference between temporal and agent neutrality that the compensation argument utilizes.

Finally, any theory that posits ubiquitous irrationality requires a corresponding error theory. An error theory for near bias would explain why we do in fact engage in, or are tempted to engage in, temporal discounting. As we will see in Section 3, thinking about error theories provides a further way to distinguish near bias from agent bias, since the most plausible error theories for near bias do not obviously apply to agent bias.

3.1.3 The View that Rationality Requires Complete Temporal Neutrality

Consider another common time bias: the bias toward the future. Near-biased agents prefer pleasurable experiences to be in the near future, but future-biased agents prefer pleasurable experiences to be in the future *simpliciter* (rather than the past). More generally, future-biased agents prefer future pleasurable experiences to past ones, and past painful experiences to future ones.

Unlike near bias, many philosophers assume that future bias is at least rationally permissible, if not rationally required.[5] One common thought is that future bias is strikingly self-evident. As Christopher Heathwood (2008, 56–7) forcefully puts it: "[A future-biased agent] is being completely reasonable in preferring that his pain be in the past. In fact, even

[5]For example, Prior (1959) and Hare (2009) assume it is rationally permissible. Other philosophers build future bias into their normative theory. For instance, Moore (1942) and Bergstrom (1966) define normatively relevant consequences to include only events in the future of an act. Agents applying theories to decision-making are thus obliged to take into account only the future costs and benefits of their acts. Brueckner and Fischer (1986) argue that the rationality of future bias shows that our prenatal nonexistence is not as bad as our postmortem nonexistence. Heathwood (2008) argues that fitting attitude theories of welfare are refuted by the fact that they do not account for the rationality of future bias. Parfit (1984, Part 2) remains neutral regarding its permissibility.

his no longer caring at all that it occurred is perfectly fitting—not at all inappropriate. Why should he care about it now? No reason—it's over and done with." In line with this, to complete the philosophical view we must add an endorsement of future bias to its rejection of near bias. We thus see that the view is a hybrid one even within the temporal domain. It is temporally neutral in relation to the near and distant future, but it is temporally biased in relation to the future and the past.

While the common concern over the philosophical view's hybridity in relation to time and agent neutrality can be overcome, I believe the real cause for concern lies in its hybrid structure in relation to near and future bias. Perhaps the sentiment conveyed by Heathwood is correct: it is more obviously arbitrary to distinguish between near and distant future experiences than it is to distinguish between past and future experiences. However, as we saw, the argument against near bias does not rest on these sorts of intuitions, but rather must crucially rely on considerations of practical upshot as featured by the compensation argument. Similar considerations can be marshaled against future bias (Section 2). Furthermore, the most compelling error theories for near bias also apply to future bias (Section 3). Given this, we have reason to think that the most compelling alternative to the economic view is complete temporal neutrality.

3.2 Future Bias and Rational Planning

Near-biased agents are imprudent planners: because they sacrifice better distant experiences for worse near ones, they end up leading worse lives than they would otherwise. Does future bias also make for imprudence? The first thought is that future bias could not possibly affect an agent's planning. This is because future bias is different from near bias in that it does not directly vindicate any relevant tradeoffs. We can trade between the near and the distant future, but not between the future and the past. It thus appears that future bias, unlike near bias, does not have a direct effect on an agent's practical reasoning.

However, future bias may still have an indirect effect on practical reasoning, through

its interaction with other principles of rational planning. As we will see, the interaction of future bias with other plausible principles places future-biased agents in strange, costly, and avoidable predicaments. Because future bias does not directly license any relevant trade-offs, but rather does so only indirectly through its interaction with other potential practical norms, we always have the option of abandoning the other norms in order to save it. However, whether we should do this depends on whether the justification for future bias is stronger than that of the principles with which it is incompatible. Let us therefore inquire into the principles with which future bias is incompatible.

3.2.1 Risk Aversion

In an interesting essay, Tom Dougherty (2011) argues that serious problems are created by the interaction of future bias and risk aversion. Specifically, Dougherty attempts to show that future-biased agents who are risk averse can be turned into "pain pumps": they would accept a series of trades that guarantees they will suffer more pain overall and be better off in no respect. Since risk aversion is often taken to be rationally permissible, if Dougherty's argument succeeds this would be a major challenge to the rationality of future bias. However, we shall see that Dougherty's argument is problematic because he relies on a nonstandard notion of risk aversion.

As Dougherty presents it, risk-averse agents are always willing to sacrifice some amount of expected value in order to reduce the amount of risk they face. Dougherty captures this idea with the following principle:

> *Every Risk Reduction Has Its Price*: If a robustly risk-averse person faces a fifty-fifty gamble, then for any reduction of the gap between the good and bad outcomes of the gamble, there is some decrease in the gamble's expected value that this person would accept in return for this reduction of the gap. (2011, 525)

In order to illustrate this, Dougherty starts by noting that a risk-averse agent would be

willing to exchange a ticket that has an equal chance of paying $20 or $0 for a ticket that pays $10 with certainty. In this case, such an agent would have received a reduction in the gap between the good and bad outcomes for free—both tickets have an expected value of $10. However, a risk-averse agent might also be willing to exchange the first ticket for one of *lower* expected value; e.g., a ticket with an equal chance of paying $8.50 or $10.50. Such a risk-averse agent would thus accept a $.50 reduction in the expected value of their ticket in order to reduce the gap between the positive and negative outcomes.

With this characterization of risk aversion in place, Dougherty presents a puzzle for agents who are risk averse and future biased that aims to show that they can be turned into pain pumps. Here is a variant of Dougherty's case:

> *Blue Pills and Red Pills*:
>
> You know that you are facing one of two equally probable situations. If you are in Situation *A*, you will experience four hours of pain on Tuesday and two hours of pain on Thursday. If you are in Situation *B*, you will just experience three hours of pain on Thursday. You do not know which situation you are in, though you always know what day it is and you always know you will be offered the following choices.
>
> On Monday you are to be offered a blue pill with the following properties: if you take the pill and are in Situation *A*, it reduces the time you suffer on Thursday by one hour. If you are in Situation *B*, it increases the length of your pain on Thursday by one and a half hours. You are risk averse, so you take the pill, intending to decrease the overall potential suffering of the worst situation: Situation *A*.
>
> On Wednesday, because you have mild retrograde amnesia, you are still not sure what situation you are in. You are offered a red pill with the following properties: if you take the pill and are in Situation *A*, it increases the length of pain you suffer on Thursday by one and a half hours. If you are in Situation

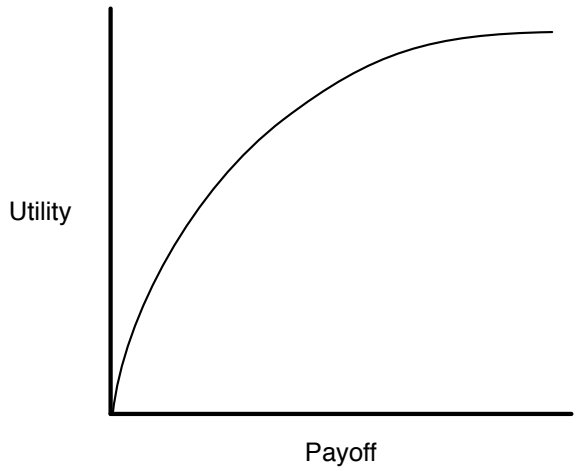

Figure 3.4: Risk Aversion for Potential Payoffs

> *B*, it decreases the length of pain you suffer on Thursday by one hour. You are future biased, and so from your standpoint the only pains that are relevant are Thursday's. And you are risk averse, so you want the worst Thursday scenario—which is now Situation *B*—to be improved. So you also take the red pill.

Because you took both pills, you are guaranteed to suffer an extra half hour of pain. If you are in situation *A*, you reduced your suffering by an hour (blue pill) but then increased it by an hour and a half (red pill). If you are in situation B, you increased your suffering by an hour and a half (blue pill) but then only decreased it by an hour (red pill). Your preferences have turned you into a pain pump.

Dougherty's argument is weakened by the fact that there is significant room to complain about the rationality of the sort of risk aversion on which he relies. Dougherty claims that his formulation is meant to be neutral between different decision-theoretic characterizations of risk aversion, but I believe it is actually incompatible with the most common characterization. According to standard rational choice theory, rational risk aversion in relation to some good requires the good to have diminishing marginal utility. Monetary

payoffs, it is thought, do indeed have diminishing marginal utility for most people. This phenomenon is usually represented by a concave utility function, as seen in Figure 3.4.

Such a phenomenon would explain the cases concerning money with which Dougherty starts, such as why a risk-averse agent would choose to exchange a 50-50 gamble between $20 and $0 for one that pays $10 for certain. Because of the diminishing marginal utility of money, the agent takes the utility of $20 to be less than twice as great as that of $10. Therefore, according to the most common approach to risk aversion, the rationality (and the ubiquitousness) of risk aversion in relation to money is explained as a rational response to money's diminishing marginal utility.

But Dougherty's patient is trading in pain, rather than money. Does pain have diminishing marginal utility? There is reason to doubt this. While the effects of a painful stimulus might dull as it is applied more frequently, an increase in actual units of pain buys you exactly that amount of experienced pain. Therefore, for any size units, two units of actual experienced pain is twice as bad as one unit of actual experienced pain, all else equal. It is different with money: for most of us, as a potential payoff increases, more units of money do not buy equivalent gains in pleasures, or equivalent reductions in pains. Nevertheless, *these experiences themselves* do not have diminishing marginal utility.

We might call risk-aversion to experiences "pure" risk aversion. If the foregoing is correct, then pure risk aversion must be motivated without appeal to diminishing marginal utility. This will require a nonstandard framework. One example of an attempt to provide such a framework—which does seem compatible with Dougherty's argument—is developed by Lara Buchak (ms).

3.2.2 Regret

Even though Dougherty's argument turns out to be limited, his focus on the preference instability created by future bias does point us in the right direction. Preference instability causes a host of problems for agents reasoning about what to do. A particularly acute problem arises from the temporary reversals in preference that might occur when an agent

is offered a vivid presentation of some tempting option. Michael Bratman provides the following example of such a case:

> *The Second Pilsner*:
>
> Consider Ann. She enjoys a good read after dinner but also loves fine beer at dinner. However, she knows that if she has more than one beer at dinner she cannot concentrate on her book after dinner. Prior to dinner Ann prefers an evening of one beer plus a good book to an evening with more than one beer but no book. Her problem, though, is that each evening at dinner, having drunk her first Pilsner Urquell, she finds herself tempted by the thought of a second: For a short period of time she prefers a second beer to her after-dinner read. This new preference is not experienced by her as compulsive. If asked, she would say that right now she really prefers to go ahead this one time and have the second drink, though she will also acknowledge that even now she prefers that she resist similar temptations on future nights. As she knows all along, this change in ranking will be short-lived: after dinner she will return to her preference for a good read. (1999, 74)

According to one view of instrumental rationality, an agent should act on whatever she most prefers at the time of choice. Ann really does prefer to have the second drink during dinner. Given this, it seems instrumental rationality dooms Ann to an unfortunate pattern of behavior: she will drink too much during dinner night after night, thereby never enjoying her book. But Ann is able to anticipate this, and she realizes that her preference for a two-beer evening will be short-lived. We might, therefore, think that she is not doomed after all. Ann just needs a way to factor temporary reversal of preferences into her rational planning.

Bratman takes cases like this to show the need for a "no regrets" requirement on rational preferences. According to Bratman, one should be committed to "taking seriously how one will see matters at the conclusion of one's plan," and thus be committed to avoiding

choices that one expects to regret (86).[6] I believe the case for a no regrets condition on rational decision making is strongest when relativized to cases in which agents have full and accurate information about the effects of the various options available to them and are able to predict with certainty their future preferences.[7] Thus, we might adopt the following principle:

> **Weak No Regrets:** If an agent has an option she foresees with certainty she will never regret, then it is irrational for her to choose an option she foresees with certainty that she will regret.[8]

This may appear to be a weak and intuitive rational constraint, but a caveat is required.

With Bratman, we should remain committed to working within a theory of instrumental rationality—a theory of what rational agents will do given their aims. There are plenty of non-instrumental reasons why an agent might choose a path of certain regret over one of certain non-regret. For example:

> *Resisting Indifference:*
>
> Jane knows she is about to develop a brain disease that will cause her to be completely indifferent to every decision. In order to prevent the disease she

---

[6]Frank Arntzenius (2008, 277) develops a similar principle, namely that "a rational person should not be able to foresee that she will regret her decisions," in order to motivate his account of rational decision-making. See also Nagel (1970, Chap 6), Rawls (1971, 421–3), Loomes and Sugden (1982), and Bratman (2006, section 8).

[7]We therefore should restrict ourselves to what Bratman calls "no-unanticipated-information cases" (1999, 79).

[8]We need agents to have at least one foreseen option that they will not regret because in cases where agents are missing crucial information relevant to their preferences, the rational thing to do might be to accept foreseen regret. For example, consider Parfit's miner case:

> *Miners*:
> Ten miners are trapped either in shaft *A* or shaft *B,* but we do not know which. Flood waters threaten to flood the shafts. We have enough sandbags to block one shaft, but not both. If we block one shaft, all the water will go into the other shaft, killing any miners inside it. If we block neither shaft, both shafts will fill halfway with water, and just one miner, the lowest in the shaft, will be killed. (Kolodny and MacFarlane, 2010, 115)

In the miner case, an unqualified no regrets condition would ban flooding both shafts, because you will for certain prefer you had chosen otherwise once you learn what shaft the miners are in. Still, flooding both shafts seems to be the rational move given your lack of information.

> must choose between two medications. The first medication prevents the disease in only a small fraction of patients. The second medication is certain to prevent the disease, but will make her a somewhat more reckless and akratic decision-maker. She will not know what the effects of the first medication would have been until many years from now.

In this case, the all-things-considered best thing to do might be to take the second medication. But this is only because it might seem that we have overriding non-instrumental reasons to prefer a life with some mistakes to a life of indifference.[9] Therefore, weak no regrets seems most plausible in cases where the agent does not have any non-instrumental reasons to criticize her future preferences. For example:

> *Surf or Turf:*
>
> Right now Lucy prefers a juicy steak to fresh lobster. But in March she is taking a cruise across the Atlantic, and she foresees with certainty that once she is at sea, she will prefer lobster to steak. She must choose now whether to buy the Turf meal plan for her voyage (which primarily features steak) or the Surf meal plan (which primarily features lobster).

Lucy ought to buy the Surf plan. Before March, her plan will have no effect whatsoever on her steak preferences. And while at sea, her lobster preferences will be satisfied. At no time will she regret her choice. It would be irrational for her to instead choose based on her current preference for steak.

Returning to Bratman's beer case, Ann the akratic drinker has no overriding non-instrumental reason to force herself into a two-beer evening. Given that she knows that very shortly she will return to her original preference for a one-beer dinner, is it rational for her to forego the second beer, and thus avoid the problems associated with this kind of temporary preference change? It seems so. Clearly, weak no regrets will not rule out all

---

[9] Cf. Parfit (1984, 327)'s Russian nobleman.

the problems that temporary preference change causes, but if there is any hope for the idea that being practically rational allows us to avoid such problems, then weak no regrets is a relatively uncontroversial starting place.

3.2.2.1 Weak No Regrets and Future Bias

Why is the weak no regrets principle relevant to thinking about future bias? For future-biased agents, reversals in preference are caused simply by the passing of time. In fact, true future-biased agents exhibit very strange patterns of regret. Consider the following case:

> *Fine Dining*:
>
> Jack wins a free meal at a fancy French restaurant on Monday morning, and he must schedule the meal for a night sometime in the next week. Given his flexible schedule, every night is equally convenient for him, and there are no other considerations that would make the meal more enjoyable on one night rather than another. Therefore, Jack schedules the meal for Monday night. As expected, it is an incredibly delicious meal. On Tuesday morning, Jack strongly prefers that his restaurant experience were in the future, rather than the past. And so he regrets scheduling the meal the previous night.

According to defenders of future bias, Jack's regret—as above, in the sense that on Tuesday he prefers that he had chosen a different option on Monday—is rational. At the time of his decision each choice available to him (the seven days in which the meal could be scheduled) involves the prospect of a future pleasurable experience. But on Tuesday, only six of the seven options includes a future meal. Therefore, Jack prefers that he had not scheduled the meal for Monday. In fact, unless Jack schedules the meal for the last possible day allowed by the restaurant, he is assured that he will come to regret his choice. Furthermore, in cases in which there is no deadline, it is not clear when Jack should schedule the meal. If we abstract away from concerns over Jack dying, losing his sense of taste, the restaurant

closing, etc., then it appears that future bias and weak no regrets require Jack to wait a very long time to have his meal. Imagine that every day Jack is asked whether he would like to schedule the meal the following day. Since Jack knows that he will come to regret scheduling it, he chooses not to schedule it each time. If Jack is unlucky enough to live forever, he may find that he never has the meal. This creates a paradox similar to that faced by agents who are constantly offered positive rates of return on their savings.[10] Call this the s*cheduling problem for future bias*.

It is worth noting that we can get to the scheduling problem from a related, but perhaps even weaker starting point. Suppose you endorse the following principle governing rational planning:

> **Weak Meta-Preference Principle:** An instrumentally rational agent will prefer a plan in which all of his present and future preferences are satisfied over a plan in which some are not.

The principle should seem intuitive: If given the option, you ought to make sure that your present and future selves have all of their preferences satisfied rather than merely a subset. In Fine Dining, assuming you are genuinely indifferent to each particular day, future bias and the weak meta-preference principle will require you to schedule the meal as late as possible. The maximally deferred meal is the only plan that could satisfy all of your preferences. But maximally deferring good experiences is absurd.

The scheduling problem is easy to miss, I believe, because it does not apply to agents who are both future biased *and* near biased. In line with the philosophical view, however, we are assuming agents who are future biased but temporally neutral regarding the near and distant future.

The problems with weak no regrets and future bias run deeper than the scheduling problem. Their combination can in fact lead to large deductions in an agent's overall happiness.

---

[10]Koopmans (1967) calls this "the paradox of the indefinitely postponed splurge." Also compare the problem to Arntzenius et al.'s *Trumped* (2004, 252). In this case, however, an agent's wealth (or, as in *Trumped*, allotted days in heaven) does not increase through time, and so it may seem even more problematic.

Consider Billy, who is offered a choice between two cookies immediately or one cookie at some point in the future. It would seem that the rational choice is clear: Billy should choose to have two cookies now. But for future-biased agents the answer is not so simple.

Assume that Billy is *absolutely future biased*: An absolutely future-biased agent assigns no value whatsoever to past experiences.[11] He must now decide between more cookies sooner or fewer cookies later. Since he is absolutely future biased, as soon as the cookies are consumed he will regret not having chosen to have fewer cookies. During the time immediately following the cookie-consumption, Billy regrets his choice. When the time at which he would have eaten one cookie passes—so that it is now also in the past—he becomes indifferent to his choice. Therefore, Billy can expect to regret choosing more cookies now, and to at no future time regret choosing fewer cookies. So the only course of action that Billy will not come to regret is choosing to wait for fewer cookies. In fact, for an absolutely future-biased agent it does not matter how many more cookies are available earlier. As long as Billy is interested in the later cookies at all, then he will be forced to choose them or end up regretting his choice. We can imagine a case, for example, in which Billy is choosing between having ten cookies now or having just a morsel of cookie later. Again, he will come to regret choosing to have the ten cookies, but he will not come to regret choosing to have the later morsel. Therefore, insofar as they accept even a weak no regrets condition on rational agency, agents with an absolute future bias will be willing to trade the most spectacular of pleasurable experiences for extremely mundane ones in cases like Billy's.[12] Call this the *meager returns problem for future bias.*

---

[11] Such an agent may of course assign value to present or future memories of those experiences. I will assume that this factor is not relevant in the case of Billy for simplicity.

[12] The discussion of Billy's case here and below focuses solely on how the temporal locations of potential cookie-eating experiences affect Billy's preferences. There may, of course, be other factors that have an affect on Billy's preferences. For example, Billy might also prefer to be the sort of person that has chosen to eat more cookies in the past. If this preference outweighs the affects of future bias then Billy may avoid choosing fewer cookies. However, there is no guarantee that such outweighing factors will always be present to mask the problems that future bias creates.

3.2.2.2 Non-Absolute Future Bias to the Rescue?

In order to avoid these problems, it might be suggested that we consider a non-absolute future bias. After all, it may be reasonable to think that many people assign some importance to past pleasures and pains, even if it is less than that assigned to future ones.[13] Perhaps, for example, a non-absolute future-biased agent might discount past pleasures 50% in comparison to future ones. This suggestion immediately fails, however, to avoid Billy's problem. In order to accommodate non-absolute future-biased agents, we simply need to adjust the difference in value between the two experiences. While an agent who discounts past experiences 50% may not regret choosing three cookies now over one cookie in the future, he will regret choosing three cookies now over *two* cookies in the future. Similar reasoning shows that regardless of the discount factor used, there will exist a case in which the agent comes to regret choosing more cookies.

As we have seen, when we discount the future we typically do not do it absolutely—we apply a discount rate. What happens if we therefore reject absolute future bias by adopting a discount rate to past experiences? As in the case of future discounting, we can imagine past discounting functions that are either exponential or hyperbolic. In either case past discounters do not avoid the scheduling problem. Figure 3.5 shows how Jack's preferences for two potential meals might evolve in *Fine Dining* if he applies a past discounting function.

If Jack schedules the meal for Tuesday, then as soon as the meal becomes past it will start to be discounted. Given that the meal is equally valuable whether it occurs Tuesday or Thursday, Jack will immediately prefer that he had scheduled the meal for Thursday, and will never prefer that he had scheduled it for Tuesday. Therefore, both weak no regrets and the weak meta-preference principle will demand that Jack schedule the meal for Thursday. Further applications of this argument show that Jack must schedule the meal on the last day possible.

It does appear, however, that future-biased agents who apply past discount functions

[13] Again, however, the importance must not be tied to the expectation of pleasant or unpleasant future memories of the experiences.

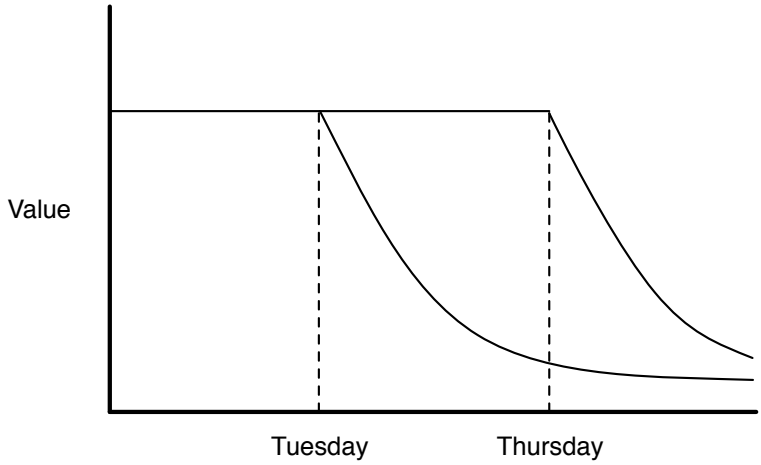

Figure 3.5: Past Discounting in *Fine Dining*

avoid the meager returns problem. If Billy applies a past discount function, then after he consumes the larger number of cookies there will be an interval—however brief—in which he prefers the past cookie experience to the future one. Due to this, no matter what Billy chooses, there will be future times in which he prefers more cookies in the past.

Therefore, exponential and hyperbolic past discounters do avoid the meager returns problem, even if they still suffer from the scheduling problem. However, this form of past discounting is deeply counterintuitive. To see this, consider one of the most celebrated motivations for future bias: Parfit's *My Past or Future Operations*:

> I am in some hospital, to have some kind of surgery. Since this is completely safe, and always successful, I have no fears about the effects. The surgery may be brief, or it may instead take a long time. Because I have to co-operate with the surgeon, I cannot have anaesthetics. I have had this surgery once before, and I can remember how painful it is. Under the new policy, because the operation is so painful, patients are now afterwards made to forget it. Some drug removes their memories of the last few hours.
>
> I have just woken up. I cannot remember going to sleep. I ask my nurse if it

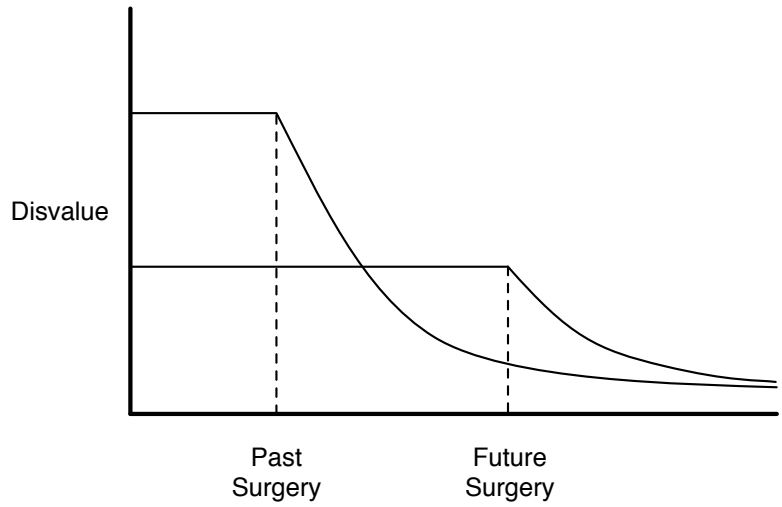

Figure 3.6: Past Discounting in *My Past or Future Operations*

> has been decided when my operation is to be, and how long it must take. She says that she knows the facts about both me and another patient, but that she cannot remember which facts apply to whom. She can tell me only that the following is true. I may be the patient who had his operation yesterday. In that case, my operation was the longest ever performed, lasting ten hours. I may instead be the patient who is to have a short operation later today. It is either true that I did suffer for ten hours, or true that I shall suffer for one hour.
>
> I ask the nurse to find out which is true. While she is away, it is clear to me which I prefer to be true. If I learn that the first is true, I shall be greatly relieved. (1984, 165)

The intuition here is clear: it is much better to have a more painful surgery in one's past rather than a less painful surgery in one's future. But for exponential and hyperbolic discounters, the judgment is not so simple. It depends on how far in the past the more painful surgery is. This is revealed by Figure 3.6. In this example, whether a past surgery is preferable to a future surgery that is half as painful depends on where the agent is temporally located between the two surgeries. At some point between the two surgeries the curves

representing the disvalue of each surgery cross. Let us imagine that the point at which they cross is about five hours after the more painful surgery. In this case, then, whether the future surgery is preferrable to the past surgery depends on whether more or less than five hours have elapsed since the more painful surgery. If it has been four hours since the more painful surgery, then the agent prefers the less painful future surgery. However, if it has been six hours, then the agent prefers the more painful past surgery. In Parfit's original example, it is supposed to appear absurd for Parfit to hope that the nurse tells him that he is the patient scheduled for the future surgery. It is equally absurd for Parfit to inquire whether four or six hours have elapsed since when the past surgery would have taken place.

The case for future bias appeals in part to our typical preferences, but the foregoing shows that past discount function models produce bizarre patterns of preference. Regardless of Parfit's past discounting function, we will be able to find cases that feature some "switch point" similar to the one just revealed. Therefore, in addition to continuing to suffer from the scheduling problem, past discount function models of future bias lack the features that motivate future bias in the first place.

3.3 Error Theories for Near and Future Bias

As mentioned in Section 3.1.3, many theorists take the practical problems created by near bias to be inadequate on their own to show that it is irrational, since we also require a corresponding theory of error for why so many people tend to be irrational. We will see that the most plausible error theory for near bias applies to future bias as well.

3.3.1 The Heuristic Model

The bias toward the near may be the result of a mistake in practical reasoning. Temporally distant experiences are typically less certain, and we have more reason to prefer more certain pleasurable experiences to less certain ones, and less certain painful experiences to more certain ones. But it would be a mistake to conclude from this that the preferability of

the experiences themselves is dependent on their distance from the present, rather than on their probabilities.

This leads to the thought that near bias is a *heuristic*. When we consider trading between experiences in our near and distant future, the two important factors to consider are the value of the experiences and their probabilities. However, probabilities are difficult to compute. Therefore, by applying a discount function to experiences based on their distance from the present we can make the problem tractable, and still roughly approximate the probabilities of the experiences.

The heuristic model is particularly amenable to accounts that see many of our preferences and emotions as the result of evolutionary forces. It is advantageous for an agent to make the proper tradeoffs between, for example, a smaller but more certain meal and a larger but less certain one. Near-biased preferences provide a way to encode how these sorts of probabilities are usually affected by temporal distance in an environment.

It is likely that emotions play a role in generating these preferences. Many of our emotions are adaptations whose purpose is to solve basic ecological problems facing organisms [Darwin (1896), Plutchik (1980), Frank (1988)]. Children in Mischel's Marshmallow experiment may indeed feel these emotions to a greater degree than adults. It is, in fact, a major area of study to show that the emotions that support the bias toward the near can be explained in evolutionary terms.[14]

This is a powerful reason to think that our emotional responses are often not perfect guides to rational preferences. Many psychologists study "emotional regulation," and I take them to be studying something real and important. As we age, we may experience the emotions associated with near bias to a lesser degree, but they do not disappear—we must learn to overcome them. Consider a common elicitor of near-biased emotions: the dentist. Imagine that you are given a choice between undergoing moderately painful dental surgery tomorrow or a slightly more painful surgery a year from now. Prudently, you choose the moderately painful surgery, and make an appointment for tomorrow. However, you may

[14]Loewenstein and Elster (1992) offer a philosophically accessible collection of some of this work.

still feel anxiety regarding the surgery and find it difficult to fall asleep that night, and this anxiety may build as the time of the appointment draws nearer. Does the fact that you feel anxiety about the surgery show that it was a mistake to schedule it sooner rather than later? Of course not, you should remain confident that you made the right decision, and continue to prefer this surgery over the more distant one.

Now imagine that right before the drill is about to go in, the dentist informs you that due to a complicating factor the surgery will need to be postponed for a year. You feel *relieved*. Does the fact that you feel relief show that it was irrational to schedule the surgery sooner rather than later? Again, the answer is no. Your relief, like your anxiety, are not good guides to your preferences. They merely show that your rational commitments often have little effect on your emotional responses.

3.3.2 The Heuristic Model Applied to Future Bias

When we switch from thinking about near bias to thinking about future bias, many of the same points seem to apply, but for some reason there is a radical shift in assessments of their plausibility. For example, Parfit (1984, 186) takes the main obstacle to rejecting future bias to be that it requires us to think that one would be "irrational to be relieved when suffering is in the past" when one learns that he is the patient that has already had the surgery.[15] But this is equivalent to the situation involving near bias just described: it is easy to imagine someone being relieved to hear that their dental surgery needs to be postponed for a year, and we need not think that this person is irrational. Rather, we might think that their emotional responses are "hard-wired" in a way that that is not under the control of their rational commitments. Why should we think otherwise in the case of future bias?

Undoubtedly, the way we are thinking about emotion will have an effect on the plausibility of Parfit's claim. Parfit may want to divorce the "feeling" of relief from the state of "being relieved that." However, the more we carve up what it means to be "relieved when

[15]Parfit, however, ultimately remains neutral regarding the rationality of future bias, see 186 and 194.

suffering is in the past" the less obvious it is that our relief provides evidence that future bias is rational.

In fact, just as with near bias, there are plausible evolutionary accounts of the emotions associated with future bias.[16] Just as it is evolutionarily advantageous to make tradeoffs that respect the relative probabilities of potential rewards and penalties, it is also advantageous for an agent to focus more attention and energy on what is within her control. This reasoning suggests a *control model for time bias*: perhaps future-based emotions and preferences evolved to track asymmetries in control. A policy of not caring very much about the past is part of a good heuristic for focusing on what is within our control; future events are sometimes under our control while past events never are.

Evolutionary accounts are a relatively recent development, but the idea that future bias is really a form of control bias goes back at least to Hume. In the *Treatise,* he writes: "There is a phaenomenon of a like nature with the foregoing, viz *the superior effects of the same distance in futurity above that in the past.* This difference with respect to the will is easily accounted for. As none of our actions can alter the past, 'tis not strange it shou'd never determine the will" (1888, Section 2.3.7.6).

The control model predicts many of our emotional asymmetries as well. We tend to feel anxiety when contemplating future painful experiences because anxiety motivates us to avoid them and painful experiences are typically harmful to survival and reproduction. But a similarly strong unpleasant reaction toward past pains would be a waste. The same applies, *mutatis mutandis*, to future pleasurable experiences: pleasant anticipation motivates us to secure pleasurable experiences, which tend to be good for survival and reproduction. But a similarly strong reaction to past pleasures would not confer an advantage.

---

[16]Maclaurin and Dyke (2002) propose an evolutionary model for these emotions, but they do not use it to criticize the rationality of time bias; rather, it is part of their defense of the B-theory against Prior's "Thank Goodness That's Over" argument. The model is also suggested by Horwich (1992, 196–8), and Suhler and Callender (2012).

There is also reason to think that intuitions in favor of future bias actually track differences in control. Consider how your judgments might change if control weren't asymmetric, as in the following case:

> *Ben's Second Day in Prison*
>
> Suppose Ben is facing a long prison sentence under harsh conditions. Every day he is to be tormented by the guards, from which he will endure a great amount of suffering. On the second day of his punishment, God offers Ben a bargain: if Ben agrees to voluntarily endure a small increase in the amount of suffering to be had later today, God will make it so that yesterday he did not suffer at all. Ben is future biased, and for this reason declines the offer. Suppose further that God offers Ben this bargain every day of his stay in prison, and Ben turns it down every time. Every day Ben suffers immensely while his time-neutral equivalent would not suffer at all.

In this case, experiences maintain their temporal properties but the asymmetry of control is removed. Is it still rational to be future biased in situations in which backward causation is in play?[17] Even though the case is exotic, it provide an illuminating test of what we care about when we consider the differences between past and future experiences. If our intuitions are responsive to temporal differences, rather than differences in control, then presumably we should still feel the same pull toward future bias when we imagine cases where agents are able to exert control over the past. But given the terrible (and seemingly avoidable) ordeal that Ben undergoes because of his future bias, it is not clear that temporal

[17]We might imagine that God enforces the deal by "looking into the future" to see if Ben accepts the bargain the following day. If we do so, then Ben's situation resembles a Newcomb problem in which an immense reduction of pain is substituted for the $1,000,000, and a small reduction in pain is substituted for the $1000. In such a case, it does appear irrational to act as Ben does. Consider Nozick, 1969, 134: "If one believes, for this case, that there is backwards causality, that your choice causes the money to be there or not, that it causes him to have made the prediction that he made, then there is no problem. One takes only what is in the second box. Or if one believes that the way the predictor works is by looking into the future; he, in some sense, sees what you are doing, and hence is no more likely to be wrong about what you do than someone else who is standing there at the time and watching you, and would normally see you, say, open only one box, then there is no problem. You take only what is in the second box."

order really predicts what we care about. If so, a better explanation of our "time-bias" intuitions is that they are control-bias intuitions, sensitive to the asymmetry of control. If we had evolved in an environment in which such cases were common—and successfully navigating them necessary for survival—then perhaps future bias would seem as strange to us as it seems natural now.

There is further evidence that emotions generated by the asymmetry of control affect our intuitions. This comes from the fact that we are resistant to future bias when considering the fate of others.[18] Suppose your friend from graduate school lives on the other side of the world. You hear through a mutual friend that he requires a painful surgery, which either occurred yesterday and was the longest ever recorded (10 hours), or it will occur tomorrow and be much shorter (about an hour). Your informant cannot remember which of these is true. In this case, it is easy to imagine preferring that your friend did not have the longer surgery, even though it would at this point be in his past.

Indeed, there is also strong empirical support for this asymmetry. Researchers at Harvard conducted a study in which they asked participants to determine fair compensation in different situations. In the first case, some participants were asked to imagine they had to do a boring data entry job one month in the future, and others were asked to imagine they had completed the job a month ago. These participants exhibited future bias, believing they should be paid 60% more for their future suffering. But when participants were asked to compensate *other* people for boring data entry work in the past or future, they recommended the same compensation regardless of when the suffering occurred.[19]

This result is anticipated in a wonderful quote from Adam Smith:

> In his steadily sacrificing the ease and enjoyment of the present moment for the probable expectation of the still greater ease and enjoyment of a more distant but more lasting period of time, the prudent man is always both supported and

[18] Introduced by Parfit (1984, 181–4).

[19] Caruso et al. (2008, 799). Suhler and Callender (2012) use this study to criticize Prior's "Thank Goodness That's Over" argument. Also see Hare (2008) for further discussion of these cases.

> rewarded by the entire approbation of the impartial spectator, and of the representative of the impartial spectator, the man within the breast. The impartial spectator does not feel himself worn out by the present labour of those whose conduct he surveys; nor does he feel himself solicited by the importunate calls of their present appetites. To him their present, and what is likely to be their future situation, are very nearly in the same manner. He knows, however, that to the persons principally concerned, they are very different from being the same, and that they naturally affect *them* in a very different manner.

According to Smith, any inclination to be near-biased disappears when we adopt the perspective of a spectator freed from the emotions that near pleasures and pains commonly elicit. What the foregoing shows is that the same is likely true of future bias as well: from an impartial perspective the inclination to endorse future bias evaporates. As Smith puts it, spectators are not "worn out" by the emotions that typically accompany future bias; i.e., temporal differences in third-person cases do not create the same levels of anxiety or excitement as that experienced in first-person cases. Without these emotions distorting our judgment we are inclined to be temporally neutral, regarding both the future and the past.

Chapter 4
Practical Reliabilism

In the previous two chapters, I have argued that agents are able to pursue their ends more effectively if they are not present biased. However, it may be claimed that a theory of rationality fails in so far as it rejects present bias. The rationality of present bias, it may be argued, is assured regardless of any practical problems it might create. Even if we accepted this thought, we may still be left feeling uneasy if it can be shown that agents who are freed from a present-biased perspective consistently experience a greater degree of success in achieving their goals. This reveals a tension between two incompatible features of our thinking about practical rationality. We may feel compelled to accept the rationality of present bias, while, at the same time, find it implausible that irrational agents should be more effective pursuers of their ends than rational ones. In this chapter, I attempt to cast this tension in an overt form. I will show that it is generated by two incompatible approaches to theorizing about practical rationality, both of which hold some appeal.

4.1 The Axiomatic Approach

The standard approach to building theories of rationality that is accepted in formal decision theory, and often in ethics, is an *axiomatic* approach. The axiomatic approach is distinctive in that it begins by trying to determine axioms or constraints on axioms. It takes these axioms and constraints to have a particular evidential status; viz., that they are self-evident or unquestionable (or, at least, not obviously disputable on reasonable grounds). The goal is to provide the weakest—and thus least objectionable—set of principles from which a full theory can be logically inferred. In this way the standard methodology for theories

of practical rationality—of first determining fundamental principles and then determining what follows from them—mirrors that of mathematics.

On technical and non-technical accounts, one common source of justification for these basic principles is their (perhaps *a priori*) intuitive appeal. Intuitive considerations, for example, generate the idea that what matters in determining the rationality of a decision is the value of acts at the time of decision. This present-directedness is then taken to be a constraint on whatever axioms we adopt. Another common source of justification is *coherence*. In addition to the kinds of preference coherence demanded by standard decision theory, many philosophers accept coherence requirements on one's intentions, such as means-end coherence, and also between one's beliefs and intentions, as represented by the *enkrasia* principle: if an agent believes that she ought to A, then she intends to A.

The desire to model *proper reasoning* may be at the heart of the axiomatic approach. Straightforward maximization, for example, seems most plausible within the more general project of modeling proper reasoning in the practical domain. When an agent faces a decision problem, the resolution of the question "what should I do?" will come in the form of a choice of one of the acts available. Thus, in evaluating what to do, we may feel that the agent should focus exclusively on what these acts would mean for his future. It seems bizarre for the agent to instead stop and spend time considering how the acts available would mesh with utility-maximizing plans at some prior time. Pausing to consider which plans would have maximized expected utility in the past, once it is clear which acts maximize utility now, seems to involve irrelevant cognition. Instead, we may feel that an agent is rational to focus solely on information relevant to what his choice would mean for his future, since this is the only period in which he has some measure of control. Thinking about the proper way to reason thus leads to the endorsement of principles like act-maximization and present-directedness.

For those working with an axiomatic approach, it is no surprise that the alternatives to

standard decision theory that theorists like Gauthier and McClennen pursue are not tempting. Any of these proposed alternatives score very low compared to straightforward maximization when intuitions about proper reasoning are taken to be of the utmost importance, since these intuitions strongly favor present-directedness.

It is possible that Gauthier and McClennen took themselves to be working within the axiomatic approach, in which case their theories are obviously unconvincing. However, I believe the reason the work of Gauthier and McClennen has failed to sway many away from standard decision theory is largely due to a failure to adequately explain how their approach to building decision theories differs from the axiomatic approach. In the next section I formulate an alternative approach that does indeed support non-classical theories of rational decision making like those offered by Gauthier and McClennen. Once properly formulated, I believe it is in many ways more appealing than the axiomatic approach.

4.2 The Reliabilist Approach

Consider an analogy to epistemic norms. In epistemic contexts it is now commonplace to note that there can be a disconnect between coherence or intuitive foundational requirements and the *goal* of epistemic reasoning: viz. to believe what is true and avoid believing what is false.[1] Reliabilism in epistemology, in its most general form, is an attempt to connect epistemic norms to this goal by emphasizing the *truth-conduciveness* of epistemically relevant factors. For the reliabilist, truth-conduciveness is the ultimate determiner of epistemic norms, and coherence or intuitive foundational requirements operate only to the extent that they are truth-conducive.

The same viewpoint is possible, I believe, in practical contexts. Newcomb problems reveal a disconnect between theories built on coherence or intuitive foundational requirements and the goal of practical reasoning. Given this, I suggest that what best motivates

[1] Perhaps more precisely, "The epistemic goal is concerned with now believing those propositions that are true and now not believing those propositions that are false" (Foley, 1987, 8).

non-classical theories of rational decision making is a form of reliabilism about practical norms. Reliabilism in practical contexts should have a similar motivation to that seen in epistemology: the desire to create a close connection between practical norms and the goal of practical reasoning. Whereas in epistemic contexts the goal is true belief, for practical norms the goal is open to a number of interpretations. How one interprets the goal in practical contexts will depend on one's theory of value. For convenience, in Chapters 2 and 3 most of the discussion assumed an egoistic theory of value. However, there is nothing about the reliabilist approach that forces this theory of value. An agent is capable of inceptively maximizing all sorts of understandings of value, just as he is capable of straightforwardly maximizing them. One might adopt a utilitarian theory of value, for example, and understand value in terms of impartial utility, or of global happiness, etc. All that is required is that we be able to identify an understanding of value that is compatible with a consequentialist approach. The reliabilist approach thus fits best with a standard interpretation of decision theory as providing answers to the normative questions that remain *after the goals have been fixed.*

Practical reliabilism therefore emphasizes the conduciveness of some practically relevant factor (e.g., decisions, intentions, or preferences) to the production of what one takes to be valuable. In comparison to the axiomatic approach, the reliabilist approach reverses the procedure for constructing theories of rational decision making. Rather than starting by trying to determine fundamental axioms, it first and foremost looks to determine how conducive a decision theory is to the production of what one takes to be valuable. *Value-conduciveness* therefore determines the success of theories, and the axioms are understood as whatever is required to infer the most successful theory. There need not be any requirements on the intuitive plausibility of the axioms; they are determined by discovering whatever supports the most value-conducive theory.

Is this an accurate way to conceptualize the motivation for non-classical decision theories, such as those offered by Gauthier, McClennen, and others? Consider some of theses theorists' motivational statements: McClennen writes in *Rationality and Dynamic Choice*:

"This is a brief for rationality as a positive capacity, not a liability—as it must be on the standard account" (1990, 118). Meacham (2010, 56) offers the plausible principle: "If we expect the agents who employ one decision making theory to generally be richer than the agents who employ some other decision making theory, this seems to be a *prima facie* reason to favor the first theory over the second". And Gauthier (1986, 182–3) takes it to be the case that "A [decision making] disposition is rational if and only if an actor holding it can expect his choices to yield no less utility than the choices he would make were he to hold any alternative disposition." I suggest that only something akin to the reliabilist approach has the potential to capture these thoughts in a metatheory that is as developed as that offered by the axiomatic approach.

4.2.1 Conditional Reliability

I have characterized reliabilism as the very general idea that normative theories should emphasize the truth- or value-conduciveness of norms. However, one might be skeptical that such an overarching theory of normativity can be maintained, due to an important difference in the *belief-dependence* of epistemic and practical norms. The reliability of any practical norm—including expected utility maximization and the instrumental principle—is belief-dependent. If M is only believed to be a means to E, but is in fact not, then following the instrumental principle may be a hindrance, rather than an aid, to accomplishing one's goals. Similarly, there is no guarantee that maximizing expected utility will lead to good outcomes, even in the extreme long run, and even when one plays simple games of chance, if the game has been rigged (and thus the relevant chances are different from what one believes). If a game is rigged, then one might lose more money in the extreme long run following expected utility theory than one would following another theory. Thus the reliability of these principles depends on the accuracy of the beliefs on which they operate.

Discussions of epistemic reliability often focus on belief-independent reliable processes like perception, but the most plausible practical norms are, at best, belief-dependently reliable. This is a major reason to be skeptical of extending the reliabilist program to cover

practical norms: if one understands practical reliabilism to concern belief-independent reliability, then the theory would seem to fail at the outset.

However, belief-dependent reliability is a standard theoretical concept. Epistemologists have, in fact, found it very useful to distinguish between *conditionally* and *unconditionally reliable belief-forming processes.*[2] Unconditionally reliable processes, like those involved in perception, are belief-independent, and so reliable regardless of the accuracy of an agent's beliefs. Conditionally reliable processes, on the other hand, are reliable on the condition that the beliefs inputted to the process are accurate. Logical inference is a common example of a process that is conditionally reliable: it is a reliable method for producing true beliefs, given the truth of the beliefs on which the inference is made. Furthermore, there is some reason to question the idea that epistemic norms should favor unconditional reliability over conditional reliability. Some epistemologists—e.g., Jack Lyons (Forthcoming)—emphasize the importance of focusing on conditionally reliable processes over unconditionally reliable ones. Among other things, focusing on conditionally reliable processes can capture what envatted agents might be doing right, even when all the beliefs they form through perception are false.

Regardless of the conditional or unconditional status of epistemic norms, the evaluation of practical norms is conditional in nature. Instead of evaluating what an agent believes (as in epistemology), we instead evaluate what an agent does *given* what she believes. This shows that conditional reliability is at the heart of practical reliabilism. With this in place, we can compare reliabilism in epistemology to the sort of reliabilism being presented here. In relation to conditional reliability, the epistemic and the practical case are the same: reliabilists take intuitions about the appropriateness of an agent's reasoning to be of lesser importance, and instead ask whether some factor is conditionally reliable in producing the targeted results**.** For epistemic reliabilists the target is true beliefs, and for practical reliabilists the target is good outcomes.

---

[2]Introduced by Alvin Goldman (1979).

4.2.2 Schelling

We can find further motivation for the reliabilist approach in the work of Thomas Schelling. It is essentially in line, I believe, with what Schelling was after in *A Strategy of Conflict* when he titled Part II of the book "A Reorientation of Game Theory," and so it may be worthwhile to discuss Schelling's project in more detail. Part of Schelling's argument was that theorizing at the time was completely blind to rational strategies for a particular type of coordination problem. A simple example asks respondents to name "heads" or "tails" with the understanding that if a partner in another room does the same they both receive a prize (1960, 56). When faced with this decision, what is it rational to do? One answer is that agents should simply pick a response at random. This is because the Heads-Heads and Tails-Tails outcomes are both in equilibrium and pareto-optimal. This strategy leads to attainment of the prize half of the time. But respondents actually tend to do much better than this. A large majority of respondents choose heads (Mehta et al., 1994). Schelling's explanation of this is that for some reason "heads" represents a point of convergence of expectations for most people, which he called a "focal point."[3] He introduces many examples to show that attention to focal points is an important tool for coordinating, bargaining, and deterring. Focal points, however, have received little attention from game theorists.[4] This cannot be due to a focus on interactions between fully rational partners, since there is no reason to think that focal points are incompatible with full rationality. Rather, the lack of attention is due to the difficulty in deriving a systematic theory of focal points from the axioms of rational choice. Nevertheless, Schelling thought it was unacceptable to leave such a gap between the success of game-theoretic agents and the success of real people, stating: "A normative theory must produce strategies that are at least as good as what people can do without them" (1960, 98). It may be difficult to derive a way of reasoning with focal points from first principles, but this consideration does not override the importance of producing

[3] In another famous example Schelling asked respondents to pick a time and a place to meet a partner in New York City without prior communication. A majority of respondents selected Grand Central Station at noon (55–6).

[4] Two notable exceptions include Sugden, 1995 and Sugden and Zamarrón, 2006.

a theory that actually results in the best outcomes.

Schelling's perspective is distinctive in game theory in that he imposes no restrictions on what a final theory must look like, instead, he is committed to finding the best theory by induction from its success. Referring to coordination games like the "heads/tails" problem, Schelling writes that his basic premise is that rational players will realize "that *some* rule must be used if success is to exceed coincidence, and that the best rule to be found, whatever its rationalization, is consequently a rational rule" (283–4). This basic premise has a large impact on Schelling's approach to game theory. Sugden and Zamarrón (2006, 618) write:

> *Strategy of Conflict* provides a tutorial in how to find focal points. The reader is taken through a wide range of examples and is encouraged to try these for herself. In each case, the author explains what the focal point is and why. In the process, general principles for identifying focal points emerge and are pointed out. Having completed the tutorial, the reader is better equipped to play coordination games with success against other people. Her increased skill in playing these games will benefit her co-players too. And she is able to make more accurate predictions about how coordination games will be played by people in general. What more can we ask? What is the point of complaining that the tutorial has not been conducted in the "right" language, or that the reader has been taught to reason from a premise which has not been derived from first principles?

Practical reliabilism departs from the standard methodology in just the way imagined by Schelling. There is no reason to think that a theory built by logical deduction from *a priori* axioms will be maximally suited for success, and indeed, the standard theory, which is built in such a way, is not. Instead, the initial focus needs to be on the outcomes of decision rules rather than their intuitive foundations. Only after this is determined can the foundations be inferred.[5]

[5]To borrow a line from William James (1896, 726): the strength of the straightforward maximizer's system

The reliabilist approach also differs from the axiomatic approach in eschewing intuitions about proper reasoning in just the way imagined by Schelling. Sugden and Zamarr?n take Schelling's work to be distinctive in that it "imposes no restrictions of validity on players' reasoning: we infer that a mode of reasoning is rational by induction from its success" (620). In line with this, the reliabilist approach conceives of proper reasoning as whatever reasoning processes allow an agent to instantiate the decisions of the most value-conducive theory. The focus is thus on the creation of value, rather than the correctness of an agent's reasoning process as revealed by independent intuitions about the correctness of reasoning.

4.2.3 "Why Be Rational?"

In taking the axiomatic approach to be the standard approach to building theories of practical rationality, I am mostly in agreement with an influential series of recent articles by Niko Kolodny (2005; 2008a; 2008b; 2008c), in which he argues that the requirements of rationality are traditionally understood as *process requirements* (they tell an agent when to retain or revise his attitudes, in other words, *how to reason*) that are fundamentally concerned with coherence. Kolodny goes on to argue that such coherence requirements on reasoning are not normative, and, therefore, rationality itself is not normative; i.e., there is no good answer to the question "why be rational?" Adoption of the reliabilist approach should incline one to accept Kolodny's conclusions regarding the normative status of coherence requirements on reasoning, but to reject the idea that rationality is all about coherence.

I suggest that the reliabilist justification of practical norms, rather than one based on coherence, is at the heart of why we think, pre-theoretically, we ought to be rational. Imagine, for illustration, that you are trying to convince a friend playing a gambling game to make a certain bet. As you explain, given the value of the various outcomes and their probabilities, the bet in question uniquely maximizes expected utility, and so, you conclude, your friend

---

"lies in the principles, the origin, the *terminus a quo* of his thought; for us the strength is in the outcome, the upshot, the *terminus ad quem*. Not where it comes from but what it leads to is to decide." Quoted in Sugden and Zamarrón, 2006, 620

should choose to make the bet. His response: "I understand that the bet maximizes expected utility, but why do I have most reason to make bets that maximize expected utility?" It is, at best, unconvincing—and at worst deeply bizarre—to respond to questions like this by trying to show how it might be incoherent of your friend not to act in this way. It is similarly strange to respond with the idea that expected utility theory is based on fundamental principles that are very intuitive. However, these are, at bedrock, the ways in which the axiomatic approach attempts to answer questions like this. And given that this is the dominant approach accepted today, it shouldn't surprise us that philosophers of practical reason have recently become more and more resigned to the idea that ultimately the conclusion we must draw is that we don't have any reason to be rational, in so far as "rationality" is understood to consist of process requirements on reasoning.

Perhaps the proper justification for following expected utility theory in simple games of chance has nothing to do with coherence or intuitiveness, but rather with reliability. The proper response to your friend is to point out that expected utility theory provides an effective tool for increasing one's winnings (on the condition that the game is as he believes it to be). He is not, of course, assured of increasing his money for any given bet, but over time the probability of his actual winnings matching the positive expectation of the bet increases. This is what makes the bet worthwhile, even when it will not be iterated.

I think the same is true of the instrumental principle on intentions. The standard line is that the justification for the instrumental principle has something to do with coherence between one's beliefs and intentions, but the standard pre-theoretical justification probably has nothing to do with that. In order to argue that the instrumental principle is a rational requirement, we first note that the means are, by hypothesis, believed to be required for the end. We then point out that those who do not intend the means should expect to not accomplish the end. Finally, we show that this explains why those who flout the instrumental principle experience extreme difficulty attempting to accomplish their goals (on the condition that things are as they believe them to be). Therefore, by this reasoning, coherence

between one's beliefs and intentions is a requirement of rationality because of the conditional reliability of being coherent in this way, and not because of the existence of reasons to be coherent for its own sake.

4.2.4 Hybrid Approaches

We have been exclusively comparing the axiomatic and reliabilist approaches to theory building, but these, of course, do not exhaust the options. They are what I believe to be the two most promising approaches to theorizing about practical norms, but hybrid approaches can, and have been, defended. An example of this is the later work of David Gauthier. As I understand it, the reliabilist approach comes closest to capturing what Gauthier was after in *Morals by Agreement*. In Chapter 2, I interpreted Gauthier as breaking from standard decision theory by endorsing the following principle: If at some time t_0 it maximizes expected utility to follow a plan that involves A-ing at some subsequent time t_1, then the agent should A at t_1 (1986; 1988/89). Later, however, in his essay "Assure and Threaten," he retreats from this claim and argues that it is only rational to follow the plan at t_1 *if by so acting one is better off than one would be had one never committed to the plan at all.*[6] This new theory seems to take elements from both the axiomatic and reliabilist approaches. It follows the reliabilist approach in so far as it recommends acting in accordance with utility-maximizing plans in order to allow for greater success in open and closed Newcomb problems that involve something akin to "assurances." An example of such a Newcomb problem is the harvester case:

> *Hume's Harvesters*
>
> My crops will be ready for harvesting next week, yours a fortnight hence. Each of us will do better if we harvest together than if we harvest alone. You will help me next week if you expect that in return I shall help you in a fortnight. (Gauthier, 1993, 692)

[6] See Gauthier, 1993.

Assuming that you are a reliable predictor of Gauthier's future willingness to reciprocate, we can predict that Gauthier can expect to do better in a case like this if he would indeed help you in return. Therefore, on reliabilist reasoning, we want a theory that recommends reciprocation. Gauthier's new theory fits the bill, because i) prior to your prediction it maximizes expected utility to follow a plan that involves reciprocating, and ii) even if he reciprocates Gauthier remains better off than he would be had you not helped him at all.

That said, the new theory seems to follow the axiomatic approach in adopting an intuitively-motivated constraint that blocks success in Newcomb problems that involve something akin to "threats." Consider a standard formula for creating a deterrence case.[7] A potential wrongdoer is about to harm someone. A defender threatens to retaliate if the wrongdoer commits the harm in a way that will make both the wrongdoer and the defender worse off. We further suppose that the defender is "transparent" in the sense that the potential wrongdoer can reliably predict whether she will follow through on the threat. Now suppose that the potential wrongdoer will only be deterred if he predicts that the defender would follow through on the threat. In this case, according to Gauthier's new proposal, deterrence is impossible. This is because a rational defender (according to Gauthier) cannot follow through on the threat, since if the potential wrongdoer commits the harm, then by retaliating the defender makes both parties worse off in comparison to the situation in which the threat was never made. Therefore, if Gauthier is right, then the potential wrongdoer can expect to harm with impunity. The addition of the counterfactual comparison condition thus prevents sincere threats from occurring—since a sincere threat usually requires a commitment to follow through even if it turns out that it would have been better not to make the threat in the first place—and in many situations those that can make sincere threats can expect better outcomes than those that cannot. Gauthier's new theory therefore seems to be the result of a hybrid approach to practical norms. Notice that he may be tempted—with defenders of straightforward maximization—to offer something close to the "binding response" (discussed in Section 2.6.1) to the problems created for his theory by

[7]Following Kavka, 1987, 15.

Newcomb problems that involve threats. He may agree that it is rational for agents to irrevocably commit themselves to following through on a threat, even though it is not rational for them to deliberately follow through. However, it is unclear why we should deploy the binding option in analyzing threats but not assurances. We might doubt that there should be any important difference between the logic of following through on a threat and following through on an assurance. Gauthier must put a lot of intuitive weight on there being such a difference, since reliabilist considerations support both equally.

In any event, Gauthier's new approach cannot be sustained even if we accept that there is an important asymmetry between the rationality of sincerely threatening and sincerely assuring. This is because a threat is not required to produce cases in which Gauthier's theory fails pragmatic tests. Consider a variant of a standard open Newcomb problem in which the predictor aims to predict whether you will refuse the $1,000 *come what may*; i.e., you will refuse the $1,000 whether or not you find $1,000,000 in the other box. If it is predicted that you will refuse the $1000 come what may, then the predictor flips a coin. If heads, then $1,000,000 is put in the box. If tails, then nothing happens. If an agent commits to refusing the $1000 come what may, but then finds no money in her box, then Gauthier's condition is not satisfied. By refusing the $1000 the agent would be worse off than if he had never committed to doing so in the first place. Therefore, those following Gauthier's new proposal would not be willing to unconditionally refuse the $1,000, and because of this they will tend to end up much poorer than those following his earlier proposal (since there will be no coin flip). All this is true while there is no threat to be found.

I take this to show that it is very difficult to find a satisfactory justification for approaches to practical norms that try to incorporate elements from both the axiomatic and the reliabilist approaches. If one takes both intuitive foundational requirements and pragmatic tests to be relevant in assessing practical norms, then it will be very difficult to create a principled account of where to draw the line between cases in which pragmatic tests apply and cases in which they don't. In addition, whatever theory one constructs will almost

certainly suffer from counterexamples in which the requirements are satisfied but a pragmatic test is failed (as is the case with Gauthier's new theory). For this reason, I believe we must choose between either holding that practical success is, fundamentally, irrelevant to determining practical norms (as with the axiomatic approach), or holding that practical success is fundamental to such norms (as with the reliabilist approach).

4.3 Defining Value-Conduciveness

The reliabilist approach to practical norms evaluates practically-relevant factors in terms of their value-conduciveness, but what does it mean for a factor to be more value-conducive than another? Let us consider this question as it relates to theories of rational decision making. The sorts of decisions an agent makes will have an obvious impact on how much value he produces. Therefore, we might think about the value-conduciveness of decision theories in terms of their influence on an agent's expected success or failure in attaining valuable outcomes. These sorts of judgments are already, in fact, ubiquitous, and they are usually easy to make without much reflection. In the standard Newcomb problem, for example, one-boxers tend to end up richer than two-boxers. The controversial point is not whether one-boxers end up better off, but rather whether one-boxing is rational, whether the predictor has rewarded irrationality, and so on. Similarly, in open Newcomb problems like *Hume's Harvesters*, it should be clear that with the appropriate conditions in place, those who reciprocate tend to end up with better outcomes than those who do not.

Of course, the fact that judgments about the expected success of those following certain decision theories are ubiquitous does not prove that there is a coherent way to judge the value-conduciveness of decision theories. After all, the decision situations an agent might face are numerous and can be very complex. Still, given any well-formed decision theory, the hope is that we can determine which act or acts it recommends for any particular situation and what these recommendations will mean for an agent's success.

That said, we must be careful to select an appropriate system for judging the success

of decision theories. We cannot, for example, simply observe which agents tend to do better than others, for there are many reasons why one agent might be more successful than another that have nothing to do with their decision making. An agent's physical and mental endowments, false beliefs, and bad luck can prevent him from attaining his goals, regardless of the decisions he might make.

However, *observability-in-principle* does play an important role. What motivates the reliabilist approach is the following sort of reasoning: we note that those following one decision theory end up much worse off than those following another (in, e.g., Newcomb problems), and we ask whether this is due to a difference in belief accuracy, cognitive and physical endowments, or luck. If it is not due to any such differences, *but instead only the decisions they have made, will make, or would make,* then we conclude that the decision theory they are following is apt to make them less successful than they could otherwise be. I suggest that the guiding principle in any attempt to define the success of decision theories should be the desire to isolate the effects of an agent's past, future, and potential decisions from the effects of belief accuracy, endowments, and luck. Once these have been isolated we should indeed be able to observe those following more value-conducive decision theories experiencing more success than those following less value-conducive ones. We shall see that in order to isolate these factors we need only take three standard decision-theoretic concepts and adapt them for the reliabilist approach.

Belief Accuracy. Isolating the effects of belief inaccuracy is a matter of moving to conditional—rather than unconditional—reliability. Consider again epistemic reliabilism. For a belief-forming process to be unconditionally reliable, like perception, it must be the case that the process reliably produces true beliefs in an agent's *actual* environment. This notion, however, is useless in thinking about practical norms. Instead, in order to respect the conditionally reliable nature of practical norms, we assess the success or failure of decision theories in the sort of environment that agents take themselves to be in. Therefore, the "environment" is defined by the agent's beliefs; the most successful decision theories are the ones that tend to produce the best outcomes in these environments. This separates

the effects of an agent's belief accuracy from the effects of his decision making. Notice that this is similar to the common practice of specifying that an agent has "perfect information," or all the information relevant to his decision.

Endowments. In order to account for differences in endowments, we must not alter the situation under consideration when comparing decision rules. When we compare one decision theory to another, we must compare the effects of altering *this agent's* past, future, or potential decisions *in this environment.* In order to accomplish this, we suppose, counterfactually, that the same agent makes different decisions—i.e., we alter the decisions he has made, will make, or would make—and then consider the likely effects of this on his success. Notice that this is the same process as that endorsed by causal decision theory in the evaluation of decision problems. Causal decision theorists suggest that rational decision making is a matter of choosing options that have the best consequences under the counterfactual supposition that the agent chooses that option.[8] Causal decision theorists thus deploy counterfactual suppositions to determine the rational options in decision problems. The proposal here is that the same notion be deployed on decision theories themselves—rather than options—to gauge the causal impact of an agent's decision theory on their success relative to an environment. This keeps the reliabilist approach well clear of recommending the dominated option in medical Newcomb problems like *The Smoking Lesion*, since in cases like this an agent's past, future, and potential decisions do not have a causal effect on the chances he will form the lesion. Instead, the connection between an agent's decisions and the formation of the lesion is merely correlational.[9]

Luck. One-boxers tend to end up richer than two-boxers in the standard Newcomb problem, but one-boxers do not end up richer than two-boxers every time. Because of this, success must be defined in terms of the outcomes that a decision theory "reliably" produces, rather than simply produces. Returning to observability-in-principle, even these sorts of tendencies can be observed if the cases are iterated to a sufficient degree. This is the

[8] See Section 2.5.1.

[9] See Section 2.5.2.

analogue of *expected* utility maximization from a present-directed perspective, but adapted for thinking about the success or failure of agents relative to an environment. Inceptive maximization provides my preferred way of connecting expected utility with reliability: it thinks of the most reliable agents as those who follow the plans that would maximize expected utility from the beginning.

To sum up: According to the reliabilist approach to practical norms, theories of rational decision making should aim for success in the sense of revealing the decisions that tend to produce the best results in the environments that an agent believes himself to be in. It is important that the relevant comparison concern the same agent in the same environment, and that the comparison be done in a way that reveals causal—rather than merely correlational—connections between a decision theory and an agent's success. As long as we specify a goal, this evaluative procedure can be applied to other practically relevant factors as well, such as intentions and preferences.

4.3.1 Operational Dominating

When David Gauthier introduced "constrained maximization" (his proposed non-classical alternative to straightforward maximization) in *Morals by Agreement*, there was an important part of his argument that has gone mostly unnoticed: the application of dominance reasoning to selecting decision theories, rather than to options in decision problems. With our understanding of value-conduciveness in hand, Gauthier's reconceptualization of the dominance requirement becomes a useful tool for theory building.

Gauthier originally proposed constrained maximization as an account of the rationality behind certain types of moral reasoning, and the situations he considers deal mostly with cooperation between agents. When an agent interacts with another, Gauthier writes, he often must choose between an "individual strategy" and a "joint strategy." Individual strategies involve the agent straightforwardly maximizing at every decision point, while

joint strategies flout straightforward maximization at one or more points and, when pursued by both agents, produce outcomes that are better than if both had pursued an individual strategy. Constrained maximizers differ from straightforward maximizers by sometimes pursuing joint strategies. Gauthier spends a considerable amount of time specifying in detail the conditions under which a constrained maximizer should pursue a joint strategy.[10] This depends on the degree to which the agents involved are transparent, and the utilities of exploitation, mutual noncooperation, and mutual cooperation. The constrained maximizer thus first determines whether mutual aid is preferable to mutual non-aid, and then considers whether the probability of detection makes pursuing an individual strategy less likely to produce a better outcome than pursuing a joint strategy. If the answer to both these questions is 'yes,' then she pursues a joint strategy. Otherwise, she pursues an individual strategy.

It is crucial to Gauthier's formulation that constrained maximizers act just like straightforward maximizers in situations in which it does not pay to be a constrained maximizer. At several points in his discussion Gauthier stresses that constrained maximizers act just like straightforward maximizers in most situations; it is only under certain narrow conditions that they act differently. They thus should be able to reap all the benefits that straightforward maximizers enjoy, without any new detriments. This sets up the weak dominance: constrained maximizers will tend to do better than straightforward maximizers when the conditions are right, and the same otherwise.[11]

As an alternative to the standard application of dominance reasoning—viz., to options in decision problems—what Gauthier has in mind might be called "operational dominance": the dominance of one decision theory, in operation, over another. A decision theory weakly operationally dominates another when an agent following it tends to do better in some environments, and the same otherwise. This concept, however, is useless without a

---

[10] See 1986, 170–87.

[11] In this light, we can better understand Gauthier (1986, 182–3)'s motivational line: "A disposition is rational if and only if an actor holding it can expect his choices to yield no less utility than the choices he would make were he to hold any alternative disposition."

specification of what counts as "doing better" relative to an environment.

Understanding the success of decision theories in terms of value-conduciveness provides us such a specification. We take environments to be specified by an agent's beliefs, and then we compare decision theories by first noting the probable outcomes for an agent who chooses (or would choose) options in accordance with one theory rather than another. We then counterfactually suppose that this agent instead chooses (or would choose) options in accordance with the other theory. Operational dominance requires that there be no environments in which the agent tends to receive more valuable outcomes following the dominated theory.

An obvious way to attempt to create a theory that weakly operationally dominates another is to start with one theory and then make a very small adjustment to it. As an illustration of this, consider straightforward maximization versus straightforward maximization*, where the recommendations of straightforward maximization* are identical to those of straightforward maximization except for a single case: straightforward maximization* recommends cooperating in *Hume's Harvesters*. Straightforward maximization* is a good candidate for a theory that weakly operationally dominates straightforward maximization, since its recommendations only differ in a case in which those who cooperate tend to do better than straightforward maximizers. That said, in the next section we will look at some concerns over the possibility of operational dominating.

Theories that operationally dominate others are clear examples of better theories according to the reliabilist approach to rational decision making. Thus, if a theory operationally dominates all others, then that is the best theory possible (such a theory would hopefully seem less ad hoc than straightforward maximization*). The reliabilist approach thus creates very clear standards for the success of decision theories. Once we determine a standard for value-conduciveness, then theories of rational decision making, consisting of a function from decision problems to acts, can be ranked, in principle, by their value-conductivity. This highlights the "top-down" thinking of the reliabilist approach: we get a grip on theorizing by attempting to determine what decision rules lead to the best results,

and then work our way back to identify the axioms of the most successful theory.

4.4 Demon Worlds

The discussion of operational dominating may raise a natural worry. One might think that operational dominance is impossible, since it will always be possible to create cases in which an agent is rewarded for following a supposedly dominated decision theory. Couldn't a demon, for example, choose to punish whatever way of making decisions he likes?

As mentioned above, one potential benefit of focusing on conditional reliability in epistemology is that it can explain what agents might be doing right in "demon worlds"—worlds that are, from the agent's perspective, indistinguishable from the actual world, but in which most of their perceptual beliefs are false. It is possible for a demon to make sure that an agent's use of modus ponens, for example, is *unconditionally* unreliable—by making sure that most of the beliefs on which the inference is based are false, or by changing the world whenever the inference is performed. However, this does not speak poorly of modus ponens, since the demon cannot make modus ponens *conditionally* unreliable.

Something similar may be true, I suggest, of reliabilist proposals like inceptive maximization. It is easy to imagine situations in which a demon makes inceptive maximization unconditionally unreliable. A demon need only make an agent's beliefs false to make inceptive maximization—or straightforward maximization for that matter—extremely unconditionally unreliable (think again of a rigged gambling game). Is it possible, though, for a demon to make inceptive maximization conditionally unreliable? In order to do so he would have to make it the case that an inceptive maximizer's choices tend to make her less successful than she might otherwise be, in environments in which she has perfect information. Let us consider our imagined demon as he attempts to do this.

The first strategy the demon might employ is to attempt to punish inceptive maximizers by punishing agents who make the sorts of decisions that he expects inceptive maximizers to make. Consider again *Hume's Harvesters.* The demon knows that inceptive maximizers

reciprocate in this case, while straightforward maximizers do not. Therefore, in an attempt to punish inceptive maximizers, the demon might plan to smite those who reciprocate. However, the demon's plan will fail. By changing the payouts associated with reciprocation, the demon has changed the case, and because inceptive maximizers are aware of the demon's plan, they will no longer choose to reciprocate. It does not make sense to reciprocate in the new case because, from the beginning, a reciprocation plan would not maximize expected utility if choosing to reciprocate would cause a smiting.

Can the demon make inceptive maximization conditionally unreliable by punishing for *potential* decisions? Perhaps he might attempt to accomplish this by predicting which agents *would* reciprocate in *Hume's Harvesters* and smiting them. Again, the plan fails. Given this information, a reciprocation plan would not maximize expected utility from the beginning, since it would cause a smiting. More precisely, if the benefits to be had from receiving help with one's crops do not outweigh the detriments of the smiting, then the inceptive maximizer will act just like a straightforward maximizer and not reciprocate.

The demon will quickly become frustrated as he finds that he cannot punish inceptive maximizers by smiting those that make or would make the sort of decisions he expects inceptive maximizers to make, since his planned punishments will simply be inputted into the calculation of what plans would maximize expected utility from the beginning. The demon's attempts here are as futile as his earlier attempts to make modus ponens conditionally unreliable. He therefore might turn to a different strategy: punishing inceptive maximizers simply for *being* inceptive maximizers, and not for their actual or potential decisions. Does this work?

One way to think about the difference between straightforward and inceptive maximizers is in terms of the temporal and modal status of the decisions they regard as important in assessing an agent's practical success. Defenders of straightforward maximization think that we should take into account the expected consequences of the decisions one *has made.* Thus, straightforward maximizers don't see it as a problem that they do not receive help in *Hume's Harvesters*. They do not receive help, after all, not because of a decision they

have made, but because of a decision they *would make*: they wouldn't reciprocate. Inceptive maximization, on the other hand, takes a more liberal approach. It attempts to take into account the expected consequences of the decisions one *has made, will make,* and *would make.* Thus, only inceptive maximization is built to handle situations like *Hume's Harvesters*, in which an agent is rewarded or punished for the decisions he will or would make.

Thinking about these theories in this way highlights what is defective about the demon's new strategy. He has found that he cannot punish inceptive maximizers for their past, future, or potential decisions. Now, he realizes that he can only punish them simply because they are inceptive maximizers—and not because they have *chosen* to be, or to continue to be, an inceptive maximizer! After all, if they knew about the demon and were given a chance to stop being an inceptive maximizer, inceptive maximization would tell them to do so [as Parfit (1984, 23–4) points out, all theories will sometimes be "self-effacing" in this way]. The demon's strategy is thus not so different from choosing to smite those with red hair (and who never had the chance to change the color of their hair). Rather than looking for agent's with a certain hair color, the demon looks for agents with certain mental properties—perhaps those whose cognitive algorithms somehow correspond to the mathematical algorithms that inceptive maximization uses to select the best options.[12]

In this demon world, there does not exist an appropriate connection between an agent's decision making and the bad outcomes he receives. I suggest, therefore, that the demon's new strategy is just as ineffective at making inceptive maximization conditionally unreliable: he cannot make inceptive maximization conditionally unreliable in the sense of causing things to go poorly for inceptive maximizers because of the decisions they have made, will make, or would make. This falls directly out of how we defined value-conduciveness

[12]However, it is not clear what "being an inceptive maximizer" entails over and above being someone who has, will, and would make decisions that inceptive maximization recommends. Inceptive maximization is not a theory of how one's cognitive systems might operate, or of whatever cognitive algorithms the demon will try to detect. There are, after all, many different such algorithms one can use to instantiate the decisions recommended by inceptive maximization. This is a further reason to think that the demon has failed to make inceptive maximization, understood as a decision theory, conditionally unreliable.

above: decision theories are value-conducive to the extent that they tend to bring about good outcomes because of the decisions that agents have made, will make, or would make. Practical reliabilism is thus insulated from the standard challenges that demon worlds pose for epistemic (unconditional) reliabilism, since it focuses exclusively on conditional reliability and the effects of actual and potential decisions.

Chapter 5
Normative Naturalism

In the previous chapter, I argued that the most promising approaches to theorizing about practical norms are *axiomatic* and *reliabilist*. In relation to decision making, the application of the axiomatic approach leads to a conception of rational agents as straightforward maximizers, while the application of the reliabilist approach leads to a conception of rational agents as inceptive maximizers. Given that there are considerations in favor of both the axiomatic and reliabilist approaches, it might appear that we are left at an impasse in deciding between the two. However, I believe that from a certain philosophical perspective there are strong reasons to favor the reliabilist approach. The reliabilist approach, as it is sometimes conceived in epistemology, is in important ways more compatible with a naturalistic approach to normativity. However, "normative naturalism" lacks a precise meaning and can perhaps be applied only to specific philosophical commitments. Both straightforward and inceptive maximization are *ontologically naturalist*—neither appeals to supernatural properties; *a fortiori*, both seem to appeal to properties that are investigated by the natural sciences. However, the reliabilist approach is *methodologically naturalist* in a way that the axiomatic approach is not. This is not immediately obvious, and this fact can perhaps explain why non-classical decision theories like inceptive maximization have not been very successful in garnering support amongst those attracted by normative naturalism. Nevertheless, I will argue that it should enjoy such support.

In thinking that the axiomatic approach fails to be methodologically naturalist I do not mean to suggest that the problem lies with the axiomatic approach's greater reliance on

intuition. Intuition plays an important role in science as well as philosophy, and methodological naturalism, as I understand it, should recognize that it can convey useful information. However, if the intuitions on which the axiomatic approach relies can be shown to be suspect—perhaps by varying in unsystematic ways, or by being explained away by psychological data—then we should be willing to consider alternatives.

In Section 1, I argue that one of the most fundamental intuitions used by the axiomatic approach—viz., the intuition in favor of present-directedness—is unstable across different types of cases. In Section 2, I propose a way to understand what generates this intuition, which gives us reason to doubt its usefulness in determining objective principles of rationality. Then, in Section 3, I reveal how a commitment to a sort of methodological naturalism compatible with the reliabilist approach—specifically, to understanding normative concepts in a way that allows us to *explain* and *predict* the world—allows us to sidestep this issue and formulate a theory of rationality that stakes a better claim to objectivity.

5.1 "Zooming In" vs. "Zooming Out"

The intuitive support for the axiomatic and reliabilist approaches varies depending on the level of generality at which questions of rationality are posed. Support for the axiomatic approach is strongest when we "zoom-in" to consider questions of rationality more narrowly, whereas support for the reliabilist approach is strongest when we "zoom-out" to consider questions of rationality more broadly. Because of this, defenders of the axiomatic approach tend to build their case by focusing on detailed concrete examples of individual decisions. Since the axiomatic and reliabilist approaches differ most notably in their treatments of present-directedness (the axiomatic approach takes present-directedness to be one of the most fundamental and important axioms of rationality, whereas the reliabilist approach rejects it) the most effective cases will highlight a situation in which the reliabilist approach counterintuitively recommends an action that violates present-directedness. In *Reasons and Persons*, for instance, Parfit adroitly builds a case that does this. He sets up

the case thusly:

> *My Slavery*
>
> You and I share a desert island. We are both transparent.... You now bring about one change in your dispositions, becoming a threat-fulfiller. And you have a bomb that could blow the island up. By regularly threatening to explode this bomb, you force me to toil on your behalf. The only limit on your power is that you must leave my life worth living. If my life became worse than that, it would cease to be better for me to give in to your threats. (1984, 22)

Parfit then asks, "How can I end my slavery?" In order to do so he need only be prepared to ignore threats: if he would ignore any threat you make, then you have no reason to make it. He is, by assumption, transparent, and therefore if he would not acquiesce to a potential threat, then you would almost certainly predict that he would not acquiesce to the threat. Therefore, you predict that you would receive no benefit—and a very large penalty given that you are a threat-ignorer—from making a threat. Of course, Parfit may think that there is a small chance that you will threaten him anyway (perhaps he is not *perfectly* transparent), but if this chance is small enough, then a threat-ignoring plan may maximize expected utility from the beginning. In such a situation inceptive maximizers would thus tend to receive a better outcome: they avoid slavery while straightforward maximizers remain slaves.

However, Parfit believes that theories that reject present-directedness are refuted by the next step in the story. He asks the reader to consider what would happen if, despite the long odds, disaster strikes:

> *How I End My Slavery*
>
> We both have bad luck. For a moment, you forget that I have become a threat-ignorer. To gain some trivial end—such as the coconut that I have just picked—you repeat your standard threat. You say, that, unless I give you the coconut, you will blow us both to pieces. I know that, if I refuse, this will

> certainly be worse for me. I know that you are reliably a threat-fulfiller, who will carry out your threats even when you know that this will be worse for you. (23)

Parfit now asks us to consider the situation he is in. He knows that you will blow him up if he does not give you the coconut. It may be the case that, given the long odds of you threatening him, people who are threat-ignorers in this sort of situation tend to be better off than people who are not. But given that he now knows that you have indeed threatened him, why does this matter? He knows that not handing over the coconut—*one measly coconut!*—will cause him to die. Parfit takes this to show that rationality requires present-directedness. It is interesting to note that in his final analysis he rejects the idea that ignoring the threat is irrational *simpliciter*; rather, he calls it "rationally irrational," since by ignoring the threat he might be acting on dispositions that it was rational of him to acquire.

John Broome (2001)'s defense of present-directedness uses the same strategy. In an effort to refute theories that reject present-directedness, he relies on *Broome's Shepherd*, which we considered in section 2.1.3. Broome holds that straightforward maximization has it right in thinking that it is irrational to hand over the juicy lamb after the flock is safe. In an attempt to show this, he asks us to take on the perspective of a shepherd actually considering whether to hand over the lamb: "The flock is safely through, and the time comes to sacrifice the lamb. Should you do so? Intuitively it seems not" (102). Again, the conclusion is that rationality requires present-directedness. Broome then suggests that the intuition in favor of present-directedness is more clear when the action in question is more unpleasant. He states: "I deliberately made it a nasty thing to do. If it is not nasty enough to convince you, I can make it nastier. Let your flock be children, and the wolf a paedophile killer" (102).

As these examples show, present-directedness seems most plausible when we are imagining a specific decision in a specific case, and is perhaps enhanced by taking the perspective of the agent in question. The intuition may also depend, and is at least increased, by the affectively-charged nature of the acts and outcomes featured by the case.

All that said, we start to see an intuitive pull in the opposite direction when we stop imagining specific decisions and "zoom-out" to consider questions of rationality more broadly. One powerful source of support for the reliabilist approach, for example, comes from thinking about *design.* Consider the perspective, for instance, of artificial intelligence researchers designing an artificial agent. Let us assume that out of a concern for safety any decision algorithm that they might choose for their agent will be freely available to the public, and because of this the agent will be almost perfectly transparent. The researchers know that transparent straightforward maximizers encounter serious problems in attaining their goals that inceptive maximizers avoid. Given this, our imagined researchers might expect that Algorithm *A*, which contains instructions for straightforward maximization, will lead to absolute disaster for a robot: it will be unable to accomplish any of its goals, and very quickly end up in a scrapheap. Algorithm *B*, on the other hand, contains decision making instructions that are expected to allow a robot to successfully navigate itself through interactions with others, even in situations in which its actions can be reliably predicted. I take it the urge is not to say that a robot programmed with *B* is irrational, or even "rationally irrational." Rather, the natural thought is that *B*-programmed robots are more rational decision makers than *A*-programmed ones. It would hardly be surprising if actual researchers focused exclusively on creating decision algorithms that allow agents to succeed, even while transparent, and were skeptical of the idea that intuitions about proper deliberation show us that the robots that fail miserably are actually the rational ones. Should we, coming from a more theoretical standpoint, think that they are misguided?

It is certainly undeniable that an agent's physical and mental endowments, false beliefs, and bad luck can prevent her from attaining her goals, but it *is* odd to think that her practical rationality should be a further hindrance. Instead, we might think that once we specify what the goals should be and eliminate all differences in endowments, beliefs, and luck (i.e., any difference not pertaining to an agent's decisions), the decision makers that tend to do the best are the ones who are most rational. This is the reliabilist approach of practical rationality: we should first and foremost look to the way of making decisions that is most

conducive to the attainment of an agent's interests—that reliably increases utility—rather than the way of making decisions that is beholden to the constraints that strike us as the most compelling in imagined cases.

This intuitive clash has clear similarities to those in normative ethics between intuitions supporting classical utilitarianism and those supporting its rivals. From a broad perspective utilitarianism appears to have a solid intuitive foundation: more of the good is better than less of the good. According to some utilitarians that is all the intuition we need; after that we should let the cases fall where they may. Practical reliabilists start with a similarly broad intuition—that being rational should not be a liability. This leads to endorsement of inceptive maximization, rather than straightforward maximization, again letting specific cases fall where they may. Where both classical utilitarianism and practical reliabilism seem to run counter to intuition is in particular concrete cases in which emotions are bound to run hot. We are then posed with the following question: should we trust our general intuitions more than the ones we have about particular cases, especially when those cases involve high affect? Practical reliabilists may pursue the same line as some utilitarians, claiming that we should trust more general and systematic intuitions, rather than letting particular cases guide the way. But for straightforward maximizers like Parfit and Broome, the fact that particular acts strike us as absolutely outrageous—like refusing to acquiesce to the coconut demand, or handing over the juiciest sheep—overrides any intuitions that we might have from a general perspective. That is why, I believe, they take the clearest refutation of theories that reject present-directedness to be given by the sort of perspective-taking in concrete situations that their cases encourage.

It is worth noting that there is some interesting psychological work that has been done on a similar topic in normative ethics concerning the debate between deontologists and consequentialists. Joshua Greene (2007) and Jonathan Haidt (2001) both influentially claim that deontological moral judgments are driven by emotionally-driven intuitions, while consequentialist judgments are supported by impassive reasoning. One interpretation of these

conclusions is that, if true, they undermine deontology and support consequentialism.[1] I do not wish to claim that this conclusion is unavoidable, but my point is that if there is a challenge here for deontological theories, then there is a challenge for Parfit and Broome's justifications of straightforward maximization, since there do indeed seem to be some deontic elements to them. Whereas reliabilists only require or forbid actions in so far as they are recommended by reliable decision theories, and take their justification to depend on that basis, straightforward maximization seems to gain much of its support from the sense of an absolute restriction on performing certain seemingly outrageous acts. Like deontology, support for straightforward maximization may thus crucially rely on a sense of irrationality owing directly to features of the action, rather than a general commitment to present-directedness. The applicability of the point is conditional, but it is important nonetheless: if one is sympathetic to consequentialist ethics, then one might have a powerful reason for questioning the grounds on which straightforward maximization is justified.

We can also find an interesting discussion of the effects of "zooming in" and "zooming out" in recent work on moral responsibility. In fact, Nichols and Knobe (2007) discuss a conflict of judgments regarding the compatibility of moral responsibility and determinism that mirrors the conflict we have found between judgments regarding present-directedness. They found that people tend to judge that moral responsibility and determinism are incompatible when the question was presented broadly, without reference to a specific moral digression. But when the question concerns a concrete case describing a heinous act the responses are reversed: most respondents judge that the agent could be held morally responsible for his determined actions. The authors also found that the amount of compatibilist responses varied in proportion to the level of detail provided by the case.

Nichols and Knobe suggest that the most plausible explanation is that the presence of emotion brought out by the concrete cases leads to performance errors:

> Our hypothesis is that, when people are confronted with a story about an agent

[1] This is Greene's contention in "The Secrete Joke of Kant's Soul" and other work.

> who performs a morally bad behavior, this can trigger an immediate emotional response, and this emotional response can play a crucial role in their intuitions about whether the agent was morally responsible. In fact, people may sometimes declare such an agent to be morally responsible despite the fact that they embrace a theory of responsibility on which the agent is not responsible. (664)

While contemporary psychology teaches us that affect plays an important role in even theoretical forms of cognition, an overload of affect can produce tainted judgments that are inconsistent with a person's underlying theory. Nichols and Knobe suggest that this phenomenon may occur when we zoom-in to consider affectively-charged concrete cases.

Whether or not the performance-error hypothesis is true (instead, we might think that emotions change subjects' opinions by serving as evidence that moral responsibility and determinism are compatible) there does seem to be something essential about the emotional reactions produced by the cases used to justify present-directedness. This is conspicuously present in *How I End My Slavery*, and it is explicitly embraced by Broome, as revealed by his comment that his case may become more convincing if the "nastiness" of the act in question is increased. I think that the comment is incisive, even if I disagree with Broome's methodology: the intuition in favor of present-directedness really does increase as the nastiness of the act in question increases. This suggests that emotion influences the intuition in favor of present-directedness. If it did not, then it is hard to explain why the nastiness of the act in question should matter at all.

To sum up: The process of zooming-in and then zooming-out creates a conflict of intuition that should cause some surprise on both sides of the debate. Defenders of the axiomatic approach should admit that it is somewhat surprising that being rational sometimes gets in the way of getting what we want. Defenders of the reliabilist approach should admit that it is surprising that the most reliable decision theory recommends that we not be present-directed. The fact that there is surprise on both sides merely reveals that there is a conflict between our intuitions at different levels of generality. We are left to decide which type of intuition is more trustworthy, and there are strong reasons to doubt the superiority

of the intuitions generated by concrete affectedly-charged cases.

5.2 The Influence of Heuristics

Let us assume, for argument, that increased affective responses play no role in generating the intuition in favor of present-directedness. Even so, the fact that the intuition is strongest in concrete cases could be explained by the increased *availability* of the features that make an act rational according to straightforward maximization, in comparison to the availability of the features that make an act rational according to theories that reject present-directedness. Let me introduce the importance of availability with a standard example. Consider the domain of route planning. Given an origin and a destination, we can distinguish between two features of potential routes: i) the route with the shortest distance, and ii) the route with the shortest time. For many travel situations these two features will coincide. Sometimes, however, this is not the case. For automobile transportation, this might not be the case when choosing between a direct route on city streets and a less direct highway route. Whenever there is a split between minimum distance and minimum time we need to know which feature is more important in order to decide which route is best. However, even if we explicitly endorse time as the most important factor, we might still have an implicit preference for routes that in reality do not minimize travel time.

When people engage in "hill climbing" they prefer taking initial steps that minimize the difference between the initial state and the goal state. Similarly, when it comes to route-planning as it relates to automobile travel, people often engage in "road climbing"—preferring routes that feature initial steps that minimize the distance between their location and their destination (Bailenson et al., 1998). People who road climb also have a preference for routes that continuously decrease the distance between themselves and their destination. The combined effect of these preferences can result in blindness to routes that minimize travel time, even for those that explicitly endorse minimizing travel time as the

most important consideration.[2]

This phenomenon is best understood as a heuristic that leads to a bias. A heuristic is a simplifying tool that is useful (especially in evolutionary-formative environments) but can lead to systematic errors in certain domains (i.e., biases). Most heuristics can be understood as *attribute substitutions*, where a readily available attribute is unconsciously substituted for an unavailable one.[3] When people road climb, they unconsciously substitute more available attributes—e.g., initially or continuously decreasing one's distance from the destination—for the less available attribute of shortest travel time. The evolution of the heuristic is easy to explain: in most environments, and with most modes of transportation, initially and continuously decreasing distance are reliable indicators of minimum travel time. Therefore, when travel time becomes difficult to compute—as with automobile travel—using these more available attributes can make the problem tractable. Nevertheless, this heuristic leads to the sort of systematic bias characteristic of road climbing.

Now consider two features of decision making strategies. The relevant comparison for our purposes is between decision theories that maximize expected utility present-directedly in an environment (such as straightforward maximization) and those that tend to produce the best outcomes in an environment (such as inceptive maximization). In most environments, a certain decision is compatible with both theories, but sometimes it is not. Given that there sometimes exists such a split, we must decide which consideration is the most important in determining the principles of rational decision making. However, regardless of which consideration is ultimately taken to be most important, we should expect to see a preference for present-directed strategies, since the attributes that make a decision compatible with a present-directed strategy are more available than the attributes that make a

---

[2]A potential example of road climbing can be obtained by viewing travel from Castroville, CA "the Artichoke Center of the World" to Los Angeles on maps.google.com. Consider two available routes. The first is to take Highway 101 to Los Angeles, which is initially a more direct route. The second involves driving over 60 miles in the wrong direction to Highway 5. This latter route is a shorter distance and takes less time, but those involved in road climbing may have an initial intuitive preference for the former. In fact, in one experiment Bailenson et al. (1998) found that road climbing caused a majority of subjects to prefer a route that was 50% longer than an alternative.

[3]See Kahneman, 2003 and Kahneman and Frederick, 2005.

decision compatible with a strategy that produces the best outcomes, and in most environments the fact that a decision maximizes expected utility present-directedly is a reliable indicator of a decision that is compatible with a strategy that produces the best outcomes. In *How I End My Slavery*, for example, the decreased availability of the attributes that make ignoring the threat compatible with a best-outcome strategy is obvious. In order to see that ignoring the threat is compatible with a best-outcome strategy, one needs to understand the history of the situation, and one needs to compute the likely outcomes that we should expect for those following one decision theory rather than another. Among other things, this involves calculating the probabilities of the various outcomes for those who would or would not ignore threats—but this information is not adequately supplied in the description of the case. In contrast, that acquiescing to the threat maximizes expected utility present-directedly *is* made clear by the case description: it is obvious that ignoring the threat will fail to maximize expected utility present-directedly (it will cause the bomb to be detonated).

Consider the effect of making more available the fact that only ignoring threats is compatible with a best-outcome strategy. In order to do this, we need to make a threat-ignorer's likely outcomes salient. The way the case is described by Parfit makes the unlikely outcome for threat-ignorers (mutual destruction) salient, and the likely outcome (freedom from slavery) barely noticeable. First and foremost, therefore, allow me to formulate a description of the likely outcome:

> *How I Probably End My Slavery*
>
> As expected, after I have become a threat-ignorer your threats cease (you are no fool, after all, and you strongly prefer not to die). Together we construct a more equitable arrangement regarding daily chores and food consumption. We find that under the new arrangement we both have plenty of time to enjoy the island's bountiful produce and temperate climate. We live out the rest of our lives in relative comfort.

In order for threat-ignoring to be compatible with a best-outcome strategy, this outcome needs to be almost certain, and yet, even so, Parfit's discussion of the case suppresses its salience. In order to give this outcome the salience it deserves, the reader needs to be presented with a representation of the case that preserves the probabilistic relationships between the two possibilities. This is important: until we have a representation of the case that preserves probabilistic relationships the intuition that it is irrational to ignore the threat is worthless. However, once we do represent the case in such a way, the intuition is reduced.

Consider such a representation: we are presented with a set of possible worlds for two separate agents, one set representing the possible outcomes for a straightforward maximizer and the other representing the possible outcomes for a threat-ignorer. In the set of possible worlds that represents the straightforward maximizer's outcomes, all the worlds are the same: in every world, the straightforward maximizer lives a slave's life that is just barely worth living. As time passes the straightforward maximizer continues to toil in the service of the threat-fulfiller, ultimately dying very unsatisfied. On the other hand, in the set of possible worlds that represents the threat-ignorer's possible outcomes, we see a striking difference: in almost every world the threat-ignorer is freed from slavery, and leads the life described by *How I Probably End My Slavery*. Out of the thousands of possibilities represented, we find a single world that is consistent with Parfit's *How I End My Slavery*. In this world, the threat-ignorer is blown up. Seen in this light, Parfit's claim might lose much of its luster. Someone considering the situation in this way for the first time, with the probabilistic relationships made clear, might ask why the worlds tend to be so much worse for the straightforward maximizer. Parfit's answer must be that things are worse for the straightforward maximizer because he is *more rational*. This is revealed, he must continue, by the one case amongst thousands in which the threat-ignorer is blown up. But Parfit's preoccupation with this one world amongst thousands may at this point appear to border on obsession.

If this representation of the case makes present-directedness less appealing, then we are given strong evidence that the intuition in its favor is being generated by inequalities in

the availability of certain features in the case description. In revising Parfit's presentation of the case, all we have done is give the possibilities the salience they deserve, as encoded in their probabilistic relationships. Therefore, it would be misguided to think that the new presentation is biased toward best-outcome strategies. The presentation has simply become less biased toward present-directed strategies.

This change can be viewed as another example of the effects of zooming-out to view the situation from a wider perspective. However, here it should be clear that the wider perspective is superior in judging the relative merits of the intuition in favor of present-directedness.

5.3 The Argument for Reliabilism from Methodological Naturalism

We have been presented with legitimate reason to question some of the basic intuitions on which the axiomatic approach relies. One therefore might be attracted to a source of justification that is more compatible with the reliabilist approach, which relies on considerations from a wide perspective and not plausibility in concrete cases. I have presented the basic intuition behind the reliabilist approach both as a claim about what rationality should not be—being rational should not hinder an agent in achieving her goals—and as a claim about what it should be—all else equal, we should expect rational agents to be more successful at achieving their goals than irrational agents.

This general idea, in both its positive and negative forms, fits nicely with a methodologically naturalistic approach to normativity. I consider the essence of methodologically naturalistic accounts of norms to be the attempt to show that a certain set of natural properties can be plausibly identified with a normative concept, which can in turn earn its keep in our theorizing by allowing us to *predict* and *explain* the world around us.[4] When we attempt to explain the success or failure of the decision makers we observe, we might appeal

[4]This is in line with, e.g., the methodological components of Peter Railton (1986, 1989) and David Brink (1989)'s influential accounts of the normative naturalist project from a realist perspective.

to several factors that can have an effect on their success. It is possible, for instance, that some agents tend to be more or less successful because they possess different information or different options than others. However, if inequalities of outcome are observed or expected even after all of these potential distorting factors are removed, then we may seek an explanation in terms of the rationality of the agents in question.

The axiomatic approach, however, is unable to explain these sorts of observations using the concept of rationality. This is because the axiomatic approach does not understand rationality in a way that allows us to predict that rational agents will tend to be more successful than irrational agents. According to the axiomatic approach, in some cases rational agents tend to be more successful than irrational ones, but in others, including *My Slavery*, rational agents tend to be less successful than irrational ones (again, even after all distorting factors are removed).

The reliabilist approach, on the other hand, provides an account of rationality that can be used to both explain and predict relative success and failure. The reason that a straightforward maximizer's outcomes tend to be so much worse than those of a threat-ignorer in *My Slavery*, according to the reliabilist approach, is that straightforward maximizers are *less rational*. Using the reliabilist approach, we can further predict that a lack of rationality will cause similar inequalities of outcome, roughly to the degree that agents fail to be rational.

I argued in Chapter 2 that the best reliabilist account of rational decision making is inceptive maximization. If this is true, then we can predict the degree to which irrational decision making will be a liability to an agent by using inceptive maximization as a normative standard: an agent's decisions will be a liability to the extent that they differ from those recommended by inceptive maximization.

5.3.1 Using Rationality to Explain Trends in the Adoption of Norms

The reliabilist account of rationality is also capable of explaining and predicting the adoption of practical norms through time. Specifically, the reliabilist approach allows us to

explain the fact that human beings tend to be reasonably effective practical reasoners (even if they have their flaws) by appealing to the pressure that most environments create toward greater rationality. The reason that human beings tend to be reasonably effective is thus explained by the fact that the decision processes and other practical relevant factors we observe in agents today are at an advanced stage of a progression toward greater rationality.[5] Using the reliabilist approach, we are also able to predict that, under certain conditions, the progression will continue, while under other conditions, the progression will be slowed or even reversed.

What is the mechanism by which most environments create pressure toward conformity to practical norms? The answer lies in what Peter Railton (1986, 171–84) calls "feedback." As Railton argues, if a potential norm of practical rationality has any chance of allowing us to make sense of what we observe, it must be capable of providing feedback: negative feedback when agents flout the norm (as given by harmful or otherwise undesirable features of their experience), and positive feedback when agents conform to the norm (as given by beneficial or otherwise desirable features of their experience). For a proposed account of rationality to have an objective basis, those who conform to the norms it proposes should tend to experience positive feedback, which tends to encourage the prevalence of such rationality. Agents who flout these norms, on the other hand, should tend to experience negative feedback, which in turn discourages the prevalence of such irrationality.

Only by possessing these features can an account of rationality serve to inform our predictions and explanations: feedback creates pressure toward greater rationality and away from irrationality, and the presence of this pressure allows us to reference an agent's rationality in explaining why they were more or less successful at accomplishing his goals. If, on the other hand, a supposed norm is completely inert in regards to this sort of influence—i.e., there is no connection between an agent's acting in accordance with the norm and the

[5]This may be especially apparent when it comes to norms relating to nonreflective intention reconsideration and nonreconsideration. Consider Michael Bratman's striking view that rationality is defined in this domain by whatever practices of nonreflective reconsideration lead to some threshold of good results for an agent (1999, 60–75). He refers to this as a "broadly consequentialist" justification (66), but as we have seen here it is better thought of as a reliabilist justification.

positive or negative results we observe—then the case for the objectivity of the norm is less clear.

To highlight the usefulness of thinking about the feedback created by norms, let us consider the instrumental principle. How might the instrumental principle provide one with feedback? As noted in Chapter 4, the instrumental principle is justified on the reliabilist approach by its importance in allowing agents to achieve their goals. This is part of what is responsible for the feedback provided by instrumental rationality. Railton elaborates: "Patterns of beliefs and behaviors that do not exhibit much instrumental rationality will tend to be to some degree self-defeating, an incentive to change them, whereas patterns that exhibit greater instrumental rationality will tend to be to some degree rewarding, an incentive to continue them" (1986, 187). This explanation supports a robust set of predictions: we should expect that such a process creates pressure in favor of instrumental rationality, and this culminates, given enough environmental variation, and enough time, in a progression toward greater instrumental rationality. This is so without the aid of any belief or attitude on our part in favor of the instrumental principle; rather, it has an effect on what we observe through processes independent of our judgments.

Return now to our two competing criteria for assessing rational decision making: i) straightforward maximization, which is supported by the axiomatic approach, and ii) inceptive maximization, which is supported by the reliabilist approach. Firstly, if we put worries over demon worlds to the side, then inceptive maximization dominates straightforward maximization in terms of the positive and negative feedback that each criterion is capable of providing. When the recommendations of straightforward maximization diverge from those of inceptive maximization, as in *My Slavery* or *Broome's Shepherd*, straightforward maximizers are left without their freedom and their flocks while inceptive maximizers possess both. Above we have seen why: straightforward maximizers are present-directed, and decision theories that require present-directedness are not reliable in cases like these. The reliabilist approach thus offers an explanation of the process that causes such agents to end up as they do. Since divergences from inceptive maximization create negative feedback, we

are able to predict that in environments which contain situations in which straightforward and inceptive maximization diverge, we should expect pressure away from straightforward maximization and toward inceptive maximization. Therefore, given enough environmental variation, and enough time, we should expect a progression toward greater rationality.

The most important variable in predicting the speed of the progression is the prevalence of cases in which straightforward and inceptive maximization diverge. This is an empirical question appropriate for the study of human psychology in conjunction with evolutionary game theory. Even if it is determined that this has yet to happen—there have not been enough cases in which the recommendations of the two theories diverge—then this only means that there has existed a kind of poverty of the stimulus, which allows for the prevalence of straightforward maximization only because of its widespread agreement with the recommendations of inceptive maximization in our environment. However, even if this is true, as further advances in technology allow for more instances of reliable prediction, the pressure away from straightforward maximization will increase.

5.3.2 Resisting Noncognitivism

A related benefit of the concept of rationality endorsed by the reliabilist metatheory is that it is more clearly independent of our evaluative attitudes. We can discover reliable norms by reflecting on possible cases, but we can also observe that they are reliable through experimentation. This can in principle be done by creating certain situations, and then exposing a large number of agents to them. The agents following reliable practical norms should tend to outperform those following less reliable ones (again, assuming we eliminate the noise created by differences in belief accuracy and other factors). Thus we can observe through experiment the conclusions we arrive at *a priori*, and confirm their independence.

These positive features of reliability highlight what the concept of rationality endorsed by the axiomatic approach lacks. Since the norms endorsed by the axiomatic approach do not provide feedback (independent of their agreement with reliabilist norms), and thus are not prediction or explanation supporting, we are quickly led to the concern that they do not

have a reality outside of our evaluative judgments. This can lead to sympathy for noncognitivist accounts of such norms, or simply further worry over the status of the intuitions used in favor of the axioms.

It is understandable that the axiomatic approach encourages sentiment in favor of noncognitivism. The application of the axiomatic approach produces a theory of rationality that is useless in making predictions, or in explaining what we observe, and for which there is considerable pressure against actual agents adopting. Further reflection allows us to predict that actual instantiations of the theory are robustly "self-effacing" and will be slowly erased from existence over time. For all these reasons, it may appear to be more of a construct of the subjective attitudes of an age, rather than the recognition of an objective and enduring truth.[6]

5.3.3 An Overarching Theory of Normativity

As argued in Chapter 4, practical reliabilists need to respect the conditional nature of the reliability of practical norms. The reliability of any practical norm—including inceptive maximization and the instrumental principle—is belief-dependent. If *M* is only believed to be a means to *E,* but is in fact not, then following the instrumental principle may be a hindrance, rather than an aid, to accomplishing one's goals. Similarly, there is no guarantee that inceptively maximizing will tend to lead to good outcomes if one's beliefs are false. Rather, these norms are conditionally reliable: they tend to produce good outcomes when combined with accurate beliefs.

Therefore, in addition to a lack of cases in which straightforward and inceptive maximization give different recommendations, an agent's belief inaccuracy can also affect the rate at which conditionally reliable decision processes are adopted. In fact, the inaccuracy

[6]Relatedly, the reliabilist approach is "future proof" in a way that the axiomatic approach is not. This is especially clear in the case of inceptive maximization. Given its stress on handling transparency, the importance of inceptive maximization will likely grow as future technology makes reliable prediction more common, regardless of whether one accepts it as an account of rationality. On the other hand, if we were to abandon any of the fundamental principles on which axiomatic theories are built, then their importance would seem to be eliminated.

of an agent's beliefs can create negative feedback that swamps any positive gain that his use of conditionally reliable decision processes may provide. Conditional reliability, however, remains prediction supporting: we should expect such reliability to produce the most positive feedback in environments in which agents are ideally informed. We should therefore expect the pressure toward conditionally reliable decision processes to increase as the unconditional epistemic reliability of an agent's belief-forming processes increases.

This leads to the final major advantage of the reliabilist approach: it holds the potential to create a unified account of normativity that transcends the divide between epistemic and practical questions. What these considerations show is that the study of conditional reliability, as it concerns practical norms, goes hand in hand with the study of unconditional reliability, as it concerns epistemic norms. Together they create a compelling picture of the normative landscape, in which positive feedback for conditionally reliable decision processes increases as belief-forming processes become more unconditionally reliable, and positive feedback for unconditionally reliable belief-forming processes increases as decision processes become more conditionally reliable. Unconditionally reliable belief-forming processes and conditionally reliable decision-making processes are thus mutually supportive. This vindicates the role of unconditionally reliable belief-forming processes in theories of epistemic justification, and shows its symbiosis with conditionally reliable decision-making processes.

Bibliography

Ainslie, George. 2001. *Breakdown of Will*. Cambridge University Press.

Andreou, Chrisoula. 2008. "The Newxin Puzzle." *Philosophical Studies* 139:415–22.

Arntzenius, Frank. 2008. "No Regrets, or: Edith Piaf Revamps Decision Theory." *Erkenntnis* 68:277–297.

Arntzenius, Frank, Elga, Adam, and Hawthorne, John. 2004. "Bayesianism, Infinite Decisions, and Binding." *Mind* 113:251–83.

Bailenson, Jeremy N., Shum, Michael S., and Uttal, David H. 1998. "Road Climbing: Principles Governing Asymmetric Route Choices on Maps." *Journal of Environmental Psychology* 18:251–64.

Bennett, Jonathan. 1988. "Farewell to the Phlogistron Theory of Conditionals." *Mind* .

Bergstrom, Lars. 1966. *The Alternatives and Consequences of Actions*. Stockholm: Almquist and Wiksell.

Bratman, Michael. 1999. "Toxin, Temptation, and the Stability of Intention." In *Faces of Intention*. Cambridge University Press.

Bratman, Michael E. 2006. "Temptation Revisited." In *Structures of Agency*. Oxford University Press.

Brink, David O. 1989. *Moral Realism and the Foundations of Ethics*. Cambridge University Press.

—. 2010. "Prospects for Temporal Neutrality." In Craig Callender (ed.), *The Oxford Handbook of Time*. Clarendon Press.

Broome, John. 1991. *Weighing Goods: Equality, Uncertainty, and Time*. Oxford: Basil Blackwell.

—. 1994. "Discounting the Future." *Philosophy and Public Affairs* 23:128–56.

—. 2001. "Are intentions reasons? And how should we cope with incommensurable values?" In Christopher Morris and Arthur Ripstein (eds.), *Practical Rationality and Preference: Essays for David Gauthier*, 98–120. Cambridge University Press.

Brueckner, Anthony L. and Fischer, John Martin. 1986. "Why Is Death Bad?" *Philosophical Studies* 50:213–21.

Buchak, Lara. ms. "Risk Aversion and Rationality." *Unpublished manuscript available at http://philosophy.berkeley.edu/people/files/208 (as of 7 August 2010)* .

Caruso, Eugene M., Gilbert, Daniel T., and Wilson, Timothy D. 2008. "A Wrinkle in Time: Asymmetric Valuation of Past and Future Events." *Psychological Science* 19:796–801.

Darwin, Charles. 1896. *The Expression of the Emotions in Man and Animals*. Harper Collins.

Dimock, Susan. 2003. "Two Virtues of Contractarianism." *The Journal of Value Inquiry* 37:395–414.

Dougherty, Tom. 2011. "On Whether to Prefer Pain to Pass." *Ethics* 121:521–537.

Elstein, D. Y. and Williams, J. R. G. manuscript. "Suppositions and Decisions." .

Elster, Jon. 1986. "Introduction." In Jon Elster (ed.), *Rational Choice*. New York: New York University Press.

Foley, Richard. 1987. *The Theory of Epistemic Rationality*. Harvard University Press.

Frank, Robert. 1988. *Passions Within Reason: The Strategic Role of Emotions*. Norton.

Gauthier, David. 1986. *Morals by Agreement*. Oxford University Press.

—. 1988/89. "In the Neighbourhood of the Newcomb-Predictor (Reflections on Rationality)." *Proceedings of the Aristotelian Society* 89:179–94.

—. 1993. "Assure and Threaten." *Ethics* 104:690–721.

Gibbard, Allan and Harper, William L. 1981. "Counterfactuals and Two Kinds of Expected Utility." In Robert Stalnaker William L. Harper and Glenn Pearce (eds.), *Ifs: Conditionals, Belief, Decision, Chance, and Time*, 153–90. D. Reidel.

Goldman, Alvin. 1979. "What Is Justified Belief?" In G. Pappas (ed.), *Justification and Knowledge*. Reidel.

Greene, Joshua D. 2007. "The Secret Joke of Kant's Soul." In W. Sinnott-Armstrong (ed.), *Moral Psychology, Vol. 3: The Neuroscience of Morality: Emotion, Disease, and Development*, 35–79. MIT Press.

Haidt, Jonathan. 2001. "The Emotional Dog and Its Rational Tail." *Psychological Review* 108:814–34.

Hare, Caspar. 2008. "A Puzzle about Other-Directed Time-bias." *Australasian Journal of Philosophy* 86:269–277.

—. 2009. *On Myself and Other Less Important Subjects*. Princeton: Princeton University Press.

Harper, William L. and Skyrms, Brian. 1988. "Introduction." In William L. Harper and Brian Skyrms (eds.), *Causation, Chance, and Credence*. Kluwer.

Heathwood, Chris. 2008. "Fitting Attitudes and Welfare." In Russ Shafer-Landau (ed.), *Oxford Studies in Metaethics*, volume 3, 47–73. Oxford: Oxford University Press.

Horwich, Paul. 1992. *Asymmetries in Time*. Cambridge: The MIT Press.

Hume, David. 1888. *A Treatise of Human Nature*. Oxford University Press.

James, William. 1896. "The Will to Believe." In J. J. McDermott (ed.), *The Writings of William James*. Random House.

Joyce, James M. 1999. *The Foundations of Causal Decision Theory*. Cambridge University Press.

Kahneman, Daniel. 2003. "A Perspective on Judgment and Choice: Mapping Bounded Rationality." *American Psychologist* 58:697–720.

Kahneman, Daniel and Frederick, Shane. 2005. "A Model of Heuristic Judgment." In K. J. Holyoak and R. G. Morrison (eds.), *The Cambridge Handbook of Thinking and Reasoning*, 267–93. Cambridge University Press.

Kavka, Gregory S. 1983. "The Toxin Puzzle." *Analysis* 43:33–6.

—. 1987. *Moral Paradoxes of Nuclear Deterrence*. Cambridge University Press.

Kolodny, Niko. 2005. "Why Be Rational?" *Mind* 114:509–63.

—. 2008a. "How Does Coherence Matter?" *Proceedings of the Aristotelian Society* 107:366–402.

—. 2008b. "The Myth of Practical Consistency." *European Journal of Philosophy* 16:366–402.

—. 2008c. "Why Be Disposed to Be Coherent?" *Ethics* 118:437–63.

Kolodny, Niko and MacFarlane, John. 2010. "Ifs and Oughts." *Journal of Philosophy* 107:115–43.

Koopmans, Tjalling C. 1967. "Objectives, Constraints, and Outcomes in Optimal Growth Models." *Econometrica* 35:1–15.

Lewis, Clarence I. 1946. *An Analysis of Knowledge and Valuation*. La Salle, Ill.: Open Court.

Lewis, David. 1981a. "Causal Decision Theory." *Australasian Journal of Philosophy* 59:5–30.

—. 1981b. "'Why Ain'cha Rich?'." *Nous* 15:377–380.

Loewenstein, George and Elster, Jon (eds.). 1992. *Choice Over Time*. New York: Russell Sage Foundation.

Loomes, Graham and Sugden, Robert. 1982. "Regret Theory: An Alternative Theory of Rational Choice Under Uncertainty." *The Economic Journal* 92:805–24.

Lyons, Jack C. Forthcoming. "Should Reliabilists Be Worried About Demon Worlds?" *Philosophy and Phenomenological Research* .

Maclaurin, James and Dyke, Heather. 2002. "'Thanks Goodness That's Over': The Evolutionary Story." *Ratio* .

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

Meacham, Christopher J. G. 2010. "Binding and Its Consequences." *Philosophical Studies* 149:49–71.

Mehta, Judith, Starmer, Chris, and Sugden, Robert. 1994. "The Nature of Salience: An Experimental Investigation of Pure Coordination Games." *The American Economic Review* 84:658–73.

Moore, G.E. 1942. "A reply to my critics." In Paul Arthur Schlipp (ed.), *The Philosophy of G.E. Moore*, 535–677. Chicago: Open Court.

Nagel, Thomas. 1970. *The Possibility of Altriusm*. Oxford: Clarendon Press.

Nichols, Shaun and Knobe, Joshua. 2007. "Moral Responsibility and Determinism: The Cognitive Science of Folk Intuitions." *Nous* 41:663–85.

Nozick, Robert. 1969. "Newcomb's Problem and Two Principles of Choice." In Nicholas Rescher (ed.), *Essays in Honor of Carl G. Hempel*, 114–146. D. Reidel.

—. 1993. *The Nature of Rationality*. Princeton University Press.

Parfit, Derek. 1984. *Reasons and Persons*. Oxford University Press.

Plutchik, Robert. 1980. *Emotion: A Psychoevolutionary Synthesis*. Harper and Row.

Prior, A.N. 1959. "Thank Goodness That's Over." *Philosophy* 34:12–17.

Railton, Peter. 1986. "Moral Realism." *The Philosophical Review* 95:163–207.

—. 1989. "Naturalism and Prescriptivity." *Social Philosophy and Policy* 7:151–174.

Rawls, John. 1971. *A Theory of Justice*. Cambridge: Belknap Press of Harvard University Press.

Schelling, Thomas C. 1960. *The Strategy of Conflict*. Oxford University Press.

Sidgwick, Henry. 1884. *The Methods of Ethics*. London: Macmillan and Co, third edition.

Smith, Holly. 1991. "Deriving Morality from Rationality." In Peter Vallentyne (ed.), *Contractarianism and Rational Choice: Essays on David Gauthier's Morals by Agreement*, 229–53. Cambridge University Press.

Sobel, Jordan Howard. 1985. "Circumstances and Dominance in a Causal Decision Theory." *Synthese* 63:167–202.

—. 1994. *Taking Chances: Essays on Rational Choice*. Cambridge University Press.

Sugden, Robert. 1995. "A Theory of Focal Points." *The Economic Journal* 105:533–50.

Sugden, Robert and Zamarrón, Ignacio E. 2006. "Finding the Key: The Riddle of Focal Points." *Journal of Economic Psychology* 27:609–21.

Suhler, Christopher and Callender, Craig. 2012. "Thank Goodness That Argument Is Over: Explaining the Temporal Value Asymmetry." *Philosophers' Imprint* 12:1–16.

Williams, Bernard. 2002. *Truth and Truthfulness: An Essay in Genealogy*. Princeton University Press.

Vita

Preston Greene

Education:

2013 Ph.D. in Philosophy, Rutgers University

2006 B.A. in Philosophy and Theater Arts, University of California at Santa Cruz

Appointments:

2009–2011 Teaching Assistant, Department of Philosophy, Rutgers University

Publications:

Forthcoming "When Is a Belief True Because of Luck?" *Philosophical Quarterly*