THESIS

INFORMING RATIONAL CHOICE THEORY THROUGH CASE STUDIES OF LOSS-AVERSION

Submitted By

Peter Rakowski

Department of Philosophy

In partial fulfillment of the requirements

For the Degree of Master of Arts

Colorado State University

Fort Collins, Colorado

Summer 2011

Master’s Committee:

Advisor: Darko Sarenac

Michael Losonsky
Stephan Kroll

`

ABSTRACT

INFORMING RATIONAL CHOICE THEORY THROUGH CASE STUDIES OF LOSS-AVERSION

The problem this thesis addresses is that there are two disparate general notions of a ‘rational decision’ and neither notion is satisfactory as the basis for a rational choice theory that can improve our lives by improving our decision-making. One is too strict, labeling too many decisions irrational, while the other is too permissive, allowing decisions to be called rational when they should not be. I attempt to outline a better version of rationality, which I call global rationality, by examining the problems with the common notions in the context of a discussion of the well-documented phenomenon of loss-aversion in decision-making. While looking at case studies of loss-aversion, I argue for two main distinguishing features of my global rationality: it should respect an internalist view so that the rigid requirements of the standard rational choice theory will often not apply (while maintaining limits regarding which consistency requirements can be disregarded), and it should respect emotional utilities—the negative or positive emotions that accompany a decision should factor into the utility calculus (with important qualifications). I conclude with suggestions as to how the skeletal global rationality I’ve outlined can be filled-out in the future, in the process also offering some insights into the dynamic nature of rationality itself.

TABLE OF CONTENTS

INTRODUCTION

The concept of rationality is central in any discussion of good decision-making. It is held up as the standard for which we should strive: the more we make our decisions in conformity with it, the better our decision-making will be, and the better our lives will go, says the conventional wisdom. As often as rationality is talked and written about, as widely recognized as its importance is, however, I think we know surprisingly little about it. In this thesis I look at case studies in order to examine problems with our current understanding of rationality, and to attempt to get a better understanding of exactly what the standard for our decision-making that we call 'rationality' is, and exactly how we should view it.

In Chapter One I detail what I take to be the most common notions of rationality today. I first consider the formalized version of rationality in standard rational choice theory. I suggest that the problem with this notion is that, due to its specialization for use in theory, by its standards too many decisions are held to be irrational when our intuitions tell us they are not. The reaction to this formal version of rationality is a more intuitive version, then, that is more permissive, allowing decisions to be rational even though they may be wrong. I suggest that a sensible place to start our global rationality is more in line with this latter, intuitive notion, but also warn that this notion will run into problems as well. In this process I also attempt to explain deeper philosophical issues that are running beneath the two opposing notions of rationality. I look at the internal-external reasons debate, and draw connections between the sides of that debate and the sides of our debate over rationality. I suggest there are parallels between the way the internal-external reasons debate is often navigated in philosophy, and the way I plan to search for a more sensible and useful notion of rationality between the two existing notions.

A set of case studies in loss-aversion is the medium I use to try to devise a global rationality. At the close of Chapter One I explain exactly what is loss-aversion—this phenomenon that potential losses looming larger than potential gains of equal magnitude leads people into what, on standard rational choice theory's view, are clearly irrational decisions. In Chapter Two I begin looking at specific, well-documented instances of loss-aversion. The main point of these case studies is to suggest that a move away from the standard rational choice theory's rationality to a more intuitive one is justified. More specifically, this intuitive notion is one that is more sensitive to the idea that rationality should be an internal question, asking only whether one is moving towards ends that she has internalized. The intuitive notion also more readily than its counterpart admits of what I call emotional utilities—positive and negative emotional responses that accompany decisions—as real factors that should be included in any utility calculus.

These internalist moves, as I call them, can be taken too far, and the intuitive notion lacks the structure necessary to prevent them from being taken too far. This is why I also impose what I call externalist checks throughout the discussion. To the first internalist move that says that internal reasons should be the sole concern of rationality, I impose the externalist check that outlaws some supposed ‘ends,’ specifically ones that directly contradict the basic tenets of what it means to be rational, from being adopted by any thinking, judging being. I open the final chapter by imposing another important externalist check to the internalist move of admitting emotional utility when I argue that some emotional utilities can be shown to be illegitimate, and leading to irrational decisions, if they are causes for a loss-averse reaction without also being a reason for that reaction.

With these internalist moves and externalist checks complete, I offer my final version of global rationality—an outlined version that I admit is bare, but is at least justified by the work in the case studies preceding. I suggest ways in which this global rationality can be further filled-out, both in substance through empirical studies, and in theory through more philosophy. I close by reflecting on the dynamic nature of rationality that my studies have helped to reveal. In our endless quest to “be rational,” it should certainly help if we remove misconceptions about rationality, and better understand the true nature of this goal for which we strive.

CHAPTER 1: WHAT IS A RATIONAL DECISION?

In this first chapter I intend to identify and explain the problem to which I hope to pose a solution with my thesis: the differing conceptions of rationality related to decision, and the confusion that such a difference causes about rationality as it relates to human nature. I will lay out the different conceptions of what a rational decision is, as well as offer what I plan to use as working definitions in my study. I will introduce the internal-external reasons debate, which will be relevant to the discussion throughout this thesis. Finally, I will introduce the phenomenon of loss-aversion, which will be the central issue in our three case studies.

§1. Rationality in Rational Choice Theory

Rational choice theory, a theory most often used in economics, holds that a rational decision is one that maximizes utility. This is simply a formalized version of the intuitive notion that given a set of options, the rational decision-maker will choose the one that will produce the best outcome. Utility is an intentionally loose term, but most often it is equated with 'good.' A unit of utility (a utile[1]) is a unit of some good, so the

[1] Martin Hollis uses this term in *The Cunning of Reason* (Cambridge: Cambridge University Press, 1987) to refer to individual units of utility. He traces the term to "the old Benthamite felicific calculus" (p. 17), but since I have not encountered the word frequently in contemporary writings on rational choice, I credit Hollis for resurrecting this useful term.

goal of the rational decision-maker, according to the rational choice theorist, is to make decisions that bring her as many utiles as possible.

Decision theory, a field closely related to rational choice theory, seeks to define rational decisions in contexts of uncertainty. For this purpose it has the concept of expected utility, which is derived by multiplying the utility associated with a certain outcome by the probability of that outcome occurring. Decision theory's rational decision-maker, then, is one who maximizes expected utility. When decision theory is viewed as a branch of philosophy, its probability-utility model easily translates into the belief-desire model of action more common in philosophy. As decision theory's rational decider is one whose goal in deciding is to maximize expected utility by considering the utility of a potential outcome in proportion to the probability that the outcome will obtain, in philosophy a rational actor is one whose goal in acting is to bring about the maximum good by considering the desirability of a possible outcome (analogous to utility) along with her belief that a certain action will cause the outcome to obtain (analogous to probability).

Thus decision theory has deep philosophical roots, and one can find traces of decision theory throughout the history of philosophy. Andre Archie has done a study of instances of decision theory in Platonic dialogues. One such example is when Socrates, seeking to determine the potential of Alcibiades as a political leader, poses hypothetical decisions to him. Socrates realizes that, with belief and desire interacting as they do, if he wants to get to the bottom of Alcibiades' beliefs and ambitions, he must strategically ask many questions. As Socrates poses various hypothetical decisions, sometimes holding the probability (or level of belief) of two options constant so that he can determine how Alcibiades assigns utility (or the makeup of his desires), Archie sees a precursor to the contribution that F.P. Ramsey made to decision theory in the twentieth century.[2]

Even more explicit connections between philosophy and decision theory can be found in the modern era, from a time still long before any formal, so-called decision theory was advanced. The following is a famous quotation from the final chapter of the Port-Royal *Logic*:

> In order to decide what we ought to do to obtain some good or avoid some harm, it is necessary to consider not only the good or harm in itself, but also the probability that it will or will not occur, and to view geometrically the proportion all these things have when taken together.[3]

Daniel Bernoulli, in his *New theory on the measurement of risk*, made a key addendum to this basic operating principle when he pointed out that one's circumstances (specifically his current wealth) will dictate whether or not a given risky decision is a good one—not the expected value of the outcome alone.[4] This too was an important contribution to decision theory long before decision theory was known as such, and is one I will later look at in greater detail.

[2] Andre Archie, "Instances of Decision Theory in Plato's *Alcibiades Major* and *Minor* and in Xenophon's *Memorabilia,*" *The Southern Journal of Philosophy* 44 (2006): 368. I don't delve very deeply into this article here, but it is a useful piece for understanding the presence of the underpinnings of modern decision theory in ancient philosophy.

[3] Antoine Arnauld and Pierre Nicole, *Logic or the Art of Thinking* (Cambridge: Cambridge University Press, 1996) 273-274.

[4] Daniel Bernoulli, *Exposition of a New Theory on the Measurement of Risk,* trans. Louise Sommer (England: Gregg Press, 1967) 25.

For all its roots in philosophy, however, the current understanding of rationality in rational choice theory is divorced from a philosophical understanding. Many phenomena observed every day are held up as examples of irrational decisions. One such example, as cited by Ori and Rom Brafman, is an overreaction to small price increases for eggs. When the price of eggs increases just slightly, "shoppers completely overreact…cut(ting) back consumption by *two and a half times."*[5]

To be able to take a real instance and analyze it in rational choice theory, however, certain assumptions must be made. The most obvious one in the egg example is that consumers' only end is to get the best bargain, or something like that. Rational choice theorists do not deny their basic operating assumptions. Still, we seldom see their conclusions worded, "Given our basic operating assumptions (like Joe Consumer being interested only in getting the most for his money), and using our specialized understanding of the term 'rational decision,' we can say that Joe Consumer is making an irrational decision." Instead, "Joe Consumer is irrational," is the conclusion most often made. Dan Ariely's book *Predictably Irrational* treats examples like these. Although Ariely certainly understands the operating assumptions he is using[6], and their sometimes questionable justification, one must wonder if his book would have made

[5] Ori Brafman and Rom Brafman, *Sway: The Irresistible Pull of Irrational Behavior* (New York: Doubleday, 2008) 18.

[6] Dan Ariely, *Predictably Irrational: The Hidden Forces That Shape Our Decisions* (New York: HarperCollins Publishers, 2008). Ariely's book is a fascinating read with many insights on human psychology and decision-making. My point here is simply that Ariely is more focused on his examples than on his definition of irrationality. He defines human irrationality as "our distance from perfection" (xix), and reflects no further on his choice of definition.

the impact that it has were it titled *Predictably Irrational, Given Certain (Sometimes Shaky) Assumptions.*

One more example, not quite as current but certainly in the same spirit, comes from D.V. Lindley in *Making Decisions*. The book purports to be a statistician's guide to applying basic principles of decision theory in order to make better decisions. Perhaps not surprisingly, then, Lindley does not focus on the operating assumptions of decision theory acting as a limitation to its practical applicability. Instead he takes decisions that appear to be irrational per decision theory as just that: "(Many studies) mostly appear to show that man does not make decisions in accord with the recipes developed here: in other words, he is incoherent."[7]

In these examples it is obvious the authors are more concerned with examining interesting examples of when we defy the prescriptions of rational choice theory than with reflecting on the concept of rationality. As our focus is to critically examine what should be called a rational decision, however, we need little time to familiarize ourselves with such supposed cases of irrationality before a natural objection should come to mind: why should these decisions be called irrational when on the surface it is not clear the agent should have even known that the decision was wrong? Isn't it important to distinguish a mistaken decision from an irrational one?[8] On the interpretation of many rational choice theorists, however, if it can be shown that Joe Consumer chose the option less advantageous, even if making such a determination

[7] D.V. Lindley, *Making Decisions* (London: John Wiley and Sons Ltd., 1971) 3.

[8] This may be the first objection that comes to mind, but it isn't the only one. That the decision might not be mistaken at all is another objection that, in various forms, will see significant treatment in this thesis.

would have been very difficult for Joe, then it is concluded that Joe made an irrational decision. This is the point where it becomes obvious that the rational choice theorist's conception of a rational decision is very different from our intuitive conception of a rational decision—there is a gap that concerns knowledge.

Under the rational choice theorist's conception, which I shall henceforth abbreviate as R_e for 'economic rationality,' an irrational decision is such because it is a choice against maximum utility, regardless of whether the agent knows it. Under the intuitive notion of rationality R_i, however, a decision is irrational only if it is a choice against what the agent considers best. One clear difference, then, between the two notions is the target in a rational decision: for R_e the target is maximum utility, for R_i the target is what the agent considers best. The difference between 'maximum utility' and 'what's best' is the springboard for many a debate (some would argue, for example, that there is no difference), but it is a distraction for us here. The difference I want to highlight, the one that attests to the epistemic gap, is that maximum utility is independent of what the agent knows, while "what the agent considers best" has agent awareness built-in. In other words irrationality under R_i carries with it the crucial feature of deciding what, at some level, you know you should not. It resembles what Aristotle called *akrasia,* or what Locke expressed when he quoted Ovid: "I see and esteem the better; I follow the worse."[9]

The difference between R_e and R_i has more subtleties, as has already been suggested, but the difference highlighted above is already sufficient to attest to a

[9] John Locke, *An Essay Concerning Human Understanding,* ed. Peter Nidditch (Oxford: Clarendon Press, 1979) 254.

disconcerting gap between the two conceptions. These two conceptions are at play in what is sometimes called the prescriptive-descriptive gap in decision theory: the oft-observed fact that people's actual decision-making often deviates from the prescriptions of rational choice theory. The prescriptions of rational choice theory assume R_e, while when we deviate from such prescriptions we don't believe we are irrational—rather, we hold that while we might be wrong,[10] we can still be rational under R_i.

Many might not agree with my assessment of this gap as disconcerting, however. The gap, as I have presented it, is not a problem that many have appeared troubled by. I surmise that they take the situation as it is and move on—the term 'rational decision' is being used in two very different senses: one highly technical, one intuitive and more general. Perhaps these two different uses of the same term are reconcilable, perhaps they are not. That not many attempts have been made implies that few have viewed a reconciliation project to be worth the effort.

Obviously I do not share that view. I have already expressed my concern above, albeit briefly, over the confusion that the different uses of the same term can cause—a confusion which I believe invites mistaken general conclusions about the nature of rationality in humanity. Any human being, and especially philosophers who concern themselves with exact definitions and human nature, should take pause when such conclusions are drawn. This thesis is the expression of my belief that a reconciliation project is worthwhile. Rather than leave the two disparate conceptions alone, I want to

[10] As was mentioned above, whether we *are* wrong will also be up for debate.

devise a global conception of rationality,[11] R_g, which can encompass the truth about rationality that both conceptions hold. Not every aspect of either conception will be admissible into R_g—some things will need to be trimmed. The goal is that an accurate global understanding of rationality will emerge from the process, and that this R_g will support a better-informed normative notion of rationality, one more applicable to our daily lives[12].

While I find it curious that attempts like this have not been more common, I do not take that as a warning sign of impending failure. Great minds have tackled questions that are relevant (albeit obliquely) to my goal. In particular, I will draw on the work of Donald Davidson often for help in solving puzzles I confront. Davidson wrote extensively on rationality, and did sometimes consider the juxtaposition of different senses of rationality. He did not address the issue by starting with the problem that I have, however, so there is some work for me to apply Davidson's thoughts to my specific question of how to devise R_g so that it can make rational choice theory work more closely with our intuitions. Where I end up with R_g will also bear some important differences from what Davidson says about rationality. But, while I will end in what I think is a novel position, at least my journey there will not be entirely solitary.

[11] 'Global rationality' is not a new term that I've invented. Robert Audi refers to it in *The Architecture of Reason* (Oxford: Oxford University Press, 2001) as the overall rationality that encompasses practical rationality and theoretical rationality, for example (195). I do think, however, that what I intend for the term to mean in this study is unique.

[12] It is crucial to bear in mind that a rationality with sufficient normativity to support rational choice theory is the goal of this study. I may sometimes make statements that some notion of rationality is inadequate or of no use. I do not mean this in an absolute sense; I mean that it is inadequate for us, given the goal of this study.

§2. Intuitive Rationality and the Beginnings of Global Rationality

I have said a few things about the intuitive notion of rationality, R_i, that draws a contrast to R_e, but more needs to be said about it. First I should address possible concerns over my choice to call this competing conception "intuitive." If I have clearly explained that I hope to expound an R_g superior to R_e and R_i, a perceptive reader might wonder if I don't plan to appeal to intuition when I am arguing away from not only R_e, but also from R_i, and of course it would be strange to appeal to our intuitions about rationality in order to show why R_i is wrong.

It is correct that I plan to appeal to intuition, but that is not all I will appeal to. I take R_i to be a rather superficial understanding of rationality that a quick intuition on the subject can yield. I don't take it to be the product of intuitions about rationality interacting with careful philosophical reflections on the concept, and this is what I hope for R_g to be (so that it can be more intuitive without losing the normativity I see as essential). More specifically, with R_i I have two basic ideas in mind. The first is the simple observation that there should be a distinction between a wrong decision and an irrational one, as was explained in making the epistemic point above. The second is that if we make the intuitive move of taking "rational decision" literally, we can understand it simply as a decision made for a reason. Hopefully in time I can show how this intuitive understanding can easily go awry.

Having clearly declared my intent to end with R_g significantly different from R_i, I think I can now safely begin a sketch of R_g with a trademark of R_i—the distinction between a rational and right decision. Martin Peterson views this as an important distinction to make in introducing his readers to decision theory, for he writes:

> "A decision is *right* if and only if its actual outcome is at least as good as that of every other possible outcome."
>
> "A decision is *rational* if and only if the decision maker chooses to do what she has most reason to do at the point in time at which the decision is made."[13]

I take making this distinction as a nod to R_i, for reasons explained above. If this distinction does exist in R_e, or what Peterson later calls "being rational in the classical sense,"[14] it might be something like a distinction between excusable irrationality and irrationality where one clearly should have known better. In any case it does not play an important role in R_e.

I will not take Peterson's definitions and use them for R_g without first discussing some important issues and making modifications. The first is an assumption I plan to make which I should put off discussing no longer: that rational and irrational decisions (and right and wrong decisions) are complementary sets. As my case studies on loss-aversion will confirm, I am often more interested in studying irrational and wrong decisions than rational and right. However, defining rational and right decisions, and letting irrational and wrong decisions follow as their complements, seems the clearest way to proceed. In most contexts it is an assumption one should not make. As Davidson points out, "The irrational is not merely the non-rational, which lies outside

[13] Martin Peterson, *An Introduction to Decision Theory* (Cambridge: Cambridge University Press, 2009) 5.

[14] Peterson 295.

the ambit of the rational; irrationality is a failure within the house of reason."[15] By calling irrational the complement of rational, I might technically be including under irrational some things that have no business being there. However, I will try to frame all of the discussions of examples so that the assumption is reasonable and harmless.

Turning attention to the definitions themselves, one should notice that a right decision depends heavily on the goodness of the outcomes. "Good" here is a highly subjective entity, and measuring goodness (or, assigning utility) is far from an exact science. Even basic versions of decision theory make no secret of this; still it is important that we note that our definition of a right decision is in no way immune to this complication. Also, our definition of right depends on the utility of every other possible (but not actualized) outcome. Therefore to know with certainty that a decision is right we need not just a complete knowledge of the agent's current subjective utilities (which, in reality even the agent herself probably won't have), but a complete knowledge of the agent's subjective utilities for every other possible outcome. Calling a decision right will at best be an approximation.

Peterson makes no attempt to specify what is to be held as "good," in his definition of a right decision, and his definition of a rational decision is even vaguer. He must know full well the consequences of getting any more specific with his definitions, and as his goal is to make this distinction and move on, one could say he does an admirable job in making this distinction in a noncommittal way. Since our task is precisely what Peterson hoped to avoid, however, we should consider the next instance

[15] Donald Davidson, "Paradoxes of Irrationality," *Problems of Rationality,* ed. Davidson (Oxford: Clarendon Press, 2004) 169.

of fuzziness. The vagueness of his rational decision definition begets two major questions, both coming from the phrase "what she has most reason to do."

The first question is whether "what she has most reason to do," *as far as she knows* is the implied meaning here. I think it is, for otherwise any wrong decision could become an irrational decision, and the distinction that we care about, and that I believe is motivating Peterson here, collapses. Thus it appears that this is one point on which Peterson could have afforded to be more specific, for it is already implied by the fact that he is making such a distinction. The second question is one, however, that could have launched Peterson into the internal versus external reasons debate, as it will now do to us.

§3. Internalism versus Externalism

The second question that "what she has most reason to do" begets is, "Given what ends?" Is it, *given the ends she has chosen*? That some end must be involved in order to even speak of rational decisions is an uncontroversial assumption supported by most reflections on the concept of rationality. Whether the end is purely a matter of the agent's choosing, however, or is something else (an objective standard of 'the good life,' for example) is a much thornier issue—one that often takes form of the internal versus external reasons debate.

I should caution the reader that an "internalism-externalism debate" can take on very different forms in philosophy. For example, in one sense this applies to the relationship between reasons and desires, and what motivates an agent. In this context the internalist is one who believes that a reason alone can motivate an agent to do some action *a*—the motivation comes internal to the reason. The externalist, by contrast, holds that a reason alone cannot motivate an agent to do *a*—it must be accompanied by a desire to do *a*. The motivation is external to the reason, then; it is contained in the desire. The internalist side here is often liked to Kant, while the externalist side is linked to Hume.

There is a different kind of internalism-externalism debate, however: one that concerns the question of what justifies an agent rather than what motivates her. In this context both sides maintain the very general view that an agent's decision is justified (and therefore rational) if the agent has a reason for the decision. The key difference is that the internalist holds that the agent must be aware of that reason, in some sense,[16] in order to have it at all, while the externalist does not. The externalist holds that a decision can be rational even if it serves an end (or, equivalently, is for a reason) that the agent is unaware of, and conversely, that a decision is irrational if it goes against an end the agent has but isn't aware of.

It should be made clear that the latter is the internalism-externalism debate that concerns this study. The question I posed that launched us into internalism versus externalism was whether the ends of the agent must be known and chosen by the agent in order to require that the agent act towards those ends. Another way to phrase this would be to ask whether the ends must have been *internalized* by the agent. The internalist is defined by the fact that he answers yes, the externalist by the fact that he answers no. The opposing sides of this debate align with the opposing notions of

[16] This "in some sense" introduces quite a bit of leeway for the internalist.

rationality that I have already detailed. R_e seems to share the philosophy of the externalist here, for the economic understanding is not concerned with what exactly the agent knows regarding how his actions lead to his ends. Additionally, it is not overly concerned with which ends the agent has adopted and prioritized; regardless of those choices there are certain external ends he should also pursue. R_e certainly affirms external standards of rationality. R_i aligns with the internalist here, for the internalist in this debate appears to be making the same claim as the intuitionist that 'rational decision' should be understood literally, thus the question of rationality should be strictly an instrumental one. It is only given certain ends, springing from internal desires, that decisions are rational or irrational. Those ends themselves are not properly evaluated by rationality.

In light of this alignment, it is natural that I expand R_e to mean, roughly, economic/external rationality, and R_i to mean, roughly, intuitive/internal rationality (thus exploiting the happy coincidence that my chosen subscripts already apply). I offer one caveat, however, that the paired notions should not be understood as identical. For example, an internalist can modify his picture so that in some respects it looks more like economic rationality than intuitive rationality, while remaining internalist at heart. As long as the fluid nature of these concepts is borne in mind, however, I promote the pairing I have described, and will use it henceforth.

I also should clarify that henceforth when I refer to internalism and externalism, I will always be using those terms in the context of the second internalism-externalism debate I have just detailed. Incidentally, I think much confusion could be avoided if the first debate were referred to exclusively as “internal versus external motivation,” and the second one exclusively as “internal versus external reasons.” In practice, however, the second name is common but the first is not, and the topics often lapse into the confusing “internalism versus externalism” label.[17]

A good example of this confusion is comes in a chapter from Martin Hollis titled “External and Internal Reasons.” From the chapter’s title we could infer that Hollis is interested in the same question about justification that we are, and he is, yet he seems to jumble this with the question of motivation. He sets up a debate between himself and Bernard Williams in which Hollis will be the externalist, arguing specifically that a person can have reasons for action that are not in his immediate, current set of desires, and Williams is the internalist taking the opposing side. The confusion enters when he follows Williams’ lead, Williams having dubbed his views “Sub-Humean,” by calling his own views “Sub-Kantian.”[18] This suggests that Kant is an externalist with Hollis while Hume is an internalist with Williams, even though, as I have explained, on the question of motivation Kant is the paradigm internalist while Hume is the paradigm externalist. Hollis could have dispelled much confusion by pointing out that whether we are considering motivation or justification makes all the difference: in the internal-external motivation debate Hume is the internalist and Kant is the externalist, while in the internal-external reasons debate the opposite holds. In fact, I believe Hollis’ crucial, yet not well-explained premise is this: Hume being an externalist about motivation leads

[17] And this is exactly what I will do in the rest of this thesis, but at least I have clarified my intentions here.

[18] Martin Hollis, *The Cunning of Reason* 74-77.

him to be an internalist about reasons and Kant's internalism about motivation leads him to be an externalist about reasons.

Despite the inherent confusion in the internalism-externalism debate, manifested in Hollis' discussion, I believe the debate, and Hollis' discussion, are worth keeping around in this study. As I have already explained how the sides in the debate align nicely with R_e and R_i, it should come as no surprise that the style of the debate will also mirror the exchange I am planning between R_e and R_i. Specifically, for each side in the internalism-externalism debate there are some test cases in which the other side more closely resembles a sensible picture of rationality, while its own side is forced into uncomfortable positions. The goal of each side is to recognize the more attractive features of the opposite side in these cases, and try to account for those features in its own side's terms. Despite the aforementioned unclear aspects of Hollis' discussion, it is a good model of this process, as he focuses on problem cases that make the internalist's notion of rationality look untenable. He considers the way Williams responds to these problems for the internalist, and while he evidently holds Williams' internalist attempts in high regard, he concludes that Williams cannot resolve the problem on internalist grounds without "giv(ing) the game to the external reasons theorist."[19] Thus I will allude to this debate and Hollis' discussion of it at times in this study.

At the very least, I hope that this dive into internal-external reasons has shown the depth of the philosophical issues beneath the answer to whether the agent's ends

[19] Hollis 86.

must be ends she has chosen. How we answer that question will determine the orientation of our initial R_g—whether it leans more towards R_e or R_i.

§4. R_g: The First Version

Without further ado, I offer my modified definitions of rational and right decisions, having taken a side on the internal-external reasons debate:

> A decision is **right** if and only if its actual outcome is at least as good as that of every other possible outcome.
>
> A decision is **rational** if and only if the decision maker, given ends she has internalized and knows, chooses to do what, as far as she knows, she has most reason to do at the point in time at which the decision is made.

As is evident from the rational decision definition, I have taken the side of internalism. My reasoning for this does not reflect a knock-down argument against externalism—rather I take it merely as a logical consequence of my commitment to a knowledge requirement ("as far as she knows") in our definitions. Since I am asserting that an irrational decision needs to involve knowledge, at least to a small degree, that the decision is wrong, it would be strange if I were to then take the externalist side and claim that the agent need not have chosen her ends, or even be aware of them. It would yield something like "what she has most reason to do, as far as she knows, in accord with ends which may or may not be of her choosing, or even of her knowledge." If one can have most reason to do *x* because doing *x* best satisfies *y, y* being some end the agent has no knowledge of, it makes no sense to specify that *x* best satisfies *y, as far as the agent knows.*

Above I mentioned that it *appeared* Peterson could have been more specific and included the knowledge requirement in his definition of a rational decision. Now that I have argued that an internalist commitment is a logical consequence of including such a requirement, however, it should be clear why Peterson did not, at least if his goal was to avoid taking a side. Recalling that the knowledge requirement is a sensible (if not logical) consequence of making the distinction between right and rational to begin with, however, I believe Peterson to be more committed to internalism than perhaps he realizes. In any case, I make no secret of my current internalist commitments, and it is fair to say that the above definitions, my working definitions for R_g, currently lean towards R_i. In fact, one might argue it just *is* R_i, as was Peterson's—the only difference with mine being that a few consequences are drawn out. Again, I readily admit that R_g as it currently stands seems to have clearly taken the side of R_i over R_e. Once our case studies in loss-aversion begin, however, the need to draw important differences between R_g and R_i should become clear.

§5. What is Loss-Aversion?

Now that R_e and R_i, have been established, and background on the internal-external reasons debate has been covered, we have many of the tools we will use in our search for a better definition of R_g. As the search will center on studies of loss-aversion, it is also necessary to get familiar with that concept. The term, coined by Daniel Kahneman and Amos Tversky, can be explained most simply by stating that "losses loom larger than corresponding gains."[20] Stated another way, in making a decision, a potential loss of *x* utiles has greater weight than a potential gain of *x* utiles. A simple

[20] Amos Tversky and Daniel Kahneman, "Loss Aversion in Riskless Choice," *Choices, Values, and Frames,* ed. Kahneman and Tversky (Cambridge: Cambridge University Press, 2000) 144.

example is the choice between a 50% chance to win $1000 (with a 50% chance to win nothing), and a sure gain of $450. Although the expected utility of the first option ($500—we make a simplifying assumption and put utility in dollars here) is greater than the second, studies show that many will pick the second option. Loss-aversion is said to be responsible for this deviation from decision theory's prescription, for people view the $450 as a potential loss if they were to take the gamble and get nothing. This looms larger than the potential gain of $1000 (expected utility $500) in the gamble, so they choose the sure thing.

Risky contexts, like the one above, are often conducive to loss-averse decisions. Daniel Bernoulli's 1738 work on risk is widely recognized as a great advancement in the prehistory of decision theory, and not surprisingly it also contains important insights on loss-aversion. In it he starts with a simple observation that in risky contexts, people often behave contrary to the dictates of a simple expected utility calculus. Specifically, he notes that people are often unwilling to take a "fair gamble:" a fifty-fifty chance of winning or losing *x* dollars. He intuits the reason to be that the negative utility of the loss of $*x* is greater than the positive utility of a gain of $*x.* Though not stated in exactly the same way, this idea present in Bernoulli's work is a precursor to loss-aversion as Kahneman and Tversky would later define it.

Bernoulli's most interesting contribution is that he did not stop at simply an intuition of this, however. He devised mathematical expressions that newly defined which gambles are rational and irrational. As an example, take one of Bernoulli's simpler expressions: $(L*G)^{.5}$, where *L* is the agent's total wealth if he loses the gamble and *G* is his total wealth if he wins. This particular expression applies only to fair gambles, but a more general form of the expression is also available for other gambles. If the expression yields an amount greater than the agent's current wealth, the rational decision is to take the gamble; and taking the gamble would be irrational if the expression yields a total wealth that is less than the agent's current wealth.[21] The expression is shown in action with a few examples below:

Table 1: Bernoulli's Expression for Fifty-Fifty Gambles

Example	current wealth	ticket price (*T*)	payout	expected value of ticket (.5*payout) (*E*)	net expected value (*E* – *T*)	wealth if won (*G*)	wealth if lost (*L*)	(*L***G*)^5
1	100000	10	20	10	0	100010	99990	99999.9995
2	200	50	101	50.5	0.5	251	150	194.0360791
3	200	75	180	90	15	305	125	195.2562419
4	200	75	200	100	25	325	125	201.5564437

Example 1 demonstrates Bernoulli's assertion, "Everyone who bets any part of his fortune, however small, on a mathematically fair game of chance acts irrationally."[22] Our example takes the form of paying $*x* for a lottery ticket that has a .5 chance of winning $2*x*, but this is simply an equivalent formulation of the mathematically fair game of chance. Example 1 confirms mathematically the intuition of loss-aversion, for

[21] Bernoulli 29.

[22] Bernoulli 29.

expected utility (zero in this example) tells us we should be indifferent, but Bernoulli's expression, which accounts for loss-aversion, tells us we should not take the gamble[23].

Examples 2 and 3 further demonstrate how Bernoulli's expression attests to loss-aversion—for by the standard utility calculus these are gambles it is rational to take, yet in both examples the last column shows a decrease from current wealth. This suggests that, despite the positive net expected value, it is irrational to take these gambles. Finally, Example 4 is a case where both the standard expected utility calculus and Bernoulli's expression dictate that the gamble be taken. With the last three examples one can get a sense of how Bernoulli has precisely defined the phenomenon of loss-aversion. The threshold is clear-cut: there are some gambles that look rational on the standard utility calculus, yet Bernoulli's expression confirms our intuition that we should not take these gambles. Other gambles, like Example 4, have higher expected utility, and while the certain amount of loss-aversion should mitigate their attractiveness, in the end they are still gambles we should take.

Certainly doubts will linger about how accurate Bernoulli's expression really is as a descriptive theory. It is hard to deny, however, that in these risky cases it is a better descriptive theory than the standard utility calculus. Thus it is a good example of how the expected utility calculus can be improved into a better descriptive theory without yielding to any mysterious, unquantifiable factors (such as emotion, some might say).

[23] In the present discussion, when I state that is 'rational' to take the gamble, or that we 'should' take it, this is operating on a simplifying assumption that increase in total wealth is our only end to consider.

For all its merits, Bernoulli's expression still faces the danger of running together the ideas of risk-aversion and loss-aversion. When Kahneman and Tversky presented Prospect Theory over two centuries later, the difference was finally drawn out. The difference can be best summarized by their key point that loss-aversion can cause people to be risk-seeking just as easily as it can cause people to be risk-averse. The following example demonstrates this point:

Choice 1: Choose between a) a sure gain of $3000 or
b) a gain of $4000 with probability .8

Choice 2: Choose between c) a sure loss of $3000 or
d) a loss of $4000 with probability .8

For Choice 1, the majority of respondents (80%) chose a), even though b) has a higher expected utility of $3200, thus displaying risk-averse behavior. Loss-aversion seems responsible for this risk-aversion, since people want to avoid the feeling of loss that would come with passing up $3000 and possibly ending up with nothing. For Choice 2, however, 92% of respondents chose d), even though c) has a higher expected utility (-$3000, over option d)'s -$3200). In Choice 2, a loss is almost inevitable, but people will risk paying more by choosing d) for the .2 probability that they will end up losing nothing. This time loss-aversion has caused people to be risk-seeking.[24] Thus, an important point that Prospect Theory makes evident is that risk-aversion is bidirectional, one might say, while loss-aversion is unidirectional. While there is an observed phenomenon of risk-seeking behavior, there is no analogue of loss-seeking behavior. All

[24] Daniel Kahneman and Amos Tversky, "Prospect Theory: An Analysis of Decision under Risk," *Choices, Values, and Frames,* ed. Kahneman and Tversky (Cambridge: Cambridge University Press, 2000) 22-23.

this might suggest that loss-aversion is more fundamental than risk-aversion, but such a point is not important for our study.

What is important for our study is that we have a precise understanding of what loss-aversion is so that we can move forward with case studies in loss-aversion. Using what we know about it, along with our background on R_e and R_i, we can move towards a better grasp on R_g.

CHAPTER 2: INTERNALIST MOVES

In Chapter One the framework for our discussion of rationality was laid out, with two competing notions, R_e and R_i, given to consider, along with the name for a target global conception, R_g. We also looked at the internal-external reasons debate, and I suggested that important parallels might exist between that debate and our discussion of R_e and R_i, which might help direct the latter discussion. Finally, I introduced the phenomenon of loss-aversion, the context in which we will examine our questions of rationality. In this chapter we will look at specific, well-documented examples of loss-aversion, with the hope that in our discussion of rationality in these cases, a clearer vision of R_g will emerge. In defining R_g along the lines of R_i, I have already implied that an internalist move away from R_e is an essential first step towards R_g; in this chapter our case studies will demonstrate why this is so.

§1. Case 1: Loss-Aversion in Risk-free Context

In the first chapter we looked at a few of the many interesting studies of loss-aversion in risky choices, but I have also made the case that loss-aversion is most easily isolated in riskless contexts, so I have defined our first case study simply by the lack of risk. Kahneman and Tversky present a case where responders are asked to imagine they currently hold a part-time job as part of their professional training, and will soon be given the choice between taking job *x* and taking job *y*. The relevant details pertaining to the overall utility of these jobs, as well as the present part-time job, are summarized below:

Table 2: The Job Choice, Part 1

	Social Contact	Daily Travel Time
Present job (training)	Isolated for long stretches	10 min.
Job *x*	Limited contact with others	20 min.
Job *y*	Moderately sociable	60 min.

In this example, 70% of respondents chose job *x*.[25] This is just the first half of Kahneman and Tversky's experiment, and they make no conclusions about loss-aversion at this juncture, but considering here whether one might be able to identify loss-aversion can be an instructive exercise. One might surmise (possibly correctly) that loss-aversion is what made job *x* more attractive to the respondents who chose it. The respondent gains moderately in social interaction but loses moderately in daily travel time. He could gain much more in social interaction by choosing job *y*, but this would come at the price of losing much more in daily travel time. Since losses loom larger than gains, job *x* is more attractive.

Thus we have an explanation via loss-aversion, but the next question of whether this loss-aversion is irrational should wait for a better example, for with this one isolated choice there is not sufficient evidence that loss-aversion even influenced the decision. Perhaps the loss-aversion explanation becomes the most likely explanation if we assume

[25] Tversky and Kahneman, "Loss Aversion" 148-149.

the respondents happen to value travel time and social contact roughly equally in the increments this problem presents, but we have no warrant for assuming that. It's plausible that the majority of the respondents values low travel time much more than high social interaction, and so the choice was simply for the lesser of two evils: a slight increase in travel time over a larger increase. Social interaction would weigh in the decision very little when compared to travel time, and the respondents simply wanted to minimize their losses in terms of travel time. It is important to note that such cases of people choosing to minimize their losses are not necessarily cases of loss-aversion—they could be just an instance of people choosing the higher expected utility on the dimension which, according to their subjective utilities, matters more to them (travel time in this example). Loss-aversion, by contrast, is when losses looming larger than gains causes agents to violate the prescriptions of decision theory, or—what would be more serious in our study—our working definition of a rational decision.

This example nicely demonstrates the trouble we should have trying to apply our definitions of right and rational decisions to isolated choices. We do not know whether the 70% majority of the respondents made the 'right' choice, for that depends on how they assign goodness. Perhaps it was right, for it is certainly plausible that they assign more goodness to saving on travel time than to gaining pleasant social interaction, but we do not have any evidence for this (unless we counted their choice as evidence, but this would beg the question). Also, we cannot say whether the decision is 'rational' because we don't know what they had most reason to do unless we know their ends. If one of their ends is to minimize travel time, however, and this takes priority over the end of increasing social interaction at work, one could easily see this decision as rational, and loss-aversion would be a non-issue.

So a case for irrationality according to our working definitions would go nowhere. What may come as a more of a surprise, however, is that it would also go nowhere according to R_e. I have presented the proponent of R_e as more ready to label a decision irrational, but even she would not do so in this case of one isolated decision. This is because of a point that she, I and the proponent of R_i all agree on: talk of rationality really makes sense only within a pattern, or set, of decisions. This is the idea Davidson has in mind when he states, "Strictly speaking, then, the irrationality consists not in any particular belief but in inconsistency within a set of beliefs."[26] R_e applies principles of rationality more stringent than anything I or the R_i proponent would apply, as evidenced by the fact that R_e is more often calling decisions irrational. But this is only because R_e's principles are formalized for application and are often accompanied by operating assumptions, and even the R_e proponent needs a set of decisions in order to apply those principles and get a case for irrationality off the ground.[27] It may seem R_e is already giving important ground to R_i on this point, but I think soon it will be clear that R_e can restrict itself to examining the consistency of a set of decisions while retaining its economic and externalist character.

[26] Donald Davidson, "Incoherence and Irrationality," *Problems of Rationality,* ed. Davidson (Oxford: Clarendon Press, 2004) 192.

[27] One might argue that the R_e proponent *would* argue that an isolated decision is irrational, in the case that it is a decision towards an end that is sufficiently absurd so that we can call the end irrational (as this move is in play for the externalist). Perhaps, but I think this objection may be conflating an irrational *decision* with an irrational *end*. In any case, we don't see rational choice theorists making charges of irrationality in this manner, and at this point I am more concerned with the charges of irrationality that R_e makes via economic principles.

So far we have stepped only halfway into Kahneman and Tversky's job choice example. Doing so has yielded some important insights: one about how choices to minimize one's losses are not necessarily instances of loss-aversion, and thus not of questionable rationality; the other about how rationality becomes an issue only when considering sets of decisions, or decisions in contexts. A look at the full example now will raise more substantial questions about rationality. Kahneman and Tversky had subjects respond to two versions of the problem. The second version differs from the first (which we have already looked at) only in the details of the present, part-time training job.

Table 3: The Job Choice, Part 2

	Social Contact	Daily Travel Time
Present job (training), Version 1	Isolated for long stretches	10 min.
Present job (training), Version 2	Much pleasant social interaction	80 min.
Job *x*	Limited contact with others	20 min.
Job *y*	Moderately sociable	60 min.

Recall that 70% chose job *x* in version one of the problem. In version two however, only 33% chose job *x*.[28] Now, within the context of two decisions, one can make a better case for an irrational decision due to loss-aversion. Let's take a typical respondent, who preferred job *x* in version one and job *y* in version two, and name him Joe. We can make a case for Joe's irrationality according to our tentative R_g because we

[28] Tversky and Kahneman, "Loss Aversion" 149.

can say Joe's choice in version one implies he has an end of saving on travel time which he ranks as more important than any end he might have to increase his social interaction at work. Therefore in version two, given Joe's ends, he had most reason to choose job *x* again, since it offered a great gain in the reduction of travel time. It would come at the cost of a loss in social interaction, but that matters less to Joe than travel time. Joe had most reason to choose *x,* and since the details of the situation are relatively simple, we assume Joe knew that. Therefore, by our definition Joe's choice of *y* in version two was irrational. There is also the possibility that Joe's choice in version two reflects his true preferences, but then that would mean his choice in version one was irrational. One way or the other, Joe has made an irrational decision.

We can make a case for Joe's irrationality according to R_g, and we can certainly make one according to R_e. In fact, much of the general reasoning is the same in both cases. R_e uses a couple of formal principles to make the case, however. The first is the principle of reference independence. Kahneman and Tversky state this principle formally as, "$x \geq_r y$ iff $x \geq_s y$ for all $x,y,r,s \in X$,"[29] meaning that one's point of reference (be it *r,* which could be the training job in version one, or *s*, which could be the training job in version two) should not affect the preference order of two options (jobs *x* and *y* in our example) that are the same in either situation. Kahneman and Tversky think that to make the charge that this is an irrational decision because it violates preference invariance, an assumption of reference independence is necessary. Preference invariance, the second principle (more widely cited in R_e than reference independence),

[29] Tversky and Kahneman 149.

"requires that the preference order between prospects should not depend on the manner in which they are described."[30] R_e's case for irrationality in ordinary language, then, is similar to R_g's: Joe's first choice reveals a preference for job *x* over job *y,* and the only difference in the second version is that his training job has changed. Jobs *x* and *y* are exactly the same, however (reference independence), so his choosing job *y* in the second version is irrational (preference invariance).

Thus we have a simple, plausible example of an irrational decision due to loss-aversion. Joe is not making a decision according to coherent preferences; instead, fear of losing what he has makes him value low travel time over high social interaction one moment, then the exact opposite the next. The time has come, however, to consider what R_i has to say on the matter.

§2. The Internalist Objection

As the case for Joe's irrationality rests on the assumption of reference independence, this is the obvious place for the internalist to make his challenge. One can imagine countless scenarios where it simply wouldn't apply. For example, Joe might say that a daily commute of sixty minutes or more would require him to buy a new, more reliable car, and he would rather just keep his old car and make shorter commutes. This is why in version one, he chose job *x*, but since in version two he was already required to buy a new car to get to his training job, he chose job *y*. In this case we wouldn't have jobs *x* and *y* equivalent in both versions. Instead we would have jobs $x_1, y_1, x_2, y_2,$ where job x_1 differs from job x_2 because one is taken coming from version

[30] Kahneman and Tversky, "Choices, Values, and Frames," *Choices, Values, and Frames,* ed. Kahneman and Tversky (Cambridge: Cambridge University Press, 2000) 4.

one's training job while the other is taken coming from version two's training job. There would be nothing inconsistent, then, about Joe's simultaneously holding $x_1 \geq y_1$ and $y_2 \geq x_2$, and preference invariance would be a non-issue.

This is an obvious case where I think even R_e would readily admit that reference independence does not apply. The R_e proponent need not give up her case for Joe's irrationality, however. She could admit that the example is bare—we know absolutely nothing about Joe except what we can infer from his two decisions. Of course we have to assume reference independence in order for the example to even be interesting, to get off the ground. She might point out that Kahneman and Tversky have identified this assumption of reference independence as an implicit, not often-stated one in standard rational choice theory: "Because the standard theory does not recognize the special role of the reference state, it implicitly assumes *reference independence.*"[31] The people who mention this assumption are the ones who want to attack it. Those in the R_e camp are well-aware of this assumption, but also know it is necessary in order to get to interesting questions of irrationality.

The internalist need not be satisfied with this reply, however. He might ask exactly when we should assume reference independence, and when we should not. Of course in the scenario above, it obviously wouldn't apply, but the point of the obvious example for the internalist is that if it didn't apply in that case, might there be other cases where it doesn't apply in more subtle ways—ways that R_e has not appreciated? Certainly getting the right answer to this question matters; just blindly assuming

[31] Tversky and Kahneman, "Loss Aversion" 149.

reference independence for the sake of theory might produce beautiful theory, but what good will that theory be in application?

The internalist *could* press the issue of reference independence thus, but at this point I am more interested in another direction he could take. I imagine our internalist granting the assumption in this case for the sake of argument. Then, becoming familiar with what is in his internalist toolbox, he turns to an even bolder objection: even if we grant reference independence, why must we accept the principle of preference invariance?

The internalist could start this objection by referencing our current R_g and its internalist flavor: it says that choosing x is rational when the agent has most reason to choose *x, given ends that he has internalized and knows*. It seems that if we want to call Joe irrational for violating preference invariance we need to establish that he has internalized and knows the end of observing preference invariance. But why should we assume this? Isn't it possible that Joe simply doesn't care about preference invariance, or whether or not he fits the rational choice theorist's conception of rational? Interviewing Joe we might find that he agrees with the reference independence assumption—that the two options are exactly the same to him regardless of which training job he has—but he chose differently in the two versions because he "felt like it," or "just wanted to." He has never adopted an end that bars his preferences from radically changing from one minute to the next.[32] Such consistency matters to the rational choice theorists, but not to him, and as such he has no reason to strive for it.

In defending Joe, the internalist has turned him into a strange character, and may have second thoughts about wanting to call him rational in any normative sense (the sense we are concerned with here). This is the logical consequence of R_i that has been waiting to be flushed out, however, and like it or not the internalist has done good work to bring it out. It points to the underlying problem with R_i—that it threatens to strip any sense of normativity from rationality. No matter how bizarre or incomprehensible someone's behavior may seem to us, we cannot call it irrational unless we can establish that he has an end the pursuit of which would preclude him from looking bizarre or incomprehensible to us. It seems any charge of irrationality can be easily escaped by an appeal to internal ends that might be difficult for others to understand, much less plug into a formal decision theory. The concept of rationality's power to shape and improve our behavior would be lost.

§3. Consistency and the First Externalist Check

Joe's notion of rationality described above is not wrong or useless—it is useful and interesting as a descriptive account of one of the many psychological manifestations of human reasoning. But given that a goal of this study is to preserve some normativity in rationality so that our final R_g can still support a formal rational choice theory, I think we must draw an externalist line here and insist on a basic consistency. It amounts to

[32] Technically, what is inconsistent is holding contradictory preferences simultaneously, as I have noted above. R_e would recognize that preferences can change over time. The more radical and rapid the change appears, however, the more likely it is a case of simultaneous contradictory preferences.

saying that rationality, as we have defined it, entails a basic consistency or coherence, one that is not optional.

This move may seem like a blow against R_i, but it should not be viewed that way. Although we cannot accept the extreme of rejecting preference invariance that it appeared R_i wanted to push, even with this externalist line drawn R_i can still exert a strong influence on the normativity-capable R_g that we want, and it need not betray itself in doing so. This issue of consistency exemplifies what I mentioned above about the intuitive and internalist notions not being identical, for perhaps the intuitive notion holds that overall consistency should be an optional end just like every other end, but no version of internalism that I have found has claimed this. In fact, it is often the internalist versions that state a consistency requirement the best.

Hollis nicely captures the consistency requirement of the paradigm internalist, Williams, when he states:

> The purpose—indeed the merit in many ways—of insisting on internal reasons is to respect the actual motivations of the actor. Respect means that there can be no further complaint about an actor who has achieved a self-conscious reflective consistency… The actor's most general project cannot be irrational, provided that it is consistent, after the usual discounting for cost, likelihood, and so forth.[33]

For Williams, consistency is not an optional end; rather it is ultimately the one criterion upon which we can judge a person's rationality.

We can find similar commitments to consistency in Davidson's work. As we noted above, Davidson means irrationality to be "the failure, within a single person, of

[33] Hollis 79.

coherence or consistency in the pattern of beliefs, attitudes, emotions, intentions, and actions."[34] In another place he declares that his present interest

> is entirely with cases, if such there be, in which the judgment that the works or thoughts of an agent are irrational is not based, or at least not necessarily based, on disagreement over fact or norm… We should limit ourselves to cases in which an agent acts, thinks, or feels counter to his own conception of what is reasonable; cases where there is some sort of inner inconsistency or incoherence.[35]

These interpretations are decidedly internalist, yet committed to an external (in the sense that it is not a matter of the agent's choosing) standard of consistency. From what we have seen, the tension with intuitive rationality's idea in the Joe case—that if we are to truly respect internal reasons, we should respect Joe's rejection of preference invariance and not call him irrational—has not been addressed. Davidson shows that he appreciates such a tension when he states, "If the agent does not have the principle that he ought to act on what he holds to be best, everything considered, then though his action may be irrational from *our* point of view, it need not be irrational from his point of view."[36] But in the end, Williams and Davidson are not on board with intuitive rationality's idea: they respect internal reasons only insofar as they conform to a standard of consistency. So far we have seen them stating that, but we have not seen an argument in support of this stance.

Throughout Davidson's work he does offer plenty that can be distilled into an argument for the consistency requirement, however. For example, his argument for the

[34] Davidson, "Paradoxes" 170.

[35] Davidson, "Incoherence" 189.

[36] Davidson, "Paradoxes" 177. The principle in question here is different from our principle of preference invariance, but I think the underlying issue is the same.

conclusion that "it is a condition of having thoughts, judgments and intentions that the basic standards of rationality have application,"[37] can be applied to our present question of consistency. I reconstruct the argument here:

> P1. A propositional attitude is constituted, in part, by the logical relations between it and other propositional attitudes.
> P2. The laws that govern these logical relations (i.e. the laws of logic) are (ideally) expressed by our basic standards of rationality.
> P3. Being a thinking, judging being entails holding propositional attitudes.
> C1. Being a thinking, judging being entails being subject to the basic standards of rationality.[38]
> P4. Many basic standards of rationality are equivalently expressed as a requirement of consistency.
> C2. Being a thinking, judging being entails being subject to a requirement of consistency.

The proposition that needs the most unpacking is P1. I take this premise from Davidson's statement that "beliefs, intentions, and desires are identified, first, by their causal relations to events and objects in the world, and, second, by their relations to one another."[39] Setting aside the first identifier (causal relations) because it is irrelevant to the issue at hand, and turning to the second, he explains that the belief that it is going to rain, for example, does not really pass for that belief if, when held in conjunction with the desire to stay dry, it does not produce some appropriate action, like taking an umbrella. In order for a belief *to be* the belief it purports to be, it must bear the proper logical relations to other beliefs, intentions and desires.[40] This is why he

[37] Davidson, "Incoherence" 195.

[38] Davidson 195-196.

[39] Davidson 196.

[40] Davidson 196.

can claim that "such (logical) relations are *constitutive* of the propositional attitudes,"[41] and this is my warrant for P1.

I think that the rest of the argument is unproblematic. With P2, certainly the question of how well the standards of rationality actually reflect these laws is up for debate (That is, in a sense, what I am doing with this thesis.), but that ideally they would is less controversial. For P3, I think making propositional attitudes a necessary condition for being a thinking, judging being is acceptable. The first conclusion, Davidson's, follows validly from the premises. If one accepts P4, which I will try to support if it is not already convincing, one should accept the second conclusion, mine, for our issue of consistency.

Davidson has a complex network of arguments supporting the central claim (C1), one to which I cannot give adequate treatment here. I think I have accurately represented a strong argument that he would give regarding our issue at hand, however. What his argument provides is a way for us to support what may have been a gut reaction as soon as the objection ("What if Joe rejects the end of observing preference invariance?") was raised—that this move, insofar as it amounts to a rejection of overall consistency, is out of play. Another way to state this would be to say that the end of overall consistency is a second-order (or perhaps we need to go to a higher-order) end, or (depending on your vertical preference) a deeper-seated end, and therefore is not subject to our choosing or rejecting, as are our first-order ends. It might be even better to state that it is improper to call the consistency requirement an end at

[41] Davidson 196.

all—rather it is a guiding principle that all thinking, judging beings, by definition, observe. All these points stand in accord with the formal argument above.

It might further clarify things if I explain that in some senses of ‘consistency,’ we can in fact choose not to observe a consistency requirement and remain rational. But obviously insofar as ‘consistency’ means something integral to rationality, we cannot violate it and remain rational. For example, consistency in the form of the principle of reference independence (that one’s preference order of two options should remain consistent, regardless of her frame of reference) is a requirement that is often questionable. I’ll argue (and have already begun to, somewhat) that rationality would actually often require us to reject it. Preference invariance is a principle that seems to align more with an overarching sense of consistency, however, assuming that we take the proper precautions and apply it only when we can establish reference independence. If, as in our Joe case, we cannot establish reference independence, preference invariance cannot apply, since the options are not really identical irrespective of reference points. But if we assume reference independence, as we did, preference invariance becomes more akin to an overall consistency requirement.

The problem with the objection becomes most apparent when we unmask a disguised consistency requirement (like preference invariance), and just call it Consistency (the capital letter denoting overarching consistency). It is certainly logically dissonant to contend that one’s inconsistency makes him Consistent, so long as inconsistency is his chosen end. Since consistency is constitutive of rationality, I think this is also akin to contending that one’s irrational behavior makes him rational, since his end is to be irrational. One cannot directly contradict the definition of what it means to be *x* in order to be *x*.

R_i was not wrong in taking our working definition of rationality and running with it; for it doesn't state explicitly that the end to violate consistency is off limits. And, insofar as R_i wants to make a rejection of a rigid consistency (like reference independence), and not a rigid rejection of consistency, it will often be correct. Nor does the fact that R_i was ultimately rebuked for this move need to be viewed as a blow against R_i, for R_i contains in its internalist aspect the solution to the problem its intuitive aspect creates. In defending Joe, the intuitive notion came to fear the monster it created, but the internalist notion can handle Joe quite easily.

An internalist like Davidson would first insist that he could find an internal inconsistency in Joe, and that this would be sufficient to call him irrational. Davidson states, "I should never have tried to pin you down to an admission that you ought to subscribe to the principles of decision theory. For I think everyone does subscribe to those principles, whether he knows it or not,"[42] which attests to the aforementioned point that the norms of rationality are not optional, as well as posits relevant information about Joe. Despite his denials, Joe, being a thinking judging being, does strive for an overarching consistency, so his present violation is (or at least certainly could be) grounds for calling him irrational.

If, however, Davidson is wrong and Joe really doesn't strive for an overarching consistency, then perhaps we cannot find the type of internal inconsistency we would

[42] Davidson 195.

like to in order to label him irrational, but this is only because we cannot make sense of him at all.[43] For Joe certainly falls outside our definition of rationality. If one objects that it is trivial that we are calling him irrational simply because he falls outside our (some might argue arbitrary) definition of rationality, we should agree and point out that the much more pressing matter is that to us Joe does not qualify as a thinking, judging being. On the issue of rationality, it would be more accurate to say that Joe is in the realm of the non-rational, for as we have seen from Davidson, irrationality is a failure within the house of reason, and Joe is not in that house.

R_i need not give in to R_e, then. It needs simply to embrace the more disciplined, internalist side of its nature. R_e should at least be able to claim, however, that R_i is making an externalist move by imposing an external standard of consistency (again, external in the sense of not being a matter of the agent's choice), and point out that R_i's quest to abolish preference invariance has failed, insofar as preference invariance is an expression of this basic consistency. The reference independence assumption, however, is still up in the air. R_i can still gain some ground on that front. This is where we will turn in the next case study.

§4. Case 2: The Endowment Effect

We have just seen R_i make a move that was successfully rebuked by R_e (or, more accurately, by internalists giving an account in their own terms for an attractive feature of R_e). R_i's mistake was that it went for too much, challenging the most basic tenets of what it means to be rational in any prescriptive sense of the word. Nevertheless, its

[43] This is really Davidson's most crucial point: if I am to make sense of your actions at all, I must assume you adhere to the basic standards of rationality.

method of questioning basic assumptions is one that can still bear much fruit. Recall that reference independence was assumed for argument's sake in the last case, but the merits of this assumption (how often it really applies) can still be challenged. A couple examples that demonstrate what is called the endowment effect provide fertile ground to consider the merits of this assumption.

The endowment effect, which is the tendency to assign higher utility to the things in our possession than we would were those things not in our possession, seems to be at work in an experiment that involves the trade of a chocolate bar for a coffee mug.[44] In the experiment three different classes at the University of Victoria were given different endowments. The first class was given a coffee mug, asked to answer a questionnaire, then shown a 400-gram Swiss chocolate bar and told they could trade their mug for a chocolate bar if they desired. Conditions were exactly the same for the second class, except the second class was initially given the chocolate bar and later given the opportunity to trade it for the coffee mug. The third class was not given one or the other immediately—they were simply offered a choice between the candy bar and the mug. The table below indicates the preferences of the three groups:

Table 4: The Mug Trade

Group	Mug over Candy	Candy over Mug
1 (76 students) Give up mug to obtain candy	89%	11%
2 (87) Give up candy to obtain mug	10%	90%
3 (55) No initial entitlement	56%	44%

[44] Jack L. Knetsch, "The Endowment Effect and Evidence of Nonreversible Indifference Curves," *Choices, Values, and Frames,* ed. Kahneman and Tversky (Cambridge: Cambridge University Press, 2000) 173.

The preferences demonstrate the endowment effect because few who were endowed with either the mug or the candy bar were willing to trade for the other. The third group served as a control group to reveal that the objective values of the mug and candy bar were roughly equal, since without initial entitlement the preferences were split 56-44. Using groups of students makes a stronger case for loss-aversion at work in the endowment effect than if this reflected the preferences of just one respondent, but for the purposes of this discussion we will take one respondent with preferences typical of this group (when endowed with a mug, she prefers it 89-11 over a candy bar, when endowed with a candy bar she prefers it 90-10 over a mug, etc.) and name her Maude. As we did in our first case with Joe, we can build a case that Maude is making an irrational decision. A rough sketch of it is that the third version reveals that Maude only slightly prefers a mug over the candy bar, but when Maude is endowed with a mug this preference becomes much stronger, and when Maude is endowed with a candy bar she suddenly has her strongest preference of all to keep the candy bar instead of trading it for the mug.

As in our first case, Maude's decisions can be viewed as irrational on both R_g and R_e if we assume reference independence. Maude's frame of reference (whether she owns one or the other, or nothing) is irrelevant to her preferences regarding the candy bar and the mug, we assume, so there is clearly an inconsistency. On the R_g interpretation, we would say that Maude's initial strong preference for the mug over the candy reveals that she holds a mugs-over-candy-bars end, yet her second decision to retain the candy bar clearly contradicts the end that we cannot assume she so quickly forgot about. Or, in the words of the R_e interpretation, we could simply say that Maude is violating preference invariance.

This example is effective for introducing the endowment effect. It is also better grounds for debating the reference independence assumption than our job example from Case 1, I believe, because this experiment involved the respondents in an actual situation and not a hypothetical one like the job choice. In the actual case, it is harder to imagine a plethora of ways that reference dependence might come in to play (though it is not impossible), since the experiment is more tightly controlled. The hypothetical case is more likely to invite considerations of reference dependence.

A weakness of our current example, however, is that like Bernoulli's fair game of chance example, it exposes loss-aversion (as the endowment effect) by identifying a case where we should be roughly indifferent, yet people consistently choose one way because losses loom larger than gains. Thus, like Bernoulli's example, it does not reveal the robustness of the endowment effect like other examples can. It is not technically accurate to assume, as we did for our discussion, that a 90-10 preference among a group means that a single person should have a preference of that strength. It could be that the endowment effect is usually negligible, but in cases like this one where the two options are roughly of equal utility, the usually negligible endowment effect is what consistently (to the tune of 90-10) tips the scales.

A similar example that Kahneman and Tversky cite does more to reveal the robustness of the endowment effect. In it a group of students at Simon Fraser University were grouped into three different classes: buyers, sellers and choosers. The good in question was a SFU coffee mug. The sellers, endowed with the mug when they entered the classroom, were given the chance to sell their mug for different prices ranging from $.25 to $9.25. The buyers, endowed with nothing when they entered the room, were surveyed on the same price intervals and range, but were asked what prices they would be willing to pay for the mug. The choosers, also endowed with nothing, were told they would be given a sum of money or the coffee mug, and were allowed to choose which they preferred at different prices within the range. The groups' responses are summarized below.[45]

Table 5: The Mug Market

Group	Average Price
1 Sellers: would not sell the mug for less than	$7.12
2 Buyers: would not buy the mug for more than	$2.87
3 Choosers: would choose cash over the mug at any amount higher than	$3.12

Here the large difference between the sellers' price and the buyers' price attests to some robustness of the endowment effect. If the choosers' price is taken to be the unbiased, true value of the mug, it appears the endowment effect is stronger for the person who has recently acquired the mug (the seller) than for the person whose endowment is the money they already have (the buyer), since the sellers' price is much farther off the choosers' price than is the buyers' price. More important for our study,

[45] Tversky and Kahneman, "Loss Aversion" 145.

however, is that the case for Maude's irrationality (Maude being someone who would have these different average prices for her value of the mug as seller, buyer and chooser) can again be made, in exactly the same manner as was done above with the mug-candy bar example. The only difference here is that since the difference in preferences is more robust, we have a stronger basis for postulating the endowment effect as responsible for Maude's change in preferences.

§5. The Internalist Move, Part Two

The internalist's earlier challenge of the assumption of reference independence was tabled, but now the time is ripe for such a challenge. Once again, such an assumption is necessary to get a case for Maude's irrationality off the ground. Maude is being irrational, the story goes, because her valuation of the mug is changing drastically depending on her state of reference (whether she is seller, buyer or chooser).[46] But the value the mug has to her should hold constant—it shouldn't depend on that irrelevant detail. If we accept this assumption, the violation of preference invariance follows.

As I mentioned above, there might be more of a reason to assume reference independence in this example than in our job example. Even in this example, where such an assumption might be *more* plausible than in the job example, I see no compelling reason to accept it. The assertion that the mug should have exactly the same value to us whether we are looking to buy or sell it is one that just does not agree with common practice (and it certainly did not agree with the above experiment's

[46] If it is puzzling why I am talking about value (and not preference order) when I am making the case that this violates preference invariance, it should be borne in mind that Maude's valuation of the mug is derivative of her preference order. A drastic change in value would indicate a drastic change in preference order.

results). I suspect most people would agree that selling a coffee mug we own for $\$x$ does not bring us quite the same utility (not quite as much) as receiving $\$x$ immediately, in lieu of ever owning the mug. It is even less like the utility of holding on to $\$x$ already in our possession and passing up a chance to purchase the mug for that price. Yet reference independence assumes these three utilities are equivalent—it insists that the reference point does not matter, when in fact we know it does. Even a formalized, prescriptive notion should be able to do better than this.

A more sophisticated version of rational choice theory does take these things into account, somewhat, when it includes costs that come with trading, factoring in negative utility that comes with the inconvenience of trading goods. R_e is generally limited to identifying such costs in concrete terms like time lost, however, and while such adjustments by rational choice theory may start to account for Maude's disparity in the amount she is willing to pay for the mug versus the amount she is willing to accept, it is hard to believe the inconvenience of trading, defined in quantifiable units like time lost, is solely responsible for the wide gap. One can imagine an experiment being designed where the concrete costs of trading are next to nothing, yet Maude's discrepancy persists. The leftover discrepancy, unaccounted for by concrete costs of trading, is Maude's irrationality, R_e would say.

Here R_i would point to our intuition that such labeling of Maude is not accurate. It seems that besides the concrete costs of trading, there might be other negative utilities for Maude involved in giving up the mug. One of the most obvious is the idea that she might have formed an emotional attachment to the mug when she received it, so that parting with it would bring her negative feelings. There is also the idea that she would fear before, during and after the exchange being shortchanged—not getting as much money for it as she should. We can imagine how R_e, holding to its charge of irrationality, might reply: if she feels negative emotions about giving up the mug, she should also feel negative emotions about passing up the opportunity to get money in exchange for the mug. Why should the negative emotion from the loss of the mug outweigh the negative emotion from the loss of the opportunity? As for the fear of getting a bad deal: if she fears getting a bad deal in exchanging the mug, she should equally fear passing up a good deal.

If R_e does reply this way, it gives hope to R_i because it seems that R_e has actually moved away from the reference independence assumption. Instead of holding that the reference point is entirely irrelevant, R_e seems to be admitting that the reference point does introduce emotional utilities—just emotional utilities that should cancel out. But the assumption that those utilities should cancel out is just as questionable as the reference independence assumption, and I believe this debate could continue in this way at length, threatening to lose the point. The main point is that the negative emotions associated with loss, what I am calling emotional utilities, *do* need to be considered, at the very least. A picture of rationality that completely ignores them and labels a decision that is apparently based on them as irrational is a picture too-narrow and not very useful. The better route for R_e is to at least entertain the possibility of incorporating emotional utility into rational choice theory's utility calculus, rather than making a premature conclusion that such emotional responses are always a force of irrationality.

Work has already begun on establishing the legitimacy of emotional utility, and although it has not always been done by traditional backers of R_e, the results should be of interest to R_e regardless. One burgeoning field that tries to assign quantified utilities to the emotions that accompany losing one's goods (or at least asserts that such an assignment should be possible) is neuroeconomics. Benoit Hardy-Valleé, in a brief paper on the topic, describes neuroeconomics as "the study of neural mechanisms involved in decision-making and their economic significance."[47] He identifies different regions of the brain connected with different aspects of decision-making, and a crucial part of the picture he paints embraces what he calls "the distributed account of utility" that he attributes to Kahneman. Instead of one general notion of utility, this distributed account posits four types of utility that comprise an agent's overall utility. The first is decision utility, which is "the expected gains and losses or cost and benefits." I take decision utility to be closest to the utility that a very simplified rational choice theory uses, and it might often be measured simply in dollars. The next is experienced utility, which is "the hedonic, pleasant or unpleasant affect" associated with the outcome. Next is predicted utility, which is "the anticipation of experienced utility," and last is remembered utility: "how experienced utility is remembered after the decision [and after the outcome, we might want to add], e.g. regretting or rejoicing." [48]

[47] Benoit Hardy-Valleé, "Decision-Making: A Neuroeconomic Perspective," *Philosophy Compass* 2 (2007): 1.

[48] Hardy-Valleé 9.

With this notion of utility and the findings of neuroeconomics, Hardy-Valleé offers the following "more precise explanation of loss aversion:"[49]

> Neuroeconomics explains loss-aversion as the interaction of neural structures involved in the anticipation, registration and computation of the hedonic affect of a risky decision. The amygdala, a structure involved in fear, emotional learning and memory modulation, registers the emotional impact of the loss; the ventromedial prefrontal cortex predicts that a loss will result in a given affective impact; and midbrain dopaminergic neurons compute the probability and magnitude of the loss (Naqvi, Shiv and Bechara; Tom et al.) Subjects are thus loss-averse because they tend to have or already had a negative response to losses (experienced utility). When they expect a loss to occur (decision utility), they anticipate their affective reaction (predicted utility). They might be also attempting to minimize their post-decision feeling of regret (remembered utility).[50]

The finer details of Hardy-Valleé's account are not as important for us as its overall implication: emotional utility (closest to what he calls "experienced utility") is a real phenomenon with a predictable corresponding brain activity. Joshua Greene found a similar connection in a study he and collaborators did involving activity in different brain areas during the making of difficult moral decisions. In this study, decisions of three classes were posed to test subjects—non-moral, moral-impersonal, and moral-personal—the three classes being designed to elicit increasing amounts of emotional engagement within the respondent. The data supported the hypothesis, as activity in brain areas associated with emotion was consistently highest in the moral-personal decisions. It was also noted that decisions that went against what the common emotional reaction would dictate (what Greene calls "emotionally incongruent"

[49] Hardy-Valleé 9.

[50] Hardy-Valleé 9-10.

decisions) consistently took more reaction time than did the more reflexive emotionally congruent decisions.[51] After further research Greene would later make a more general conclusion that the emotion-connected areas of the brain are especially influential in making quick (presumably relatively non-reflective) decisions that involve emotional engagement. In most cases this mechanism works fine, Greene suggests, but in some cases, usually in uncommon circumstances, it leads to decisions of questionable rationality.[52]

R_e might correctly point out here that these studies do not make the case that such decisions, with emotional aspects, are in fact rational. R_e can maintain just as easily as before that the loss-averse decisions in question are irrational. To isolate an area of the brain active during the decision does not mean that the decision is not irrational; it could just as easily be seen as finding the culprit, the part of the brain that, when activated, leads us to irrational decisions.

R_e's objection is useful in bringing us back to the point, but again it seems that if we fully accepted R_e's objection we would be prematurely dismissing an important issue that warrants much further study. As more studies are done, the emotional responses look increasingly like normal human responses, and we should be cautious in labeling such responses irrational, lest we reaffirm what is unfortunately the common practice of calling people consistently irrational. Certainly other possible explanations for this behavior should be considered, not the least of which is an evolutionary hypothesis for

[51] Joshua Greene, R. Brian Sommerville, Leigh E. Nystrom, John M. Darley, Jonathan D. Cohen, "An fMRI Investigation of Emotional Engagement in Moral Judgment," *Science Magazine* Vol. 293, no. 5537 (2001): pp. 2105-2108, 2106.

[52] "The Deciding Brain," *The Charlie Rose Show,* PBS, September 30, 2010.

loss-aversion, which would say that evolution has favored those who seek to keep the things they have. In other words, the advantageous nature of loss-aversion has caused it to evolve as a normal human response, and this is why we are habitually loss-averse.

Seeing loss-aversion as rational *qua* evolved behavior certainly requires a zoomed-out, larger-scale view on what makes a decision rational. Decisions that seem irrational when isolated can be part of a broader pattern of behavior that over a course of time—perhaps a lifetime, perhaps even several generations—proves to be advantageous. The idea of loss-aversion being advantageous in decision theory is analogous to the well-documented phenomenon in game theory of apparently irrational decisions in *The Ultimatum Game* proving to be advantageous over the long run. In *The Ultimatum Game*, player one is given an amount of goods (typically $10) to split between herself and player two. Player two, when given player one's offer (typically anything between $0 and $10 in $1-increments), has the choice to take it, in which case player one keeps the rest, or leave it, in which case neither player gets anything.

By rational choice theory, player one should offer $1 to player two (thus keeping $9 for herself), and player two should accept the offer. Actual results are quite different, however, as player two will consistently reject low offers, even though rational choice theory tells her getting something is better than nothing. Player two disregards that maxim, and wants to punish player one for what she perceives as an unfair offer. When this game is iterated with the same players, a norm of fairness will develop, unfair offers are made with decreasing frequency, and both players might be said to benefit.[53] Player one might make less money than if she had repeatedly made only $1 offers, but only if her $1 offers to player two were not consistently rejected leaving her with nothing. By making fair offers she gets a decent amount (around $5 or $6) consistently and with much more certainty. Player two clearly benefits because she can accept $4 or $5 each time instead of just $1.

The analogy between loss-aversion and this behavior in game theory and its long-term advantages is just one more point on the side of admitting emotional utility into the utility calculus. The actual work of this incorporation would undoubtedly require researchers and theorists to take this zoomed-out view, and this is a view that R_e is not well-equipped for. R_e's strength is looking at isolated decisions and judging their rationality on a micro-level, not considering what might be rational given long-term considerations of overall well-being. I see R_e as having to make a choice, however, between on the one hand trying to incorporate emotional utility and on the other hand doggedly insisting that emotional utility is an illegitimate factor in rational choice theory, taking a hard-line in order to preserve questionable assumptions like reference independence, keeping a neat theory whose applicability steadily decreases as we learn more about human behavior.

It's obvious what I think the better option is for R_e, but hopefully I have offered some good support for that claim. I hope to have shown, specifically, that the reference independence assumption is often inappropriate because it excludes the emotional utilities—which I have argued are legitimate—that are often very dependent on a

[53] Brian Skyrms, *Evolution of the Social Contract* (Cambridge: Cambridge University Press, 1996) 25-26.

person's reference point. Insofar as I have been convincing, we can claim a major victory at this point for R_i. It appears that we *should* leave room in the utility calculus for emotional utility, and that R_i *can* successfully challenge the key assumptions of R_e.

This victory for R_i does not come without its dangers, however, and just as R_i went too far after internal reasons were admitted by questioning overall consistency, it could once again go too far with emotional utility. Whether R_i makes the move itself, in its familiar self-destructive fashion, or R_e points us to the problem, the problem itself remains the same: it appears that once we admit emotional utility into the utility calculus we could be issuing a free pass for the rationality of any decision where emotion is involved. This is in part because we currently know so little about quantified emotional utility, but the more basic issue is that it is not clear how we should objectively quantify emotional utility at all. Also it's unclear on what grounds we could ever deny the legitimacy of an emotional utility. The consequence is that even emotional utilities that seem outlandish to most, like ones resulting from an extremely, disproportionately strong emotional attachment to an object, could be the bases for decisions we must call rational. Again the normativity we want for rationality is threatened. A goal of the next chapter will be to look at examples that highlight this problem, and to impose a final important externalist check.

CHAPTER 3: TO A GLOBAL RATIONALITY

In the first two chapters I aimed to show that when starting with the principles and structure of R_e, the way to R_g necessarily includes important internalist moves away from R_e. Contrary to the manner in which R_e is often applied, we should be considering what an agent knows when making a decision, the balance of the agent's internal reasons (while holding a basic consistency requirement on this matter), and the agent's emotional utilities. Overall these moves appear to affirm R_i over R_e. But we have also seen that important externalist checks are necessary along the way if the concept of rationality is not to become a trivial one, stripped of its normativity and its power to help us recognize and correct irrational behavior. In Chapter Three one more important externalist check will be added to our conception of R_g. I will then step back and consider the implications R_g has for the future of rational choice theory and decision theory, as well as what it means for the individual seeking to be rational.

§1. Case 3: Framing Effects

The third case study will present the strongest case we have seen for an assumption of reference independence, and consequently that some common decisions are irrational. In cases of this type, the only significant difference between two sets of options is the language in which they are described, or "framed." Since emotional utilities have been admitted, it is possible that such utilities should really depend on and vary according to the framing of an option, thus negating reference independence and the case for irrationality. This idea should certainly be challenged, however.

A simple example presented by Kahneman and Tversky demonstrates loss-averse decisions due to framing effects. In it, respondents are asked whether they would be willing to participate in two separate lotteries as described below:

> 1: Would you accept a gamble that offers a 10% chance to win $95 and a 90% chance to lose $5?
>
> 2: Would you pay $5 to participate in a lottery that offers a 10% chance to win $100 and a 90% chance to win nothing?[54]

We should note that risk is involved in this example, unlike our first two case studies. By this point we should be familiar enough with loss-aversion that we should be able to focus on that phenomenon without risk-aversion or risk-seeking interfering. This is especially true because the two questions seem to pose equally risky prospects, so again the element of risk is controlled.

The curious result this example produced is that of the 132 undergraduates questioned, 42 rejected the first gamble but accepted the lottery in 2.[55] If we take one of those 42 respondents and call him Frank, we can again make the case that Frank is making an irrational decision. As usual we would have to assume reference independence to make that case, but that looks like a safe assumption here, for the only difference between the two propositions is the language used to describe them. As their final outcomes go, they are identical. Nothing about Frank's point of reference is changing between the two problems; since it is the framing of the problems that is

[54] Kahneman and Tversky, "Choices" 15.

[55] Kahneman and Tversky 15.

different, we might say the only issue to debate is not the one of reference independence, but what we might call "framing independence."

Whatever we want to call it, however, the assumption seems reasonable, and the case for Frank's irrationality is that in the first problem he rejected the gamble, revealing his preferences, yet in the second problem he accepted the equivalent lottery, which proposed the exact same net results with the same probabilities. Frank is clearly violating preference invariance, and the loss-aversion hypothesis says the presence of the word "lose" in the first problem caused him to give a higher weight to the possible loss than in the second, where he is asked to "pay" a cost up front for chances to "win" different amounts.[56] Frank is displaying an inner inconsistency that produces a decision we can call irrational on R_e's view, or potentially irrational according to our current R_g.

The reason we call Frank's decision only potentially irrational by R_g, of course, is the knowledge requirement. Unless we know that Frank was fully aware of the equivalence of the two problems, this decision would not fit our definition of a full-fledged irrational decision. While the equivalence in this example is fairly transparent, it is certainly plausible that Frank did not see and appreciate this equivalence, especially if

[56] Although I stated above that I think the element of risk is mostly neutralized in this experiment, there is one difference in wording that may be significant: the first problem is called a "gamble" while the second is called a "lottery." I can't rule out that this difference may have elicited slightly different responses from people depending on their attitudes towards risk. It's unfortunate for my purposes that they didn't just frame both problems as "lotteries," though I understand it's technically incorrect to call the first problem a lottery. Anyway, I maintain that the most influential difference in the framing comes from the "loss-cost" difference, as do Kahneman and Tversky (15).

As an interesting side note, Kahneman and Tversky postulate that this example demonstrates what they call a "cost-loss discrepancy," meaning that not only do losses loom larger than gains (the standard definition of loss-aversion) but that losses loom larger than costs too, even if the difference between the two is nothing more than the word chosen to describe the loss (15).

his thought process was disrupted by feeling put on the spot or compelled to answer quickly. We can imagine a scenario where Frank comes to realize the equivalence of the two problems and sees the error in his ways. He realizes that if he did not want to accept the first gamble, he has no reason for wanting to accept the second, equivalent lottery, so he changes his answer to no for the second case as well. This would be a predictable, sensible story about Frank's loss-averse reaction causing a potentially irrational decision, but when Frank has sufficient opportunity to reflect the loss-aversion dissolves, and he can make consistent and rational decisions.

While the above scenario nicely demonstrates how loss-aversion in itself is not irrational, only that it has the potential to cause irrational decisions, it is not the most interesting scenario to consider. I have implied that were Frank's loss-aversion not to dissolve and he maintained his original choices, then we would certainly call him irrational. In light of the emotional utilities we have admitted into R_g, however, it is not so clear that this is the case. While we are assuming in this scenario that the knowledge requirement has been met, even then, if Frank's loss-aversion persists regardless, can R_g really call him irrational? We can imagine Frank explaining that he does see that the two cases are equivalent, still he doesn't care. The mere presence of the word "lose" in the first problem causes a loss-averse emotional utility in him, and he feels he must account for that in order to make the right choice. There is the brute fact of the negative emotional utility for the first problem that he can't help, regardless of his knowledge that there is really no difference, he explains. This is the scenario that exposes the problems that come along with admitting emotional utility into R_g.

§2. The Uncomfortable Internalist Commitment

R_g's current predicament is due to the fact that it moved away from R_e and towards R_i in accepting emotional utilities as legitimate. R_i may try to assuage R_g's concerns by pointing out that it has intuition on its side (and in its name) and offering the intuitive answer. We know that Frank who adjusts his decisions (Frank$_a$) is being more rational than Frank who persists in his original choices (Frank$_p$) because Frank$_a$ is turning away from an irrational emotion whereas Frank$_p$ is clinging to it. The loss-aversion is an irrational emotion in this example because reflection uncovers that it has no good reason behind it—it is merely the result of an illusion created by language.

While R_i's intuitive answer does hold some promise, as it stands R_g must still be troubled, for it doesn't suggest a way to distinguish between legitimate and illegitimate emotional utilities other than this seemingly *ad hoc* method. R_g has admitted emotional utilities, but now it sees that without some structured approach to evaluating the legitimacy of these emotional utilities, the process threatens to become an arbitrary mess. What would stop R_e from making the case that all emotions are irrational by definition, putting us back where we began with no legitimate emotional utility in the utility calculus? It looks like R_g may need to accept all emotional utilities, even the dubious ones. It might have to accept Frank$_p$'s argument that once we have the brute fact of the emotional reaction and thus the emotional utility, the decision based upon it can be considered rational; the question of the grounds for the emotion is outside the scope of rationality, as we are treating it.

I offer one more example that really highlights the problem. In it we have two respondents, Mindy and Cindy, each with just a one-dollar bill currently in her pocket. They are both offered the chance of a lifetime to pay that one dollar immediately to play a lottery that offers a 99% chance of winning $1 million, and a 1% chance of getting nothing. Both women consider the prospect and decline to play. As we know by now, we certainly should not call them irrational at this point before doing some investigating, attempting to expose an inner inconsistency. When we question Mindy we learn that she is a “normal” person in every respect, and would welcome winning $1 million if it cost her only a dollar. However, the particular dollar bill in Mindy’s pocket is an extremely rare one that she could sell for $2 million whenever she pleases. Suddenly her decision looks like a very good one—loss-aversion is not an issue, and it’s simply a choice for the higher utility.

Cindy is a more curious case, however. Cindy says she declined the lottery just because she “really, really likes” the particular dollar bill in her pocket. It is not a rare one like Mindy’s—it is really worth only $1. Furthermore, Cindy admits that the $1 million is not unattractive to her. If she played and won the $1 million, she’d be very happy she played. The 1% chance that she’ll give up her dollar bill and end up with nothing is just too great, however, and such a loss would absolutely devastate her, so she is confident that passing up the lottery is the right decision for her.

Is R_g, because it admits emotional utilities and wants to do so consistently, doomed to a commitment to Cindy’s rationality? Cindy has an extremely high loss-aversion connected with her dollar bill, which produces an extremely large negative emotional utility that should factor into her decision. Is R_g forced to say this is a rational decision regardless of the fact that when questioned about her special dollar bill, Cindy states that she just likes it and would really hate to lose it, and that she really can't explain it further?

§3. Reasons, Causes and the Next Externalist Check

Intuitively, what distinguishes rational from irrational decisions in the examples above is that in the cases that seem irrational, something is lacking in the explanation for the decision. Frank$_p$'s account, and especially Cindy's when compared to Mindy's, do not represent the kind of thorough explanation we think is necessary for a decision to be rational. As I have argued above, however, R_g cannot stand pat with simply asserting that this intuitive difference is what will determine our questions of rationality—that what seems rational to us is rational and what does not is not—for this would introduce a host of problems. Some kind of structure and formal definitions will be necessary to bolster this intuition. Recall that on the question of consistency, a temptation was to simply decree that rejecting the end of overall consistency is out of play. While this stipulation was added to R_g in the end, it was not put there by mere decree. Instead we were able to find an argument from Davidson that the stipulation was a necessary consequence of the fact that rationality deals with thinking, judging beings. I think that with our current issue we can also do better than a mere decree.

One possibility is that with the consistency requirement admitted, we already have enough at our disposal to call Frank$_p$ and Cindy irrational. It would simply be a matter of probing deeply enough into their preferences and ends to expose an internal inconsistency. I think that internalists like Williams, and especially Davidson, would advocate this route. If the hypothesis is true that Cindy is "normal" in all respects except those regarding this particular dollar bill, we can understand why the internalist might be confident that he can expose an inner inconsistency. In fact, given the interconnectedness of propositional attitudes, the hypothesis is unrealistic, and such a loss-aversion would accompany other strange propositional attitudes until either there is a contradiction or we begin to question Cindy's status as a thinking, judging being. If we suppose for argument's sake, however, that a probe fails to reveal that Cindy's loss-aversion for this dollar bill contradicts any of her other ends, I don't think this means our only option is the mere decree, defining rationality so it excludes her. There is another way we can focus on Cindy's loss-aversion for the dollar bill, independent of most (if not all) of her other ends, and call it irrational.

The intuitive idea of a sufficient explanation *will* play a central role in what I take to be a better solution, but I want to offer a bit from Davidson to fill out that criterion. In a passage reminiscent of our earlier passage about the constitution of propositional attitudes, Davidson states,

> In standard reason explanations…not only do the propositional contents of various beliefs and desires bear appropriate relations to one another and to the contents of the belief, attitude, or intention they help explain; the actual states of belief and desire cause the explained state or event.[57]

Again we see that propositional attitudes necessarily entail logical relations to each other and causal relations to the world. In a reason explanation, then, both the reason and cause for the event need to be present. "In the case of irrationality, the causal

[57] Davidson, "Paradoxes" 179.

relation remains, while the logical relation is missing or distorted…there is a mental cause that is not a reason for what it causes,"[58] he explains, offering what I believe is a crucial distinction that is the beginning of a solution to our present problem. His statement that "many desires and emotions are irrational if they are explained by mental causes that are not reasons for them,"[59] perfectly characterizes the underlying trouble with the explanations given by Frank$_p$ and Cindy. They can explain what causes them to decide as they do, but they cannot give a good reason for that cause. In a sense, both Mindy and Cindy have the same cause for their decision not to play the lottery—the mental event that is their desire to keep their dollar bill. But only Mindy's explanation counts as a reason explanation, for she has a cause with a good reason, while Cindy has merely the bare cause.

§4. R_g: Current State and Future

With our latest and final externalist check in place, I again offer R_g, in what will be, for us, its final form:

> A decision is **right** if and only if its actual outcome is at least as good as that of every other possible outcome.
>
> A decision is **rational** if and only if the decision maker, given ends* she has internalized and knows, chooses to do what, as far as she knows, she has most reason** to do at the point in time at which the decision is made.
>
> * *End* here is not to be construed so broadly that it could include an end to defy overarching consistency, or—what that amounts to—an end to be irrational. Such constraints of consistency and rationality are preconditions for thinking, judging beings, rather than ends that one can choose to adopt or reject.
> ** *Reason* here, while it respects the internalist perspective that one must know about the reason to have it, is not to be construed so broadly that it includes

[58] Davidson, "Paradoxes" 179.

[59] Davidson 179.

bare causes, i.e. mental events that explain how a decision is caused without actually being a reason for the decision.

The main text of R_g has not changed at all from the form in which we left it at the end of the first chapter. As we have seen, however, the devil has been in the details of clarifying further how crucial words are to be read. The first note comes from our first externalist check that outlawed the end to be inconsistent. As for the second note, we have seen that 'reason' can be even more loaded than we initially assessed, as now we see the need for specifying that a bare cause must not be mistaken for a reason.

R_g stands clearly distinguished from the R_e and R_i that we began with, for it has cast off the unrealistic rigid requirements of R_e, while it has kept hope for normativity by staying away from the problematic extremes that R_i runs towards. The question of which original notion, R_e or $R_{i,}$ has the most influence on R_g can be debated, I think, just as the sides of internalism and externalism debate each other when they modify their stances to accommodate intuitions. Externalists may claim that by incorporating my externalist checks I have given the deciding nod to them. We find Hollis making a similar claim when addressing the question of how, on the internal reasons account, desires that are not immediate can possibly override immediate desires. He examines Williams' answer, his definition of the self as "a construct marked by a project,"[60] and points out that this project must be sufficiently rigid, lest it be subject to the whims of immediate desire and thus no help at all. The more rigid it is, however, the more "the shape of this overarching project…sounds uncommonly like the guidance of external reasons."[61]

[60] Hollis 85.

[61] Hollis 86.

Just as Hollis concludes that “the idea of a grand project gives the game to the external reasons theorist,”[62] he might argue that my stipulations added to R_g have given the game to R_e. R_g is the same in spirit as R_e, he might say. It is only a minor difference that R_g insists we consider more complex utilities, and R_e can readily do that.

We should note at this point, then, that for what it is worth, we might have the same end results for R_g without the stipulations being stated explicitly. This seems clear at least for the second stipulation, for as I mentioned above, the internalist might be able to resolve all actual problem cases by finding an inner inconsistency. For the cases of Frank_p and Cindy, for example, if we can assert that all thinking judging beings necessarily have the end to “make decisions for good reasons,” that might be all the internalist needs—Frank and Cindy seem to be at least somewhat aware that they are violating that end. The problem with this is that it assumes an understanding of “good reason” which relies on the reason-cause distinction, which is precisely the distinction my explicit stipulation attempts to draw.

Is my explicit addition of the reason-cause distinction necessary, then? Does the answer to this question determine whether external or internal reasons wins the day? Luckily I think this question matters only to those who have a vested interest in the internal-external reasons fight, and I do not. As far as I am concerned, adding these explicit stipulations makes R_g much more ready for real use, and it bothers me little if the theoretical debate has not been settled. I thank the internal-external reasons debate for modeling for me how each side can attempt to account in its own terms for

[62] Hollis 86.

intuitively attractive features of the other side, as well as how the most sensible and viable answer lies in the middle.

I say that R_g is more ready for real use now that my stipulations have been added, but clearly there is a mountain of work to be done before R_g can reach its full potential in applicability. Much of that work will be in the fields of behavioral economics and psychology, for if an accurate utility calculus requires that emotional utilities be incorporated, as I have argued, then it is clearly important that we learn more about these utilities through the experiments of behavioral economists and psychologists. Empirical studies will need to be done to quantify emotional utilities. The mug example we examined in Case 2, where the endowment effect's strength was suggested by the dollar value differences between the owned mug and the not-owned mug, offers a good model of the kind of experiment that might quantify emotional utility. I imagine that more experiments where the threshold is sought—the point at which the person decides he should discard the emotional attachment and accept the surprisingly large sum of money for the mug (and perhaps marvel that the previous, also large, offer was not enough to pry the mug away from him)—would do much to put loss-aversion in quantifiable units.

Another type of experiment that could quantify loss-aversion due to framing effects would be similar to the lottery choice in Case 3, except it would seek to find the robustness of this effect. The example in Case 3 offered equivalent options and loss-aversion produced inconsistent preferences, but I imagine that loss-aversion is often robust enough that people will often choose the option phrased in non-loss-language over a *superior* option that happens to be phrased in loss-language. Again, finding the threshold, exactly how inferior an option can be before even a framing advantage cannot make people choose it, would help us in quantifying loss-aversion, and thus emotional utilities.

It may also be apparent by now that not only will such experiments offer guidance on how to quantify emotional utilities, but they can also help define a baseline amount of loss-aversion—the “normal” amount. This will be crucial information because what’s irrational will, in part, be an empirical question. There will be the theory of R_g to help us know, but we will also need to heed the empirical clues we can gather. The question of how far one can deviate from normal emotional utilities and still be considered rational will need to be addressed, and we cannot do so without some idea of what the normal amount is.

Finally, two more empirical matters that also need to be addressed are how and in what quantities loss-aversion is advantageous over a lifetime. In Chapter 2, I suggested one of the main justifications for the rationality of the loss-averse reaction is that it may prove to have long-term benefits, hypothesizing that if the behavior evolved in us in the first place, it must have been advantageous in the past. Investigating the advantages of loss-aversion over a lifetime in today’s world is certainly a tall task that would probably necessitate many long-term, complex studies. I can only offer as consolation that at least I am offering plenty of work to keep researchers busy.

There is also important philosophical work to be done for R_g which, compared to the empirical work I have just suggested, looks entirely manageable (though obviously I did not deem it so here). What I have offered so far is a philosophical answer about loss-aversion. According to what I have argued, loss-aversion in itself is not irrational, first because that would not respect the epistemic point I have made about irrational decision entailing a knowledge in the decision-maker that the decision is wrong. There are certainly cases where a person is unaware that their loss-aversion is leading them to the wrong decision, and these are different from truly irrational decisions. Even if one *is* aware that loss-aversion is leading him to make a decision that on some accounts might be called inconsistent, however, we have seen with the admission of emotional utilities that the loss-aversion still may be leading him to the correct decision. In cases like these where the loss-aversion is justified, or there is good reason for the loss-averse reaction, the loss-averse decision is not irrational. In fact the opposing decision, which would disregard the loss-averse reaction as automatically unfounded and illegitimate, would be at least a mistaken decision made upon an unsophisticated and mistaken rational choice theory, if not an irrational decision.

I have also argued, however, that loss-aversion still has the potential to lead to irrational decisions in many cases, namely ones where the agent sees that the loss-averse reaction is groundless—is merely a mental cause rather than a cause *cum* reason—yet persists in the loss-averse behavior. I've offered the beginnings of a theory that can help answer the crucial question of when loss-aversion is justified and when it is not: the reason-cause distinction. But as it is this idea is not ready for use; it is just the rudiments of a structure. A large remaining philosophical task, then, is filling out the details of this distinction in a way that is relevant to rational choice theory. For example, some characteristics of the typical cause *cum* reason might be offered alongside some characteristics of the typical bare cause.

I believe there will always be some problem cases, some gray areas regarding a possible reason-cause schism. It will be debatable sometimes whether an emotional utility is just a bare cause or whether it has at least a weak reason attached, and I don't think any philosophical reason-cause theory, no matter how filled-out it is, will be able to give a definitive answer for every case. This is why, as I mentioned above, we will need empirical clues: an emotional utility that looks outlandish compared to the norm is more likely to be a bare cause. Nevertheless, a philosophical reason-cause distinction filled out to the best of our ability will still be a valuable tool in answering questions of rationality, and it represents a definite improvement over the intuitive approach of calling some explanations good and others bad.

Overall, then, the future work I am suggesting necessarily involves much interaction between the empirical and philosophical. The empirical will need guidance from the philosophical to avoid calling the behavior it observes rational simply by virtue of its being normal human behavior. The philosophical will need guidance from the empirical to prevent it from overstepping its bounds and defining a rationality before it is even clear what rationality should be. I see this interaction as the way in which society as a whole moves towards a better definition of what is rational.

§5. Self-Knowledge and the Individual's Quest for Rationality

I argued early on that a decision should be called irrational only if we know that the agent knows in some capacity that the decision he is making is wrong. This epistemic point has remained mostly in the background throughout our case studies, however. Certainly I have offered plenty of reminders that by my account, this knowledge is necessary if a decision is to be properly called irrational. The point also saw a slight improvement when, in light of the discussion of consistency, I could modify the issue from the agent "being aware that he's wrong" to "being aware of the inner inconsistency." My main focus in the case studies, though, was whether things were even potentially irrational. I asked if it is even potentially irrational to have inner inconsistency, because one might reject the end of inner consistency. I also asked if it is even potentially irrational to make a decision based on emotional utilities, especially if we establish legitimacy for emotional utilities. If, despite these questions, the potential irrationality was established, then actual irrationality was a mere matter of the agent becoming aware of the problem and refusing to adjust accordingly.

As this study comes to a close, however, I think it is important to revisit this epistemic point, and explain why I felt it necessary to constantly remind my readers that inconsistency needs to be recognized as such for a case of true irrationality to occur. In short, it is because my conception of the individual's quest for rationality needs it. I begin my conception with an idea from Davidson (which he says he borrows from Freud), the idea of semi-independent structures in the mind. "If parts of the mind are to some degree independent, we can understand how they are able to harbour inconsistencies, and to interact on a causal level,"[63] he states, and makes it clear that such a hypothesis is necessary if we are to make sense of irrationality. Davidson cannot

[63] Davidson, "Paradoxes" 181.

conceive of how a person can actually believe "*p* and not-*p*," yet this is what every internal inconsistency reduces to. If "*p*" and "not-*p*" are harbored in different parts of the mind, however, he can understand how they can coexist.[64]

Up to this point, I like Davidson's explanation and the spatial metaphor it suggests, where contradictory beliefs are kept a safe distance from each other. He often suggests that the mere coexistence of contradictory beliefs can be called irrationality, however: "If someone has inconsistent beliefs or attitudes, as I have claimed (objective) irrationality demands, then he must at times believe some proposition *p* and also believe its negation."[65] Granted Davidson does not state that this is *sufficient* for irrationality, but I am also unable to find him specifying, as I do, that the agent's awareness of the inconsistency is also necessary.

I think that Davidson's way of speaking makes rationality too strict, as even the most obscure contradictions would be irrationalities. This hearkens back to my argument for making the epistemic point in the first place: without it, many decisions that are just 'wrong' become 'irrational.'[66] Davidson states, "In the possible case, of simultaneously, and in some sense actively, believing contradictory propositions, the thinker fails to put two and two (or one and one) together."[67] To be sure, it is not incongruous to call the failure to put two and two together irrational, but would we

[64] Davidson, "Incoherence" 198.

[65] Davidson 198.

[66] This is also evidence for my earlier claim that Davidson was not addressing these issues starting from the problem that I am treating in this thesis, for if he were he wouldn't have made it so easy to run together 'wrong' and 'irrational.'

[67] Davidson 198.

want to say the same about, say, failing to take 54 to the 17th power? The answer is just as certain, if not as accessible, as putting two and two together, and technically if we are to call a failure to do the addition irrational, we must also call the failure to exponentiate irrational. This point is most demonstrable in framing examples. The example we studied in Case 3 was relatively transparent, but we could lure Frank into the exact same inconsistency by way of a much more complicated formulation.[68] Rather than try to draw a line between easy operations that should be done, and difficult ones for which he have an excuse, I think it is much more natural to simply require awareness of an inconsistency before we have irrationality—until then it is only potential.

This distinction upon which I insist, the epistemic point, allows us to truly appreciate the dynamism of the individual's quest for rationality. By my conception, she is rarely, if ever, at full-blown irrationality, and the same holds for rationality. The more self-reflective she is, however, the more often she will be forced to adjust her propositional attitudes. For it is self-reflection that puts *p* and not-*p* on the collision course, and if she wants to avoid the fiery crash of irrationality she must either adjust (by rejecting either *p* or not-*p*) or stop the process of self-reflection.

Here I should admit that on my account, one who never engages in self-reflection and thus can never become aware of inner inconsistencies is in no danger of irrationality. This is hardly an endorsement of the one who refuses to reflect, however.

[68] See Tversky and Kahneman, "Rational Choice and the Framing of Decisions," *Choices, Values, and Frames,* ed. Kahneman and Tversky (Cambridge: Cambridge University Press, 2000) 213, for a good example of a choice between two fairly complex lotteries. The respondents consistently choose the lottery of *lower* expected value, and it is understandable because given the complexity, they have little to go on for their decision other than the inferior lottery being phrased in more favorable terms.

While it is true that technically I don't call her irrational, this certainly does not mean she is better than the self-reflector who occasionally lapses into irrationality. Her absolute refusal to reflect on any of her propositional attitudes (if she can have any) makes her more like Joe from Case 1, for it causes her to be a stranger to the house of reason. As implied, I do think most if not all human beings do engage in this self-reflection. It is a good thing, because by my account it is only by this process of self-reflection and averting the collisions of irrationality—this hurtling oneself towards irrationality so that he can weed it out—that the individual can move towards rationality.

On the larger, societal scale, the same process is mirrored. Society cannot simply theorize its way to rationality. Like the individual, it needs to reflect on itself, in this case in the form of empirical studies, so that it can see what should be held as rational. Where the theory and empirical data conflict, an adjustment must be made to avoid the crash of irrationality. This characterizes the process for judging the legitimacy of emotional utilities, as well as the process for defining reason-cause criteria.

Finally, the individual cannot engage in his own quest without society also engaging in its quest. The individual often needs to defer to what society tells him about human beings for help in determining whether he has exposed an inner contradiction, and whether he might be acting on a mental cause without a reason.

I am just touching on what is probably becoming increasingly apparent: a complex network of dynamic processes that together, in time, can determine a better and more correct notion of rationality. I hope that what I have done here has at least drawn attention to some serious problems with current conceptions of rationality, and offered a way to begin understanding a global rationality. In response to the question, "Are people rational?" Kahneman suggests that "the time has perhaps come to set aside the overly general question,"[69] and this seems especially wise in light of all I have argued. If one insists upon asking the question still, however, I reply that it takes a mistaken notion of rationality to even venture an answer to the ill-formed question. By one notion, which imposes a rigid theory of rationality in spite of rationality not being fully understood, people are often irrational—in fact so often that one might begin to wonder why we should want to be rational. By another notion, people are almost always rational, in an almost trivial way. It defers to our ignorance about rationality and insists that "there must be some reason, even if we don't understand it." But if rationality is to be properly understood (or if we are to *begin* to properly understand rationality) we need to ask whether people *move towards* rationality. I hope I have shown to that question we can answer that if we are engaging in a process whereby we are trying to make better decisions and improve the way we get through life; if we are holding ourselves to some inflexible standards of consistency while also recognizing that we need to learn more about ourselves, because there might be more to us than we understand; if we are admitting that there can be reasons for our decisions that we do not fully understand, without resigning ourselves to a metaphysical truth that there are always reasons for our decisions which we will never understand, then we are engaging in, or chasing after, rationality.

[69] Kahneman, "New Challenges to the Rationality Assumption," *Choices, Values, and Frames,* ed. Kahneman and Tversky (Cambridge: Cambridge University Press, 2000) 774.