UNIVERSITY OF OKLAHOMA

GRADUATE COLLEGE

ESSAYS ON THE INTERPRETATION OF DECISION THEORY

A DISSERTATION

SUBMITTED TO THE GRADUATE FACULTY

in partial fulfillment of the requirements for the

Degree of

DOCTOR OF PHILOSOPHY

By

ZIMING SONG
Norman, Oklahoma
2019

ESSAYS ON THE INTERPRETATION OF DECISION THEORY

A DISSERTATION APPROVED FOR THE
DEPARTMENT OF PHILOSOPHY

BY

Dr. Stephen Ellis, Chair

Dr. Gregory Burge

Dr. Hugh Benson

Dr. James Hawthorne

Dr. Zev Trachtenberg

Table of Contents

Acknowledgements

I am indebted to more than whom I can acknowledge here. I am thankful to my undergraduate philosophy education at Wuhan University and the professors there that fostered my intellectual curiosity, so that I had continued my doctorate in philosophy. I am grateful to the philosophy community at the University of Oklahoma, a place that has allowed me to develop not only my intellectual interest but also the various skills and virtues of a being a professional philosopher. Dr. Wayne Riggs, who stepped in as Chair of the department the same semester that I came in, as well as Dr. Linda Zagzebski, Dr. Sherri Irvin, Dr. Martin Montminy and Dr. Amy Olberding, among other faculty, whose classes I have tremendously enjoyed, have been essential to a congenial and supportive community for my professional growth and intellectual development. I am especially indebted to my dissertation committee members, Dr. Hugh Benson, Dr. James Hawthorne, and Dr. Zev Trachtenberg. Under their guidance, it has been pure joy for me to read and analyze classical philosophical works, and to express thoughts in mathematically elegant ways. You have been, and will continue to be, the philosophical mentors that I respect and admire.

Lastly but most importantly, none has been more influential on me than my dissertation advisor and mentor, Dr. Stephen Ellis. The initial ideas of my dissertation topic emerged from his Philosophy of Social Science class, and then developed through many discussions with him. My area is fairly new in the discipline and the territories are not as clearly defined. For quite some time before this dissertation had taken shape, my research felt like walking in the fog. Looking back, I realize how invaluable Steve's guidance was in shaping this project during its most critical period, turning it from rough ideas to fine writing. The quality of this dissertation has

been incredibly enhanced by the many hours we spent working through multiple drafts. The final product has benefited greatly from his detailed, sentence-by-sentence suggestions on all its parts. Steve has also been of tremendous support in my professional development as a teacher and an all-round academic. Thank you for your care and oversight, and I will continue to look up to you as a role model.

My gratitude extends to my fellow colleagues and friends whom I met both in graduate school and at conferences. I especially thank Dr. Madeline Martin-Seaver, Dr. Andrew Chau, Dr. Patrick and Kelly Epley, Dr. Jonathan Rutledge, Wenhui Xie, Dr. Stacey Goguen, and people at the OU Writing Center, for making graduate school and dissertation writing considerably more cherishable for me. Finally, I am most thankful to my mother, Huifang Li, and my father, Honghua Song, whom I lost along the way, for their compassion and encouragement so that I can pursue what I love. I am also thankful to my extended family members for their continuous love. This dissertation is dedicated to my father, and to my cat, Leah.

Abstract

Decision theory (DT) aims at explaining and predicting rational choices. But ample empirical evidence suggests that descriptively, people's actual choices do not conform to its predictions in various ways. Some of the counter-evidence, such as cases proposed by Amartya Sen, challenges even the normative adequacy of DT. My dissertation defends DT as a normative theory while also offering a novel explanation of why DT fails descriptively. Contrary to the Sen-style critique, I argue that DT is unimpeachable as a logic of choice. Human decisions do not range over the kinds of objects that philosophers and economists normally assume that they range over. Instead, they range over much finer grained outcomes, what I call "holistic outcomes." A holistic outcome reflects everything that affects a decision, including the full range of one's values and the particularities of a choice context. Appreciating that decisions are made over holistic possibilities has important consequences for how we understand DT both as a normative and descriptive theory. My dissertation aims at establishing that human decisions range over holistic outcomes and then tracing out the consequences of this insight.

Chapter 1

Introduction

Decision theory is a multidisciplinary subject. From a philosophical perspective, it is an account that tries to explain human actions and helps us evaluate choices. It makes more precise the folk-psychological insight that we tend to choose options that best realize our ends given our beliefs – and ought to do so. The idea that people are instrumentally rational in this way is utilized for both normatively evaluating and descriptively explaining choice behavior.

Decision theory (DT hereafter) formalizes our folk-psychological intuitions about instrumental rationality, and thus carries over the normative elements of folk psychology. It is defensible as a norm of choice – or so I argue – because it provides a coherent account of preference revision and identifying optimal courses of action in new situations. Empirical evidence, on the other hand, suggests that decision theory is not a good descriptive account - people's actual choices do not conform to its predictions in various ways. My dissertation research explores the implications of disconfirming empirical data on normative decision theory, in particular. I concede that decision theory is not an adequate descriptive-explanatory account, but I hold that, properly understood, it succeeds as a normative-evaluative account.

In most cases, counterexamples to descriptive decision theory are not also normative counterexamples: where actual behavior diverges from the account, the behavior seems mistaken. Explaining the way in which the behavior deviates from the model will often suffice to

convince even the agent that she has chosen poorly. I see such cases as evidence against the descriptive adequacy of DT, although not its normative adequacy.

There is a range of cases, however, where descriptive counterexamples seem to have normative importance. In such cases, decision-theoretic patterns (at least seem to) fail but the behaviors observed (at least seem) to be perfectly reasonable. Amartya Sen, for example, considers a case where someone would have chosen an apple over having nothing when faced with a basket full of apples, but chose to have nothing over having any apple when faced with a basket containing only one apple. This behavior seems perfectly sensible – when in a group, the person doesn't want to be rude by taking the last piece of fruit. Still, it seems to violate a standard principle of decision theory: if one set of options is a subset of another and the best option in the larger set is available in the subset then that option should still be the best option.

My solution, basically, is to say that the violation of decision theory is merely apparent: choosing apple *x* when there are many other apples is really a different action than choosing apple *x* when it is the only apple left. The latter involves what Sen calls "rule[s] of good behavior" in a way that the former doesn't. More generally, problems appear to arise only when decision theory is wrongly applied. Basic outcomes need to be distinguished finely enough to differentiate anything that would make a difference to the agent's evaluation. I call this the ***holism*** view.

In Chapter 2, the first essay of this dissertation, I begin with the connection between decision theory and an account of instrumental rationality. The central task of understanding instrumental rationality is understanding what it means to behave consistently with one's own evaluations,

choosing to realize some valuable states of affairs rather than others. Decision theory is rooted in folk psychology that contains normative intuitions about consistent choosing. I interpret the formalism of decision theory as an account that evaluates choice behavior (*actions*) in terms of evaluation of *prospects* which is grounded in the foundational evaluation of *outcomes*. I distinguish my interpretation from the revealed preference interpretation popular in economics, and then develop my theory as a logic of choice. The standard formalism of decision theory is taken to be consistency conditions on rational choosing, and those consistency conditions are thought to apply across contexts. In other words, the consistency conditions are internal to the formalism and are independent of external factors and particularities in context. I consider Sen's critique that the formalism of decision theory fails to be internal and sufficient to govern rational choices across contexts. Although Sen's critique is originally framed as a critique of the internal consistency conditions of its revealed-preference interpretation, I argue that it challenges my logic-of-choice interpretation of decision theory as a realist, normative view as well. The rest of the chapter is devoted to articulating the holism view as an addition to my logic-of-choice interpretation and how it successfully answers the Sen-style challenge. In a nutshell, the holism view "fine-grains" the objects of choice and thus defends decision theory as a norm of choice from Sen-style critiques. More generally, my view is that our folk-psychological intuitions, as formalized by decision theory and understood in light of my holism strategy, serve as an adequate normative account of instrumental rationality.

Chapter 3, the second essay, develops the holism interpretation by comparing it with Buchak's alternative account of decision theory with respect to attitudes toward risk in particular. The "fine-graining" of the objects of choice I advocate seems like an obvious move, but it is opposed

by several authors on a variety of grounds. Most of those arguments are intended to safeguard decision theory as a descriptive-explanatory tool. Buchak argues that risk-weighted utility theory is a superior alternative to fine-graining for certain sorts of cases. I consider this challenge by arguing that risk attitudes can be accommodated by my holistic strategy. I show how my approach handles Buchak's problem cases – including the classic Allais paradox. Moreover, I note that Buchak's theory is subject to Sen-style counterexamples, just like standard decision theory, if the basic outcomes are not holistic.

In Chapter 4, the third dissertation essay, I extend my arguments to game-theoretic models. In strategic interactions we notice that an agent (player) forms her preferences of game outcomes based on a number of factors, some of which seem specific to the concrete game context. This feature suggests extending the 'fine-graining' strategy to game theory. Building on Bicchieri (1993) and Skyrms (1988), getting to the solution of a game (a Nash equilibrium) requires one to correctly figure out one's opponent's preferences and probabilities. The requirement that basic preferences evaluate holistic outcomes suggests that it will sometimes be unrealistically difficult to figure out one's opponent and so that using descriptive game theory to get to the solution of a game is sometimes also unrealistically difficult. Similar to my view, Guala (2006) argues for the failure of descriptive game theory. But my 'fine-graining' strategy defends game theory instead as a normative view of rational choice in strategic interactions: it is good advice when a game is correctly represented, that is, the strategic outcomes are fine-grained enough to be evaluated unequivocally by the players.

Chapter 2

Decision Theory as a Logic of Choice

1. Chapter Overview

Consider the following case from Amartya Sen.

The Last Apple

> Suppose the person faces a choice at a dinner table between having the last remaining apple in the fruit basket (*q*) and having nothing instead (*p*), foregoing the nice-looking apple. She decides to behave decently and picks nothing (*p*), rather than the one apple (*q*). If, instead, the basket had contained two apples, and she had encountered the choice between having nothing (*p*), having one nice apple (*q*) and having another nice one (*r*), she could reasonably enough choose one (*q*), without violating any rule of good behavior. (Sen 2002, p. 129)[1]

The apparent choice pattern here is inadmissible according to Decision Theory (DT): if *q* is preferred to *p* where *p*, *q*, and *r* are options, then the elimination of the unchosen option *r* shouldn't change that.[2] Sen concludes that DT is inadequate as a general account, then, because it cannot deal with such cases with its own resources. To explain what is going on, a theorist must appeal to *external* factors in a concrete choice context that go beyond DT - the particulars of a decision maker's motives, objectives, and values, social norms, etc. (Sen 2002)

---

[1] I have altered Sen's formal notation to be consistent with my own. The need for this move will be apparent below.

[2] There are various ways to characterize this sort of principle - Independence of Irrelevant Alternatives, Weak Axiom of Revealed Preference, Basic Contraction Consistency, etc. - but something along these lines is required for appropriate choosing (Sen 2002).

Certain social scientists - mostly economists and those in affiliated disciplines - see DT (also known as Rational Choice Theory, RCT) as a good *explanatory* or *predictive* account. Others are more impressed with the empirical evidence mounting against such interpretations. My interest is distinct from this debate.[3] DT, I argue, is an account of *instrumental rationality* in that it provides formal standards of consistency for navigating choices, independent of what in particular a decision maker prefers or believes. Understood in this way, DT provides a *logic of evaluation/choice*: it tells you that if *these* are your basic preferences over, and beliefs about, states of affairs then *this* is what you ought to want, and so to choose, in order to be rationally coherent.

Sen's Last Apple case and those like it pose a problem for a logic-of-evaluation view. Decision theorists have uncovered a large amount of experimental data over the last few decades that is inconsistent with the explanations and predictions of standard DT. Some well-documented violations of DT principles seem like cognitive *errors*: explaining the way in which the behavior deviates from the model suffices to convince even the agent that she has chosen poorly. In Tversky's and Kahneman's famous "Asian disease" case, for example, two sets of subjects are presented with the same choice options described using different phrases, and so 'framed' in different ways. Subjects' choices differed systematically with the different frames.[4] This framing

[3] I think the critics of descriptive RCT have already prevailed.

[4] "Problem 1 …: Imagine that the U.S. is preparing for the outbreak of an unusual Asian disease, which is expected to kill 600 people. Two alternative programs to combat the disease have been proposed. Assume that the exact scientific estimates of the consequences of the programs are as follows: If Program A is adopted, 200 people will be saved. … If Program B is adopted, there is a one-third probability that 600 people will be saved and a two-thirds probability that no people will be saved. … Which of the two programs would you favor? … Now consider another problem in which the same cover story is followed by a different description of the prospects associated with the two programs: Problem 2 …: If program C is adopted, 400 people will die. … If program D is adopted, there is a one-third probability that nobody will die and a two-thirds probability that 600 people will die. … Which of the two programs would you favor?" (Kahneman and Tversky 1984, p. 341-350)

effect cannot be justified in any normative sense.[5] In the Last Apple case, however, not only does the DT behavioral pattern fail to hold; in context, choosing to forego a tasty apple seems to be perfectly reasonable. In such cases an agent may remain convinced that she acted rationally even after someone shows her the expected utility calculations. A normative decision theorist cannot get away with simply pointing out that the choices in those cases run counter to her favored theory.[6] This result undermines normative views of DT.

A critic of Sen is free to argue that there is something wrong with his analysis of The Last Apple case. Arguably, treating *q* - having a nice apple - as the same option in both decisions mischaracterizes the agent's preferences: she cares about "behav[ing] decently" and "rule[s] of good behavior" and those interests are only implicated in one of the choice situations. Taking a particular nice apple when there is another one available is just a different action than taking that apple when it is the last one. I propose an account of this sort: the ***holism*** view. A correct representation of a decision maker's most basic preferences requires a correct representation of the outcomes in that ranking: it must be specific with regard to everything that they care about. DT does a good job as a normative account, I argue, if the basic outcomes over which a person has preferences are specified clearly enough to account for all of the values that would

[5] The Allais paradox is another widely-discussed case among philosophers. It stands somewhere between the Asian Disease case and the Last Apple case in that there is disagreement about whether the typical responses are reasonable. In the Allais case you are invited to choose among two pairs of lotteries. The first is a choice between a lottery that gives you a 33% chance at $5 million, a 66% chance at $1 million, but a 1% chance of getting $0 (lottery A) and a guaranteed $1 million (lottery B) and a lottery that gives you some chance at $5 million but a small chance of getting $0 (lottery A). In the second pair of lotteries, you have a small probability of getting $0 in both lotteries - 34% in lottery C and 33% in lottery D - in both lotteries, but lottery C gives you a 66% chance at $5 million while lottery D gives you a 67% chance at $1 million. When tested, most people choose B over A and C over D. But when the expected utilities of these lotteries are calculated, DT tells you that a person should choose A over B if and only if you choose C over D: u(A) > u(B) iff 0.33u($5 million) + 0.66u($1 million) + 0.01u($0) > u($1 million), and so u($0) > 34u($1 million) - 33u($5 million); likewise, u(C) > u(D) iff 0.33u($5 million) + 0.67u($0) > 0.34u($5 million) + 0.66u($0), and so u($0) > 34u($1 million) - 33u($5 million). (See Resnik 1987)

[6] Although Leonard Savage responded to Allais's challenge by conceding that, on reflection, he saw his original choices (A and C) were irrational, most people who choose as Savage did continue to think that their choices are perfectly rational. That is why the Allais paradox is often considered a counterexample to normative DT: it seems to allow for choices that are inconsistent with DT but are nonetheless instrumentally rational. (Floris 2015)

make a difference to their evaluation. The fact that the normatively compelling apparent violations of DT are not among the clearest descriptive counterexamples gives the defender of normative DT an opportunity to argue that they not violations at all.

To make sense of what follows, we will need a brief overview of the formal apparatus of RCT. I'll go over all of this in greater detail later in this chapter, but this will suffice to get us started. According to DT, a decision maker evaluates a set of *outcomes*: states of affairs represented, following Jeffrey (1965), as propositions like 'I spend a vacation in Columbia,' 'I get the big job promotion,' 'my friend is sentenced to 10 years in jail,' etc. Let Z be the set of basic outcomes. The decision maker has *preferences* over the set of outcomes. Her preferences rank each outcome against all of the others, giving her a 'list' of outcomes, from best to worst, allowing for ties. The preference ranking is a binary relation on Z, denoted by $\mathbf{R}^*$ ($x\mathbf{R}^*y$ is interpreted as x is at least as good as y). A numerical representation of preferences $\mathbf{R}^*$ - a *utility function*, denoted by U - assigns larger numbers to better outcomes in order to represent a preference ranking. U is unique up to positive monotonic transformation.

The decision maker's beliefs about how likely a choice option would be to yield an outcome is modeled by a probability distribution. Let L denote the set of *prospects*, where a prospect offers the elements of Z with certain probabilities. These *gambles* or *lottery tickets* represent uncertain states of affairs. The agent's choice options A can be understood as a subset of L.

The choice-worthiness of an option is usually indexed by the *expected utility* of the option. The existence and uniqueness of the expected utility function is guaranteed by a set of axioms regarding the agent's preference ranking. The axioms allow us to go from a preference ranking

over outcomes **R*** to a ranking **R** over lotteries L, and so over actions A. And they guarantee that the ranking over lotteries can be represented by an *expected utility* function u, unique up to positive linear transformation. The expected utility of a choice is the sum of the utilities of the prizes, weighted by the probability of getting those prizes.[7] An action $a \in A$ is a better choice option than $b \in A$ just in case $u(a) > u(b)$ (the expected utility of a is greater than that of b). (Resnick 1987; Kreps 1988, Ch. 4; Kreps 1990, Ch. 3) Regardless of any technical variations, the basic idea of DT is that of folk psychology: a decision maker should (and so often will) choose the option that best realizes her ends given her beliefs about her options and states of the world.

Sen's critique of DT is formulated in terms of *choice functions*.[8] As above, $A \subseteq L$ is a set of options for a decision maker. Let Pow(A) denote the set of all non-empty subsets of A. Formally a *choice function* for A is a function c: Pow(A) → Pow(A) such that for all $X \subseteq A$, c(X) $\subseteq X$. (Kreps 1988, p. 12) The intuitive idea of a choice function is that it is the set of options that are preferred in a circumstance, and so chosen, by an agent. Sen thinks that any choice function should satisfy the consistency axiom he calls Condition α**:** If $x \in B \subseteq A$ and $x \in$ c(A) then $x \in$ c(B). This axiom says that if a particular option is chosen from a set of options then it will also be chosen from a subset of the larger set if it is a member of that subset. (Sen 2002, p. 128) Expected utility functions obey Condition α: if $x \in B \subseteq A$ is such that $u(x) \geq u(y)$ for all $y \in A$ then $u(x) \geq u(z)$ for all $z \in B$.

---

[7] For example, suppose that a is an action that results prospect p with Pr($p|a$) and in prospect q with Pr($q|a$) = [1 - Pr($p|a$)]. The expected utility of the choice a (viz., a lottery over p and q) can be calculated as $u(p)$Pr($p|a$) + $u(q)$ [1 - Pr($p|a$)].. Generally, the expected utility of prospect p is a function u: $L \rightarrow \mathbb{R}$ such that

$u(p) = \sum_{z \in Z} u(z)$ Pr($z|p$), and

p is preferred to q iff $\sum_{z \in Z} u(z)$ Pr($z|p$) > $\sum_{z \in Z} u(z)$ Pr($z|q$).

$\mathbb{R}$ is the set of real numbers.

[8] It was originally framed as a critique of the internal consistency conditions of its revealed-preference interpretations, and since the revealed preference theory is a behaviorist project, the consistency conditions are stated in terms of behavior.

Sen isolates the 'recipe' for generating counterexamples like the Last Apple case. There are two option sets - $A = \{p, q, r\}$, and $B = \{p, q\}$ - and following choice pattern:

$\{p\} = c(\{p, q\})$;

$\{q\} = c(\{p, q, r\})$.

When q is the choice from the larger set A, this choice does not remain the same as the option set contracts from A to B: when the menu only offers p and q, the agent would choose p. The trick for evaluating *normative* DT is to find cases where choosing p from B and q from A makes sense. Sen's own example is supposed to be such a case. The fact that the choices in Sen's scenario is both rational and seem to be inconsistent with DT suggests that decision theorists need to modify even their normative theory.

There are no Sen-style counterexamples to DT for a certain range of cases, namely where a person's basic preferences **R*** rank a set of outcomes Z, each element of which is *holistic*. The term "holism," which is used as a technical term, is adopted from Frederick Schick. An outcome is *holistic* for an agent just in case they are indifferent among all of the ways of bringing it about. A proposition is *not* holistic for an agent just in case there are multiple, finer ways for it to be true *and* the agent prefers one of those ways over others. To use Schick's example, several candidates running for some office are bald. You want a certain one of them to win. Here *a bald man wins* is not holistic for you. (Schick 1984, p.18) The idea is that an apple is not just an apple. If taking the last apple means violating a social rule to which the decision maker is committed then the outcome should be redescribed or individuated in a more fine-grained way. In other words, if the difference between 'I have an apple' and 'I have the last apple' makes a difference to the decision maker's evaluation then the outcomes should be distinguished properly. Having the outcomes at the foundation of the DT formalism have constant utilities for an agent avoids

Sen-style cases - if an agent would evaluate bringing about the same proposition in distinct circumstances differently then she should start with a more fine-grained understanding of outcomes in the first place. The degree of specification required is fixed by the agent's own values.

Here, then, is how the holism strategy handles Sen's Last Apple case. Initially we seem to be looking at two outcomes,

p: I have nothing.

q: I have a nice apple.

When applying the holism strategy, the outcome "having a nice apple" should be distinguished as (at least) two holistic outcomes. One is having a nice apple when taking it is not behaving decently, and the other is having a nice apple when doing so does not violate any social norm. The decision maker clearly prefers the latter to the former. More explicitly, let *q'* and *q"* designate these two holistic outcomes,

q': I have a nice apple when doing so would be indecent.

q": I have a nice apple when doing so does not violate any social norm.

One might quite reasonably prefer having an apple when it doesn't violate a social norm to not having an apple, but I likewise prefer having no apple to having an apple when it would be rude. It should turn out, then, that $u(q') < u(p) < u(q'')$. There is no difficulty with the following choice pattern:

$\{p\} = c(\{p, q'\})$;

$\{q''\} = c(\{p, q'', r\})$.

Standard DT accompanied by the holism strategy yields the correct result that the decision maker would not choose to take the last apple.

A critical reader might worry that this holism solution is trivial. My solution to the Last-Apple type of cases begins with the obvious thing to say: if what seems like one outcome means something different for an agent in different contexts then it should be treated like more than one outcome.[9] In my view, the obviousness of this approach should tell us something about the normative status of RCT. But the apparent triviality of this 'redescription' strategy leads many to reject it as *too* simple.

One main reason for rejecting fine-grained outcomes, mentioned by both economists and philosophers, is that it would make RCT *tautological*. There could never be any violation of normative DT if a decision theorist could always redescribe and refine outcomes to make behavior consistent with DT. But such a consequence would be undesirable since it makes DT normatively "empty." (Guala 2006, pp. 248-51) Guala provides a further elaboration of this concern. He acknowledges that in order to save DT from Sen-style anomalies, a decision theorist has to refine an outcome even up to the point where the description includes the entire causal history of an outcome. He argues, however, that descriptions tied to very specific contexts cannot be accommodated by technical requirements of DT ("the Savage measurement procedure," as Guala calls it). Therefore, although the redescription strategy seems obvious, it makes each agent's outcome set so fine-grained that it is unable to give any content to their preference ranking.

[9] See for example, Guala (2006), Joyce (1999), Chapter 2.2, Rubinstein (2012), pp. 27-8, and Luce and Raiffa (1957).

My view avoids worries about refinement of outcomes by interpreting DT as a normative account, and so basically abandoning positive ambitions. As a logic, DT is independent of the contents of its inputs: it only insists that those inputs have certain formal features. The holism point sets a requirement on the set of basic outcomes *Z* that ultimately ensures that preferences over gambles over Z are free from any contamination of Sen-style anomalies.

Interpreted as a logic of choice, DT has both a principled end point for any refinement process and a sufficient structure to accommodate any redescription of outcomes. On the one hand, it is *not* the case that every possible proposition would have a unique utility value, making the refinement of outcomes an endless procedure. The holism requirement sets out DT with a preference ranking over basic outcomes *for a particular decision maker*. Given how people actually are, agents will have particular values, objectives and goals that are definite enough to avoid endless refinements. The particularity of the values of a decision maker yields a definite way of carving up the space of possible states of the world, which gives decision-theoretic analysis a definite starting point. *Ad hoc* redescription is ruled out by principle. The question is not how a modeler can redescribe actions, but rather what particular values and preferences a decision maker has.

On the other hand, DT places no restriction on how the basic outcomes are described - even when a description refers to the specifics of a choice situation or when the description includes an entire causal history of that outcome, nothing prevents such outcomes to enter a decision-theoretic analysis. It is virtually impossible to determine what is holistic for a decision maker, so descriptive and predictive uses of DT become, at least at times, unrealistically difficult. Of course, there is no general way for a modeler to figure out which propositions a given agent sees

as holistic from a third-person point of view, at least given the data to which such a modeller would have access. But the issue isn't epistemological: where an agent has a set of beliefs and preferences, RCT constrains their further evaluations and actions. What instrumental rationality requires of someone given her set of mental states can be clear even if it may be practically impossible for a third party to read a person's mental states off of her behavior.

I argue, then, that DT is defensible as a normative account of decision making. As a logic of choice, DT is a mechanism of evaluating the desirability of options, taking as inputs one's preferences over basic outcomes and beliefs about how actions are related to outcomes and providing as an output preferences over actions. This is analogous to Bayesian inductive logic, which gives you normative guidance on what new beliefs to form, given your prior beliefs and evidence. Putting it this way helps clarify the need to be precise about inputs. The normative elegance of Bayesian updating is that it allows one to go from one belief state to another based on some starting priors. The normative value of DT is that it shows you how to go from one preference state to another as choice options and beliefs change. What underwrites the whole structure, however, are preferences over basic - which is to say holistic - outcomes. The only Sen-style-counterexample-free way in which DT can justify choices is to begin with a set of holistic outcomes. Given a person's preference over a set of holistic outcomes, DT produces a rationally consistent ranking of options.

2. Instrumental Rationality

I believe that standard DT serves as a normative logic of choice: it picks out the instrumentally rational actions that go with particular desire-belief inputs. This only holds true, however, when these inputs are appropriately understood. Just as deductive logic stumbles over equivocal

propositions, RCT stumbles over certain assignments of value. In this paper, I defend the ***holism*** view about correct preference representation. The outcomes evaluated by an agent's most basic preferences must be characterized by everything that they would care about, i.e., make a difference to their evaluation. If they care about, say, the path by which they arrive at a certain end state, that consideration should distinguish as different outcomes the different paths. If everything that the decision maker cares about has been represented by their preferences over basic outcomes, then RCT picks out rational courses of action.

Philosophers have long been interested in the problem of *instrumental rationality*: deliberation about taking proper means in order to achieve some desired ends. Hume, for example, gives us a straightforward picture of human psychology to investigate this issue. There are two distinct categories of mental states, namely, desires and beliefs. Beliefs represent the way the world is, and desires represent the way a person wants the world to be. Beliefs can be either true or false, but desires are not truth-apt. Desires, Hume claims, are distinct because they are not subject to rational criticism: it does not make sense to judge the having of certain desires as rational or irrational.

> Where a passion is neither founded on false suppositions, nor chuses means insufficient for the end, the understanding can neither justify nor condemn it. `Tis not contrary to reason to prefer the destruction of the world to the scratching of my finger. (Hume 2000, Book II)

Hume's view is controversial, but whether desires themselves are subject to any sort of rational criticism, it is surely meaningful to ask Hume's question as to whether a person is choosing actions in such a way that they realize their desires, i.e., whether their actions are instrumentally rational. Hume seems to think all rationality is instrumental rationality. Even he is wrong about

that, however, it makes sense to consider what he is considering. Instrumental rationality evaluates how desirable an action is as a means to the realization of some end(s) in a given situation. Note that there are three elements involved in a judgment about instrumental rationality: end(s) given; the situation in which a choice action occurs and the agent's beliefs about it; and the connection between the action and the desire-belief complex.

Certain desires are evaluable with respect to instrumental rationality as well as actions. I might, for example, see longevity as valuable for its own sake, and so desire maintaining a healthy diet as a means to attaining longevity. Further, my desire to maintain a healthy diet might lead me to see eating a lot of whole grains as valuable if I thought whole grains were part of a healthy diet. Clearly, my beliefs about the the connections among whole-grain consumption, healthy diets, and human lifespans are relevant here. If I thought a healthy diet *didn't* lead to greater longevity then I might not want to maintain a healthy diet. Some evaluations, then are more primary and basic than others: my interest in longevity *transmits* desirability to eating a healthy diet, and some of that desirability is transmitted to eating whole grains. Some evaluations are more secondary and derived: my interest in eating whole grains is (at least partially) *derived* from my interest in maintaining a healthy diet, and this latter interest is derived (at least partially) from my desire for longevity.

We notice three things about human desires from the foregoing. First, since derived desires are evaluated as a means to some more basic ends, they are conditioned on beliefs. Secondary desires are derived from basic desires insofar as they are believed to be an effective means to basic desires. I must believe that maintaining a healthy diet is effective in attaining longevity (to at least some degree) in order to derivatively desire a healthy diet. Second, desires are

complicated and derived desires can often be in conflict, given different beliefs in different situations. Someone who desires better family relationships might hope both to get promoted at their job to better support the family and to spend more time with their family. Given some obvious beliefs, there might be tension between their desire to work late into the night and to spend the evening with family. Third, since derived desires are belief-mediated, they can be ill-formed due to false beliefs. We can form beliefs in deficient ways that affect the effectiveness of attaining a desired end. Someone might choose to perform an action by looking into a crystal ball that she thinks would forecast the consequence of the action, without realizing that the forecast of a crystal ball provides little evidential support. By sheer luck, the result might turn out to be good, but her choice is not instrumentally commendable.

The central task of understanding instrumental rationality is understanding what it means to behave consistently with one's own evaluations, choosing to realize some valuable states of affairs rather than others. The normative ideals that fall out of DT - evaluation by expected utility and maximization of expected utility in decisions - are minimal, formal requirements on valuing and choosing consistently. They impose no constraints on a decision maker's preferences and beliefs. To the extent that it makes sense to limit desires and beliefs, that is a task for another (set of) norm(s). RCT is supposed to tell you how to get from *your* preferences over outcomes and *your* beliefs about the world to some action that will realize the best states of affairs given those beliefs.

This account of instrumental rationality is part of the Humean tradition but goes beyond it. Modelling instrumental rationality with RCT resembles the way standard formal logic models good deductive reasoning. To claim that *modus ponens* is a valid argument form is to claim that

substituting any concrete, univocal sentences for the placeholders in the argument form will result in a valid argument, such that the conclusion must be true if the premises are true. In much the same way, the normative ideal of instrumental rationality, as captured by DT, is formal and concerns the consistency of preferences over states of affairs and subjective probabilities on the one hand, and evaluations of prospects on the other.[10]

3. Decision Theory and Its Roots

Common sense holds that people act in order to achieve their own ends to the extent they think possible. From a first person perspective, I act to bring about good things. From a third person perspective, this amounts to the claim that people make choices based on what they see as valuable (i.e., their wants, desires, preferences, etc.) and how they understand the world (their beliefs, credences, etc.). I decide to get up and get a drink of water because I am thirsty and I see that the water is right there. Sam chooses to apply to law school based on her preferences over career paths and her beliefs about law school being a good means to more desirable careers.

Further, we *explain,* and sometimes even *predict*, choice behavior - that of others and ourselves - by appealing to an agent's desires and beliefs. We say, for example, that I drank water *because* I wanted to quench my thirst and I believed that there was a bottle of water over there on the counter. We anticipate that Sam will apply to law school because she wishes to become a lawyer and she believes that attending law school is a good means to that end. These sorts of platitudes

[10] In Bayesian epistemology, similarly, you begin with some prior beliefs and you need to appraise them against new evidence and information. To do that you use Bayesian updating framework, such as Bayes' Rule, and form new or posterior beliefs. In this process of evaluating new information from prior beliefs, Bayesian framework offers a set of rules that get you from your prior beliefs to posterior beliefs. The role those rules play is simply to give you normative guidance on what new beliefs to form, given your priors. The focus again is to ensure epistemically rational consistency.

about how people choose, and about how people explain and predict choices, are part of "folk psychology," the common-sense account of the mental process by which people make decisions.

Unlike, say, folk medicine, folk psychology is not just conventional wisdom developed as human behavior patterns are observed. Folk psychological explanations appeal to, and depend on, a conception of instrumentally rational action. Rational creatures interpret one another as acting intentionally; to interpret an action as intentional is to attribute a *reason* to it. Folk psychological explanations accomplish this by attributing a desire-belief pair as the reason for the action that also serves as the cause. The desire-belief-as-reason style of explanation conforms to commonly recognized principles of instrumental rationality. Donald Davidson argues, for example, that the principle of instrumental rationality is constitutive of having desires, beliefs and other intentional attitudes. (Davidson 2004, Ch. 12) As rational beings, we make sense of one another by seeing people - others and ourselves - through the lens of folk psychology. I see you as (instrumentally) rational, not because you have my values or beliefs, but because I would make the same decision and behave exactly the same if I had your desires and beliefs. Folk psychology serves as a normative gauge of instrumental rationality.

Over the last several decades, economists and other social scientists have developed RCT as an account that characterizes decision-making in a formal and axiomatic way. DT is not an alternative to folk psychology, however. Rather, it is an interpreted formalism that *starts* with folk psychological principles. David Lewis nicely characterizes the origin of RCT and the role it plays:

> Decision theory (at least if we omit the frills) is not esoteric science, however unfamiliar it may seem to an outsider. Rather it is a systematic exposition of the consequences of

> certain well-chosen platitudes about belief, desire, preference and choice. It is the very core of our common-sense theory of persons, dissected out and elegantly systematized (Lewis 1983, p. 114).

Choice behavior is explained (described, predicted, advised, etc.) via bridge principles that spell out the formalism in terms of desires/preferences, beliefs/credences, and choices. As such, DT carries over the normative element from its folk psychological roots.

Key formulations of Rational Choice Theory include von Neumann and Morgenstern (1944), Savage (1954), and Jeffrey (1965). My account draws on these classic expected utility formulations of DT, although I follow Resnik (1987) most closely. My exposition proceeds in two stages, first considering simple decisions under conditions of subjective certainty and then treating cases of uncertainty. What follows is a very standard account, but organized to highlight the way in which folk psychology provides the starting point for DT. I hope to bring out certain elements of the view that are important for handling problems that arise for normative DT.

Decision makers care about *outcomes* - how things turn out. Again, following Jeffrey, I represent outcomes as propositions.[11] The decision maker's evaluations of these propositions - what she *wants* - is captured by a *preference ranking* over the set of outcomes. These preferences rank each outcome against all the others, giving her a list of outcomes, from best to worst, allowing for ties. The preference list reflects the order in which she would choose to take them, all things considered.

[11] The treatment of these propositions figures importantly in my view, but I will leave it at this level of analysis for now.

Formally, let Z be the set of outcomes. *Preference* is then defined as a binary relation on Z, denoted by **R***, that satisfies the following consistency requirements,

Completeness: For each $x, y \in Z$, x **R*** y or y **R*** x.

Transitivity: For each $x, y, z \in Z$, if x **R*** y and y **R*** z then x **R*** z.

R* captures the idea of "at least as good" for the decision maker: x **R*** y when your attitude is that you think that x is at least as good as y or that x is not worse than y. This is sometimes characterized as *weak preference*. From **R*** we can define **P*** (*strict* or *strong preference*) such that x **P*** y iff not y **R*** x and **I*** (*indifference*) such that x **I*** y iff x **R*** y and y **R*** x. Completeness requires that all alternatives in Z are comparable; transitivity requires that preferences be *acyclic* (the agent would not rank z over x if x is at least as good as y to you and y as good as z for them). DT often employs a numerical representation of preferences - a *utility function* - that assigns utility values (larger numbers to better outcomes) to represent her preference ranking.

$U: Z \rightarrow \mathbb{R}$, such that x **R*** y *iff* $U(x) \geq U(y)$.

Any numbering scheme that gets the order right does as well as any other, so a numerical representation U is unique only up to positive monotonic transformation. (Resnick 1987; Kreps 1988 Ch. 4; Kreps 1990 Ch. 3)

The foregoing is how RCT starts to formalize the desire part of a folk psychological account. The belief part is modeled, at this initial stage, by a set of actions that the decision maker thinks they can take and a function that picks out an outcome that the agent believes will result from each action. In terms of behavior, at least, this is how the decision maker represents the situations they inhabit. Formally, let A be the set of actions the agent believes available and $B^*_t: A \rightarrow Z$ be a function that represents their beliefs about how actions are related to outcomes at time t.

Regarding choices, the basic idea of DT remains that of folk psychology: a decision maker should (and so often will) choose the action that she believes best satisfies her preferences given her beliefs about her actions and how they would lead to outcomes. Formally, an agent chooses an action $a_i \in A$ at time t such that $B^*_t(a_i)$ **R*** $B^*_t(a_j)$ for all $a_j \in A$. In terms of utility, the decision maker chooses an action $a_i \in A$ at time t such that $U(B^*_t(a_i)) \geq U(B^*_t(a_j))$ for all $a_j \in A$.

This first-stage story about choices above is too simple. To start, the account of belief isn't nearly flexible enough - agents are rarely certain (even subjectively) about the outcomes that go with their available actions. In general, a given action might result in a number of possible outcomes. In order to handle this sort of uncertainty, DT appeals to subjective probability. Each agent has a probability function that represents their beliefs about how likely a choice option would be to yield various outcomes.

Formally, let L be a set of prospects, where a prospect can be understood as a lottery ticket or gamble that offers elements of the outcome set Z as 'prizes' with certain probabilities. L contains all of the probability distributions over Z.[12] I use p, q, r etc to denote the elements of L. Each uncertain state of affairs faced by a decision maker can be understood as such a gamble. Each decision maker has beliefs in the form of a belief function B_t: $A \rightarrow L$ which associates each action an agent thinks is available with a prospect (gamble) over outcomes, at time t. This is a better account of belief,[13] but it severs the tight connection between actions and outcomes that

[12] Each probability distribution maps the outcome set Z to the interval [0, 1], and the values of all the distributions add up to 1. In other words, a probability function p in L is defined as
p: $Z \rightarrow [0,1]$ such that $\sum_{z_i \in Z} p(z_i) = 1$.

[13] Probabilism is the view that degrees of belief should obey the probability calculus. The Dutch Book Argument justifies this view. See for example, Christensen 1996, Hájek 2008.

makes the simple account of choice plausible. We know how to rank outcomes; if actions are tightly connected with the outcomes then they simply inherit those rankings. In cases of uncertainty, however, actions are associated with prospects. What we could use, then, is a method for forming a preference ranking over such prospects (gambles).

It is a well-known result in RCT that there are axioms which allow us to go from a preference ranking over outcomes (relation **R***) to a preference ranking over gambles (relation **R**). These axioms (and the lemmas derivable from them) can be picked and presented in different ways. They are formal propositions, but they are derived from, and can be interpreted as, hypotheses about human behavior that are reasonable enough to make the analysis of lotteries relevant to human behavior. We begin with the completeness and transitivity conditions: they apply to preference rankings over gambles as well as preferences over outcomes, as mentioned above. We assume that lotteries *p* and *q* can form compound lotteries such as $ep + (1 - e)\, q$ for $e \in [0, 1]$.[14] One axiom is often known as the *independence* axiom, the *substitution* axiom, the *sure thing principle* or *separability*. The idea is that when two lotteries are compared, $ep + (1 - e)\, r$ and $eq + (1 - e)\, r$, the evaluation should be determined by the difference between the *p* and *q* alone. The irrelevant alternative *r* with probability $1 - e$ that occurs in both lotteries should have no effect on the evaluation of the two. This axiom makes good sense in an analysis of human behavior. (Resnick 1987; Kreps 1988 Ch. 4; Kreps 1990 Ch. 3)

Another axiom is the *Archimedean* axiom, or *continuity* condition. It says that with three lotteries *p*, *q*, and *r* and you rank *p* greater than *r*, and *q* between *p* and *r*, no matter how great *p* is (say, heaven) or how bad *r* is (say, your death), there is always some probability *e* that makes up a

[14] $Z \subseteq L$ since each $z \in Z$ can be understood as a degenerate lottery that gives the 'prize' z with probability 1.

mixture of p and r such that you are indifferent between the mixture lottery $e_1p + (1 - e_1)\ r$ and q. Intuitively, this seems plausible enough: people are willing to drive themselves to someplace - taking a risk at losing their lives - in order to receive a large enough gift. Likewise, someone might submit themselves to *some* risk of temptation - imperiling, however slightly, a heavenly reward - to achieve some other good. From these two axioms, two other conditions can be derived, and in some formalism they are all treated as axioms. The "better prize" condition says that you will prefer the lottery ticket that gives the better prize, given two otherwise identical lotteries. The "better odds" condition says that, given two otherwise identical gambles, you will prefer the lottery that offers the better chances at the higher prize. (Resnick 1987)

Once we have a preference ranking over gambles, choice looks familiar: an agent chooses an action $a_i \in A$ at time t such that $B_t(a_i)$ **R** $B_t(a_j)$ for all $a_j \in A$. This account involving preference rankings of gambles is probably less familiar than the numerical representation of the account. The set of axioms guarantee a preference ranking over lotteries guarantee, by way of proof, that the preference ranking over lotteries can be represented by an *expected utility* function, unique up to positive linear transformation. It should be clear that preferences over lotteries, **R**, can be numerically represented in the way preferences over outcomes, **R***, can be. The axioms that allow us to go from **R*** to **R** provide enough structure to allow for a more powerful numerical representation. The expected utility of an action is the sum of the utilities of the prizes, weighted by the probability of getting each prize. For example, suppose that x is the utility of a prize that results from an action, y is the utility of the prize from an alternative action, r is the probability of realizing prize x, and 1-r is the probability of realizing y. The expected utility of the choice (viz., a gamble over x and y) can be calculated as $xr + y(1 - r)$. Generally, the expected utility of a choice p is a function $u : L \rightarrow \mathbb{R}$ such that

$u(p) = \sum_{z \in Z} u(z) \Pr(z|p)$, and

p is preferred to q iff $\sum_{z \in Z} u(z) \Pr(z|p) > \sum_{z \in Z} u(z) \Pr(z|q)$.

In words, p is a better choice option than q just in case the expected utility of p is greater than that of q. (Resnick 1987; Kreps 1988 Ch. 4; Kreps 1990 Ch. 3) With desires and beliefs represented by a utility function and a belief function respectively, the choice-worthiness of an action is then indexed by the expected utility of the action.

The foregoing captures - perhaps in a tedious fashion - the usual connections between folk psychology and DT. It is worth going back to the beginning to see how the full apparatus of DT fits with its folk psychological roots. It is important to focus, in particular, on the distinctions between *outcomes*, *prospects* (*gambles*), and *actions*, which are identified by propositions. Outcomes, prospects, and actions are all evaluable according to DT. A person's preferences over prospects depend on her rankings of outcomes, and her preferences over actions depend, ultimately, on her preferences over outcomes *and* her beliefs. It is important to see that the same (sort of) action can be associated with a different prospect if the agent's beliefs differ. (see Fig. 1.1)

Evaluand	***Symbolization***	***Evaluation***	***Properties***
outcomes	$Z = \{x, y, z ...\}$	**R***; indirectly evaluated by **R**	foundational evaluation; invariant with respect to belief
prospects (lottery tickets, gambles)	$L = (p, q, r, ...\}$	**R**	dependent evaluation; invariant with respect to belief
actions	$A = \{a, b, c, ...\}$	not evaluated directly; indirectly evaluated by **R**	dependant evaluation; varies with respect to belief

Figure 1.1

Consider an example, borrowed from Swoyer and Ellis, that illustrates some of these distinctions. (2005) Suppose that Tom is choosing whether to take a flu shot or not. There are

two possible states in the world – there being a flu outbreak or not. Following Swoyer and Ellis, suppose that the best *outcome* for Tom is that there is no flu outbreak and he doesn't get a flu shot (shots hurt!), which gives him a utility of 3. The next best thing, according to Tom, is that there is an outbreak but he has gotten a flu shot which gives Tom a utility of 1.[15] If Tom gets the shot and there is no outbreak, he experiences regret - painful shot for no reason! - so his utility is -1. The worst thing for Tom is when there is a flu outbreak but he didn't get a flu shot, since he will be sick for a week; this gives him a utility of -6. With the expected utilities of the relevant outcomes as follows

u(Flu shot|Outbreak) = 1

u(Flu shot|No outbreak) = -1

u(No flu shot|Outbreak) = -6

u(No flu shot|No outbreak) = 3

we can calculate that

u(Flu shot) > u(No flu shot) iff

u(Flu shot|Outbreak)β + u(Flu shot|No outbreak)(1-β) >

u(No flu shot|Outbreak)β + u(No flu shot|No outbreak)(1-β) where β = p(Outbreak) iff

(1)β + (-1)(1-β) > (-6)β + (3)(1-β) iff

β > 4/11 ≈ 0.36.

Tom's choice	**Outbreak** (w/ probability β)	**No outbreak** (w/ probability 1-β)
Shot	1	-1
No shot	-6	3

Figure 1.2

[15] His utility here is affected by the trouble he takes to get the flu shot but apparently not the suffering of others!

Suppose that Tom now believes that the probability of a flu outbreak is 40%. In that case, as Swoyer and Ellis note, *u*(Flu shot) = -0.2 and *u*(No flu shot) = -0.6. If Tom comes to change his mind and decides that there is now a 25% of a flu outbreak then *u*(Flu shot) = (1)(0.25) + (-1)(0.75) = -0.5 and *u*(No flu shot) = (-6)(.25) + (3)(0.75) = 0.75. The value for Tom of 'I get a flu shot' changes, then, with his beliefs about flu outbreaks, even if the values of the various flu-related outcomes (e.g., 'I get a flu shot and no flu outbreak happens') stays constant, (as does the value of a gamble like 'I get a flu shot when there is a 40% chance of a flu outbreak').[16]

The proposition Tom entertains when he considers his action - 'I get a flu shot' - can pick out distinct prospects/gambles, depending on his beliefs at the time. The same can also be said for certain outcomes, however. In another circumstance, Tom might come to believe that getting a flu shot involves a risk of receiving tainted vaccine, and so outcomes such as 'I get a flu shot and no flu outbreak happens' are also gambles with expected values that depend on probabilities. In this case, there are beliefs not represented in the decision problem, and so not involved in calculating the expected values of the action 'I get a flu shot', that can change the expected value of the action. This suggests, in turn, that RCT has clear results only as long as the values of the relevant outcomes - the basic building blocks - remain invariant.

4. Revealed Preferences

[16] The outcomes here are complete specifications of how things turn out: Tom gets a flu shot and there is a flu outbreak; Tom gets a flu shot and there is no flu outbreak; Tom doesn't get a flu shot and there is a flu outbreak; Tom doesn't get a flu shot and there is no flu outbreak. The prospects are the 'lottery tickets' one can construct by considering probability distributions over outcomes. Each action picks out a particular prospect: u(shot) = Pr(shot & outbreak | shot) + Pr(shot & no outbreak | shot) + Pr(no shot & outbreak | shot) + Pr(no shot & outbreak | shot); u(no shot) = Pr(shot & outbreak | no shot) + Pr(shot & no outbreak | no shot) + Pr(no shot & outbreak | no shot) + Pr(no shot & outbreak | no shot)

The current mainstream interpretation of the formalism of DT in economics - at least in principle - is known as *revealed preference theory*.[17] On this account, the notion of preference is identical to patterns in choices. A thing *x* is revealed-preferred to *y* just means that *x* is chosen when *y* is also affordable. Each time an agent faces a set of options that contains both *x* and *y* and the agent chooses *x*, economists conclude that the agent prefers *x* to *y*. To prefer *x* to *y* is simply to always choose *x* over *y* when both options are available; preferences are an agent's consistent choice behavior.

The motive behind the move to reduce preference to behavioral patterns stems from behaviorist, and ultimately positivist, scruples about mental states. Following Samuelson (1938), revealed preference theory has become a standard approach to consumer choice behavior in economics. As economist Ian Little wrote, "a theory of consumer's demand can be based solely on consistent behavior," and this "new formulation is scientifically more respectable [since] if an individual's behavior is consistent, then it must be possible to explain that behavior without reference to anything other than behavior." (Little 1949) Economist John Hicks stated that "the econometric theory of demand does study human beings, but only as entities having certain patterns of market behavior, it makes no claim, no pretense, to be able to see inside their heads." (Hicks 1956) On this view, there are no mental states being referred to in decision theoretic models. Preferences are patterns in past choices. Revealed preference theorists adopt an instrumentalist interpretation of a decision maker's beliefs and desires without committing to their underlying reality. Merely looking at behavior pattern and choice actions, they do not have to be committed to people actually having certain beliefs and desires that cause them to actually maximize their expected

[17] Actual economic work doesn't seem to rely very much on revealed preference interpretations, but it shows up in more theoretical work and in the economics classroom.

utility. RCT, on this view, is a predictive tool that provides consistent regularities between observed and unobserved choice patterns.

To explain or predict choice behavior, economists begin with a consumer's behavior regarding items in a relevant set and then construct a preference ranking of those items - tease out a pattern to project - by generalizing from the behavior. Suppose that when a consumer comes to a supermarket to buy milk, she exhibits some behavioral patterns, and so some preferences, over milk options. Her preferences are often associated with particular qualities - price, nutrition, brand, etc. - of milk. After observing a large enough sample of data, an economic modeler would extract a ranking of the consumer's most preferred to the least preferred type of milk, allowing for ties. Suppose that the modeler observes that this consumer always chooses fat free milk with Vitamin D compared to 2% fat milk, that she never chooses Walmart milk when there is another option, and that otherwise, when confronted with milk products that have the same qualities, she always chooses the one with the lowest cost. These observations give the modeler enough information to assign some structure to her preferences over milk products. She might, for instance, like fat free milk with vitamin D better than 2% fat milk, prefer Meijer to Walmart milk brands, and prefer lower prices to higher. Suppose that this time, the consumer is facing the choice between milk X and milk Y. X contains vitamin D and is fat free, is 'Meijer' brand, and costs $1.89. Milk Y contains 2% fat, is 'Walmart' brand, and costs $2.50. The economic modeler can predict that this consumer will choose X in comparison to Y, given her preference ranking. The modeler can also explain - in some sense - why she made that choice by appealing to the pattern.

The identification of choice of action and preferences over actions was seen as harmless, according to Paul Samuelson (among others.)[18] Samuelson shows that the two notions can be co-extensive if we assume axioms of preference such as the weak and strong axioms of revealed preference, and basic contraction consistency requirements such as Sen's Properties α and β.[19] Two sets are defined, c(S) and c^R(S), signifying the choice set and preference ranking, respectively. With the Weak Axiom of Revealed Preference (WARP), it is proved that c(S) = c^R(S). So there is a theorem – "the revelation theorem," as Daniel Hausman calls it (Hausman 2000) – showing that there is a one-to-one correspondence between the choice set and the preference ranking. This theorem, then, seems to entitle economists to dispense with any talk of the preference ranking, since it seems that anything that can be said in terms of preference ranking can also be said in terms of choice. Past choice patterns that are consistent are treated as projectable into future situations, as governed by axioms such as the basic contraction axiom (Sen's Property α) and the weak axiom of revealed preference. Those axioms are thus viewed as consistency conditions, governing a consistent pattern from past to future choices.

There is a lot to criticize in the revealed preference project. (See, for example, Sen 1973; Sen 2002; Hausman 1992; Hausman 2000; Guala 2006; Reiss 2013; Beshears et al. 2008) Preference over belief-mediated outcomes and preference revealed by choice are distinct. Philosophers, in the first place, have long abandoned the behaviorist and logical positivist projects: sometimes scientists need to postulate unobservable entities. Preferences are not merely behavioral patterns. Rather, we have reason to think that they are mental states or attitudes that give rise to behavior.

[18] I. Little 1949, John Hicks 1939 and H. Houthakker 1950.

[19] It is provable that choice set satisfies the weak axiom of revealed preference iff it satisfies Properties α and β.

Further, the revealed preference view doesn't work in its own terms. The patterns it can detect are patterns in behavior, and we have very good reasons to think *those* patterns aren't projectable. Suppose that a revealed preference theorist, who only has access to your choice behavior, observes that you have chosen to take flu shot every year over the last decade. But this year, you didn't. All the revealed preference theorist can say about this "inconsistent" choice is that it is an outlier in the data and it doesn't mean anything significant. This misses a lot of what is going on. Suppose your preference over outcomes are those for Tom (above, Figure 1.2) and that you usually don't have any idea about how likely it is that there will be a flu outbreak. You 'guesstimate' that there is a 50/50 chance of an outbreak and that's why you have been choosing to take flu shots: with that belief, getting flu shots gives you higher expected utility than not getting flu shots. This year, however, you gained some insight from a medical expert and came to believe that Pr(Outbreak) < 0.36. So you choose not to take flu shot. A revealed preference theorist cannot recognize the result of your belief change and must see this year's choice as inconsistent, and therefore unintelligible. But in fact your choice this year is perfectly intelligible given what we know about your preferences over outcomes (and utilities) and your beliefs. Revealed preference theory cannot account for any atypical choice behavior that is due to belief change and information updating. The problem with revealed preference theory is that revealed preferences concern only actions and not outcomes. But preferences over actions are belief-mediated. Whenever a belief changes, a revealed-preference patterns about to change and can no longer be consistently projected into the future. RCT aims to show how preferences over action are formed from preferences over outcomes and beliefs about relevant information, and therefore simply cannot take the pattern exhibited by past actions as already given.

Finally, behavior may not reflect attitudes in obvious ways. A person's choosing to take an orange rather than an apple in a particular situation, for instance, may not reveal a general preference for oranges over apples. For example, when making a decision under risk and uncertainty, one might choose the normally less preferred option (as gathered from past choice patterns), just because one believes that the normally more prefered option is basically unobtainable in a particular situation.[20] RCT simply cannot get off the ground if a preference ranking is just a pattern in past choice behavior and not a matter of real attitudes. DT should begin with a proper representation of preference as a relation over basic outcomes. In particular, for the purpose of giving sound decision-theoretic advice, a normative decision theorist has to assume the reality of the underlying beliefs and desires.

5. Decision Theory as A Logic of Choice

Instrumental rationality is part of our folk psychological account of decision making in that it (partially) characterizes how desires and beliefs interact. Agents sometimes see outcomes as desirable in themselves. More often, however, they see states of affairs or actions as desirable as means to or parts of other desirable outcomes, prospects, or actions. I want to make it to downtown by 5:30 PM on Tuesday because I want to be on time for the show there at 6:00 PM that I desire to attend. If I ceased to want to go to the concert, I would cease wanting to make it to downtown by 5:30 PM. Likewise, If I stopped believing that the show was on Tuesday, I would quit wanting to be in downtown at 5:30 PM on Tuesday. The desirability of being in downtown by 5:30 PM on Tuesday is *derived* from the desirability of going to the 6:00 PM

[20] As with game theory in a Rock, Paper, Scissor game, for instance, one's past strategy patterns may suggest a mixed strategy of randomly choosing one out of the three gestures, but this is only so because information of one's belief has already been assumed, namely the belief that one's opponents also play randomly. This same player, however, might choose to stick to paper this time just because she acquires new information that this opponent she is playing against always plays rock. Strategies or actions suggested from past choice patterns are doomed to miss such rational choices that result from updated information and belief.

concert on Tuesday and my belief that being in downtown by 5:30 facilitates going to the 6:00 PM concert. In other words, my desire to go to the show at 6:00 PM *transmits* its desirability to being in downtown 30 minutes early by way of my beliefs. I may also cease to desire an object because I believe it hinders some other end that I value. I would cease to want to be in downtown at 5:30 PM on Tuesday if I thought that being there would interfere with my desires to sit with my friend when they are having a health crisis.

RCT captures all of these insights formally. Let p be 'I go to the 6:00 PM concert on Tuesday' and suppose that $u(p) > u(\sim p)$. Suppose also that I think I can bring about q, 'I make it downtown by 5:30 PM on Tuesday'. Initially, I believe that if I bring about q then p is 95% likely: $Pr(p|q) = 0.95 > Pr(p)$. Accordingly, q is associated with the gamble $[0.95p; 0.05\sim p]$ and $u(B_1(q)) = 0.95u(p) + 0.05u(\sim p) < u(p)$. If I change beliefs in a way that makes q independent of p (i.e., $Pr(p|q) = Pr(p)$) then $u(B_2(q)) = Pr(p)u(p) + (1 - Pr(p))u(\sim p) < u(B_1(q))$. If the value of $\sim p$ increases such that $u(p) < u(\sim p)$ then the value of bringing about q also changes.

DT formalizes the idea of instrumental rationality by making information about beliefs, desires, and the connections among them precise. It does not follow that DT is an empirical psychological theory. Instead, it (partly) characterizes the concepts of 'belief' and 'desire'. RCT is a formal system that articulates and makes precise folk psychological platitudes; it spells out a certain conception of how desirability, in particular, is distributed over propositions. That holds true even if actual reasoning doesn't work that way. Philosophers such as David Hume divide human inquiry into "relations of ideas" and "matters of fact." (Hume 2000, Book I) This distinction is fairly clear in mathematics and deductive logic. Mathematical principles are about

how we should reason *a priori*. A theory of mathematics is quite different from a theory of actual mathematical reasoning that might be the subject of an anthropological study. Scientific theories often play multiple roles in such situations. When our reasoning in fact conforms to algebra or *modus tollens*, for example, theorists may see themselves as both articulating the norms of how we should reason and at the same time describing some of the ways in which we actually reason. The fact that RCT is studied in multiple disciplines illustrates this joint role. DT *can* be taken as a (limited) account of the ways actual decisions are made. But this does not automatically define and exhaust the role that DT plays. RCT, I argue, is more successful playing the role of spelling out "relations of ideas" with respect to evaluation and choice. Analogous examples abound in science. Theorists think they are describing the way the world is, but scientific theories are often cognitive devices used to understand the world. When a theorist claims, for example, that water freezes at 0 degrees celsius, that does not mean that the celsius temperature scale is something we discover.[21] Still, it tracks some real patterns in the world. Such patterns are the focus of scientific theories. But theories themselves are formal devices that theorists generate to track certain patterns.

RCT can be viewed as a norm for (re)distributing judgments about desirability over propositions, taking preference rankings of outcomes together with credences over propositions as inputs, and outputting preference rankings over prospects and actions. Consider, for example, Sam, who likes candy bars. Her preferences over the various candy bars sold in town can be represented by a ranking or a numerical representation thereof (utilities). Sam also has beliefs about the different stores carrying the candy bars she likes, which is represented by a probability distribution

[21] Likewise, arguably, with using wave functions in describing quantum mechanics.

governed by the probability calculus. The expected utility theorem then generates a ranking of the desirability of her choice options to go to certain stores for candy. Suppose that Sam knows that store *a* carries candy bar *x*, store *b* carries candy bar *y*, and store *c* carries candy bar *z*. If she likes candy bar *x* better than *y* and *y* better than *z*, then she should prefer going to store *a* over going to store *b* and going to *b* over going to *c*. Sam's preferential judgements about various candy bars are indexed by utilities. With both utilities and credences, decision theory produces an index of the expected utilities of going to various stores for different candy bars. In this process, an initial distribution of utilities for each outcome transmits desirability to each action by way of means-ends beliefs. Sam's beliefs about which store carries which kind of candy bar connect the action of going to a certain store to the desired outcome of getting a certain kind of candy bar. With actions and outcomes connected in a means-ends way, decision theory transmits an index of the desirability of outcomes (i.e., utilities) to an index of the desirability of actions (i.e., expected utilities). Going to a store is only desirable to Sam because it is an effective means to obtaining an desirable end. It would no longer be attractive to Sam if going there no longer facilitates getting that particular kind of candy bar. Taking the action is of instrumental value to the agent and decision theory transmits instrumental desirability.

Since DT formalizes the transmission of instrumental evaluation, it sets formal constraints on the coherence of choice. There is a reason for an agent to choose a certain action because that action would bring about some valued outcome. Choosing that action would be a rationally coherent choice only if that outcome is desirable. Decision theory redistributes the initial utility judgments according to how effective the choice actions are to bring about those utilities. Going to store *a* is the most effective means to getting candy bar *x* which is at the top of the initial ranking of utilities. Going to store *a* is, therefore, assigned the highest ranking in the index of expected

utilities (at least as long as no other valued outcomes are implicated). Decision theory preserves the desirability of an outcome in the transmitted evaluation of an action. It maintains consistency of instrumental rationality in the judgments of desirability.

While DT articulates formal constraints of consistency, it does not constrain what is desirable for an agent, the contents of their desires. RCT abstracts from what an agent happens to value and provides the form of rational consistency. Given an initial utility ranking and a set of credences, decision theory determines a certain ranking of desirable actions. The claim of the theory is conditional - if *these* are your preferences and beliefs then *this* is what you ought to choose in order to be instrumentally rational/coherent. The particular interests, values and beliefs that an agent come with might be seen as external facts of the world. And the consistency claims of decision theory are constraints internal to the theory. Given evaluations of desirability and credence, decision theory redistributes and dictates another set of evaluations of desirability and credence as situations change.

Thus, I take decision theory to be a ***logic of evaluation/choice*** because it shares some crucial characteristics with two paradigms - deductive and inductive logics. Logic aims to characterize appropriate reasoning and inference. Deductive logic characterizes validity; inductive logic characterizes strength of evidential support. A deductive system picks out a set of axioms and rules of inference (or just rules of inference) in order to deduce all and only true sentences from a given set of true sentences.[22] For example, suppose that $(P \vee (Q \rightarrow P)) \vee P$ and Q are two true sentences. Since there is a deduction from these two sentences to the sentence P, P is guaranteed to be true. Note that with an initial truth assignment to the two sentences, deductive logic

[22] Only deductive systems that are complete will deduce all true sentences. Incomplete systems cannot exhaust all true sentences following from a given set of true sentences.

transmits that truth assignment to another one which assigns *P* to be true. Deductive logic necessitates this redistribution of truth values by way of deduction. The claim is a conditional one, that *if* the initial truth assignment is as such, *then* the redistributed truth assignment has to be so-and-so, in order to be logically consistent. Thus, deductive logic is said to be *truth-preserving*. The logic itself is a mechanism that inputs and outputs assignments of truth values in a way that preserves consistency so that true sentences can and can only be deduced from true sentences.

Moreover, the consistency constraint is purely formal and abstracted away from content. Substitute any particular sentences for the placeholders in the inference, the transmission of truth values will hold invariably. What matters for the logic to work is the truth value of a sentence. The logic guarantees to output a particular set of truth values given a particular input. The logic imposes no constraint on what those sentences are. As long as a truth value assignment is fixed as input, a fixed truth assignment is determined as output. If an input sentence changes its truth value, the logic will produce a different result. The characteristic of being a formal mechanism and abstracting away from content frees the logic from problems of equivocation. The fact that different outputs are produced due to varied inputs does not mean that the logic is compromised. Instead, the logic works the same way. Suppose that in the inference from *P* to *P v Q*, *P* is instantiated as "this piece of chocolate tastes sweet," and *Q* is instantiated as "5+7=13." *Q* has truth value 'False,' and *P* is assigned value 'True.' Deductive logic produces value 'True' for *P v Q*. But the instantiated sentence *P* can be ambiguous since the chocolate may or may not taste sweet to different people. When someone assigned 'False' to the sentence "this piece of chocolate tastes sweet," deductive logic would yield value 'False' to *P v Q*, instead of 'True.' The changing values of the input would vary the value of the output, but the logic itself does not suffer from this problem of equivocation. As long as the same truth values are instantiated for

every instance of *P* and *Q*, the transmission of truth values will hold and the logic guarantees that *if* the instantiated premises are true *then* the instantiated conclusion must be true.

The case of inductive logic proceeds in much the same way. We begin with a set of credences, or degrees of evidential support of some hypothesis. Through a process governed by Bayesian updating rules, we move to another set of credences or degrees of support. The probability calculus requires certain axioms, for example, that my credence of *P&Q* be smaller than my credence of *Q*. Bayesian confirmation theory models how empirical evidence confirms or disconfirms hypotheses. The degree to which a hypothesis is confirmed by some evidence is determined by the degree to which the hypothesis is supported prior to the evidence (known as the "prior probabilities of hypothesis") and the degree of how likely the evidence is to occur given the hypothesis (known as the "likelihood"). The Bayesian Convergence Theorem shows that the posterior probability of a true hypothesis will be driven to the top of the list of alternative hypotheses, as new evidence and rival hypotheses continue to be tested. For our purposes, it is important to note that this process of redistributing degrees of support is constrained by the Bayesian inductive logic such that the posterior probability of a hypothesis that has stronger evidential support will increase and the posterior probability of a hypothesis that have weaker support will decrease. Thus, inductive logic is said to be *truth-indicating*. (Hawthorne 2011) The Bayesian mechanism counts as a *logic* of hypothesis evaluation because it maintains consistency across degrees of support. Inputting different sets of probabilities will output different sets of probabilities. If we begin with a fixed set of likelihoods and prior probabilities then the posterior probabilities have to turn out as the Bayes's Theorem claims. Inductive logic also has the formal features of a logic. It is irrelevant to the logic what particular hypothesis gets tested. As long as

coherent probabilities are given as inputs, the same sort of probabilities will be given as outputs. There should be no problem of equivocation as long as the probabilities are kept straight.

DT provides conditions of rational coherence that govern the (re)distribution of evaluations of desirability. Evaluative consistency here is based on intuitions about what constitute effective means to desirable ends. DT formalizes and makes precise the intuition about desirability transmission canvassed above. DT can be said to have the virtue of being *effectiveness-indicating* in the way that deductive logic has the virtue of being *truth-preserving* and inductive logic has the virtue of being *truth-indicating*. Let us streamline this comparison between deductive and inductive logics and the logic of evaluation/choice in the following chart.

	Inputs/Outputs	Process	Virtue
Deductive logic	Truth value assignment over s for sets of propositions	Redistributing truth values by way of deduction	Consistency in validity; truth-preserving
Inductive logic	Set of credences (degrees of support)	Redistributing credences (degrees of support) by way of Bayesian updating/confirmation	Consistency in strength of inductive support; truth-indicating
Decision Theory as logic of choice	Indices of desirability	Redistributing judgments of desirability	Consistency in judgments of desirability; effectiveness-indicating

Figure 1.3

6. Sen's critique

Although Amartya Sen's critique of DT has broader application, it was originally aimed at the internal consistency conditions of revealed-preference interpretations. In this section, I will first

describe Sen's critique. Then I will explain the more significant application it has to DT interpreted as a realist and normative project. Even if we get past the revealed-preference instrumentalist interpretation, we will still face the challenge from the Sen-style anomalies.

Sen argues that the axioms of DT are not plausibly understood as conditions that capture the internal consistency of past-to-future, observed-to-unobserved choice patterns. Since revealed preference theory is a basically behaviorist project, its consistency axioms are stated in terms of *choice functions*. As before, $A \subseteq L$ is a set of options for a decision maker. Let Pow(A) denote the set of all non-empty subsets of A. Formally a *choice function* for A is a function c: Pow(A) $\rightarrow$ Pow(A) such that for all $X \subseteq A$, c(X) $\subseteq X$.

The formal definition of a choice function signifies the intuitive idea of a set of options that are preferred and chosen. When A is a set of options from which you make choices, the items that you better prefer and choose should constitute a subset of A. The requirement in the formal definition is thus very intuitive and basic.

Sen's Condition α. If $x \in B \subseteq A$ *and* $x \in c(A)$, then $x \in c(B)$.

This axiom says that the choice from a set of options must remain the same when choosing from a subset of that larger set. If x is chosen from a larger set A then x must be chosen from a subset of A that contains x. Sen states the basic idea of this condition with an analogy, "If the world champion in some game is a Pakistani, then he must also be the champion of Pakistan."[23]

[23] Kreps 1988, p.13.

Condition α - also known as the "basic contraction consistency"[24] signifies a sense of consistency where choice *x* should not change as the set of options contracts.

Sen's critique proceeds by way of examples. We've considered *The Last Apple* case. Sen also considers a companion case:

Cocaine

> To illustrate, given the choice between having tea at a distant acquaintance's home (*p*), and not going there (*q*), a person who chooses to have tea (*p*), may nevertheless choose to go away (*q*), if offered by that acquaintance a choice over having tea (*p*), going away (*q*), and having some cocaine (*r*). (Sen 2002, pp. 130-1)[25]

In both cases, there is a pattern of one sort of option being preferred to another (an apple preferred to nothing; not having tea with an acquaintance preferred to having tea) but the pattern turns out to not hold in the circumstances discussed. Not only does the behavioral pattern fail to hold; in context, there can be perfectly good reasons why someone might deviate from her past choice patterns.

Since revealed preference theory only looks at past choice patterns, the particular choices that violate Condition α must be understood as 'noise' in the data. Adopting a realist view about preferences (i.e., holding that they are actually mental states) offers a chance to save DT from the Sen-style counterexamples. The agent's choice in the Last Apple case is explained by her interest in conforming to a social rule outweighing (or perhaps even *overwhelming*[26]) her interest in consuming apples. In most apple-eating situations, such social rules are not at stake. When they

---

[24] Or else, "the Chernoff condition" and "the independence of irrelevant alternatives."

[25] To keep my notation consistent, I have changed Sen's letters here.

[26] See F. Schick or S. Ellis on understandings/tunnel-vision. Frederic Shick (1991). *Understanding Action: An Essay on Reasons*. New York: Cambridge University Press.

are, however, they might be decisive, at least for some individuals. In the Cocaine case, having tea at an acquaintance's home is a different prospect once doing cocaine is an explicit possibility - one's beliefs about the acquaintance are likely to shift.

The challenge - for my interests, at least - is that while it is aimed at revealed preference views, Sen's critique still creates issues for realist, and ultimately even normative, DT. In the previous milk-purchase scenario, there is nothing that a revealed preference theorist could say if in fact, the consumer bought milk Y instead of milk X that seemed to have a higher expected utility except that this datum must be understood as noise. Even for realist, descriptive interpretations of RCT, the choice would be irrational and so unexplainable. It posits a pattern among actual mental states that doesn't seem to hold. These sorts of failures are not merely 'noise' in a pattern, but evidence that perhaps some other pattern fits the data better. Decision theorists have uncovered large amounts of experimental data over the last few decades that are inconsistent with the explanations and predictions of standard DT. The most widespread counterexamples - among philosophers, anyway - include the Allais paradox[27] and the Sen-style cases.

Of course, one reply to such cases might be that they mischaracterize either consumers' preferences or beliefs. It would be easy, for instance, for a decision theorist to 'lose track of' an overriding preference: a consumer might love any product that features her favorite celebrity, and her choice of milk Y would be based on the fact that Y was endorsed by that celebrity. Or a

[27] In the Allais case, is where you are invited to choose among two pairs of lotteries. The first is a choice between a lottery that gives you a 33% chance at $5 million, a 66% chance at $1 million, but a 1% chance of getting $0 (lottery A) and a guaranteed $1 million (lottery B) and a lottery that gives you some chance at $5 million but a small chance of getting $0 (lottery A). In the second pair of lotteries, you have a low-ish stand a small probability of getting $0 in both lotteries- 34% in lottery C and 33% in lottery D - in both lotteries, but lottery C gives you a 66% chance at $5 million while lottery D gives you a 67% chance at $1 million. When tested, most people choose B over A and C over D. But when the expected utilities of these lotteries are calculated, decision theory tells you that you should choose A over B if and only if you choose C over D: u(A) > u(B) iff 0.33u($5 million) + 0.66u($1 million) + 0.01u($0) > u($1 million), and so u($0) > 34u($1 million) - 33u($5 million); likewise, u(C) > u(D) iff 0.33u($5 million) + 0.67u($0) > 0.34u($5 million) + 0.66u($0), and so u($0) > 34u($1 million) - 33u($5 million).

decision theorist could fail to predict the choice of Y because she fails to see that the consumer thought the particular bottle of milk X was outdated. The idea that the company had outdated product on shelf might have greatly lowered the consumer's confidence in milk X. In other words, the choice of buying milk Y instead of X could be a result of a preference change, a belief change, or a preference change driven by a belief change rather than a counterexample to DT.

In addition to thought experiments like the Last Apple case and the Allais paradox, psychologists and behavioral economists have also studied the ways in which real people deviate from the standard expected utility theory through lab experiments and empirical observation. Consider, for example, the Asian Disease case from above, where two sets of subjects are presented with the exact same choice options described using different phrases, and so 'framed' in different ways. (Tversky and Kahneman 1981) Subjects' choices differ systematically with the different frames.[28] There are also a number of heuristics and biases that we - mostly unconsciously - use in making decisions. We attach higher weight to low probabilities than their actual significance while underweighting high probabilities. (Camerer 1999) We can be too focused on a single aspect of a situation and develop a tunnel vision that makes us cognitively blind to other aspects needed to be considered.[29] We are often subject to loss aversion and see loss as worse than we see gain as good.[30] (Hayden and Ellis 2007)

[28] See footnote 4.

[29] A non-trivial number of tourists at the Great Canyon are documented to unconsciously step back and fall to the canyon when they take photos. Credit to Stephen Ellis.

[30] Evidence shows that more stock trades happen with winning stocks than with losing ones, even though the winning stocks promise higher expected utilities. Lab experiments also suggest that once a subject sees a commodity as her own, she expects higher selling price since selling one of her own would realize a loss and she sees a loss as more valuable than its actual material value.

All of the foregoing cases demonstrate behavior that deviates from what RCT models predict/suggest. In some, choices are clearly irrational: explaining the way in which the behavior deviates from the model will often suffice to convince even the agent that she has chosen poorly. But there are other cases - the Allais case and Sen's cases, for example - where the choices are *not* clearly irrational: someone who refuses to take the last apple, for instance, may remain convinced that she acted rationally even after someone shows her the expected utility calculations. In such a case, a normative decision theorist cannot get away with simply pointing out that the choices in those cases run counter to her favored theory. Although Leonard Savage responded to Allais's challenge by conceding that, on reflection, he saw his original choices (A and C) were irrational. Most people who choose as Savage did, however, continue to think that their choices are perfectly rational. (Heukelom 2015) That is why the Allais paradox is usually considered a counterexample to *normative* DT: it seems to allow for choices that are inconsistent with DT but are nonetheless instrumentally rational. Similarly, Sen's Last Apple case poses a challenge for even a normative decision theorist: refusing to take the last apple seems inconsistent with expected utility maximization but it still seems like a perfectly reasonable thing to do for agents with certain goals.

Again, a theorist is free to argue that there is something wrong with any of the foregoing cases. If they accept - as I do - that these behavioral anomalies are counterexamples to what the models, however, there are two ways to react. One is to admit that while these behaviors show that descriptive uses of RCT are limited (since it can only handle rational behavior), they are still irrational, and so normatively deficient. The other approach tries to defend the non-model behavior as rational, implying that there is something wrong with DT as a normative guide. In other words, if the descriptive adequacy of RCT is called into question, there remains the issue of

its normative adequacy. There seem to be clear cases of irrational choice - framing effects and tunnel-vision examples, for instance. I see such cases as evidence against the descriptive adequacy of DT, although not its normative adequacy. There also seem to be cases - including the Allais case, the Last Apple case, and the Cocaine case - where the choices made are indeed rational, given a commonsense, folk psychological understanding. The fact that the normatively compelling anomalies are not among the clearest descriptive counterexamples to RTC gives the defender of normative DT an opportunity to argue that the compelling "anomalies" are not anomalies at all. Anomalous behavior is relevant to the normative model of rationality, but the the critic must really make the case.

As noted before, Sen has a 'recipe' for generating relevant counterexamples to RCT. What they have in common is that there are two option set - $A = \{p, q, r\}$, and $B = \{p, q\}$ - and following choice pattern:

$\{p\} = c(\{p, q\})$;

$\{y\} = c(\{p, q, r\})$.

When the menu only offers p and q, the agent would choose p. But when a third option r is added to the menu, the agent's choice would change to q. And just as Sen says,

> This pair of choices violates many of the standard conditions of internal consistency – not only the weak (and of course, the strong) axiom of revealed preference, but also the even weaker requirements of binariness of choice and basic contraction consistency (Property α). (Sen 2002, p. 129)

The trick for evaluating normative DT, of course, is to find cases where choosing p from B and q from A makes intuitive sense. But as Sen's own examples show, such cases can be found. The choices in Sen's scenarios are indeed rational, and the fact that they seem to be inconsistent with

RCT suggests that decision theorists need to modify even their normative theory. Sen concludes that normative decision theory, which gives rules of consistency for navigating choices, is inadequate because it cannot deal with such cases with its own resources. He argues that the theory has to appeal to external factors such as the particulars of a decision maker's motives, objectives, and values, social norms, etc. in the concrete choice context. (Sen 2002)

7. Response to the Sen-Style Critique: The Holism Strategy

I believe that there are no Sen-style counterexamples to DT where a person's basic preferences **R*** are formed over a set *Z* of holistic outcomes. The term "holism" is adopted from Frederick Schick, who has inspired the view. On this account, the appropriate representation of basic outcomes - the most fundamental objects of evaluations - and a person's preferences over them must include literally everything that would make a difference to the decision maker's evaluation. An outcome is holistic just in case the agent is indifferent among all of the ways of bringing it about. My holism approach resembles but is arguably more sophisticated than what is known as the *redescription strategy* or the *individuation strategy* that is discussed by philosophers of DT. The basic idea of them all is that an apple is not just an apple. If taking the last apple means violating a social rule which is costly to the decision maker, then the outcome should be redescribed or individuated in a more fine-grained way. In other words, if 'I take an apple' and 'I take the last apple' makes a difference to the decision maker's happiness according to her values, then they should be distinguished properly. The holism view follows Jefferey in treating outcomes as propositions. I will use possible world semantics to characterize propositions and give the set of basic outcomes a more fine-grained structure shortly.

I think that Sen is right that if we do not provide the DT mechanism the right inputs, it will not yield the right normative output. My holism view insists that DT modeling and explanation must *begin* with accurate inputs of one's preferences over holistic propositions. RCT itself is simply a logic of how to adjust one's preferences over all sorts of propositions in a rationally consistent way given their preferences over basic outcomes, i.e., holistic propositions. Once the inputs are appropriate, we can trust the logic to provide good normative guidance on consistent choosing. The counterexamples to DT do not refute the normative theory as a logic of choice. They work only to the extent that we had an inadequate account of preferences of basic outcomes. Therefore, my view allows me to agree with Sen that his counterexamples pose problem for DT. But I think they only pose a problem for the theory as interpreted through the lens of revealed preference lenses, which has been the mainstream interpretation in economics, because it can't distinguish non-holistic propositions. My reply to Sen is that RCT can be defended by breaking it free from the revealed preference interpretation and reinterpreting it as a holistic, realist and normative theory.

In what follows, I argue that the choices in Sen's cases are problematic because they are results from belief change in the scenario. My solution to the problem will be based on this diagnosis. I argue that decision theory has a way to take belief change into account. People's choices are always conditioned on their beliefs about the circumstances, and their choices differ if their beliefs change.

What Sen goes on to say about the Cocaine Scenario seems to be an anticipation of a potential objection to his counterexample argument. He writes,

> It is, of course, true that the chooser has different information even about p (i.e., having tea with the acquaintance) when the acquaintance gives him the choice of having cocaine with him, and it can certainly be argued that in the "intentional" (as opposed to "extensional") sense the alternative p is no longer the same. But an "intentional" definition of alternatives would be, in general, quite hopeless in invoking inter-menu consistency, especially when (as in this case) the intentional characterization changes precisely with the alternatives available for choice (i.e., with the menus offered). (Sen 2002, p.131)

The objection is that the alternative $p \in A$ is not the same alternative as p in B. The presence of prospect $r \in A$ reveals information about the acquaintance that makes a relevant difference in the choice. One way to understand the relevant difference of p in the two choice problems is to think that p has different meanings in the two cases. Sen's "intentional sense" may suggest understanding the problem as a result of not keeping the meaning of p constant. Sen would happily accept this *prima facie* diagnosis. But he continues to doubt the viability of a solution, saying that a solution "would be, in general, quite hopeless." In order to assess Sen's doubt, we need get clear on the diagnosis.

At first glance, the problem seems to be that we did not specify p carefully enough to capture the relevant difference in the two choice problems A and B. The suggestion is that the $p \in B$ is not really the option of 'accepting the offer to go to the acquaintance's home;' instead, $p \in B$ should be described as 'accepting the offer to go and have tea,' or 'accepting the offer to go without being offered cocaine,' etc. And the $p \in A$ should also be described differently, as 'accepting the

offer to go and probably having cocaine.' The solution based on this diagnosis is that there would be no problem for the context-independence of Condition α if we adequately and properly described the option so that the description captures relevant differences. In the Cocaine Scenario, if the prospects $p \in B$ and $p' \in A$ were kept apart, then there would be no counterexample.

This is sometimes known as the "problem of relevant description of options," "redescription" or "individuation problem."[31] The solution to this problem emphasizes on describing and individuating the options more finely, so that the axiomatic internal properties in choice theory would not be violated. Case by case, this solution may seem viable. But as a theoretical move, it needs be accompanied by some principle that determines in a general way how options should be individuated. But identifying such a principle appears to be difficult.[32]

Though the redescription or individuation strategy is a good start on the right track, a more complete diagnosis has to take into account the belief change in the scenarios. Take the Cocaine Scenario. With the basic framework of decision theory we saw in Section 1, we can look at the case more closely. When the decision maker is offered only two prospects in B, viz. p (accept the offer to have tea) and q (refuse to visit), the offer is very likely a normal, friendly offer. The chooser can expect that if she chooses p, there is a relatively high probability that she will enjoy a pleasant catch-up with the friend, say an 80% probability that she will have her favourite green

[31] See for example, Verbeek, Bruno (2001). "Consequentialism, Rationality and the Relevant Description of Outcomes." *Economics and Philosophy*, 17, pp.181-205. And see Buchak, Lara (2013). *Risk and Rationality*, chapter 4, pp.114-145.

[32] Buchak surveys three different principles, as she characterizes them: individuation by justifiers (John Broome), individuation by actual preferences (Jamie Dreier), and individuation by preferable properties (Philip Pettit). Verbeek in his paper argues against Broome's principle.

tea and cake ($x \in Z$). Also associated with the option p is her confidence that there is an extremely low probability, say 10%, that her host will do cocaine ($y \in Z$). Assume that this constitutes the lottery ticket p. And let u represent her preferences. Her expectation of the choice is

$$u(p) = \sum_{z_i \in Z} u(z_i) \Pr(z_i \mid p).$$

But when she is offered cocaine (r), the probability distribution associates with choosing p changes. With the information revealed about the friend, a choice of p to accept the offer is now associated with a very different probability distribution. The outcome y that, her host will do cocaine, increases greatly. Let $p' = B_1(a)$ be this new lottery. $B_0(a) = p$ is not p'. Since we are dealing with the same decision maker, we can assume that her preferences remain the same, and so does u. Now her expectation of the same choice x is different:

$$u(p') = \sum_{z_i \in Z} u(z_i) \Pr(z_i \mid p').$$

We see that the change in choice out of A and B is due to a change of prospects - p to p' - associated with the same action but changed beliefs. Intuitively these lotteries may be understood as reflecting the decision maker's degrees of belief on the relevant states of the world and the resulting consequences. One specific lottery ticket p involves one's credences in one specific situation. But information that is specific to a context may well alter her relevant beliefs about the consequences of an action, and thus alter the expected value of a choice. The formal model handles this by supposing that agents choose actions ($\in A$) that are linked to (ranked) prospects ($\in L$) by way of a belief function B_t: $A \rightarrow L$. As beliefs evolve, so does the action to prospect connection. So the lesson from the above analysis is that without determining all the relevant

information about a choice, the decision-theoretic verdict of which choice to make is unreliable, because the verdict has to be based on the expected value of that choice but the expected value can change when beliefs change. And since the relevant information about a choice is context-sensitive, it follows that decision theory cannot reliably determine, in a context-independent way, what choice to make. Instead, decision theorists have to incorporate the context-specific norms, objectives, and values. Or so argues Sen.

Recall that the function u: $L \rightarrow \mathbb{R}$ is defined as a representation and index of rational choice: $u(B_i(a)) = \sum_{z_i \in Z} u(z_i) \Pr(z_i|a)$. Prospect $p = B_i(a)$ involves beliefs and utility function u represents values and desires. The expected value of an action a, represented by $u(B_i(a))$, depends on u and $B(\cdot)$. We have seen that belief changes can lead to changes in the prospect evaluated when an agent chooses an action.The problem for decision theorists, then, is how to capture the variability of an action a.

One way of getting rid of the variability involves using DT only when a is tied to a particular prospect. To use RCT, then, you would need to know the relevant beliefs such as "I'm a guest and the host and other guests have not picked their fruit yet," "other people may well want that last apple too and I will probably be seen as socially awkward if I take the last apple." Once these relevant beliefs are determined, the prospect p is determined by the action a. Then the expected value of a choice, $u(B_i(a))$, has one specific value, $u(p)$. For example, in the Cocaine scenario, you need to have the information regarding whether the acquaintance will offer you cocaine or not, in order to use decision theory to decide if you want to accept their tea offer. Because once the relevant beliefs are settled and the variability of a is gone, the expected value of your choice would not fluctuate, thus the index u would give you a reliable verdict.

My worry with this solution, as shared by many other authors, is that it makes RCT too context-dependent. Philosophers of DT, when discussing the redescription or individuation strategy, often reject the strategy because it would make decision theory "vacuous" (to use their phrase). For example, Lara Buchak writes, "what an agent chooses in one decision scenario will tell us little or nothing about what he will or should choose in any other scenario." (Buchak 2013, p. 138) This solution makes the use of DT a matter of case study, since every case of decision making will be different. I believe that RCT is powerless if thought to be merely a cost-benefit analysis tool that one uses under a particular situation. Instead, the descriptive power of DT draws from the consistency of an agent's choices and then predicts what one should choose given the desires and values that are manifest in their preferences and choices. Normative decision theory works on the assumption that agents choose in a consistent and rational pattern. But it does not make any substantial assumption on what gets chosen. It is supposed to be a logic of good choosing. Making decision theory too context-dependent would hurt what the theory is supposed to achieve.

The holism solution that I develop is inspired by a notion from Frederic Schick, what he calls "holistic propositions." (Schick 1984) Essentially, we want the basic outcomes to have a constant utility for an agent. According to Schick, we can accomplish this if we *fine-grain* outcomes to the point where the agent does not care about any further fine-graining. Schick uses the idea of "a proposition's being holistic for a person" to capture this idea. Outcomes in decision theory can be described as propositions. Jeffery's version of decision theory is developed under the presumption that objects of desires and beliefs are propositional attitudes. According to Schick, a proposition is *not* holistic for an agent just in case it is a disjunction of more than one ways to

specify an outcome and the agent cares more about one way than another of carrying out that outcome. To use his example, several candidates running for some office are bald. You want a certain one of them to win. Here *a bald man wins* is not holistic for you. (Schick 1984, p. 18) Another example that illustrates the idea of holistic propositions is from Stephen Ellis. Suppose that I want to get a car for myself, but only if it is Mazda Miata. Otherwise, I do not have to have a new car. Let P be the proposition 'I get a car,' and Q 'I get a Mazda Miata.' I like $P\&Q$ a lot, but I do not care very much for $P\&\sim Q$. Here P is a more general proposition, and $P\&Q$, $P\&\sim Q$ are more specific, finer particularization of P. P is not a holistic proposition for me, since it has further particularization that I care for differently.

To illustrate the holism solution, suppose that you want to go to a concert that begins at 8 pm and you want to be seated a few minutes before 8 o'clock. As long as you are seated a few minutes earlier, you do not care if you are seated at 7:55, or 7:56, or 7:57. Let z be the outcome that you are seated at 7:55, z' be the outcome that you are seated at 7:56, and z'' be the outcome that you are seated at 7:57. These three different outcomes do not affect your happiness in any way in the situation. In other words, (z or z' or z'') is holistic for you. A neat consequence of holistic outcomes is that they give you exactly same utility, i.e., $U(z) = U(z') = U(z'')$. Suppose that z, z' and z'' are the only members in your outcome set Z in this situation. Therefore we have the following result,

$$u(p) = \sum_{z_i \in Z} u(z_i) \Pr(z_i|p)$$

$$= u(z) \sum_{z_i \in Z} \Pr(z_i|p)$$

$$= u(z)$$

Notice that $u(p)$ now has a constant value. It is equal to $u(z)$, the utility that you have with the holistic outcomes. No matter how your beliefs change which would result in changes in the probability distribution p, the expected value of $u(p)$ would remain unaffected. This is how holistic outcomes can block any change in expected values of your choice alternatives resulting from belief changes.

The holism strategy can handle Sen's Last Apple case. Initially we seem to be looking at two outcomes,

p: I have nothing.

q: I have a nice apple.

When applying the holism strategy, the outcome "having a nice apple" should be distinguished as (at least) two holistic outcomes. One is having a nice apple when it is the last remaining one, and the other is having a nice apple when doing so does not violate any social norm. The decision maker clearly prefers the latter to the former. More explicitly, let q' and q'' designate these two holistic outcomes,

q': I have a nice apple when doing so would be rude.

q'': I have a nice apple when doing so does not violate any social norm.

I quite reasonably prefer having an apple when it doesn't violate a social norm to not having an apple, but I likewise prefer having no apple to having an apple when it would be rude. It should turn out, then, that $u(q') < u(p) < u(q'')$. And let $\Pr(q'')$ be the probability of the state of affairs "taking the last apple would not make me look rude." Given the context of the scenario, a common sense judgment is that the action of taking the last apple - b - in this circumstance makes the probability that I would not look rude extremely low: $\Pr(q''|b) \approx 0$.

Now we evaluate action *b* by calculating the expected utility of the gamble it is associated with,

$$u(B(b)) = u(q'')\Pr(q''|b) + u(q')\Pr(q'|b) \approx U(q') < U(p)$$

Standard expected utility theory accompanied by the holism strategy yields the correct result that the decision maker would not choose to take the last apple.

On my account, the outcome set is treated as one that is composed of holistic propositions. The idea to treat the objects of preference and of belief as propositions is attributed to Richard Jeffrey. Let us introduce the set Ω of all possible worlds, and utilize the familiar notion of a proposition as a set of possible worlds where that proposition holds. A proposition *N* is represented by a subset of Ω which contains possible worlds where it is the case that *N* obtains. If we think of a proposition as a way of partitioning the space of possible worlds, then the finer the partition, the smaller the set of worlds. The singleton {ω}, then, contains a world that is maximally specified and it represents a proposition that cannot be further partitioned. We say that an agent S *believes that N* when S thinks that some $\omega \in N$ is actual. *S desires that N* means that S desires that some $\omega \in N$ is actual.

We can illustrate how to construct a holistic outcome set *Z* for a certain situation. Consider some non-holistic outcome, *p*, for example,

p = "I take the last apple,"

There are different ways to realize *p*. With the possible worlds account, these different ways can be represented as the intersection of *p* and some other proposition p_1, p_2, …, or p_n. Two of these other propositions may be the following,

p_1: "I'm a guest and the host and other guests have not picked their fruit yet, and they may well want that last apple too, so I will probably be seen as socially awkward if I take the last apple."

p_2: "I'm a guest and other people have all picked their fruit, so I will not look socially awkward if I take the last apple."

Since p_1 and p_2 are incompatible, the intersection of them must be empty. But there are other finer particularizations of p, in other words, some intersection sets of p with these other propositions that are not empty. Suppose that for some particular individual decision maker, Sam, some finite ways of realizing p exhaust all the different ways that can affect Sam's evaluation of p. Let $p_1, \ldots, p_m$ be these propositions. The intersection of p and $p_1, \ldots, p_m$ is a holistic proposition for Sam. And all subsets of this intersection are holistic for Sam. As a result, they should have a constant utility for Sam. For each w in the power set of the intersection of p and $p_1, \ldots, p_m$, $u(w) = c$ where c is a constant real number. Let the intersection $p \,\&\, p_1 \,\&\, \ldots \,\&\, p_m$ be denoted by z. We have the following result,

$$u(z) = \sum(w \in \mathrm{Pow}(z))\, u(w)\, \Pr(w)$$
$$= u(w) \sum(w \in \mathrm{Pow}(z))\, \Pr(w)$$
$$= c,$$

This explains how holistic propositions can free decision theory from the Sen-style counterexamples. A decision maker's belief changes no longer have any effect on the expected value of a choice among holistic prospects, as $u(z)$ is shown to be a constant value.

A consequence of the holism strategy is that the utilities of holistic outcomes are a constant value. One way we can look at the effect of context-sensitive motives, values or social norms is to locate them in the change of probability distributions. But regardless of belief change, the

expected utility of a choice stays constant if we apply holistic outcomes. Therefore, holding every proposition in the outcome set to be holistic can block any change in the expected utility of a choice that is due to a change in beliefs.

8. Objects of Preference as Holistic Propositions

The holism solution seems similar to but is different from the ideas of a redescription strategy that are commonly mentioned in the literature. Those ideas were commonly mentioned but rarely developed in details. In his recent paper "Has Game Theory Been Refuted?" (Guala 2006), Guala discusses the question of whether game theory is tautological. His discussion is placed with game theory, treated as an extension of rational choice theory, and thus I will transfer his points and make them with decision theory.

According to Guala, one might question the idea of redescription and fine-graining because doing so would seem to define away any violation of the theory. The worry is that allowing theorists to redescribe the outcomes whenever an action appears to be inconsistent with DT would make DT unfalsifiable. Building the fine-graining strategy into DT gives it the power to help itself whenever challenged by counter-evidence. The holism strategy compels DT to make the following claims, which are claimed to be tautological:

1. Rational agents always choose the option that maximizes expected utility.
2. Action *a* maximizes expected utility for the agent.
3. Therefore, the agent always chooses action *a*.

Guala correctly notes that a tautological claim is one that is true sheerly in virtue of its logical terms, and the set of (1), (2) and (3) is an argument and not tautological propositions. Related to the worry of being tautological, other authors complain that implementing the fine-graining

strategy in normative decision theory costs the theory of making vacuous claims like (1), (2) and (3). But the set of (1), (2) and (3) is a valid deductive argument. A scientific hypothesis accompanied by auxiliary claims about the background conditions in which the hypothesis is applied may deduce a set of data where the hypothesis applies. The existence of such data resulting from the hypothesis and bridge claims renders the hypothesis falsifiable. The formalism of DT together with the expected utility theory bridging the formalism to empirical content make predictions about rational choice. When the predicted choice is irrational, contrary to the DT prediction, DT is falsified. For example, a significant number of tourists at the Great Canyon are documented to unconsciously step back and fall into the canyon when they take photos. We can be too focused on a single aspect of a situation and develop a tunnel vision that makes us cognitively blind to other aspects needed to be considered. Given these people's unwillingness to die while they travel, DT would predict that they not back themselves into the canyon, and yet these documented instances falsify the DT prediction.

It is a separate question whether an instance of counter-evidence is a successful instance of counter-evidence, which should be distinguished from the falsifiability issue. This other question, which seems a more critical worry to the fine-graining strategy, questions the adequacy of the structure of DT to reflect external motives, objectives and values. As Guala points out, "[s]uppose that subjects cared about something else besides their own monetary gain, but that game theory was not flexible enough to accommodate this 'something else' within its framework." (2006, p. 251) He argues that the standard treatment of utilities and beliefs in the DT formalism does not provide enough flexible structure to accommodate certain external factors such as reciprocity, which shows that the standard normative decision theory is inadequate. He writes,

> According to our everyday understanding, any observed pattern of behavior can in principle be rationalized by imputing an appropriate—perhaps highly sophisticated—structure of beliefs and desires. … [But] the formal theory of rational choice is less flexible than folk psychology, and as a consequence has more limited explanatory resources. This applies to the theory both in its normative and in its descriptive version … (Guala 2006, p. 251-2)

While my holism strategy demonstrates that DT successfully captures the essence of folk psychological principles of instrumental rationality, Guala rejects the refinement strategy because he thinks that the Savage measurement procedure is too restrictive to accomplish some refinement such as that of reciprocity.

The utility measures of preference developed by von Neumann and Morgenstern, Ramsey and standardized by Savage have occurred in the textbooks of decision theory. Guala specifically refers to the "Savage measurement procedure" as "standard in contemporary game theory, noting that "according to 'kosher' game theory, the application of the theory should start with a Savage measurement, that is, with the identification of players' utilities and beliefs." (2006, p. 246) The Savage formalism defines a set of outcomes, X, and a set of states of the world, S. It is essential for this formalism to keep the states and the outcomes separate, so that all the uncertainty is confined to the states and values are associated with the outcomes. The set of actions (or 'alternatives,' 'prospects'), A, is constructed by mapping from S to X. For example, the action of 'booking a flight in January' assigns outcome 'the flight is canceled' to the state 'there is heavy snow;' the action of 'taking a nice apple' maps from the state 'there is a nice apple available' to the outcome 'I have a nice apple.' A preference relation, or "betterness" relation in

J. Broome's term, $\succsim$, is defined on $A \times A$ such that for $a, b \in A$, $a \succsim b$ iff $u(a) \geq u(b)$. And $\succsim$ is an ordering (i.e., a reflective, transitive and complete relation). Since $A: S \rightarrow X$, we can specify the preference relation $\succsim = \{<<s, x>, <s', x'>> \mid s, s' \in S \ \& \ x, x' \in X\}$. Guala argues that the fine-graining of reciprocity is inconsistent with what is called the "rectangular field assumption" attributed to Broome. For each state of the world, s_i, take any possible outcome, x_i for that state. Call the set of all such pairs X_i, that is, $X_i = \{<s_i, x> \mid s_i \in S \ \& \ x \in X\}$. The rectangular field assumption, according to Broome, is that $X_i \times X_j = \succsim$. The Cartesian product that is constructed by picking out any outcome for a given state of the world is the same set as the preference relation. In other words, actions that are arbitrarily picked out from S and X will occur in the preference relation. As Broome describes the assumption, remembering that he uses "prospects" for "actions,"

> Take the set of all outcomes: the set of all the possible outcomes that any prospect may lead to. Now let us go through the states of nature one by one, and to each assign, quite arbitrarily, some outcome from this set. This operation will define an arbitrary prospect: the prospect that delivers, in each state, the outcome we have assigned to that state. The assumption says that *any* arbitrary prospect constructed this way has a place in the preference ordering. This is the rectangular field assumption as I described it in Section 4.4, with the added assumption that any outcome can appear in any state of nature. (Broom 1991, pp. 115-6)

The more intuitive idea of the assumption is that every function mapping from the set of states of the world to the set of outcomes must be a meaningful action, meaningful in the sense that a

preferential attitude can be expressed toward that action so that a preference relation can be formed over the set of all actions. As Sugden notes,

> So to say that a person has any kind of preference relation between two acts f and g is to imply that it is possible to confront that person with a choice between those two acts. This feature of acts - that any pair of acts must be capable of constituting a meaningful choice problem - is clearly required by Savage's approach, in which preferences are defined in terms of observable choice behaviour. (Sugden 1991, pp. 761-2.)

Guala argues that the rectangular field assumption which is implied in the standard Savage measurement procedure conflicts with the fine-graining of reciprocity which produces prospects that cannot be offered as choice, either because those prospects do not make sense or they depend on counterfactual paths. We will first see his argument made in the case of the sequential version of the Prisoner's Dilemma game. Player A can choose Up where player B may then choose to cooperate (Left), giving both A and B the optimal outcome overall, while running a risk of getting the worst outcome if B does not cooperate. B has incentive to defect (Right) because playing Right gives B a better outcome than the one resulting from cooperating. Or player A can choose Down where A is guaranteed better outcomes, while leaving B with the worst and sub-optimal outcomes. A makes a choice before B responds. In this game, when we look at B's choice, B would choose to cooperate if A had chosen the cooperative move, due to the fact that B values reciprocity. Suppose that the valuing of reciprocity causes B to prefer to cooperate. This preference for reciprocity and cooperation must be conditioned on A's playing the cooperative move first. Reciprocity consists in responding nicely to the other's previous trust.

However, as Guala argues, the Savage measurement procedure would present B with the prospects Left or Right by themselves and not as a subsequent move after A. The fine-grained reciprocity-valuing outcome would not make up a prospect that has a place in B's preference ordering. When an agent values reciprocity highly and chooses an action that results in her reciprocating a cooperative move, refinement of such an outcome has to make reference to the specific choice context, and therefore cannot fulfil the general purpose of outcomes in the formalism of DT. (Guala 2006, pp. 252-6) Guala acknowledges that payoffs are not completely determined by the attractiveness of outcomes in themselves, but are also affected by an agent's beliefs associated with the choice situation. As Guala writes,

> The identification of the preferences associated with each consequence requires an examination not just of what will happen, but also of what might have happened. To put it another way, the utilities are context-dependent: they do not just depend on the outcomes taken in isolation, but on the whole structure of the game. (p.265)

But when the description of an outcome has to refer to an entire causal history that leads to that outcome, arbitrarily constructed actions from such an outcome would make no sense to an agent to express a preference over. In short, the Savage model of rational choice requires that preference be expressible among arbitrarily constructed acts. But this requirement places restriction on what goes into the outcome set. Outcomes that are fine-grained to include reference to the causal paths will be illegitimate in the Savage decision model.

To better understand Guala's critique, let us look at the Last Apple scenario. The set of outcomes, X, consists of the following two outcomes.

x: I have nothing.

y: I have a nice apple.

The set of states of the world, S, consists of

s_1: I am in a personal space where my choice would not affect others.

s_2: I am in a social setting with social norms present.

The set of prospects, A, can be constructed as functions from S to X. A = {<s_1, x>, <s_1, y>, <s_2, x>, <s_2, y>}. The rectangular field assumption requires that the preference relation be the same set as the one constructed by picking out outcomes arbitrarily from X for each member in S, as follows.

$\succsim = X_1 \times X_2$ = {<<s_1, x>, <s_2, x>>, <<s_1, x>, <s_2, y>>, <<s_1, y>, <s_2, x>>, <<s_1, y>, <s_2, y>>},

where X_1 = {<s_1, x>, <s_1, y>} and X_2 = {<s_2, x>, <s_2, y>}.

The fine-graining of the outcome y into y' and y'' creates problem for the current model.

y': I have a nice apple and feel bad for being rude.

y'': I have a nice apple without hard feelings.

With the more refined X, the prospect set should be modified: A^* = {<s_1, x>, <s_1, y'>, <s_1, y''>, <s_2, x>, <s_2, y'>, <s_2, y''>}. The rectangular field assumption insists that any prospect that is a subset of A^* has a place in the agent's preference relation, including a^* = {<s_1, y''>, <s_2, y''>}. This prospect describes the action *Take a nice apple without hard feelings whenever - no matter whether I am in a personal space or in a social setting*. Prospect a^* makes no sense because the reason why I am without hard feelings is that my behavior is not rude in a social setting. Taking the last apple in a social setting makes it impossible for me to have no hard feelings. Thus, I cannot have a preference over prospect a^*, which contradicts with the rectangular field assumption.

Among other philosopher and economist, Guala and Broome worry that the fine-graining strategy produces prospects that either do not make sense as actions for an agent to have preference over, such as a^*, or ones that involve counterfactual causal paths and cannot be offered as actual choice. But the standard Savage measurement procedure requires that prospects be actually choosable, which makes it unacceptable for them to adopt a full-blown fine-graining strategy.

While the Savage model defines the objects of choice as prospects, my formal model defines them via a set of gambles L and a belief function B. A gamble in L is a lottery ticket that distributes some probabilities to some basic outcome in Z. The belief function then narrows down the set of L to the actual choices that the agent thinks are available. Formally speaking, this model still hinges on a version of the rectangular field assumption. It is a different version because, instead of every actual prospect having a place in an agent's preference, it states that every gamble in L has a place in the preference ranking over gambles **R**. The original rectangular field assumption requires that any function that picks out certain outcomes from certain states of the world must make sense as an action and must be presentable as an actual choice for the agent to have a preference over. Fine graining certain outcomes creates problem for the original assumption. However, in my version, the assumption would require that every member in L has a place in **R**, which is exactly a result of the formal definitions.

In my case, respecting a rectangular field assumption does not create conflict with fine-graining because all refinement occurs in the set Z of holistic outcomes. It is no surprise for the holism model that some gambles in L can never actually happen and thus cannot be offered as choice. Any option that counterfactually depend on some other path in the past of the agent's history

would be inadmissible as actual choice. A philosopher can never choose between which plane she will fly had she decided to be a pilot at 17. Guala expresses this worry that if an outcome is counterfactually dependent then an action leading to it cannot be presented as actual choice. My holism model avoids this problem because it allows the objects of preference to be counterfactually dependent and it does not compel actions to be actual the way the Savage model does. It is certainly true that some gambles in L cannot be offered as actual choice, just as some of the problematic prospects pointed out by Guala and Broome. My holism model needs not that all gambles be actual choices. It only requires that they be evaluable. A philosopher can certainly evaluate the prospect that she chooses to fly one plane rather than the other had she became a pilot. The set of evaluable options is much larger and richer than the set of options that can be actually offered as choice. L from the holism model represents the set of evaluable options, whereas the Savage model restricts preference to actual choices.

In the holism model, it is the belief function B that narrows down L to actual, realistic actions. The action set A is composed of the actual actions that the agent thinks are available in the actual choice situation. B maps A to L so that $B(a)$ (for $a \in A$) is presentable to the situation as an actual choice. Thinking in terms of how we actually choose, we do not operate on a rectangular field of actions that we have figured out before we encounter any real situation. Instead, actual choices are almost always belief-mediated. We believe that certain actions conduce to certain desirable outcomes under certain conditions, and then we choose them under those conditions. Those conditions are represented by B.

The fact that Guala and Broome, among other philosophers and economists, insist that preference can only be tied to actual choices may be explained by a revealed preference

"hangover." The revealed preference project equates the notion of preference to patterns in actual choice. Hence the requirement that any arbitrarily constructed action must have a place in the preference ordering and all preferences are tied to actual choices.

This discussion of Guala and Broome's worry about fine-graining also manifests a significant distinction of my holism strategy to the normative DT from the Savage framework. The Savage model has to exclusively separate beliefs and utilities, limiting the prospect of fine-graining outcomes. But that is a problem for the Savage model and not for my holism account. Holistic outcomes are by definition "loaded" outcomes. To make that clear, let us distinguish *consequences* and *objects of preferences*. Consequences are the end results of some act. They are the familiar end results such as "I have an apple for dessert," "I get $10," "my friend is sentenced to jail," etc. in the Savage model. The Savage model takes the objects of preference to be consequences. What I label as "holistic outcomes," or the proper objects of preference, are more than just end results. This is because people's preferences are context-sensitive and path-dependent, as a matter of fact. The objects of preference may include and be influenced by values like knowledge, pleasure, peace, kindness, justice etc. People's actual preferences are always more complex. A proper representation of those preferences should reflect the complexity and path-dependency.

Pettit (1991) describes four common phenomena that suggest this idea that our preferences are often complicatedly entangled in specific aspects of a choice situation. The first phenomenon is that we experience internal conflict in desires. Pettit gives the example of a conflicting choice between attending an important departmental meeting and seeing his son perform in the school play. He says that after he formed the desire or preference to attend the meeting, he continues to

feel the conflict. He argues that on decision theory, once the preference for the meeting is in play and the desire to see his son is overthrown, there is no reason that he should still feel a conflict in his desires. Pettit's own explanation is to appeal to a distinction between preferring a prospect and preferring a property, what he calls the assumption of desiderative structure. According to Pettit, he feels a conflict in his preferences because there is in him a distinct desire of the property of being able to see his son on stage, even though he prefers to choose the other prospect.

The second phenomenon is the distinction between what Pettit calls "desire simpliciter" and "prima facie desire". Prima facie desire is sometimes also described as "desire in so far as something is true, desire in a certain respect." He says that this distinction is taken as a primitive in decision theory, but can be explained in more basic terms. One comes to desire that p, period, only in so far as she desires that p, qua some property F. This is naturally explained by saying that one forms a prospect-preference only in so far as she identifies it "as the bearer of certain properties" that she desires.

Third, Pettit notes the fact that linguistic desire-contexts are not extensional. Suppose that an agent desires that p, and p is equivalent to q. We cannot substitute p for q and say that the agent desires that q. To use his example, John desires to go to the movies, but going to the movies is equivalent to disappointing his mother tonight. He has the desire for the movie, but we do not want to say that he desires to disappoint his mother tonight. To explain this non-extensional feature, we can say that when we use sentence p to pick out a prospect, we also pick out a certain property displayed by p, for example, the pleasure in going to the movies. But another sentence q

can pick out the same prospect but with a different property, the property of disappointing one's mother. It would be misleading to substitute p for q in a preference context.

The fourth phenomenon is that sometimes we are requested, and able, to give reasons for our choices. According to the standard decision theory, there is only one kind of reason, that is that the option chosen best realizes one's preferences given one's beliefs. But sometimes more specific reasons are asked for with a particular aspect of a choice. For example, we ask questions like, "How could you want anything so cruel?" "How could you desire such a comparatively unfair outcome?" and "How could you ignore the self-destructive aspects of your decision?" To answer these questions, it seems that we need a more refined structure of desire and preference.

These phenomena show that human preferences are clearly complicated and context-sensitive. My approach of fine-graining the objects of preferences is based on acknowledging this fact of our preferences. Having a nice apple is an end state. But your preference is different towards taking an apple when you are in your personal space and taking an apple when it is the last fruit on table and you are with a group of people. Getting $0 is a different outcome when you could have got one million. Your preference toward being given $2 can differ depending on whether you think you are being treated unfairly or it is the result of an fair split. Consequences have been treated as the objects of preference in traditional rational choice theory. But as it is increasingly recognized, the objects of preference are highly intertwined with specific contexts and choice paths.

In the Savage framework of rational choice, actions are functions constructed by mapping outcomes to states of the world. The set of outcomes and the set of states are exclusively distinct from each other, so that actions can be constructed from arbitrarily assigning some outcome to

some state. The rectangular field assumption requires that actions thus constructed must present as meaningful options toward which the decision maker can express a preferential attitude. The difficulty is that outcomes that include a description of the specific choice context cannot be freely assigned to construct meaningful actions across contexts.

My holism strategy avoids such technical difficulties because it does not rely on the same rectangular field assumption. Unlike the separation of consequences and states of the world, my model makes it unnecessary to exclusively maintain uncertainty in the set of states. Outcomes and beliefs are not exclusively distinct categories since people's preferences are affected by their beliefs. A representation of outcomes should reflect this belief-loadedness of outcomes. The set of holistic outcomes sets out a representation of the objects of preference by reflecting the complexity of belief-mediated utilities in terms of holistic descriptions of outcomes. A preference ranking **R*** is first defined on the set of holistic outcomes, before the expected utility transmission from **R*** to ranking **R** on the set of lotteries. This is different from the Savage model where a preference relation is defined on functions from states of the world to consequences. The Savage procedure measures the utilities of prospects x, y and z by comparing the agent's preference for y and a lottery of x and z. The holism model accomplishes such measurement procedure for basic outcomes by representing them in the set of holistic outcomes Z. Thus it can avoid the difficulties that the Savage model has with fine-graining, and saves the DT expected utility model as a good norm.

When making choices, the agent's beliefs about the situation pick out the gambles to which actions would lead. And the gambles are composed of holistic outcomes and conditional probabilities of those outcomes, as shown in the following relation.

$$B_i(a_i) = L_i = [p(z_1|a_i)\, z_1, \ldots, p(z_n|a_i)\, z_n]$$

Unlike the Savage model, choice options are picked out by the assignment of gambles according to the agent's belief function B_i at the time of decision making. The holism account thus significant differs from the Savage framework and avoids its tension with fine-graining outcomes. Instead, the refinement of outcomes is an integral part of the holism account.

9. Holism in Sequential Choice

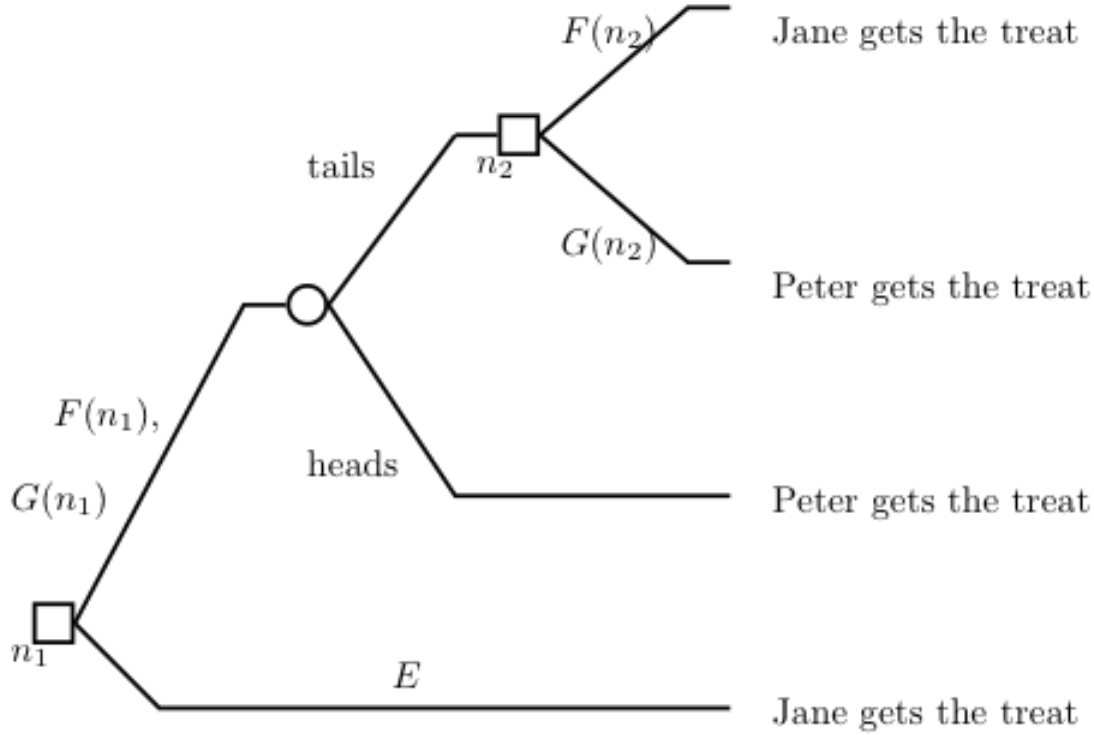
A more refined structure of preference should capture the fact that decision-theoretic evaluations are path-dependent. Holistic approach to decision theory allows path-dependency to be captured in the outcomes, and thus maintains the folk-psychological framework of decision theory as an evaluative tool. In the following example, which I call "the Mom case," discussed by Verbeek (2001) and adopted from Diamond (1967), we can see path-dependency matters and how - unlike the traditional Savage model - my approach captures the difference in preference as a result of path-dependent outcomes.

The Mom Scenario

Mom has a treat that she wants to give to her two children, Jane and Peter. The treat is indivisible so only one of them can receive it. Mom may choose to either give the treat to Jane, or to give it to Peter. Or Mom may use a fair coin to decide. The decision problem that Mom faces can be expressed in the following table, and as a dynamic choice in a decision tree.

	Coin lands heads	Coin lands tails
Give to Jane	Jane gets the treat	Jane gets the treat
Give to Peter	Peter gets the treat	Peter gets the treat
Let coin decide	Jane gets the treat	Peter gets the treat

Figure 1.4



Decision tree specifies a choice situation where you make multiple choices over time. Such a situation is also known as a `dynamic choice' or a `sequential choice.' A *decision tree* is a collection of choice nodes and chance nodes. A *choice node* is a point where you make a choice, represented by a square. A *chance node* is where a chance event happens, such as flipping a coin or rolling a dice, and a chance node is represented by a circle. A tree begins with a choice node, meaning that you make up a plan and make a choice to follow that plan at the outset. The end-point (the *terminal nodes*) of a tree is the *outcome* of the choice(s) -- namely the consequence resulting from a series of choice and chance events. We will also need the notion of a `truncated tree.' A *truncated tree* is the part of a tree that begins with a node *n* and ends with the terminal nodes following *n*. Finally the chooser has *plans* that specify her choices at every choice node that she can possibly reach.

Mom has three options (plans) to choose from, denoted by italic capital letters:

E Take 'down' at n_1 (*viz*. take $E(n_1)$),

F Take $F(n_1)$, then take $F(n_2)$ if 'tails',

G Take $G(n_1)$, then take $G(n_2)$ if 'tails'.

The outcome of plan *E* is that Jane gets the treat (unfairly). The outcome of plan *F* is that Jane gets the treat if 'tails' and Peter gets the treat if 'heads,' which is the fair outcome that Mom prefers. The outcome of *G* is that Peter gets the treat no matter what. Since the Savage model of decision theory has to keep distinct states of the world and end results of an action plan, it cannot reflect the difference between an end result and a path-dependent end result. (Guala 2006; Verbeek 2001) But in the Mom scenario, some end results are clearly path-dependent. The end result 'Jane gets the treat' could either be obtained from Mom giving it to Jane directly, or from the result of a fair coin flip, and Mom's preference is clearly different toward the two. Verbeek shows that the Savage model relies on "consequentialism" principles that prevent the refinement of path-dependency.

Chapter 3

Decision Theory and Risk Attitudes

1. Introduction

Decision theory (DT) purports to capture the notion of instrumental rationality by formally representing the way people's desires and beliefs lead to their actions. Expected utility theory, the standard normative articulation of DT, claims that if a decision maker's beliefs about and preferences over outcomes are consistent with its axioms then her optimal choice would be to maximize her expected utility.

Expected utility theory is faced with a critique, however, that it makes mistaken judgements about some choices that are pre-theoretically rational. Philosophers and decision theorists discuss counterexample scenarios where an agent's choices seem to make good sense - pre-theoretically - but they are judged to be irrational according to DT. Counterexamples that are widely discussed include the Allais paradox, the Diamond paradox (which I call "the Mom case" in the first paper), and Amartya Sen's apple case, to name a few. Note that the focus here is not descriptive but rather normative. The critique I discuss here is *not* that people in ordinary lives actually make irrational choices - we know that there are such choices that result from various psychological biases and heuristics. Instead, the focus here is on *normative* DT. The critique is that DT should, but fails, to recommend certain choices as rational.

In Chapter 2, I defended DT from a range of these counterexample arguments by giving a diagnosis and solution inspired by Frederic Schick's idea of **holistic** propositions. We cannot universally assume that a decision maker's preferences of available outcomes are internally consistent with the axioms of the theory because an actual decision maker's preferences of outcomes are always ambiguous. The "holism strategy" I proposed insists that outcomes be

'fine-grained' enough to reflect everything that matters to the decision maker and that this keeps the objects of preference univocal and unambiguous. I showed that if outcomes are univocal in this way then internal consistency of preferences, and so choices, can be assured. Thus, I defended the idea that is central to the decision-theoretic model – the idea that we can model a decision maker's choice behavior with expected utility maximization and can recommend her best choice in future situations on the condition that her preferences are consistent with the axioms of the model.

In her recent book *Risk and Rationality*, Lara Buchak proposes her "risk weighted decision theory" as an alternative to traditional DT which, she argues, cannot accommodate a rational agent's attitudes toward risk. She employs several counterexamples where the agent seems to have a rational attitude of risk-aversion, but that traditional DT considers irrational. Before reaching her final conclusion that in order to accommodate risk attitudes we need to update from the traditional theory to her risk-weighted version, Buchak considers and rejects two main defenses from traditional decision theorists. One of the main defenses, the "individuation strategy" as Buchak calls it, will be the focus of this paper. On this view, traditional DT mistakenly judges certain choices to be irrational because the relevant outcomes are not specified finely enough, so that outcomes that are in fact distinct appear as one single outcome. Confusing distinct outcomes is what causes trouble for traditional DT. The individuation strategy then argues that traditional DT can be defended by adequately 'fine-graining' the relevant outcomes.

While Buchak eventually rejects it, I endorse and further develop the individuation strategy. It is easy to see that my holism strategy is in the same spirit. According to the holism strategy, traditional DT can accommodate risk attitudes if preferences are based on holistic outcomes --

namely, outcomes fined-grained to the extent that everything that matters to the decision maker is captured. However, Buchak thinks that this way of defending DT ultimately fails, and that her risk weighted expected utility theory is the correct view.

I will divide Buchak's argument into three steps. First, she introduces several examples where decision makers make choices that are intuitively, pre-theoretically rational, but are judged to be irrational according to traditional expected utility theory. The decisions in the counterexamples all involve considerations of the risks associated with the gambles. Second, she makes a distinction between local and global sensitivity to risk. Sometimes when people express their preferences for outcomes they pay special attention to the riskiness of a particular outcome. These agents are "locally sensitive to risk" in Buchak's phrase. At other times, people pay special attention to the risk-relevant features of an overall prospect: how much the minimum return is, the spread between the maximum and the minimum, how much volatility there is, etc. The attention here is not on features of the particular outcomes by themselves, but rather on the overall structural features of a gamble. This is what Buchak calls "global sensitivity to risk." Buchak argues that traditional expected utility theory, with its utility function and probability function that model preferences and credences respectively, cannot give a full story of global sensitivity, and therefore should be supplemented with a third element – risk function on probabilities, in order to capture the unique phenomena of global sensitivity to risk. In the third step of her argument, Buchak gives some reasons for thinking that the risk function is not an ad hoc amendment because it naturally captures real phenomena that are recognized as global sensitivity to risk.

In order to properly deal with local sensitivity to risk, Buchak needs to appeal to the holism strategy anyway. As we will see, an essential step in her risk function analysis is to have a

preference ranking of basic outcomes. But, as I have argued elsewhere, such a ranking cannot be obtained without proper individuation of the outcomes. For example, in Sen's Apple Case, an apple is not simply an apple when it is the last one on the table at a social gathering and more than one guests may want it. A decision maker in this case will not consider the apple as a plain apple that is free to take. Context-specific social norms matter. Likewise, one million dollars guaranteed may be your optimal choice in the Allais Case, but you may not want it if it is associated with something that compromises your integrity. You may not want it if you cannot get it in an honest way, for instance. As I argued elsewhere, properly analyzing a rational decision requires properly individuate the basic outcomes. Buchak's risk analysis cannot get around a preference ranking of properly individuated outcomes, giving us further reason to see if the holism strategy can handle global risk as well.

I agree with Buchak that there is genuine sensitivity to risk that concerns the global features of a gamble, but I do not think that decision theorists need to appeal to a third ingredient other than the basic framework of preference and credence. I think that the global sensitivity of risk is best understood as a feature of decision makers' preference rankings and that appropriate descriptions of outcomes on which the preference ranking is based help us see that. I think that outcomes can be fine-grained enough so that a decision maker's concern for risk, be it local or global, can be captured in preference rankings of holistic propositions. The holism strategy suggests that outcomes be described specifically enough to reflect anything – any objectives, motives, values, or context-specific norms – that matters to the decision maker. There is nothing special about risk attitudes that should separate it from a unified treatment of all preferential

attitudes. Nothing prevents us from "loading everything into the consequences."[33] In response to Buchak, I argue that since every preferential attitude can be loaded into the preference ranking of holistic outcomes, the holism strategy can give an explanation of the global sensitivity to risk that is manifest in Buchak's counterexamples. I argue that my explanation equally naturally captures the correct rationale behind those rational choices. So there is no good reason to look further for a third ingredient besides the basic framework of preference and credence.

2. Buchak's Counterexample Argument and the Global Sensitivity to Risk

2.1 Alice and Bob

The first example from Buchak is a case where she thinks DT cannot explain why two agents make the same choice for distinct reasons.

Alice and Bob

> Consider two people who collect stamps. Alice is only interested in obtaining one Elvis stamp (she likes her collection to be diverse): once she has one, a second Elvis stamp is next to worthless. Bob, on the other hand, has an insatiable appetite for Elvis stamps (he does not become "saturated" with respect to Elvis stamps), but he does not like to take risks. In general, he thinks that the *possibility* of getting something good does not make up for the *possibility* of getting something bad, when the bad thing is as bad as the good thing is good, and there is an equal chance of each. He might even dislike taking risks so much that he always just wants the gamble with the best worst-case scenario. When offered a choice

[33] Hampton (1994) uses the phrase "loading up the consequences" to describe this sort of idea. Although Hampton rejects this strategy because she thinks that this strategy does not capture the difference between "a preference for a consequence" and a preference for "the state in which the consequence will occur." (Hampton 1994, p.233)

between one Elvis stamp on the one hand, and a coin flip between two Elvis stamps and none on the other hand, both Alice and Bob will choose the former. (Buchak 2013, p.25)

Alice chooses the one Elvis stamp option without taking a risk, because two Elvis stamps has the same value to her as one stamp and DT gets it right about Alice's choice since the one-stamp choice offers her better expected utility than the two-stamp choice. u_A(1 Elvis stamp) = u_A(2 Elvis stamps) > u_A(no stamps) so

u_A(1 Elvis stamp) > 0.5u_A(2 Elvis stamps) + 0.5u_A(no stamps).

In Bob's case, taking one Elvis stamp or taking a 50/50 chance at two stamps each offers the same expected utility. u_B(no stamps) = 0; u_B(2 Elvis stamps) = 2u_B(1 Elvis stamp) so

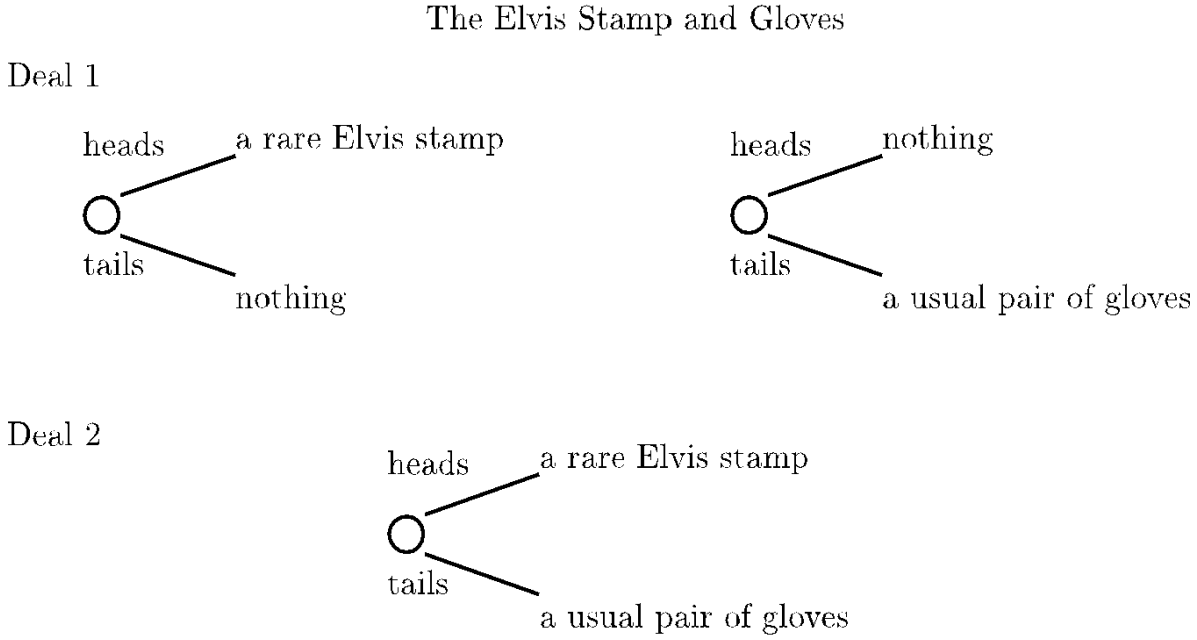
u_B(1 Elvis stamp) = 0.5u_B(2 Elvis stamps) + 0.5u_B(no stamps).

Bob definitively chooses not to take a chance despite the fact that he should be indifferent. Thus, DT cannot explain Bob's choice, as Buchak argues. However, given Buchak's careful description of Bob's rationale for his choice, he seems to have a perfectly legitimate reason to not take the gamble. Bob has distaste for gambles in general. He cares about the "global" feature of his options – he "cares about how outcomes of particular value are arranged across the possibility space," to use Buchak's words.[34] Bob does not choose the risky option due to the fact that it is risky and he dislikes taking risks, even though the risky option offers the same expected utility as the other option. Buchak argues that since DT dictates that the optimal choice should always follow expected utility, DT cannot explain why Bob is making a rational choice. But Bob's global concerns for the riskiness seem real and reasonable.

[34] Buchak 2013, p.25.

2.2 Elvis Stamp and Gloves

I will call the second counterexample scenario from Buchak the *Elvis Stamp and Gloves* scenario, as shown in the following figure. In this scenario, Jeff is offered a choice of one of two deals. A fair coin is used. With Deal 1, two consecutive coin flips decide which prize Jeff would receive. If the first coin flip lands heads, Jeff gets an Elvis stamp. The stamp is considered to be precious and Jeff would love to get it. If the first coin flip lands tails instead, Jeff gets nothing. If the second coin flip lands heads, Jeff also gets nothing. If the second coin flip lands tails, Jeff gets a normal pair of gloves. He likes the gloves and prefers to have the pair than nothing. Deal 2 offers a simple one-time coin flip. Jeff gets a rare Elvis stamp if the coin turns up heads, and he gets a usual pair of gloves if tails.



Here is Buchak's elaborated assumptions on Jeff's preferences.

> Jeff values the two goods *independently* in the sense that having one does not add to or decrease from the value of having the other; they are not like, say, a left-hand glove and a right-hand glove. Receiving a prize does not have any value

> apart from the value of the prizes themselves: Jeff does not like winning for its own sake. And he is not the sort of person who experiences regret or disappointment when he might have received something but did not: he only cares about what he actually has. He decides that [Deal 2] is worthwhile – it would be nice to guarantee that he gets something no matter what – so he decides to pay a few cents to allow the first coin to determine both prizes. … Jeff prefers Deal 2 to Deal 1. (Buchak 2013, p.11)

Buchak assumes that the two prizes are independent, so we know that the probability of getting the Elvis stamp as a result of the first coin landing heads is 1/4, and so are the other probabilities of the events on each branches of the decision tree, as shown in the figure of Deal 1. We also know that the prospect of getting nothing does not lower Jeff's happiness, since it is assumed that Jeff does not experience "regret or disappointment" when he could have got something better – it is assumed that "he only cares about what he *actually* has." This means that the value of getting nothing should equal to the status quo, i.e., the utility of getting nothing = 0. u_J(Elvis stamp) > u_J(gloves) > u_J(nothing) = 0; u_J(Elvis stamp and gloves) = u_J(Elvis stamp) + u_J(gloves) so

u_J(Deal 1) = 0.25u_J(Elvis stamp and gloves) + 0.25u_J(Elvis stamp) + 0.25u_J(gloves) + 0.25u_J(nothing)

= 0.5u_J(Elvis stamp) + 0.5u_J(gloves) and

u_J(Deal 2) = 0.5u_J(Elvis stamp) + 0.5u_J(gloves) so

u_J(Deal 1) = u_J(Deal 2).

This is inconsistent with Jeff's actual attitude in the scenario.

What should be highlighted from this scenario, as with the Alice and Bob scenario, is that the structure of gambles matter. Although two gambles involve the same prizes, they can differ in

utility value to Jeff if they differ in global structure and feature. Both Deal 1 and Deal 2 involve some possibility of an awesome prize and of an okay prize. The way the prizes are arranged is different in the two deals. The arrangement in Deal 1 makes it the case that there is a small probability that Jeff will end up with nothing, whereas Deal 2 is arranged so that Jeff gets something no matter what. In this sense, Deal 1 is risky while Deal 2 is not as risky. And notice that the riskiness comes from the way the deals are arranged, or in other words, the structural and global feature of the deals. Jeff has different preferential attitudes toward the two deals. His reason for preferring Deal 2 to Deal 1 is solely based on his concern for the way they are arranged. In other words, Jeff's preference is explained by his concern for the riskiness of Deal 1, which is a global and structural feature of the gamble. Jeff is said to be globally sensitive to risk.

2.3 Allais Paradox

The third counterexample from Buchak that I will discuss is the famous Allais Paradox, proposed by the economist Maurice Allais.[35] To begin with, you are given two lottery tickets, A and B. With A, you have .1 chance of getting $5 million, .89 chance of $1 million, and .01 chance of $0. With B, you get $1 million for sure. The case is shown in the following table:

	If 0 is drawn	If 1-10 is drawn	If 11-99 is drawn
A	$0	$5,000,000	$1,000,000
B	$1,000,000	$1,000,000	$1,000,000

[35] Allais (1953) and Allais (1979).

Although with A you have a small chance of getting a larger prize, you have to take the risk of ending up with nothing. Compared with $1 million for sure, most people choose B over A. Now you are presented with another pair of lotteries, C and D. Lottery C gives you .1 chance of $5 million and .9 chance of $0. Lottery D gives you .11 chance of $1 million and .89 chance of $0:

	If 0 is drawn	If 1-10 is drawn	If 11-99 is drawn
C	$0	$5,000,000	$0
D	$1,000,000	$1,000,000	$0

With C you have a small chance for $5 million and a very large risk of getting $0. With D, the risk reduces a tiny little bit, .89 as compared to .9. And your chance of getting a prize is very slightly increased, from .1 to .11. But with virtually the same chance of getting a prize and virtually the same risk of getting nothing, C offers you a much bigger prize. In this case, most people's choice is C over D.

To see the paradoxical result that DTis accused of, let us calculate the expected utilities of options A, B, C and D:

$$u(\text{A}) = .01\ u(\$0) + .1\ u(\$5\text{m}) + .89\ u(\$1\text{m})$$

$$u(\text{B}) = u(\$1\text{m})$$

$$u(\text{C}) = .9\ u(\$0) + .1\ u(\$5\text{m})$$

$$u(\text{D}) = .11\ u(\$1\text{m}) + .89\ u(\$0)$$

Then calculate the difference between u(A) and u(B), as well as between u(C) and u(D), and we get:

$$u(\text{B}) - u(\text{A}) = .11\ u(\$1\text{m}) - .1\ u(\$5\text{m}) - .01\ u(\$0)$$

$$u(\text{D}) - u(\text{C}) = .11\ u(\$1\text{m}) - .1\ u(\$5\text{m}) - .01\ u(\$0)$$

The difference is the same. Thus, DT says that B is chosen over A if and only if D is chosen over C. But this result goes against most people's choices. People choose B over A but C over D, and these choices seem to make sense. So DT fails to account for a legitimate rational choice.

As Buchak describes the decision maker's rationale, in the case of lotteries A and B, "[h]e reasons that the minimum amount that he stands to win in [B] is a great deal higher than the minimum amount he stands to win in [A]," and in the case of lotteries C and D, "[h]e reasons that the minimum he stands to walk away with is the same either way, and there is not much difference in his chances of winning *some* money." (Buchak 2013, p.12) Such descriptions of the decision maker's rationale for his choices refer primarily to the global and structural features of the lotteries, not the particular outcomes involved in the lotteries. This is what Buchak calls "the global sensitivity to risk."

2.4 Risk Weighted Expected Utility Theory

With her examples, Buchak hopes to show that decision-theoretic explanations are incomplete. DT formalizes instrumental rationality. Desires for results are represented in the agent's utility function. Options are actions that the agent can choose from and are means to obtaining the end results. A probability function represents the agent's beliefs about how likely an option will have certain results. Buchak argues that our action-directed thought is not fully captured by this means-end reasoning. She argues that there must be a third element involved. All her examples are designed to show that in addition to the outcomes and beliefs about probabilities, the agent also considers the global features of a gamble. As Buchak writes,

> The global sensitivity interpretation of preferences says that even holding fixed how much one likes various outcomes, and thus holding fixed the average utility

> value of a gamble, which outcomes constitute the gamble and how they are arranged matter. (p.30)

According to Buchak, global features of a gamble include its minimum value, its maximum value, mean value, and matters about which outcome obtains under which possibility. (pp.28-30) "These are considerations that necessarily depend on the particular gamble in which the outcome is embedded." (p.136) For instance, the minimum value of Deal 1 in the Elvis stamp and Gloves scenario is the status quo and the minimum value of Deal 2 is a pair of gloves. The maximum value of Deal 1 and 2 is the Elvis stamp; the maximum value of Deal 2 is an Elvis stamp and a pair of gloves. And Deal 1 and 2 share the same mean value. In the Alice and Bob scenario, although he likes two Elvis stamps better than just one, he cares if the one Elvis stamp obtains as a result of sure gamble. The structure of a sure gamble has the prize in every possibility, and that is a global feature of the gamble.

Buchak argues that even if we hold fixed the utilities that the agent assigns to the various outcomes and her credence about how likely different choice options lead to the outcomes, we still cannot uniquely decide which option is optimal, contrary to what the standard DT claims. The agent may have different views on the global risk features of the gamble, that is, on "how values are arranged across the possibility space." In other words, when two agents share the same utility and belief functions and thus the same expected utility function, their optimal choice may differ, if, for instance, one agent can tolerate a highly volatile gamble while the other cannot. As Buchak illustrates the point,

> How should an agent trade off the fact that one act will bring about some outcome for sure against the fact that another act has some small probability of bringing about some different outcome that he cares about more? This question will not be

> answered by consulting the probabilities of states or the utilities of outcomes. Two agents could attach the very same values to particular outcomes (various sums of money, say), and they could have the same beliefs about how likely various acts are to result in these outcomes. And yet, one agent might hold that his preferred strategy for achieving his general goal of getting as much money as he can involves taking a gamble that has a small chance of a very high payoff, whereas the other might hold that he can more effectively achieve *this same general goal* by taking a gamble with a high chance of a moderate payoff. Knowing they can only achieve some of their aims, these agents have two different ways to structure the potential realization of them. (Buchak 2013, p.35)

People often show distinct attitudes to these different risk features, and we believe that they are legitimately rational. So if DT cannot accommodate these distinct risk attitudes, then it is inadequate, and that should point to supplementing a third element to the standard theory.

Adding a supplementing third element gives Buchak an alternative theory of instrumental rational choice. She calls it "Risk-Weighted Expected Utility Theory." (*REU* for short.) Buchak's template to calculate the risk-weighted expected utility of a gamble $G = [p_1, x_1; p_2, x_2; \dots ; p_n, x_n]$, where $u(x_1) \leq u(x_2) \leq \dots \leq u(x_n)$, $0 \leq r(p) \leq 1$ for all p, $r(0) = 0$, $r(1) = 1$, and r is non-decreasing, is as follows:

$$REU(G) = u(x_1) + r(\Sigma_{i=2 \dots n}\, p_i)\,(u(x_2) - u(x_1)) + r(\Sigma_{i=3 \dots n}\, p_i)\,(u(x_3) - u(x_2)) + \dots + r(p_n)\,(u(x_n) - u(x_{n-1}))$$

For example, the risk-weighted expected utility of a simple gamble $A = [p, x; (1\text{-}p), y]$ where $u(y) \leq u(x)$ is $REU(A) = u(y) + r(p)\,(u(x) - u(y))$.

To illustrate the idea of this template, let us first calculate the standard expected utility of the simple gamble $A = [p, x; (1\text{-}p), y]$. According to the standard theory, $EU(A) = p\,u(x) + (1\text{-}p)\,u(y)$

= $u(y) + p\,(u(x) - u(y))$. Looking at the transformed equation, the expected utility of A can be seen as the utility of y, in addition to the possibility of getting the difference between the utilities of x and y. As Buchak says, "Taking [x] to be weakly preferred to [y], this is equivalent to taking the minimum utility value the gamble might yield, and adding to it the potential gain above the minimum – the difference between the high value and the low value – weighted by the probability of realizing that gain." (p.48) This simple two-outcome gamble can be generalized to multiple-outcome gambles, so that the expected utility of a gamble, in general, can be seen as the minimum utility value, in addition to the probability of the gain from the next highest utility value, up until the very highest utility value, as follows,

Let $G = [p_1, x_1; p_2, x_2; \ldots ; p_n, x_n]$, where $u(x_1) \leq u(x_2) \leq \ldots \leq u(x_n)$, and $\Sigma_{i=1\ldots n}\, p_i = 1$,

$EU(G) = u(x_1) + (\Sigma_{i=2\ldots n}\, p_i\,(u(x_2) - u(x_1)) + (\Sigma_{i=3\ldots n}\, p_i\,(u(x_3) - u(x_2)) + \ldots + p_n\,(u(x_n) - u(x_{n-1}))$

Notice that the only difference of this template of expected utility from Buchak's template of REU is the risk function $r(\cdot)$ added to the probabilities.

2.5 Risk Function: "the Third Ingredient"

In addition to the standard framework – a subjective utility function representing desires and a subjective probability function representing beliefs, Buchak argues that there is a "third ingredient" of instrumental rationality – a subjective risk function, $r(\cdot)$. Formally, the risk function is defined on the probabilities. It represents how much weight you assign to the probabilities.

Buchak argues that the information about desires and beliefs is not enough pick out best choices, because agents have distinct attitudes about the global risk features of the available choice options. Different decision makers have their own personal risk attitudes. Some people

prioritize options or gambles that have the best worst-case scenario, whereas other people will go for gambles that have the most promising best-case scenario.

Buchak points out that your sensitivity to risk as a global and structural feature of a gamble actually reflects ideas in common sense folk psychology. As she nicely explains,

> Two competing goals we have are to ensure that the worst-case scenario is as good as possible: to choose a course that is certain to go fairly well, and to choose a course that might go very well. And being drawn to each of these goals corresponds to ordinary virtue-words in the English language: a man can be prudent, or he can be venturesome. These two goals are in competition when deciding how much risk to tolerate: to the degree we care about being prudent, we must reject gambles that have some possibility of turning out very well; and to the degree that we care about being venturesome, we must accept gambles that might turn out very poorly. (Buchak 2013, pp. 55-6)

The folk psychology of practical wisdom involves this familiar idea of weighing how much risk to tolerate in exchange for more potential gain. Facing uncertain outcomes, we find ourselves constantly trading off between risk and return.[36] Buchak also points to some psychological findings about this global sensitivity to risk. Findings show that if you attach more weight to the best-case scenario being as good as possible – in other words, if you are risk-seeking, this

[36] Not only in the realm of instrumental rationality, we do similar trade-offs as a part of epistemic rationality. As Buchak points out, people as epistemic agents may have different standards for what they can accept as knowledge and truth. A serious scientist may have set a high level of acceptability, while an ordinary person will not insist on investigating every possible alternative theory before deciding to believe a hypothesis. Buchak thinks that the situation can be modeled similarly as with instrumental rationality. Epistemic agents are taking risks when deciding which hypothesis they accept as truth. Setting a high standard promises less mistakes but carries the risk of missing out more truths. Setting a lower standard of acceptability lets more truths in but also possible errors. Thus, a trade-off has to be made between the two ends. And the point is that it is a trade-off of structures, regardless of what the particular outcomes or truths are. The focus of the decision maker or epistemic agent is global, not local.

attitude is associated with hope, and if you attach more weight to the worst-case scenario being as good as possible – if you are risk-averse, this attitude is associated with fear and anxiety. (p.54) So, the global attitudes toward risk not only find its place in common sense folk psychology, but also in psychological studies.

Buchak believes that the utility functions and probability functions that represent a decision maker's desires and beliefs are exclusively focused on the particular outcomes – whether it is the values brought about by the particular outcomes, or the particular beliefs about how likely those outcomes are to be obtained. Thus, desires and beliefs are exclusively local concerns. Yet, the phenomena of global concerns for trading off risk and return are widespread and genuine. Therefore, there is no way the utility functions and probability functions can capture the global sensitivity to risk.

3. How Individuation and the Holism Strategy Work: Reply to Buchak's Counterexample Argument

In Section 1, I have described Buchak's counterexample scenarios where the standard expected utility theory gets the result wrong. Buchak argues that those scenarios reveal that there are these genuine and widespread phenomena of global sensitivity to how one trades off risk and return, but since the standard apparatus of utility and probability functions only captures local considerations of particular outcomes and beliefs, a third ingredient – risk function – must be supplemented to the standard model.

In what follows, I will explain where and why I disagree with Buchak. I argue that once the standard model is revised through the holism strategy (in Buchak's terms, the individuation strategy), her distinction between local and global considerations of risk breaks down. According to the holism strategy, any global considerations can be loaded into the decision maker's

preference ranking, and thus become local. I will show that the third ingredient is unnecessary, because the same good story can be told by the holism strategy which correctly explains the counterexamples equally well. Furthermore, the risk function approach is more restricted and will turn out to need to appeal to the individuation strategy that Buchak rejects, in order to apply to a wider range of cases.

In sum, here is Buchak's main argument from Section 1.

Buchak's Argument against Decision Theory

(Assumption 1) DT's judgment of the choice-worthiness of gambles is indexed by the expected utilities of the gambles.

(Assumption 2) The expected utility of a gamble is solely determined by the decision maker's utility function and probability function.

(Hypothesis) There are global risk features of a gamble that can affect its choice-worthiness.

(Premise) Global risk features of a gamble cannot be captured in a decision maker's utility function or probability function.

(Conclusion) DT is inadequate, and a supplementing third element – the risk function – is necessary.

I agree with (Assumption 1), (Assumption 2) and (Hypothesis), but I disagree with Buchak on her (Premise). In other words, I agree with Buchak that her counterexamples pose difficulty to traditional DT. But I disagree that we have to introduce a third component to the traditional theory in order to accommodate those rational choices. Instead, I argue that the decision maker's evaluation of how to structure the obtaining of outcomes can be captured by their utility function that reflects their preferential attitudes toward outcomes. If the outcome "winning an Elvis stamp as a result of certainty" means something different for a decision maker from the outcome "winning an Elvis stamp as a result of 50/50 chance," then the difference should be reflected in

the description of the outcomes, and thus in the utility function of the decision maker. This is true regardless of whether this is a general point about *any* outcomes with certainty versus *any* outcomes that result from a 50/50 chance or a local point about Elvis stamps in particular. These two outcomes should be treated as two distinct outcomes that can be compared and ranked differently in the decision maker's preference ranking. This is what the individuation strategy and the holism strategy claim. In Sections 2.1 – 3, I show how Buchak's counterexamples can be accommodated by the holism strategy.

3.1 Alice and Bob Explained

We can specify an outcome set Z for Bob. Let z_1 be the proposition that Bob gets one Elvis stamp risk-free. Let z_2 be that Bob gets two Elvis stamps as a result of 50/50 chance. Let z_0 be that Bob gets no Elvis stamp as a result of 50/50 chance. Also included in Z might be z_3, that Bob gets two Elvis stamps risk-free, and z_4, that Bob gets no Elvis stamp risk-free, etc.[37] Note Bob's preference ranking: z_3 is strongly preferred to z_1, z_1 to z_2, z_2 to z_0, and he is indifferent between z_0 and z_4. Without loss of generality, let $u_B(z_0) = 0$; $2u_B(z_1) > u_B(z_2)$ since more stamp Elvis stamps are better for Bob, but risk is bad. With this preference ranking, we can compare the two choices A = "take one Elvis stamp risk-free," and B = "take a 50/50 gamble at two Elvis stamps or none."

$u_B(A) = 1\ u_B(z_1)$,

$u_B(B) = .5\ u_B(z_2) + .5\ u_B(z_0)$ so

$u_B(z_1) > 0.5u_B(z_2) + 0$.

[37] To put it in terms of possible worlds, let Ω be the set of all possible worlds, and utilize the familiar notion of a proposition as a set of possible worlds where that proposition holds. z_1 is {w∈Ω | it is true at w that Bob gets one Elvis stamp risk-free}. z_2 is {w∈Ω | it is true at w that Bob gets two Elvis stamps as a result of 50/50 chance}. z_0 is {w∈Ω | it is true at w that Bob gets no Elvis stamp as a result of 50/50 chance}. z_3 is {w∈Ω | it is true at w, that Bob gets two Elvis stamps risk-free}. z_4: {w∈Ω | it is true at w, that Bob gets no Elvis stamp risk-free}. P = {<z_1, z_2>, <z_1, z_0>, <z_3, z_1>, <z_1, z_4>, <z_2, z_0>, <z_3, z_2>, <z_2, z_4>, <z_3, z_0>, <z_0, z_4>, <z_3, z_4>}.

Thus, DT correctly yields the result that Bob chooses option A over B.

3.2 Elvis Stamp and Gloves Explained

Move to the Elvis Stamp and Gloves scenario where the stamp-collector Jeff is faced with two options: Deal 1 gives him a rare Elvis stamp if it turns out heads on the first coin flip, nothing if tails on the first coin flip, nothing if heads on the second coin flip, and a usual pair of gloves if tails on the second coin flip. Deal 2 involves just one coin flip and gives him the Elvis stamp if heads while the pair of gloves if tails. Jeff chooses Deal 2 over Deal 1, but DT yields the result that they have the exact same expected utility and Jeff should be indifferent.

But according to the holism strategy, since Jeff is sensitive to how the various outcomes are arranged in the gambles, that is, he cares about where an outcome is in the decision tree, we should thus reflect that in his utility function by redescribing and individuating the various outcomes to reflect an outcome's particular place in the decision tree. In other words, we should find holistic propositions as the basic outcomes by specifying how that outcome is obtained. The following propositions should be holistic to Jeff,

z_1: Jeff gets nothing as a result of the first coin flip landing up tails and gets nothing as a result of the second coin flip landing up heads, via Deal 1.

z_2: Jeff gets nothing as a result of the first coin flip landing up tails and gets a usual pair of gloves as a result of the second coin flip landing up tails, via Deal 1.

z_3: Jeff gets a rare Elvis stamp as a result of the first coin flip landing up heads and gets nothing as a result of the second coin flip landing up heads, via Deal 1.

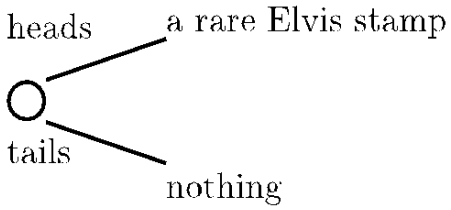
z_4: Jeff gets a rare Elvis stamp as a result of the first coin flip landing up heads and gets a usual pair of gloves as a result of the second coin flip landing up tails, via Deal 1.

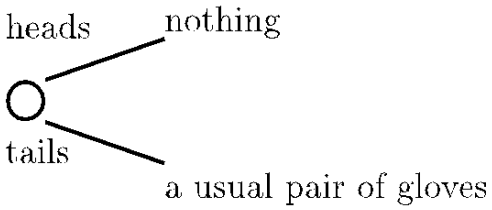
z_5: Jeff gets a usual pair of gloves as a result of a coin flip landing tails and the alternative is a rare Elvis stamp, via Deal 2.

z_6: Jeff gets a rare Elvis stamp as a result of a coin flip landing heads and the alternative is a usual pair of gloves, via Deal 2.

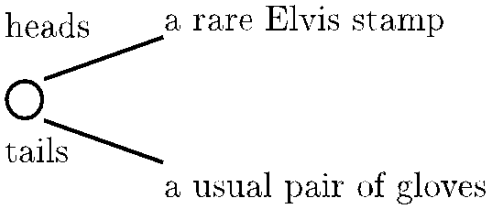
The Elvis Stamp and Gloves

Deal 1





Deal 2



Recalling the story that Buchak told about Jeff's preferences, we can draw a few conclusions about his preference ranking. The process of forming this ranking is a crucial step and we need to go through it carefully. Jeff wants the stamp more than the gloves and either more than nothing. Other risk features being equal, the value of getting both is just equal to the value of each added together. Without loss of generality, let $u_J(z_1) = 0$. Also note that, since the features of Deal 1 make it worse than the features of Deal 2, so $0 < u_J(z_2) < u_J(z_5)$ and $u_J(z_2) < u_J(z_3) < u_J(z_6)$. Finally, $u_J(z_4) = u_J(z_2) + u_J(z_3)$.

$$u_J(\text{Deal 1}) = .25\, u_J(z_1) + .25\, u_J(z_2) + .25\, u_J(z_3) + .25\, u_J(z_4)$$

$$= .25\, [0 + u_J(z_2) + u_J(z_3) + (u_J(z_2) + u_J(z_3))]$$

$$= .25\, [2\, u_J(z_2) + 2\, u_J(z_3)]$$

$= .5\ u_J(z_2) + .5\ u_J(z_3)$

$u_J(\text{Deal 2}) = .5\ u_J(z_5) + .5\ u_J(z_6)$

$u_J(\text{Deal 2}) - u_J(\text{Deal 1}) = .5\ u_J(z_5) + .5\ u_J(z_6) - [.5\ u_J(z_2) + .5\ u_J(z_3)]$

$= .5\ [(u_J(z_5) - u_J(z_2)) + (u_J(z_6) - u_J(z_3))]$

> 0 since $(u_J(z_5) > u_J(z_2)$ and $u_J(z_6) > u_J(z_3)$.

Therefore, we conclude that $u(\text{Deal 2}) - u(\text{Deal 1}) > 0$, and that Jeff should indeed choose Deal 2 over Deal 1 given his personal preferences and risk attitudes.

3.3 The Allais Paradox Explained

Now let us look at how the holism strategy can accommodate the Allais paradox. Recall that the result of standard DT says that lottery B is chosen over A if and only if D is chosen over C. But this result goes against most people's choices, which seem plausibly rational. People choose B over A but C over D. The holism strategy needs to tell a good story where it is rational for people to choose B over A, and at the same time, C over D.

	If 0 is drawn	If 1-10 is drawn	If 11-99 is drawn
A	\$0	\$5,000,000	\$1,000,000
B	\$1,000,000	\$1,000,000	\$1,000,000

	If 0 is drawn	If 1-10 is drawn	If 11-99 is drawn
C	\$0	\$5,000,000	\$0
D	\$1,000,000	\$1,000,000	\$0

According to the holism strategy, if a decision maker cares differently for two results then they should be properly distinguished. Notice, for example, that the \$0 in the first choice - between A and B - means something different for some agents compared with the \$0 in the

second choice - between C and D. Although they are equal in numerical monetary value, the $0 in option A means you missed $1m *for certain* by not choosing B. $0 is a likely outcome no matter what you do in the second choice. Appeal to this distinction among cases - by itself - suffices to explain the popular evaluation of the choices offered. Let $u(\$0!)$ be the utility of ending up with $0 in choice A. In the first decision, B is a reasonable choice over A when $u(B) > u(A)$, which is to say

$$u(\$1M) > .01\ u(\$0!) + .1\ u(\$5M) + .89\ u(\$1M)$$

$$.11\ u(\$1M) - .1\ u(\$5M) > .01\ u(\$0!).$$

In the second decision, C is a reasonable choice over D just in case $u(C) > u(D)$, which implies

$$.1\ u(\$5M) + .9\ u(\$0) > .11\ u(\$1M) + .89\ u(\$0)$$

$$.01\ u(\$0) > .11\ u(\$1M) - .1\ u(\$5M).$$

There is trouble, of course, if $u(\$0) = u(\$0!)$, but the structure of the first deal suggests $u(\$0) > u(\$0!)$ here, which is precisely what is required to make sense of the case here.

4. More Arguments for the Holism Strategy

Buchak's examples are supposed to show that DT is inadequate to capture a decision maker's choice of a gamble when their rationale involves global attitudes toward the risk. This argument leads to Buchak's risk weighted expected utility theory which tries to overcome this inadequacy by supplementing standard expected utility theory with a risk function. I agree with Buchak that there is indeed a genuine phenomenon of global sensitivity to risk which standard versions of DT cannot adequately capture. But I believe that DT can be augmented and defended by the holism strategy, and so that there is no need for a "third ingredient" risk function. In Section 3, we saw how the holism strategy properly individuates the basic outcomes so that they induce a fine-grained preference ranking that not only successfully explains the rational choices in the

scenarios but also captures the right rationale of global sensitivity to risk that is behind those choices. I conclude that standard DT augmented by the holism strategy handles Buchak's cases as well as risk weighted expected utility theory: everything that can be explained by the latter can be explained by the former.

In Section 4.1, I will develop this conclusion further and argue that the holism strategy fares better than the risk weighted expected utility theory, since the story told by the holism strategy can capture the accurate rationale involving not only the global risk attitudes but also any other considerations behind a choice that may or may not be characterized as risk attitudes. There is nothing special about risk attitudes that should separate themselves from all other preferential attitudes and considerations. In Section 4.2, I will respond to objections from Buchak. I will show that maintaining the basic outcomes to be holistic is faithful to decision makers' actual preferences, and that being a realist about the preferences has an advantage. Finally, in Section 4.3, I will argue that in order to properly deal with local sensitivity to risk, Buchak needs to appeal to the individuation and the holism strategy anyway. That gives us further reason to endorse the holism strategy.

4.1 Loading the Consequences: The Global is the Local

The holism-strategy-augmented expected utility theory handles risk attitudes at least as well as the risk weighted expected utility theory. We can see this point most clearly if we have a general way to translate between the two devises. Recall Buchak's template to calculate the risk-weighted expected utility of a gamble of a general form $G = [p_1, x_1; p_2, x_2; \dots ; p_n, x_n]$, where $u(x_1) \leq u(x_2) \leq \dots \leq u(x_n)$, $0 \leq r(p) \leq 1$ for all p, $r(0) = 0$, $r(1) = 1$, and r is non-decreasing, is as follows:

$$REU(G) = u(x_1) + r(\Sigma_{i=2 \dots n}\, p_i)\,(u(x_2) - u(x_1)) + r(\Sigma_{i=3 \dots n}\, p_i)\,(u(x_3) - u(x_2)) + \dots + r(p_n)\,(u(x_n) - u(x_{n-1}))$$

To illustrate how we can translate between the risk weighted expected utility and the expected utility of a gamble, use a simple gamble A = [p, x; (1-p), y], and assume that $u(y) \leq u(x)$. By Buchak's template,

$$REU(A) = u(y) + r(p)\,(u(x) - u(y)).$$

Grouping by the utility values, this is equivalent to

$$REU(A) = r(p)\,u(x) + (1 - r(p))\,u(y).$$

Compare this with the expected utility of A with holistic outcomes.

$$EU(A) = p\,u(x{\sim}A) + (1\text{-}p)\,u(y{\sim}A).$$

A general schema to have REU(A) and EU(A) say the same thing is to let the same utility term be equal. That is, let

$r(p)\,u(x) = p\,u(x{\sim}A)$, and let

$(1 - r(p))\,u(y) = (1\text{-}p)\,u(y{\sim}A)$.

Solving for the holistic terms, we have

$$u(x{\sim}A) = (r(p) / p)\,u(x)$$

$$u(y{\sim}A) = [(1 - r(p)) / (1 - p)]\,u(y).$$

This is what you get when you transform Buchak's risk function - one that operates over increments of improvement - to a risk function that applies to outcomes. Note that the expressions on the left of both equations still just involve attitudes/utilities and probabilities/credences. When $r(p) < p$, as we can see from the equations, $u(x{\sim}A) < u(x)$ and $u(y{\sim}A) < u(y)$.

The example above shows how to translate from REU to EU formula with gambles that have only two possible outcomes. Generally, we can transform the formula as follows, for a gamble of

a general form $G = [p_1, x_1; p_2, x_2; \ldots ; p_n, x_n]$, where $\Sigma_{i=1 \ldots n}\, p_i = 1$, $u(x_1) \leq u(x_2) \leq \ldots \leq u(x_n)$, $0 \leq r(p) \leq 1$ for all p, $r(0) = 0$, $r(1) = 1$, and r is non-decreasing,

$$REU(G) = u(x_1) + r(\Sigma_{i=2 \ldots n}\, p_i)\,(u(x_2) - u(x_1)) + r(\Sigma_{i=3 \ldots n}\, p_i)\,(u(x_3) - u(x_2)) + \ldots + r(p_n)\,(u(x_n) - u(x_{n-1}))$$

$$= [1 - r(\Sigma_{i=2 \ldots n}\, p_i)]\, u(x_1) + [r(\Sigma_{i=2 \ldots n}\, p_i) - r(\Sigma_{i=3 \ldots n}\, p_i)]\, u(x_2) + [r(\Sigma_{i=3 \ldots n}\, p_i) - r(\Sigma_{i=4 \ldots n}\, p_i)]\, u(x_3) + \ldots + [r(p_{n-1} + p_n) - r(p_n)]\, u(x_{n-1}) + r(p_n)\, u(x_n)$$

$$= p_1([1 - r(\Sigma_{i=2 \ldots n}\, p_i)]/p_1)\, u(x_1) + p_2([r(\Sigma_{i=2 \ldots n}\, p_i) - r(\Sigma_{i=3 \ldots n}\, p_i)]/p_2)\, u(x_2) + p_3([r(\Sigma_{i=3 \ldots n}\, p_i) - r(\Sigma_{i=4 \ldots n}\, p_i)]/p_3)\, u(x_3) + \ldots + p_{n-1}([r(p_{n-1} + p_n) - r(p_n)]/p_{n-1})\, u(x_{n-1}) + p_n(r(p_n)/p_n)\, u(x_n).$$[38]

Now we can see that holistic utilities can be restated in terms of regular utilities and risk functions:

$$u(x_1 \sim G) = [1 - r(\Sigma_{i=2 \ldots n}\, p_i)]/p_1\, u(x_1),$$

$$u(x_2 \sim G) = [r(\Sigma_{i=2 \ldots n}\, p_i) - r(\Sigma_{i=3 \ldots n}\, p_i)]/p_2\, u(x_2),$$

$$\ldots$$

$$u(x_{n-1} \sim G) = [r(p_{n-1} + p_n) - r(p_n)]/p_{n-1}\, u(x_{n-1}),$$

$$u(x_n \sim G) = r(p_n)/p_n\, u(x_n).$$

This shows that the holism strategy augmented expected utility theory is at least as effective as the risk weighted expected utility theory. In Section 3, we saw that the holism strategy successfully accounts for global risk attitudes and tells a story that is as satisfactory as the risk weighted theory. Now we have seen that more generally, anything that can be said with the risk weighted theory can be said with the holism strategy, because non-holistic outcome utility terms can be translated into holistic outcome utility terms.

[38] Thanks to James Hawthorne for the suggestion of this general transformation.

We also notice that nothing in the holism strategy confines it to considerations of risk only. The holism strategy requires that the basic outcomes must be holistic in the sense that any more particular, further fine-grained way of realizing an outcome would not alter its utility value to the decision maker. This claim is extremely broad and widely applicable, since it does not require what kinds of considerations you have, whether they are your personal values, motives and objectives in a certain situation, or specific social norms in a certain context.

There is nothing special about risk attitudes that single out, on the holism view. Risk is just one of the many features of an outcome, such as the color of a pen, the tastiness of an apple, the volatility of a trade etc. These features may or may not make a different to your happiness if the outcome obtains. There is no reason to make risk an exception. This implication is backed by our previous argument that the holism strategy offers an equally satisfactory explanation of the risk attitudes in the counterexample scenarios, and that it is unnecessary to appeal to a third ingredient other than preference and credence. Every preferential attitude, including risk attitude, can be loaded into the consequences (viz. the outcomes).

Therefore, there is no fundamental distinction between the local and global considerations of risk. According to the holism strategy, any global considerations can be loaded into the decision maker's preference ranking, and thus become local. And as we will see in the end, Buchak's risk analysis will turn out to have to appeal to the individuation strategy that she rejects, in order to deal with local risk attitudes in a wide range of cases.

4.2 Global Individuation: Objection and Response

Buchak anticipated a similar line of argument, which she calls "global individuation." She has a formal way of translating between her risk function analysis and the expected utility analysis supplemented with global individuation. But she rejects the global individuation strategy because

of a few problems she sees, especially what she calls "the problem of proliferation." She thinks that the global individuation strategy gives rise to *too many* utility functions and probability functions. However, as she eventually admits, the proliferation problem is only an issue if preferences and credences are matters of 'modeler's choice'. If preferences and credences are real attributes people actually have then they aren't too many - just an epistemic issue of determining which of the candidates is the correct one. Recourse to a wide variety of evidence might be required to determine that.

In chapter 4 of her book *Risk and Rationality*, Buchak considers and rejects the individuation strategy as a solution to the problems of standard DT. Her discussion of the individuation strategy does not go as far and explicit as the holism strategy; instead, the view she discusses is developed mainly from John Broom (1991) and Philip Pettit (1991). At the end of the chapter, however, what she calls the "global individuation" strategy looks a lot like the holism strategy. It shares a common claim with the holism view that outcomes should not just be viewed in isolation, but should rather be viewed in the context of the gambles they figure in. She then raises three problems with the global individuation strategy, which can be seen as potential objection to the holism strategy. In this subsection, I will respond to those problems.

The first problem, Buchak says, is that "outcomes that we might have initially wanted to count as equivalent will be differentiated." (p.137) But this is exactly the holism point. The holism strategy urges that we should not have wanted to count as equivalent outcomes that figure differently to a decision maker's evaluations. If a decision maker cares differently about some outcome, though it may seem like the same outcome, we should count it as different in the different ways that it affects the decision maker's welfare.

Buchak is aware of this holism point. As she writes on behalf of the standard expected utility theorists accompanied by global individuation, agents "can claim that the state of affairs in which one takes a deal one disprefers and gets an Elvis stamp is worse than the state of affairs in which one take a deal one prefers and gets an Elvis stamp." (p.137) But Buchak thinks that "there is an odd double counting going on." As she points out, the way DT is supposed to represent instrumental rational choice is that we begin with a decision maker's preferences about outcomes, things they desire, and then use that information to determine how choice-worthy various gambles are. It seems odd, she argues, to include the decision maker's evaluation of how much they like the gamble in the forming of their preferences. She also thinks that including the evaluation of a gamble would make the gamble valuable in itself, when it should not be but rather instrumentally valuable – namely, "it does a better job of getting you the things you do value in themselves." (p.138).

An act of choice, a gamble, or a lottery ticket is indeed a means to an end, and not an end in itself. But that does not make it necessarily instrumentally valuable. In fact, we often care a lot about how we reach some end, and we care intrinsically the way in which it is obtained. Someone may value a car of a particular make and model in itself. But they may also value a lot about how they got that car. Purchasing the car and stealing the car makes a great difference in the utility of the outcome. Someone may value in itself the honor of winning a world champion, but do not care very much if they achieved it by cheating. The point is that the particular way in which an outcome is obtained can and often is valued by the decision maker. So it is natural and reasonable to include it in forming their preference ranking of outcomes that are obtained through particular ways that they care about. Features of the gamble itself can instantiate those

path-dependent properties. If, for example, you like to think that you *earned* certain outcomes then you might not like those outcomes as much when they result from low-probability gambles.

"The second problem with global individuation," Buchak says, "is that what an agent chooses in one decision scenario will tell us little or nothing about what he will or should choose in any other scenario… And there will be nothing to constrain his preferences … as the result of (even slightly) different gambles." (p.138) She also shares with David Velleman (2000, p.163) the critique that "the more we allow that the outcomes of a decision-maker's various choices are different from each other, the less intelligible the decision-maker becomes." And,

> if decision theory does not rule out that an Elvis stamp might have a different value in every context in which it appears, then there is no pressure from prescriptive decision theory to take a unified stance on the stamp's value. And things seem even worse for the interpretive theorist, since facts he learns about the agent in one context will tell him nothing about the agent in other contexts. (p.138)

My response is to say that it is true that the same Elvis stamp may have different utility values in different contexts, but it does not follow that we cannot learn anything intelligible from a choice in a context. For example, Jeff chooses the less risky Deal 2 over Deal 1 where he might get nothing. This tells us something about his risk-averse attitude. In some parallel context where he needs to decide between a particular insurance that guarantees the same minimum – a utility value that is similar to the value of the pair of gloves, compared with a gamble whose minimum is null, a prescriptive decision theorist will be able to predict that Jeff will choose the former insurance deal. Properly distinguishing outcomes that are in fact different in utility values does not make a decision maker less intelligible. The fact that close outcomes aren't identical because

of of distinguishing path dependent features doesn't mean we can't learn anything from similar situations. As a normative point, recognizing that evaluations of outcomes are path-dependent does not make DT invalid. It only fine-grains the outcomes and makes decision-theoretic evaluations of them more accurate. An outcome, such as an apple, is not evaluated in isolation, but rather in context. Recognizing the fact that some agent prefers not to take the last apple gives a modeler more information about the agent's preferences. A decision modeler can gain information about a decision maker's relevant desires and beliefs given their choice, and that information will be much more accurate. The modeler can then use the information in a similar context. Making the outcomes context-specific does make it more difficult for a modeler to transfer the information, but that should be the reliable and desirable thing to expect. We should not expect an easy transfer of information if that information is not accurate or reliable in the first place.

The third problem that Buchak sees with the global individuation strategy is what she calls the "problem of proliferation." She develops the strategy formally and shows an equivalence theorem between her risk weighted expected utility theory and the standard expected utility theory whose gamble options are based on "relabeled" outcomes. She shows that $REU(A) \geq REU(B)$ if and only if $EU(A^*) \geq EU(B^*)$. The problem with the formal apparatus, however, is that the choice is representable with more than one, and in fact, proliferative utility functions. She thinks that there is a lot of freedom in setting the utility function for the decision maker.

If DT is a purely formal apparatus, Buchak is correct. But it is important to note that the holism stories of rational decisions are not merely formal stories. A holism story is built on actual preferential attitudes that a decision maker actually has. A preference ranking over outcomes, on which a utility function is defined, is a ranking for a particular decision maker. So

keeping it real leaves no uncertainty at all in settling a utility function for a decision maker's preferences. Given an actual individual's particular values, goals and objectives, we can avoid Buchak's worry of proliferative utility functions. *Ad hoc* utility functions ruled out by principle. In fact, Buchak briefly acknowledges this advantage of being realist about preferences. She writes,

> The non-constructivist realist EU theorist has a stronger leg to stand on: she can claim that there is a privileged utility function (of new outcomes) and that it is a function relative to which the agent is an EU maximizer, rather than a maximizer of some other quantity. (p.145)

A modeler might worry that she cannot fully access a decision maker's particular values and preferences, and thus cannot avoid the proliferation problem because when preferences are uncertain, they can be represented by multiple utility functions. But the particularity of a certain decision maker's holistic preferences results in a definite way of carving out the space of possible worlds. The normative question is not whether a modeler can definitely describe one's holistic preferences.

4.3 The Last Apple

In Buchak's counter-argument to DT, she makes a key distinction between local and global sensitivity to risk, and uses the fact that people have globally sensitive risk attitudes toward gambles in addition to their local preferential attitudes about outcomes to show that a third ingredient of risk function should be supplemented to traditional DT. Again, local attitudes about an outcome are attitudes about how desirable the outcome is in itself. For example, how much utility value an apple or a vacation presents to you. Global attitudes are ones that are not

confined to the outcomes themselves, but rather concerned with the overall structural characteristics of a gamble, including its minimum (or the worst-case scenario), maximum (the best-case scenario), and how widespread it is between the two. A decision maker may have an attitude just about having the worst-case scenario above a certain threshold, which is a global risk attitude.

I agree with Buchak that people do have such global risk attitudes, but I think that they can all be loaded into attitudes of outcomes, into the decision maker's preference ranking. Therefore, every global risk attitude is in fact also a local attitude, because the global attitude is built into an attitude of outcome.

Since there is nothing special about risk attitudes to separate them from any other preferential attitudes, the holism strategy will be more flexible and widely applicable than the risk weighted expected utility theory. In order for the risk analysis to apply to other cases, such as Amartya Sen's famous Apple Case, it will have to appeal to finely individuated outcomes after all. This demonstrates that the holism strategy has an advantage over the risk analysis on its explanatory power.

Sen makes an important observation that points to the holism strategy, which is that an outcome is rarely just that outcome by itself, instead, what it means in terms of utility value to a decision maker is almost always affected by things context-sensitive. For example, an apple is not just an apple in the following scenario.

The Last Apple

> Suppose the person faces a choice at a dinner table between having the last remaining apple in the fruit basket (q) and having nothing instead (p), forgoing the nice-looking apple. She decides to behave decently and picks nothing (p),

> rather than the one apple (q). If, instead, the basket had contained two apples, and she had encountered the choice between having nothing (p), having one nice apple (q) and having another nice one (r), she could reasonably enough choose one (q), without violating any rule of good behavior. (Sen 2002, p.129)

Initially we seem to be looking at two outcomes,

p: I have nothing.

q: I have the last apple.

And the choice to be made is whether to take the last apple (q) or not (p). In order to apply the risk analysis to this case, we can characterize it as involving the risk that it may or may not be socially okay to take that apple, depending on, say, there may or may not be someone else at the table who also would like to have the apple. Let p be the probability that it socially okay to take that apple. $0 < \alpha < 1$. So here is the gamble S: [p, 1-α; q, α].

If we want to apply the risk analysis, the first thing we must have is a ranking of the utility of p and q. But clearly in this case, whether the decision maker prefers having nothing the having that apple depends on whether taking the apple will be socially okay or not. She would love to have the apple if she would not turn out to be indecent, but would rather have nothing if there is a clear social norm to not take the last dessert on the table. Without a clear presence of social norm, we the modeler or the decision maker cannot rank $u(p)$ and $u(q)$. Thus, the risk analysis simply cannot be applied. It has to fine grain the outcomes to get a ranking before it gets off the ground. It has to appeal to the holism point that the outcome "having an apple" should be distinguished as two holistic outcomes. One is having an apple when doing so does not violate any social norm, and the other is having an apple when it is the last remaining one. The decision

maker clearly prefers the former to the latter. More explicitly, let *q'* and *q''* designate these two holistic outcomes,

q': I have a nice apple when doing so would be indecent.

q'': I have a nice apple when doing so does not violate any social norm.

We know that $u(q') < 0 < u(q'')$. And let p be the probability of z. Give the context of the scenario and the fact that the apple is the last remaining one, a common sense judgment is that p should be extremely low: $p \approx 0$.

Now we calculate the expected utility of the gamble $A = [\alpha, q''; 1-\alpha, q']$,

$$u(A) = \alpha\, u(q'') + (1-\alpha)\, u(q') = u(q') + \alpha\,(u(q'') - u(q')) \approx u(q') < 0$$

Standard expected utility accompanied by the holism strategy yields the correct result that the decision maker would not choose to take the last apple (i.e. to take the gamble A), and it adequately explains why the choice.

5. Conclusion

DT has been interpreted both as a descriptive account that predicts choice behavior and as a normative and prescriptive account of explanation. The counterexamples that philosophers discuss mostly are counterexamples to DT as a normative and prescriptive account. DT purports to capture what it is to choose rationally. In Chapter 2, I have argued that DT succeeds as a norm/logic of rational choice, and defended my normative interpretation from the Sen critique that DT makes mistaken judgements about some choices that are pre-theoretically rational.

Philosophers and decision theorists have also focused on counterexample scenarios where an agent's choices seem to make good sense - pre-theoretically - but they are judged to be irrational according to DT. In this Chapter, I have considered Buchak's counterexamples where the

decision makers' risk-averse attitudes are rational yet it seems that they cannot be accommodated by DT. Buchak amends the standard formalism of expected utility theory (EU, or normative DT) by replacing the probability function with a risk function on the probabilities. By introducing a third element - the risk function that represent risk attitudes about tradeoffs - to the standard formalism of utilities and credences, Buchak proves a representation theorem from a weaker set of axioms on preference. Her risk-weighted expected utility theory (REU) is then shown to accommodate her counterexamples.

Buchak's counterexample scenarios nicely illustrate the fact that people do not only evaluate outcome prizes in isolation, but are also concerned with risk features associated with a certain gamble. Her distinction between local and global sensitivity to risk nicely differentiates different cases of risk. One is where the utility of some outcome is affected by the fact that the outcome is a risky outcome - the case of local sensitivity to risk. The global case is where the "location" of some outcome in a gamble matters in terms of agent's utility evaluation of it. An outcome that may look the same may be evaluated differently as appearing in different places in a gamble. While Buchak thinks that such context-dependent risk attitudes cannot be captured by standard DT, I have argued that they can, because both local and global sensitivity to risk are cases of path-dependency.

I have applied my holistic account of DT to Buchak's counterexamples, showing that they are cases of path-dependency. The holism story is at least as good as the REU story for such cases. With the Allais Paradox, for example, the evaluation of an outcome 'getting $0' is different when you could have got $1 million instead. What makes this outcome risky is the fact that it occurs in a gamble where the outcome is realized along a particular path and the agent cares about the features of that path. I have shown that any path-dependency should and can be loaded

into the description of outcomes as sets of possible worlds. The REU formula can be transformed to a EU formula that restates the utilities $u(x)$ in terms of holistic terms - $u(x$ via some gamble) - which makes my holism account consistent with the REU theory.

When any risk dependency has been figured into the basic outcomes, Buchak's scenarios are no longer counterexamples to EU (normative DT). Buchak's REU analysis begins with a ranking of utilities of outcomes, but such a preference ranking is vulnerable to Sen-style critique if those outcomes are not holistic propositions. In response to Buchak's argument that only REU, but not EU, can handle global considerations of risk, I have also argued that there is no fundamental distinction between the local and global because any global considerations can be loaded into a decision maker's preference ranking over holistic outcomes, and thus become local.
Buchak considers but rejects what she calls the "global individuation" strategy which resembles my holism view. She objects that "there is an odd double counting" to include one's evaluation of a gamble in and as a part of one's evaluation of basic outcomes. In response, I have argued that as a matter of fact, people's preferences of an outcome seldom depend solely on the outcome in itself, but rather are path-dependent. We often care a lot about certain features of how we reach some end result, and not just the end result itself. Therefore, it makes sense for a decision theorist to sensitively model what people care about exactly. Although it is true that building path-dependent features into the basic outcomes would make any descriptive use of DT unlikely (see Chapter 2), we can still learn something of the decision maker, for example, that she is risk-averse and that normatively speaking, she would make an irrational choice if she does not take into account her risk-averse attitudes. As to what Buchak calls the "problem of proliferation" of utility functions, I have replied by identifying the utilities with particular decision makers. A

holism story is built on actual preferential attitudes, and an actual person's particular values, motivation and objectives provide DT with a determinate starting point.

Chapter 4

The Game Theoretic Norm: What and Why

1. Introduction

Imagine the following bargaining situation. Player I and Player II have $4 to divide. Player I can just impose an equal split on Player II or she can offer a division where she keeps $3 and gives Player II $1. If Player I plays *Offer $1*, Player II can accept that division and get $1 (allowing Player II to keep $3), or he can reject it, in which case each Player gets $0. This situation is summarized in the following game tree. This is a partial version of what is known as an *Ultimatum Game*.

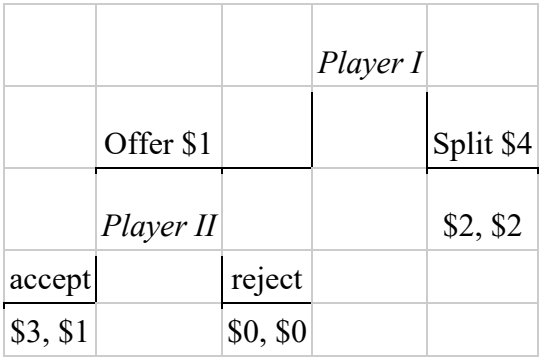


Figure 3.1

Assume for the moment that both players prefer more money to less in this situation, so that the monetary amounts represent the players' utilities. Standard Game Theory (GT) offers two "solutions" to this game: one is when Player I chooses *Offer $1* and Player II *Accept*; the other is when Player I chooses *Split $4*.[39]

Despite GT's solutions, it could easily turn out that Player I makes an unequal (unfair?) offer and Player II chooses to reject it. When GT is challenged by this *prima facie* disconfirming evidence, several questions can be asked. First, is the behavior a genuine instance of *rational* choice? GT is

[39] The former is usually seen as more likely. I will discuss this below.

supposed to predict or advise rational agents. When Player I offers $1 to Player II and he rejects it, they both get $0. GT will fail to predict this result if Player II really prefers $1 to $0 in this circumstance because Player II's choice is not rational. Players will want to change their choices to rational ones if they realize they can get better payoffs. Further, does the datum really challenge the theory? GT may also fail to make the correct prediction if it is wrongly applied. The problem would lie in the application of a theory that could be impeccable. It supposes that players are instrumentally rational, but is it committed to the view that monetary amounts adequately represent player utilities? If not, does it place *any* constraints on player preferences?

GT has largely been developed as a tool for explaining and predicting choice behavior; these are descriptive projects. Even though there are multiple players in a game situation, however, their decisions can still be seen from a first person, normative perspective. When you play a game, you are still facing a decision problem: which strategy to choose. You start with a set of outcomes that interest you and a set of beliefs about how those outcomes can be reached. Information about other players in the game will influence your belief set. From such a perspective, classical GT retains the basic structure of the RCT model, and so is a special case of DT. In the first paper of this dissertation, I defended DT as a logic of evaluation and choice, arguing that DT captures the notion of instrumental rationality, which is about consistently deriving preferences over choice actions from preferences over outcomes and beliefs. Given the connection between DT and GT, we can distinguish normative interpretations of GT which treat it as the standard and norm of instrumental rationality for strategic choices -- a criterion we can use to test whether certain situations are handled rationally.

Among the ways that GT can go wrong, some are ways it fails descriptively, that is, when its prediction of what a player will choose does not match people's actual choices. Experiments can put game-theoretic judgments to test and disagreements between its predictions and actual behavior can refute descriptive interpretations of the theory. Some of those experimental failures can be accounted for by the irrationality of people. But in other cases, behavior seems to disagree with GT but people still seem to have behaved rationally. In these cases, the normative adequacy of the theory is challenged. In this paper, I will defend GT as an unimpeachable normative account of rational choice in strategic situations, and so that the normative challenges can be met. I will focus on a set of cases where people make what seem to be reasonable choices but GTy does not seem to recognize them as such. I will argue that these cases are genuinely rational choices and that they do not refute GT because, appearances to the contrary, it can recognize them as such. It is not the fault of GT if it is incorrectly applied.

2. Game Theory as An Extension of Rational Choice Theory

GT, at least in its classical version, is an extension of decision theory (DT). DT is concerned with individual choice behavior while GT is concerned with situations where multiple agents make choices in strategic (interactive) environments. DT studies how an individual person makes decisions. It assumes that individuals are rational in the sense that they choose their best options based on their preferences over outcomes and on how they believe their options will lead to desirable outcomes. For a simple example of individual decision making, suppose you had to decide whether to have fried chicken or a bowl of salad for lunch. Suppose further that the values relevant to your choice are health and flavor, that you value having healthy food more than

having tasty food, and that you believe, while the chicken is somewhat tastier, the salad is much healthier. DT says that you would - and should - choose to get a bowl of salad for lunch.

Here we have a model of individual rational choice. We begin with a set of propositions that, based on your values, you have some interest in. Following the argument of Chapter 2, some of those propositions can be understood as basic outcomes and others might be best understood as gambles over basic outcomes. All of these propositions are evaluable in a preference ranking which is numerically represented by an expected utility function (unique up to linear transformation). Further, you have an action set that contains the propositions that you believe you can bring about, things like 'I buy a bowl of salad for lunch at this restaurant today.' Information from nature, the external environment, etc. allow you to form beliefs that link each action with a gamble and so ultimately a set of outcomes. You look, for example, at the menu of a restaurant and see salad offered. You form a belief that if you choose to buy salad here, you are likely to get the desired outcome of having salad for lunch. You should (and likely will) choose the action that is associated with the highest expected utility, i.e., the most preferred gamble.

GT studies decision making in strategic settings, those where the expected utility of each individual is a function of what each chooses to do. Two people, for example, might choose whether to participate in a project. Each person's success - in terms of their own preferences - depends not only on their own actions, but on how the other person chooses to engage. In the language of GT, the individual parties are called "players," a player's choice options are called "strategies," and a "payoff" is an expected utility value that represents a player's evaluation of the desirability of the outcome/gamble that results from the combinations of strategies the players choose.

Strategies in GT are a little more complicated than *actions* in DT. Games can involve multiple 'turns' so an action 'plan' is required. A strategy for a player is a plan of action for each opportunity they have to act in the game. This is required for strategic analysis because a player's choices will often depend on what they think another player would do at a possible choice node. In the game situation in Figure 3.1 from the beginning of this Chapter, what makes sense for Player I to do depends on what Player II would do at a choice point he might never reach. Player II's strategy will spell out what he would do if he reached his decision node after Player I plays Offer $1, even if Player I were to choose her strategy Split $4.

Consider a simple bargaining situation where Alice will decide to split $4 and share it with Bob. The offer can only be made in dollar bills, which means that Alice can choose either to offer $1, $2, or $3 to Bob. When offered, Bob will decide either to accept or reject the offer. Accepting will result in Bob getting the amount offered and Alice getting the remaining amount. Rejecting will result in both people getting nothing. This is a simplified version of what is known as an Ultimatum Game, which will be the main illustration for the arguments in this paper. (Figure 3.1 shows an even more limited *partial* Ultimatum Game.) To describe the situation as a game, we need to know the players, players' strategies, the expected payoffs to outcomes/gambles, and the information structure of the game (roughly, the how it is played). Our simplified Ultimatum game is a two-player game, where Alice is an allocator and Bob is a recipient. A strategy is an action plan consisting of the actions a player can choose to take. Alice's strategies are

Offer $1

Offer $2

Offer $3.

Bob has eight distinct strategies:

Accept $1, Accept $2, Accept $3

Accept $1, Accept $2, Reject $3

Accept $1, Reject $2, Accept $3

Accept $1, Reject $2, Reject $3

Reject $1, Accept $2, Accept $3

Reject $1, Accept $2, Reject $3

Reject $1, Reject $2, Accept $3

Reject $1, Reject $2, Reject $3.

Picking one strategy from Alice and one from Bob gives us a strategy profile: (e.g. <*Offer $1*; *Reject $1, Accept $2, Accept $3*>). A decision problem is distinctively a game problem when each decision maker's utility payoffs depend on the other players' actions. In the formal representation of a game, this interdependence is reflected by defining a player's payoff function over strategy profiles rather than over their own actions or strategies. A player's preference (payoff) function is defined over the set of strategy profiles in the game. For example, Alice will receive $3 if she chooses to offer $1 and Bob accepts it. This is a favorable outcome for Alice, so we assign a higher number to it in Alice's payoff function. It has been common practice among economists to equate the monetary amount with the payoff number.

A game can be represented either with a game tree or a table. Representation with game tree is also known as the extensive (or sequential) form of the game. Our simplified Ultimatum game has a temporal structure that goes in two steps. The Allocator proposes an offer first; the Recipient hears the offer and 's answers with Accept or Reject. The game tree is in Figure 3.2.

				Player I				
	Offer $1			Offer $2			Offer $3	
	Player II			*Player II*			*Player II*	
accept		reject	accept		reject	accept		reject
$3, $1		$0, $0	$2, $2		$0, $0	$1, $3		$0, $0

Figure 3.2

The game begins from the initial decision node, representing Alice's choice. It then branches into three choice nodes, representing Bob's possible choices. In this game, Bob would know which branch he is at by the time he makes a choice. Finally, each terminal node represents the outcome yielded under that strategy profile. As mentioned, monetary amounts in the outcomes have been commonly taken as the players' payoff numbers for the outcomes.

A normal (or strategic) form game matrix is in Figure 3.3. Alice's strategies are the actions to offer $1, $2, or $3. Bob's strategies are action plans that cover every possible choice of action. That is, Bob needs to make a choice of Accept or Reject at each of the three decision points he might end up with. For example, one such strategy of Bob's is *Accept $1, Reject $2, and Accept $1* (ARA). This strategy is shown in the third column. Finally, a cell represents an outcome resulting from Alice's and Bob's strategies. The cell in the first row and the third column, ($3, $1), says that Bob gets $1 and Alice gets $3 given the strategies played pointing to that cell.

	AAA	AAR	ARA	ARR	RAA	RAR	RRA	RRR
Offer $1	$3, $1	$3, $1	$3, $1	$3, $1	$0, $0	$0, $0	$0, $0	$0, $0

Offer $2	$2, $2	$2, $2	$0, $0	$0, $0	$2, $2	$2, $2	$0, $0	$0, $0
Offer $3	$1, $3	$0, $0	$1, $3	$0, $0	$1, $3	$0, $0	$1, $3	$0, $0

Figure 3.3

3. "Solving" Games: Rationalizability and Nash Equilibrium

Classical model of GT (also known as *epistemic GT*) develops solution concepts based on the rational-choice assumption that players are instrumentally rational, i.e., they act to maximize their expected utilities.[40] We say that players in a game have *common knowledge* of rationality when every player knows the game structure and knows that everyone is a utility maximizer, every player knows *that*, every player knows that every payer knows, etc. The GT 'solution concept' of *rationalizability* is the result of treating strategic situations as a straightforward decision problem.[41] In a decision problem, an agent has preference over a set of outcomes, beliefs about available actions and how they are associated with gambles over outcomes, and she makes her choices over actions by choosing one that is associated with the most preferred gamble over outcomes. The reasoning is from the perspective of one player as a decision maker. Uncertainty about what the other players choose is part of the decision maker's conjectural beliefs. Those beliefs, however, must now reflect the fact that each player is instrumentally rational, i.e., each player chooses a *best response* to what she thinks the other players' strategies will be, given that they will all do the same. A player's best response to some opponent strategy is a strategy that gives the player the highest payoff given what the opponent plays. A

[40] Evolutionary GT, a recent alternative model to the classical GT, drops the rationality assumption and assumes instead that successful strategies (in terms of non-utility payoffs) will drive out less successful strategies over time. Much of recent GT research concerns the evolutionary version, but I am interested here in the classical account.

[41] These are standard results in GT. See for example, Osborne & Rubinstein (1994), and Heap & Varoufakis (1995).

rationalizable strategy for a player is one that is consistent with some internally coherent story about each player's beliefs. Usually it is easier to see which strategies are not rationalizable. Take the game in Figure 3.4 as an example.

	C1	C2	C3
R1	100, 99	1, 0	99, 100
R2	1, 0	0, 1	1, 0
R3	99, 100	1, 0	100, 99

Figure 3.4

The Row player ('Row') has three strategies: *R1*; *R2*; *R3*. The Column player ('Col') has strategies: *C1*; *C2*; *C3*. The first number in each cell designates Row's utility, and the second number Col's utility. Given that Col plays *C1*, *R1* is Row's best response. Given *C2*, both *R1* and *R3* are best responses for Row. Given *C3*, *R3* is Row's best response. *R2* is not a rationalizable strategy because it is not a best response to any of Col's strategies. In the language of GT, it is "dominated" by both *R1* and *R3*: for any strategy chosen by Col, each of *R1* and *R3* is a better option than *R2* for Row. *C2* is Col's best strategy against *R2*, but if Row will never play *R2* then *C2* is not rationalizable either since it is not a best response to either *R1* or *R3*. *R1*, *R3*, *C1*, and *C3* are all rationalizable strategies. *R1* is Row's best response to Col's *C1*; Col's *C1* is a best response to Row's *R3*. Row can rationalize playing *R1* if she thinks Col will play *C1* because he thinks she will pay *R3* in response to his playing *C3*, etc. Note that Row here thinks Col has erred in predicting her strategy. Likewise, *C1* is rationalized by Col by his belief that Row will

play *R3* because she thinks he will play *C3* in response to his belief that she will play *R1*, etc. Again, Col here assumes that Row gets him wrong.

Dominated strategies can't be rationalizable since they are never best responses. As above, however, some strategies cease to be best responses to another player's strategies once dominated strategies are eliminated from considerations. Game theorists have shown that a rationalizable strategy is one that survives the *iterated elimination of dominated strategies*. (Osborne & Rubinstein 1994, pp. 53-63). Sometimes each player only has one rationalizable strategy. Take the game in Figure 3.5 as an example.

	C'1	C'2	C'3
R'1	10, 4	1, 5	99, 3
R'2	9, 9	0, 3	98, 2
R'3	1, 99	0, 100	100, 98

Figure 3.5

Notice that *R'2* is never a best response to any of Col's strategies - it is always better to play *R'1* compared with *R'2*. Similarly, when we look at Col's strategies, we find that *C'3* should be eliminated as a dominated strategy. It is dominated by both *C'1* and *C'2*. Having eliminated *R'2* and *C'3*, the game structure can be simplified as in Figure 3.6.

	C'1	C'2
R'1	10, 4	1, 5

R’3	1, 99	0, 100

Figure 3.6

To continue the elimination of dominated strategies, we notice that *R’3* is dominated by *R’1* in this table. Row has only *R’1* left as a rationalizable strategy. Given that Row will play *R’1*, *C’2* maximizes Col's utility. So we can predict that (*R’1*, *C’2*) is the ‘solution’ of this game.

The reasoning that allows us to identify rationalizable strategies is just the reasoning of DT. In general, however, game theorists go beyond this reasoning. A well accepted - indeed, the primary - ‘solution concept’ in GT is *Nash equilibrium*. In a Nash equilibrium, each player plays a strategy that is a best response to the actual strategies played by other players. This is an equilibrium because no player has an incentive to deviate when doing so - they would only risk lowering their payoffs. Nash equilibrium strategy profiles consist of rationalizable strategies for each player. Further, in a Nash equilibrium each player has *true* beliefs about which strategy each player will play - that is, they settle on actual best responses to the strategies others actually choose to play. In that case, we say that the players’ beliefs are *consistently aligned*. Rationalizability does not require that players’ beliefs converge to be consistently aligned. In other words, you can play a rationalizable strategy and have a false belief about what you conjecture the choice behavior of the other player. Nash equilibrium strategies require that every player’s conjectures of one another turn out to be true. (Bicchieri 1993)

The existence of Nash equilibria does not guarantee that the players get the highest payoff in the game model; it merely requires that no one would regret their choice, because no one would get a higher payoff deviating from their part of a Nash equilibrium strategy profile. Appeal to Nash

equilibrium 'solutions' raises an important question about *how* players come to form correct beliefs about each other. There is no generally applicable story about how players converge on a Nash equilibrium. Skyrms (1988) argues that if players start with common prior beliefs about each other then they can correctly figure each other out through a process he calls "deliberational dynamics." Common prior beliefs are not the usual case, however. Further, Skyrms' deliberational dynamics are quite computation intensive, not the sort of procedure most people actually use. When Nash equilibrium play is observed, it is usually the result of unmodelled belief formation processes.

In Figure 3.3, the normal form of the simplified Ultimatum game, (*Offer $2*; *Reject $1, Accept $2, Accept $3*) is a Nash equilibrium *if* both player's utilities are increasing in dollars because the rationalizable story for both Alice and Bob does not involve either of them believing falsely about the other. When Alice believes that Bob will reject her offer of $1 but accept her offer of $2, her best response to this expectation - offer $2 - is a rationalizable strategy for her. Likewise, if Bob thinks Alice will offer $2, *Reject $1, Accept $2, Accept $3* is a best response. And since Alice's and Bob's beliefs about each other for both their rationalizable strategies are correct, they are in a Nash equilibrium.[42]

4. Disagreement Between Theoretical Prediction and Actual Behavior

[42] There are a number of Nash equilibria in the simplified Ultimatum game of Figure 3.3 where utilities are increasing in money: (*Offer $1*; AAA), (*Offer $1*; AAR), (*Offer $1*; ARA), (*Offer $1*; ARR), (*Offer $2*; RAA), (*Offer $2*; RAR), and (*Offer $3*; RRA). Not all of them are equally plausible.

In the simplified Ultimatum game represented in Figures 3.2 and 3.3, Alice (Player I) could reason that Bob (Player II) will accept her (unfair?) offer of $1. But it is also reasonable to think that Bob will reject an unfair split - money isn't everything. Nothing in the game structure or the rationalizability concept prevents Bob from rejecting $1. Though Nash equilibrium concept requires that Alice gets Bob right, it is far from clear how a Nash equilibrium could be reached in this circumstance. Alice may reasonably think that Bob will reject the offer unless it is at least an even split. If Bob says no when she offers $1, Alice's payoff would be $0. If Alice expects this she would be better off offering $2 so that Bob will accept and she gets a payoff of $2. Alice still reasons in a dollar-maximizing way here. Furthermore, Alice does not necessarily fail to see Bob as a maximizer if she can reasonably believe that Bob values fairness more than mere dollar amounts. If Bob places getting a fair split higher in his preference ranking than getting $1 when his partner gets $3, then rejecting Alice's offer of $1 would be utility maximizing for Bob. As long as $u_B(\$0) > u_B(\$1)$ in this circumstance, the set of Nash equilibria shrinks to (*Offer $2*; *Reject $1, Accept $2, Accept $3*), (*Offer $2*; *Reject $1, Accept $2, Reject $3*), and (*Offer $3*; *Reject $1, Reject $2, Accept $3*).

	AAA	AAR	ARA	ARR	RAA	RAR	RRA	RRR
Offer $1	$u_A(\$3)$, $u_B(\$1)$	$u_A(\$3)$, $u_B(\$1)$	$u_A(\$3)$, $u_B(\$1)$	$u_A(\$3)$, $u_B(\$1)$	$u_A(\$0)$, $u_B(\$0)$	$u_A(\$0)$, $u_B(\$0)$	$u_A(\$0)$, $u_B(\$0)$	$u_A(\$0)$, $u_B(\$0)$
Offer $2	$u_A(\$2)$, $u_B(\$2)$	$u_A(\$2)$, $u_B(\$2)$	$u_A(\$0)$, $u_B(\$0)$	$u_A(\$0)$, $u_B(\$0)$	$u_A(\$2)$, $u_B(\$2)$	$u_A(\$2)$, $u_B(\$2)$	$u_A(\$0)$, $u_B(\$0)$	$u_A(\$0)$, $u_B(\$0)$

Offer $3	$u_A(\$1)$, $u_B(\$3)$	$u_A(\$0)$, $u_B(\$0)$	$u_A(\$1)$, $u_B(\$3)$	$u_A(\$0)$, $u_B(\$0)$	$u_A(\$1)$, $u_B(\$3)$	$u_A(\$0)$, $u_B(\$0)$	$u_A(\$1)$, $u_B(\$3)$	$u_A(\$0)$, $u_B(\$0)$

Figure 3.7

If the dollar amounts are taken as payoff numbers for both players, GT predicts that Alice picks the rationalizable strategy which maximizes her payoff. Looking at the game tree, choosing to offer $1 maximizes her payoff. And when Bob is offered $1, which places him in the left tree branch, he will maximize by choosing to accept the offer so that he gets $1 instead of $0. GT then predicts that the game would end up in the left branch.[43] In more general Ultimatum games, GT would predict that the allocators offer as little as possible, and that the recipients take any positive offers.

Game theoretic predictions of the ultimatum game and its variations have been put to test in the last several decades. Actual experiment results deviate significantly from these theoretic predictions. The first set of experiments of a simple ultimatum game concludes that the average offer proposed by the allocators is close to an even split. (Guth, Schmittberger, and Schwarze 1982; Thaler 1988) The average offer is reported to be 37% of the total amount to be split - significantly higher than the minimal amount. And a few positive offers got rejected. The first round was played by subjects who are new to the game. After one week of deliberating on the game, those subjects were asked to play it a second round. The average offer was slightly lower but still 32% which is significantly above the minimum amount. And the rejection rate went

[43] While (*Offer $1*; *Accept $1, Accept $2, Accept $3*) is not the only Nash equilibrium when utilities are increasing in monetary value, it is the only subgame perfect Nash equilibrium, i.e., one where each agent acts like a utility maximizer at each choice node and not just with respect to the game as a whole. Where $u_B(\$0) > u_B(\$1)$, (*Offer $2*; *Reject $1, Accept $2, Accept $3*) is a subgame perfect Nash equilibrium.

higher - 5 out of 21 offers were rejected. These experiments could suggest that people have a taste of non-monetary values such as fairness (or at least for not being taken advantage of unfairly). In other experiments when the allocator is thought to have earned the right to allocate, the average offer decreases and the recipients are more likely to accept a lower offer. (Hoffman, McCabe, Shachat, and Smith 1994, pp. 346–380.) These empirical results show that actual choice behavior Ultimatum games differ from the usual game theoretic predictions. Such experimental results have been taken as evidence that tends to falsify explanatory and predictive GT.

I believe that people actually have more fine-grained preferences than preferences that are (sometimes simplistically treated as purely) driven by monetary considerations. Game theorists often specify payoffs by equating them with the monetary payoffs offered in the game. A game is identified by specifying the players, their strategies, their payoffs, and the information structure of the game. Game theoretic analyses begin only when a game is clearly specified. Being clear about this allows us to see how apparent anomalies are not evidence that falsify GT but are cases where GT is mistakenly applied. Binmore argues that game-theoretic modeling of players must be sensitive to the players' explicit thinking process. Such modeling will inevitably go beyond abstract mathematical representation and cannot be done solely from an armchair. GT with sensitive modeling will then be able to explain anomalies by modifying the model. (Binmore 1987, 1988) Experiments about people's actual preferences do not falsify GT as a normative account of strategic interaction. Instead, they show that the normative account should take fine-grained preferences more seriously. Applications of GT are a priori, formal models; they need to capture relevant features of cases to be applicable, however.

5. Guala's argument against fine-grained outcomes in game theory

In his paper "Has Game Theory Been Refuted?" Francesca Guala argues that descriptive GT has been refuted by empirical anomalies. Although he explicitly states that his argument is confined to descriptive GT, at various points he says things that would lead one to think that it also applies to a normative interpretation of GT. When describing his critique, Guala writes, "This applies to [game] theory both in its normative and in its descriptive version, but given the focus of this paper, let us phrase it in a descriptive idiom." (Guala 2006, p. 252)

In GT, the proximate objects of preference are strategy profiles that include every player's strategy. In our simplified Ultimatum game scenario, for example, Alice prefers strategy profiles where she offers Bob $1 and he accepts the offer to strategy profiles where she offers $1 and Bob rejects such offers. But monetary gain isn't the only thing drives preference judgments; people can derive utility from a variety of sources of value. Alice can take into consideration that Bob dislikes, and so is likely to reject, what he sees as an unfair offer. Alice has the following preference ranking, expressed via a utility function $u_A(\cdot)$.

u_A(I offer $1 and Bob accepts) >

u_A(I offer $2 and Bob accepts) >

u_A(I offer $3 and Bob accepts) >

u_A(I offer $1 and Bob rejects) = u_A(I offer $2 and Bob rejects) = u_A(I offer $3 and Bob rejects).

Bob has at least the following preference evaluations, expressed by $u_B(\cdot)$

u_B(Alice offers $2 and I accept) >

u_B(Alice offers \$1 and I reject) >

u_B(Alice offers \$1 and I accept) >

u_B(Alice offers \$2 and I reject).

Also,

u_B(Alice offers \$3 and I accept) >

u_B(Alice offers \$1 and I accept) and

u_B(Alice offers \$3 and I reject) >

u_B(Alice offers \$1 and I reject).

Under this sort of preference assignment, (*Offer \$2*; *Reject \$1, Accept \$2, Reject \$3*) is a Nash equilibrium and it is the sole Nash equilibrium if u_B(Alice offers \$3 and I reject) > u_B(Alice offers \$3 and I accept). Guala admits that people can derive their preferences from a variety of values. He also notes that some game theorists have recently made explicit the point that observed empirical anomalies do not automatically refute GT as descriptively false. He quotes James Cox and Jorgen Weibull,

> In their seminal work on game theory, von Neumann and Morgenstern (1944, 1947) thought it necessary to simultaneously develop a theory of utility and a theory of play for strategic games. In contrast, much subsequent development of game theory has focused on analyzing the play of games to the exclusion of utility theory. In the absence of a focus by game theorists on utility theory, it is understandable that experimentalists testing the theory's predictions have typically assumed that agents' utilities are affine transformations of (only) their own monetary payoffs in the games. This interpretation of game theory incorporates the assumptions that agents do not care about others' (relative or absolute) material payoffs or about their intentions. There is a large experimental

> literature based on this special-case interpretation of the theory ... But this does *not* imply that the observed behavior is inconsistent with game theory, which is a point that has not generally been recognized in the literature. (Cox 2004, pp. 260-81)

Most game theorists now acknowledge the fact that assuming that utilities are derived only from monetary payoffs is too restrictive. On top of that, Cox and Weibull make clear the point that game theorists need to pay sensitive attention to which values influence utility judgments. This accords with Sen's worry that there is no internal consistency in rational choice theory *without* taking in external motives, objectives and values of an agent. (Sen 2002)

Guala, however, argues that the mere possibility of refining outcomes does not save GT from empirical anomalies. The existence of expected utility functions is at the center of rational choice theory; the existence proof relies on what is called the "rectangular field assumption." Refining outcomes across contexts can violate this assumption. (p. 256-8) Guala's discussion of GT assumes Savage's framework for utility, the most widely adopted account. The Savage framework defines a set of outcomes and a set of states of the world. Then the set of acts (choice options) is constructed by mapping from states to outcomes. The act of 'booking a flight in January' can assign outcome 'the flight will be canceled' to the state 'heavy snow.' A preference ranking over the set of acts is thus a ranking over functions from states to outcomes. The name "rectangular field assumption" is attributed to Broome (1991, pp.115-7) The idea of the assumption is that every function from the set of all states of the world to the set of outcomes must be a meaningful act, meaningful in the sense that a preferential attitude can be expressed toward that act so that a preference ranking can be formed over the set of all acts. The notion of

having a preferential attitude here is that of the revealed preference view: an agent prefers A to B just in case they choose A when A and B are her actual options. As Sugden puts it,

> So to say that a person has any kind of preference relation between two acts f and g is to imply that it is possible to confront that person with a choice between those two acts. This feature of acts - that any pair of acts must be capable of constituting a meaningful choice problem - is clearly required by Savage's approach, in which preferences are defined in terms of observable choice behaviour. (Sugden 1991, pp.761-2)

Because of the rectangular field assumption, the Savage model of rational choice requires that preferences be defined over all combinations of of states of affairs and outcomes. The behavioral account of preference, inherited from revealed preference views, imposes a further constraint - no preference without a possible choice. Guala argues that these requirements place restriction on what goes into the outcome set, and that outcomes that are too fine-grained will be illegitimate in a rational choice model.

When an outcome refined, the refinement could either be about the outcome "taken in isolation" (Guala 2006, p.256) or it could also include structural features of a gamble or game.[44] An outcome of a certain description, can be individuated not only by its own features but also according to the global features of the game in which it occurs. For example, in Ultimatum

[44] Buchak makes an analogous distinction between what she calls the "local and global features of a gamble." (Buchak 2013, ch.4) When an agent is concerned about the risk of a choice option (a gamble), for example, she may be concerned with how risky an outcome is by itself, and she may be concerned with how the outcome bears on the gamble overall. The local concern about the gamble is placed on a particular outcome - if there will be heavy snow on the day I book my flight, for example. The local concern is about how likely that outcome is to obtain, taken in isolation. On the contrary, the global concern of the risk of having my flight canceled is about how the outcome stands in the gamble that I am choosing. Is the outcome the only way my situation can turn out? Or are there multiple forms of transportation that I can take even if my flight gets canceled? In the latter case the missing of a flight does not seem as risky as in the former case. The global concern of risk is not about how likely the outcome will obtain by itself, but rather about how the outcome can affect the whole gamble. It is about the overall structural features of an option.

games, an outcome of ($3, $1) where the allocator offers $1 and the recipient accepts may occur in two different situations. In a game where the allocator has *earned the right* to make the offer, this outcome might be accepted as a fair offer by the recipient. In a simple Ultimatum game, the recipient might see the offer as unfair but value monetary gain more than the fairness at stake (or vice versa). The recipient might also just see fairness as irrelevant. The two games are different contexts. When individuated by the features of a particular game, the redescription will make reference to those features of the structure of that game. Guala argues that if the identity an outcome contains features of a game, the outcome cannot be used to generate rankable actions by way of the rectangular field assumption. In his words,

> To sum up: when too much is included in the description of outcomes, the consequence itself remains tied to the specific game and cannot be used to construct arbitrarily other acts (or functions from states of the world to consequences). ... The Savage measurement procedure is not flexible enough to neutralize reciprocity counterexamples. (Guala 2006, pp.257-8)

On the one hand, Guala recognizes the need to distinguish consequences and outcomes. Consequences are the end states of some action. But outcomes are the objects of preference that contain more than end states. As a matter of fact, people's preferences are context-sensitive and path-dependent. Eating an apple is an end state. But your preference evaluation of eating an apple by yourself might be different than that of eating an apple when it is the last fruit on table and you are with a group of people. Getting $0 is a different outcome when you could have gotten $1 million. Your preferences toward being given $2 can differ depending on whether you think you are being treated unfairly or it is the result of a fair split. Mere consequences have been

treated as the objects of preference in traditional rational choice theory, but more and more economists recognize - as does Guala - that the objects of preference are highly intertwined with specific contexts and choice paths.

On the other hand - and this is the real challenge to GT - Guala argues that it is too restrictive to recognize the difference between consequences and outcomes. The rectangular field assumption says that the act formed by arbitrarily picking out some states and outcomes must have a place in the agent's preference ranking. Broome points out that the proof of the existence of expected utility functions relies on this assumption, yet refining outcomes poses challenge to the assumption. In the Allais paradox, for example, the outcome of "getting $0" is not so much about the monetary amount, but is rather loaded with disappointment when you could have got $1 million dollars. However, redescribing the outcome and redescribed it as "getting $0 with disappointment" makes trouble for the Savage framework because the feeling of disappointment is a result of specific gamble, a particular state of the world. But Savage assumes that the set of outcomes and the set of states must be independent of each other. It should be possible, for example, to associate "getting $0 with disappointment" with states of the world that aren't disappointing at all. In other words, Savage needs to separate outcomes from states conceptually to make sense of preference comparisons. But individuating outcomes requires that some outcomes be closely tied to certain sorts of states.

While Guala admits that it is crucial to begin with a correct representation of a player's utilities and beliefs, he agrees with Broome's critique of Savage and argues that GT can be too restrictive to accommodate a correct representation that requires refining and tying outcomes to particular game contexts. If, for example, Bob from the simplified Ultimatum game sees being offered $1

by Alice in a different light when he knows there is $4 available to split then the latter prospect is not really detachable from the from the game context. You could never offer Bob that prospect in a state of the world where fair dealing isn't an issue.

Broome and Guala are right in thinking that the Savage framework is too simple to capture the complexity of preferences. But Broom also acknowledges that the rectangular field assumption is too strong and unnecessary in a proof of the existence of expected utility function (Broom 1991, p.81; p.117). This leaves room for rational choice theory without the assumption. DT is fundamentally folk psychology systemized. Unlike folk medicine, say, folk psychology is more than conventional wisdom that we have observed from human behavior patterns. Folk psychological explanations depend on the assumption that all rational creatures share a common principle of interpreting one another as acting intentionally. According to Donald Davidson, to interpret an action as intentional is to attribute a reason to it. We interpret someone's bodily movement of getting up and getting a drink of water as something that has a reason behind it, that the person acted with some belief and end in mind. The belief and desire constitute the reason for an instance of otherwise meaningless physical movement. Davidson argues that questioning an action, desire, or belief as irrational must presuppose that the agent subscribes to some normative principle of rationality, and it is because their beliefs or preferences fail to comply with the normative principle that we say it is irrational. (Davidson 2004, Chapter 12.) Expected utility maximization, the idea that people act according to their beliefs and preferences and they choose the option that best realizes their most desired outcome given relevant beliefs of states of the world, serves as such a normative principle of rationality.

My defense of normative decision theory and game theory insists that outcomes can and should be as fine-grained as possible, to the extent that no further refinement affects an outcome's utility. I call such outcomes "holistic outcomes" since they fully characterize the arguments of an agent's utility function, no matter what values she happens to have. This can include state/path dependent features. The set of all lotteries over such outcomes will include some weird gambles (e.g., a 50% chance at getting choice B in the Allais case and a 50% chance at accepting the $1 offer in the simplified Ultimatum game). Since a holistic outcome is a disjunct in the specification of a non-holistic outcome, the utility of a non-holistic outcome will be a lottery over all of the holistic outcomes. Let *P* be a proposition that describes outcomes. An agent might like *(P&Q)* a lot but not care for *(P&~Q)*. In that case, *P* is non-holistic for her and the utility of a non-holistic outcome can change depending on whether *P* is associated with *Q* or ~*Q*. But once outcomes are fully fine-grained, there are no other ways of individuating them so that their utilities change. As a result, the utility of a holistic outcome is a constant value. All of this shows that my account involves a version - or at least an analogue - of the rectangular field assumption: I think agents (can) evaluate all lotteries over holistic propositions. Only some of those lotteries are picked out by possible courses of action, however. (See Chapter 2 of this dissertation at pp. 52-53).

Guala's challenge depends on a particular fact of the Savage framework. The acts that are functions from states to outcomes must be meaningful as actual options so that a behavioral preference can be expressed about it. When outcomes are too fine-grained, the gambles formed from them may not be meaningful in this way. However, the Savage framework is just too simple to account for the complexity of preferences. The basic structure of DT begins with a preference ranking over outcomes, as individuated by what the agent cares about. weighed by a

probability distribution of states of the world, and then induces a preference ranking over choice options. To produce a correct ranking of choice, the theory must have correct inputs which are holistic outcomes. A holistic outcome is defined formally as a disjunction of all alternative ways of realizing an outcome. If one way makes a difference to the outcome's utility, individuate it. Once every possible way of individuating an outcome is specified, the outcome is fully fine-grained and the disjunction is called 'holistic.'

It is important to note that individuating outcomes is agent-specific. Every decision maker comes with her own preferences, values and objectives. We take her initial preferences as most basic and real, and those are the ultimate source and authority of how fine-grained an outcome should be for that agent. The individuation for an agent who has real preferences will be coarser than the maximal specification of the set of all possible worlds. If the agent has no interest whatsoever in whether the number of leaves on a tree is odd or even, then an outcome does not need to be individuated by that aspect.

Guala's challenge to GT does have some force in descriptive terms. If preference is understood in revealed-preference terms then there may be gambes which are not evaluable. Even if, as I urge, preferences are to be understood as real psychological states, evidence about such psychological states may be hard to come by if no actual choices provide the data. Guala's challenge, however does not really apply to my normative framework, which involves holistic outcomes and places no restriction on what gets included in an individuated description of an outcome. It only requires that people's real preferences which are actually fine-grained be authentically represented by constant utilities of holistic outcomes. Guala's complaint could be an epistemological point that what outcome is holistic for an agent can hardly be discovered. But

this point must be distinguished from my more metaphysical point that holistic outcomes can be used to model an individual's actual preferences. Normatively speaking, holistic outcomes gives GT a solid foundation, even though they can be epistemologically inaccessible. The logic of evaluation and choice need not guarantee epistemic access to the inputs it considers any more than traditional first-order logic needs a method for vouchsafing the truth-values of the propositions over which it operates.

6. Conclusion

GT deals with rational decision making in the context of strategic interaction where one's opponents' moves partly determine what one chooses. Since GT is DT with the same basic framework extended to a more complicated context, my normative interpretation of DT applies similarly to GT. In this chapter, I have discussed apparent failures of GT to account for certain cases as counterexamples to normative Rational Choice Theory, and argued that they are no longer counterexamples if GT is interpreted as a norm through the lens of my holism point. I have focused on the Ultimatum game as an illustration of how people make reasonable choices but those choices seem to be inconsistent with game theoretic analysis. One may reasonably choose to hurt one's own wealth as part of an act that retaliates for an unfair offer. This does not mean that one is not choosing rationally, but rather that there is more to the evaluation of outcomes/payoffs than money. A game theoretic analysis should begin only when it is clear which game is played, that is, when the elements of the game, including payoff numbers, are determined. In the previous two chapters, I have argued that the basic objects of preference are complicated and context-laden, and thus a modeler should specify the basic outcomes as holistic propositions that represent all path-dependent features of an outcome. In other words, when the outcomes in a game are not holistic, the payoff numbers cannot be settled, and it follows that a

game theoretic analysis cannot get off the ground. I attribute the apparent violations of GT to the fact that game payoffs rarely reflect holistic outcomes.

Further, I have explained the difference between the two main solution concepts of GT: Rationalizability and Nash Equilibrium. The concept of Nash Equilibrium, which is the most popular solution concept among game theorists, goes beyond treating strategic situations as straightforward decision problems. Building on Bicchieri (1993) and Skyrms (1988), getting to the Nash equilibrium of a game requires that a player correctly figure out one's opponent - namely, to have true beliefs about which action strategy the other player would choose. The fact that outcomes must be holistic for GT to have any normative traction leaves little epistemic room for application of game theoretic techniques in particular cases. It is unrealistic for a player to know her own preferences in all cases, let alone for her to know someone else's payoffs so she can deliberate about her opponent's play. Since Nash equilibrium solutions can be reached only if all players' beliefs about each others' strategies turn out to be true, the epistemic issues raised by Guala suggest that players will have a difficult time reaching Nash equilibria.

Bayesian GT adds some sophistication to traditional game models than might help with this issue. It handles situations where a player does not know the payoffs of other players by distinguishing the other players as different types and assignment a subjective probability to each type. Although the tools are still in need of development, the Bayesian game model suggests a way to more sensitive modeling. Bayesian games are sensitive, to some extent, to the fact that a player is uncertain about the other player's specific values and preferences that give rise to his/her holistic outcomes. For some highly constrained cases where distinguishing the types of a player exhausts the aspects of an outcome that would make a difference to the player's payoff, a

Bayesian game model can begin with payoffs of holistic outcomes. In those cases, using the rational-choice theoretic analysis would successfully predict Nash equilibriums.

Chapter 5

Conclusion

I am interested in the explanation and evaluation of human actions and choices. Decision theory purports to be the current 'scientific' account of choice behavior but it faces stiff empirical challenges. Aside from empirical disputes, I believe that insights from folk psychology about instrumental rationality carry normative merits. There are better and worse ways of choosing based on one's desires and beliefs. I argue that decision theory, which makes precise the idea that we tend to choose options that best realize our ends given our beliefs, provides a norm of rational choice. The model precisely distinguishes reasonable choices given one's basic values and beliefs about situation, from irrational, bad choosing. Decision theory provides a norm of choice by providing conditions of consistency between the evaluation of one's existing values and those in a new choice to make.

The Sen-style cases, where a choice seems to be reasonable but fails to fit the model, challenges a normative interpretation of decision theory. I believe that such challenges are best met by interpreting the basic objects of choice as holistic propositions, in a sense made precise in Chapter 2. On my view, all substantial evaluative inputs can be captured by a preference ranking over holistic outcomes. Inputs to the model must be specified finely enough to differentiate anything that would make a difference to the output. While the Sen-style cases refute decision theory either as a normative or a descriptive account, my holism view preserves it as an adequate normative account of instrumental rationality.

Decision theorists are concerned with the formalization of consistency conditions of rational choice. Refutation of the standard formalism have prevailed in recent decades. The Apple case, Allais paradox, and similar cases proposed by Buchak, lead to amendments to the standard axioms of the theory. My interpretation of standard decision theory offers an alternative defense. The holism interpretation shares similarity with the redescription strategy in the literature which is accused of being "normatively vacuous." In response, I have argued that any path-dependent feature of an outcome can be represented by a difference in holistic propositions, and thus the standard consistency conditions successfully models the mechanism that inputs evaluation of holistic outcomes and updates to evaluation of choices. Since decision theory is a logic of choice, a preference updating mechanism, it is independent of particular preference inputs. I have also argued that the particularity of an actual decision maker's evaluative attitudes provides a definite starting point to the use of the mechanism, itself a norm/logic of consistency.

But there comes with the side effect of making any descriptive use of decision theory unrealistically difficult in a range of cases. It is almost certainly too difficult for social scientists to be able to divine the holistic options that humans actually choose between – they are simply too fine grained. This gives us some new insight into why decision theory fails to successfully predict human behavior. In economics and related social scientific disciplines, decision theory is used as a descriptive-explanatory account that models how people actually make choices. Revealed preference theory has been a mainstream economic interpretation. It is an instrumentalist interpretation which stresses on regularity (correlation) among past and future behavior patterns, without committing to preferences and beliefs as real mental states. I side with philosophers' rejection (Sen and Hausman, for example) and adopt a realist view. A consequence

of my argument is that decision theory is best interpreted as a normative rather than descriptive account. The naïve, straightforward use of decision theory without holism fails either as normative or descriptive account. My view weakens the theory as a predicative account, but strengthens it as a reliable normative gauge of rational choice.

The observation that decision theory can be preserved as a good norm despite the fact that people often fail to act instrumentally rationally provides important insights into applied ethics. In many ethical situations, the problem is not (merely) that people do not have the right values, but rather that people fail to consistently align all their values to make a coherent decision. For example, someone might be in Sen's Apple situation and choose to take the last apple: she doesn't think about how rude it is because she is focused on her hunger. This kind of scenario is closely connected with the empirical evidence about the failures of decision theory as a descriptive view. Conscious consumers should pay attention to several features of the products they buy, for example, but in the store, the information captured by prices looms larger than it should. Understanding decision theory as a normative account of instrumental rationality helps us disentangle applied ethics problems that are essentially problems of poor decision-making. (Ellis 2008)

Finally, I have extended my arguments to choices in the context of strategic interactions. A game structure is not determinate unless the payoffs in the game are determinate. An accurate modelling of a game structure must depend on accurate payoffs that represent holistic outcomes. But as implied by my arguments, such modeling must be sensitive to concrete and path-dependent features in a game context with regard to particular players' evaluative attitudes,

which is unrealistically difficult especially from a modeler's perspective. Moreover, building on Bicchieri (1993) and Skyrms (1988), getting to the solution of a game (Nash equilibrium) requires a player to correctly figure out one's opponent. As a result, my argument focuses attention on Bayesian game theory, since Bayesian games handle situations where a player does not know the payoffs of other players by distinguishing the other players as different types and assignment a subjective probability to each type.

References

Allais, Maurice (1953). "Le Comportement de l'Homme Rationnel devant le Risque: Critique des Postulats et Axiomes de l'Ecole Americaine," *Econometrica*, 21(4): 503-546.

Beshears, J., Choi, J., Laibson, D., Madrian, B. (2008). "How are Preferences Revealed?" *Journal of Public Economics*, 92: 1787-1794.

Bicchieri, Cristina (1993). *Rationality and Coordination*, Cambridge University Press.

Binmore, Ken. (1987). "Modeling Rational Players Part I," *Economics and Philosophy* 3: 179-214.

Binmore, Ken. (1988). "Modeling Rational Players Part II," *Economics and Philosophy* 4: 9-55

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

Buchak, Lara (2013). *Risk and Rationality*, Oxford: Oxford University Press.

---- (2014). "Risk and Tradeoffs," *Erkenntnis* 79(6): 1091-1111

Camerer, Colin (1999). "Behavioral Economics: Reunifying Psychology and Economics," *Perspective*, Vol. 96.

Christensen, David (1996). "Dutch-Book Arguments Depragmatized: Epistemic Consistency for Partial Believers," *Journal of Philosophy*, 93: 450–479.

Cox, James (2004). "How to Identify Trust and Reciprocity," *Games and Economic Behavior*, 46: 260-281.

Davidson, Donald (2004). "Incoherence and Irrationality," *Problems of Rationality*, Chapter 12, Oxford University Press.

Diamond, Peter (1967). “Cardinal Welfare, Individualistic Ethics, and Interpersonal Comparison: Comment,” *Journal of Political Economy*, 75: 765-66.

Ellis, Stephen (2006). “Multiple Objectives: A Neglected Problem in the Theory of Human Action,” *Synthese* 152:2, pp. 313-38.

---- (2008). “Market Hegemony and Economic Theory,” *Philosophy of the Social Sciences*, 38(4), 513–532.

Guala, Francesco (2006). “Has Game Theory Been Refuted?” *Journal of Philosophy*, 0305: 239-63.

Guth, W., Schmittberger, R., and Schwarze, B. (1982). “An Experimental Analysis of Ultimatum Game,” *Journal of Economic Behavior and Organizations*, 3: 367-388.

Guth, W. (1995). “On Ultimatum Bargaining Experiments - A Personal Review,” *Journal of Economic Behavior and Organization*, 27: 329-344.

Hájek, Alan (2008). “Arguments for – or against – Probabilism?” *British Journal for the Philosophy of Science*, 59: 793–819.

Hampton, Jean (1994). “The Failure of Expected Utility Theory As A Theory of Reason,” *Economics and Philosophy*, 10: 195-24

Hausman, Daniel (1992). *Essays on Philosophy and Economic Methodology*, Cambridge University Press.

---- (2000). “Revealed Preference, Belief, and Game Theory,” *Economics and Philosophy*, 16: 99-115.

Hawthorne, James (2011). “Bayesian Confirmation Theory,” *Continuum Companion to the Philosophy of Science*, S. French & J. Saatsi (eds.), Continuum Press.

Hayden, Grant. & Ellis, Stephen (2007). "Law and Economics After Behavioral Economics," *Kansas Law Review*, Kansas Law Review Inc. April, 2007: vol. 55(3).

Heap, Shaun P.Hargreaves & Varoufakis, Yanis (1995). *Game Theory: A Critical Introduction*. London: Routledge.

Heukelom, Floris (2015). "A History of the Allais Paradox," *British Journal of the History of Science*, 48(1): 147-69.

Hicks, J. R. (1939). *Value and Capital: An Inquiry Into Some Fundamental Principles of Economic Theory*, Clarendon Press.

---- (1956). *A Revision of Demand Theory*, p.6, Oxford University Press.

Hoffman, E., McCabe, K., Shachat, K., and Smith, V. (1994) "Preference, property rights and anonymity in bargaining games," *Games and Economic Behavior*, 7, pp. 346–380.

Houthakker, H. S. (1950). "Revealed Preference and the Utility Function," *Economica*, 17: 159-174.

Hume, David (2000). *A Treatise of Human Nature*, Oxford University Press.

Jeffrey, Richard C. (1990). *The Logic of Decision*, 2nd ed., Chicago: The University of Chicago Press.

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

Kahneman, D. and Tversky, A. (1984). "Choices, Values, and Frames," *American Psychologist* 39, 341-350.

Kreps, David M. (1988). *Notes on the Theory of Choice*. Westview Press.

---- (1990). *A Course in Microeconomics*, Chapter 3.

Lewis, David (1983). *Philosophical Papers Volume I*. Oxford University Press.

Little, I. M. D. (1949). “A Reformulation of the Theory of Consumer's Behaviour,” *Oxford Economic Papers*, 1: 90, 97.

Luce, Robert D. & Raiffa, Howard (1957). *Games and Decisions*. New York: Dover Publications.

Osborne, Martin J. & Rubinstein, Ariel (1994). *A Course in Game Theory*. The MIT Press.

Pettit, Philp (1991). “Decision Theory and Folk Psychology,” *Essays in the Foundations of Decision Theory*, Michael Bacharach & Susan Hurley (eds.) Blackwell. pp. 147-175

Reiss, Julian (2013). *Philosophy of Economics: A Contemporary Introduction*, New York: Routledge.

Resnik, Michael D. (1987). *Choices: An Introduction to Decision Theory*, the University of Minnesota Press.

Rubinstein, Ariel (2012). *Lecture Notes in Microeconomic Theory: The Economic Agent*. Second edition, Princeton University Press.

Savage, Leonard (1954/1972). *The Foundations of Statistics*. Dover: John Wiley and Sons.

Sen, Amartya (2002). “Internal Consistency of Choice” in *Rationality and Freedom*. Cambridge, Massachusetts: The Belknap Press of Harvard University Press.

Schick, Frederic (1984). *Having Reasons: An Essay on Rationality and Sociality*, Chapter 2, pp. 18-9. New Jersey: Princeton University Press.

---- (1991). *Understanding Action*. Cambridge University Press.

Skyrms, Brian (1988). “Deliberational Dynamics and the Foundations of Bayesian Game Theory,” *Philosophical Perspectives*, 2: 345-367.

Sugden, Robert (1991). “Rational Choice: A Survey of Contributions from Economics and Philosophy,” *The Economic Journal*, 101: 751-785.

Swoyer, Chris and Ellis, Stephen (2005) “Rational Choice,” *New Dictionary of the History of Ideas*, ed. Maryanne Horowitz, pp. 2006-8, Charles Scribner's Sons.

Thaler, Richard (1988). “Anomalies: The Ultimatum Game,” *The Journal of Economic Perspectives*, 2: 195-206.

Verbeek, Bruno (2001). “Consequentialism, Rationality and the Relevant Description of Outcomes,” *Economics and Philosophy*, 17: 181-205.

Von Neumann, John, and Oskar Morgenstern (1944). *Theory of Games and Economic Behavior*. Princeton, NJ: Princeton University Press.

Weibull, Jorgen W. (2004). “Testing Game Theory,” in Stephan Huck, ed., *Advances in Understanding Strategic Behavior*. New York: Palgrave, pp. 85–104, especially pp. 85–86.