I'll Have What She's Having: Reflective Desires and Consequentialism

by

Jesse Kozler

A thesis submitted in partial fulfillment
of the requirements for the degree of
Bachelor of Arts with Honors
Department of Philosophy
in the University of Michigan

2019

Advisor: Professor James Joyce
Second Reader: Professor David Manley

Acknowledgments

This thesis is not the product of solely my own efforts, and owes its existence in large part to the substantial support that I have received along the way from the many wonderful, brilliant people in my life. First and foremost, I want to thank Jim Joyce who eagerly agreed to advise this project and who has offered countless insights which gently prodded me to refine my approach, solidify my thoughts, and strengthen my arguments. Without him this project would never have gotten off the ground. I want to thank David Manley, who signed on to be the second reader and whose guidance on matters inside and outside of the realm of my thesis has been indispensable. Additionally, I want to thank Elizabeth Anderson, Peter Railton, and Sarah Buss who, through private discussions and sharing their own work, provided me with inspiration at times I badly needed it and encouraged me to think about previously unexamined issues.

I am greatly indebted to the University of Michigan LSA Honors Program who, through their generous Honors Summer Fellowship program, made it possible for me to stay in Ann Arbor and spend my summer reading and thinking intentionally about these issues. I am especially grateful to Mika LaVaque-Manty, who whipped me into shape and instilled in me a work ethic that has been essential to the completion of this project. Additional thanks goes to Megan Taylor for her generous suggestions and helping me learn to explain the subject matter of my thesis in a less dense, esoteric way. Thanks also to my peers in the fellowship cohort for sharing their own amazing research and giving feedback on mine - their passion and drive inspires me greatly.

I would also like to thank my peers in the philosophy department who have also undertaken honors senior thesis projects. Their solidarity has been a constant source of support. Especially important in this process were Arooshee Giroti, Josh Greenburg, and Kelsey Muniz, without whom I am not sure I would have survived. Thank you for the laughs, the late nights in Tanner, and the endless support.

Finally, I would like to thank my mother and father, who gave me the immeasurably valuable gift of a lifelong love of learning, and my sister Anna who has always been my biggest cheerleader.

Abstract

The aim of this thesis is to treat critically a particular problem in rational choice theory that I believe has been overlooked. Decisions come in roughly two flavors; static and dynamic. Dynamic choices require a successive sequence of decisions to be made in order to reach some final outcome. A large portion of the economic and philosophical literature on dynamic decision making has focused on the problem of dynamic inconsistency. This terminology refers to the phenomenon which occurs when an agent at some point in the sequence deviates from their original plan in pursuit of an alternative outcome. The general assumption in the literature has been that dynamic inconsistency is a problem in rational choice theory.

Economists have demonstrated that such inconsistency typically leads to a failure of intertemporal utility maximization. Importantly, this conclusion relies on several assumptions about the integrity of the agent's rationality at the initial stage of the decision process at which time the plan is adopted. In a sense, these assumptions present an impediment to a properly thorough treatment of the problem. They limit the scope of the discussion to a narrow and perhaps unrealistic subset of all potential cases of dynamic inconsistency. Philosophers, while also concerned with maximizing expected utility, have been attentive to other considerations such as an individual's obligations to self-commitment and consistency, qua rational agent.

There are additional, related philosophical concerns about these self-interactions across time. What exactly does it mean to will one's future well-being, even when the person one will become has very different preferences and beliefs than one currently holds? What does it mean to have obligations to one's past self and their intentions for their future? These are all topics relevant to dynamic choice specifically that philosophers have examined elsewhere.

In the first section of this thesis, I will survey the existing literature on the problem of dynamic inconsistency. The responses to the problem can be broadly categorized into two theoretical camps, consequentialism and its detractors. I will examine both critically, and argue that expected utility maximizing consequentialism is the superior framework, though I take only one specific variant of consequentialism to be relatively problem free and suitable for general use. Then, with this framework in mind, I will investigate what the point of rationality is; what it is for someone to be rational in a more abstract sense, above and beyond expected utility maximization.

Most critical to this work, the area of this literature that I believe remains underdeveloped concerns the question of how a rational agent ought to respond to her expectations of dynamic inconsistency. As stated, the literature has mostly assumed that such preference shifts and fluctuation of choice behavior is prima facie irrational, but I am not convinced this is the case.

The question this thesis project seeks to answer in the second section is how rational agents ought to respond to expectations of their future self undergoing changes such that they are no longer inclined towards the same choices as their current self. In the case of changes of belief, a reflection principle has been advocated for by some philosophers which aims to regulate an agent's beliefs by mandating synchronicity between current and future beliefs. I argue that this rational rule is plausible under conditions in which our future self can be credibly assumed to be in an epistemically privileged state relative to the present moment. Adherence to the reflection principle prevents susceptibility to Dutch books and also meshes well with the conception of rationality for which I advocate.

But what of cases where the changes to the agent's identity rearrange the structure of their basic desires as opposed to their beliefs? Preferences have two component pieces, beliefs and desires, meaning dynamic inconsistency can result from shifts in one or both internal structures. I investigate whether or not the framework of the epistemic reflection principle can be used as the basis for an analogous norm in the case of desire shifts. In many respects, this case is more difficult than the cases of epistemic shifts. Desires, unlike beliefs, are not truth evaluable by appealing to facts about the world which concern the propositions believed. Therefore additional, normative assumptions or claims must be made to justify the evaluative stance towards a particular disposition of desires.

Ultimately, I argue that the analogous desire version of the reflection principle can be successful and does provide a degree of rational motivation for conforming one's current desires to match one's future desires, assuming that the agent has reason to believe her future self's desires constitute an improvement to her own and that such conformity will on the whole improve the agents well-being. However, the strength of arguments from analogy diminishes proportional to how dissimilar it is to the base case. Because the case of reflective desires can not benefit from non-pragmatic support and relies on a tenuous type of authority, I argue that reflection in the case of desires is not rationally motivating to the same degree as the case of reflective beliefs.

Finally, I will briefly examine the compatibility of this proposed principle with the consequentialist paradigm that I take as my starting point in this work. I will also briefly suggest some potentially interesting areas of further exploration.

Contents

1 Introduction to Dynamic Choice Problems

Decisions are rarely made in isolation from the future and past. What I prefer to eat for lunch at noon is to some extent informed by what I plan to eat for dinner later this evening. Deciding on which university to attend was partially determined by my expectations of future job prospects at each option. Whether or not I pass an upcoming exam will depend heavily on how diligently I study in the days leading up to it. When it comes to assessing the merits of possible actions and choices, our preferences are informed by past events, present happenings, and expectations of the future alike.

What these situations have in common is that they are dynamic or diachronic - the decision-making takes place in an environment that changes over time, either due to the previous actions of the agent or due to events that are outside of the agent's control. Because an agent's preferences are malleable over time, an experience or good might be worth some fixed amount of utility to her at the present moment, but its value may fluctuate as time passes. Diachronic choice problems are not exactly a recent exploration in economics. Economists have studied intertemporal decision making since the early twentieth century, mostly focusing on optimizing an agent's consumption strategy in terms of the quantity and mix of goods they consume at different time points.[1] But in recent decades, there has been a growing body of philosophical literature on the problem of dynamic inconsistency.

Dynamic inconsistency is the phenomenon which occurs when an agent's current assessment of which future course of action best and her assessment of the same action in the future time are inconsistent. For example, as I sit here typing this out, I think that the best course of action for my future self at 6 o'clock tomorrow morning will be to wake up and go for a run. However, at 6 am, I may adopt the opposite perspective and consider staying in bed to be the much more attractive option. This mismatch between what my current and future selves judge to be the best action is definitive of dynamic inconsistency.

This chapter presents a survey of the academic literature on dynamic choice and dynamic inconsistency. In it, I analyze the leading paradigm in rational choice theory, the standard expected utility model, and discuss what it can teach us about how approaching dynamic choices as well as its limitations in dynamic contexts. I will examine the elephant in the room as it were - the problem of dynamic inconsistency and discuss what makes certain types of choosers inconsistent, then survey the possible solutions that have been proposed. I believe that the dominant model of sophisticated choice is the superior option and I will do my best to respond to expected utility theory's critics, though I also concede that there are ways which it is unsatisfactory or incomplete as a prescriptive account of dynamic choice. I am aware that not every reader will be familiar with this problem in rational choice theory so I intend this chapter to be a useful exposition of the problem to aid in making my thesis argument intelligible to a wider audience.

1.1 The Standard Model, in Brief

Before proceeding to dynamic choice, I first examine the theoretical underpinnings of simpler static decisions. These decisions are characterized by a singular and irreversible choice point for the agent in question. The paradigmatic method of rational choice assessment comes

[1] Introduced and developed by Rae (1905), Böhm-Bawerk (1889), and Fisher (1930).

from Savage,[2] who envisioned all decision problems as situations in which an agent must choose among a fixed set of actions ($a_1, a_2, ... a_n \in A$), the outcomes of which are risky or uncertain. The set of all possible outcomes ($o_1, o_2, ... o_m \in O$) depends on both the action chosen by the agent and the state of the world which obtains after or concurrent with the agent's choosing the action. These states of the world ($s_1, s_2, ... s_k \in S$) are the vehicles of uncertainty in these situations because they obtain independently of the agent's actions.

An agent, qua rational, acts as if she is assigning a probability distribution to the members of S, accounting for all available evidence, prior to choosing an action. Note that this is an 'as if' statement; it is that agents act as if they are assigning a probability distribution, not that agents *do* in fact. Most people do not explicitly perform the mathematical calculus needed to produce an accurate probability distribution when evaluating their own decisions. Rather, we tend to think in terms of loose approximations and estimations. Some events are 'more likely' while others are 'less likely' or 'as likely as not.' We do engage this process when we deliberate in choice situations, but it is rapid, unconscious, and imprecise.[3]

As an aside, some in the philosophical community, uncomfortable with the ontological vagueness of what types of things the members of A, S, and O are, have opted to interpret these features of Savage's model in the way that Jeffrey first suggested - as different types of propositions.[4] On his view, the various options in A are represented as propositions detailing behaviors that the agent can make true in virtue of completing that action. The members of S are propositions detailing certain features of the world which are outside of the agent's control. Each outcome then constitutes a proposition which describes the final situation for the agent and the effect it has on the agent's happiness or well-being.

The benefits to this type of thinking are twofold. First, it simplifies the system by replacing three completely different types of things with one central thing (propositions) which has three distinct flavors. That simplicity is useful for rational choice theorists. Second, it allows philosophers to apply logic to different features of the decision problem. For example, we can discuss the probability of the disjunct of different possible states (e.g., $p(s_1 \vee s_2)$) or apply conditional statements (e.g., $a_1 \wedge s_1 \rightarrow o[a_1, s_1]$).[5] This distinction, while perhaps irritatingly esoteric for some readers, is worth mentioning briefly here but will not be pertinent to future discussions.

Returning to Savage's basic model, the relationship between these three types of variables present in decision problems can be visualized in a simple decision matrix like the one below:

[2] Savage (1972).

[3] I suppose this stance broadly classifies expected probability calculations an example of a "System 1" process which is fast, autonomous, and unconscious as opposed to "System 2" processes which are conscious and effort-intensive. See Kahneman (2011) for a more fleshed out characterization of this distinction.

[4] Jeffrey (1990).

[5] This style of notation has been adopted for outcomes and indicates in the subscript the combination of action and state of which the outcome in question is a function. Presenting each individual outcome in the form $o[a_n, s_k]$ more easily illustrates the unique combination of action and state which produces it for reference.

	s_1	s_2	s_3	$\cdots$	s_k
a_1	$o[a_1, s_1]$	$o[a_1, s_2]$	$o[a_1, s_3]$	$\cdots$	$o[a_1, s_k]$
a_2	$o[a_2, s_1]$	$o[a_2, s_2]$	$o[a_2, s_3]$	$\cdots$	$o[a_2, s_k]$
a_3	$o[a_3, s_1]$	$o[a_3, s_2]$	$o[a_3, s_3]$	$\cdots$	$o[a_3, s_k]$
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
a_n	$o[a_n, s_1]$	$o[a_n, s_2]$	$o[a_n, s_3]$	$\cdots$	$o[a_n, s_k]$

Table 1: Standard Model Decision Matrix Structure

To put this in context, Maria may be debating between taking the freeway or the normal roads to her office in the morning. She knows that the freeway is the fastest option when there is no traffic, but during rush hour the traffic slows her down appreciably enough that she will be late for work. To make this decision she may consult a traffic app on her phone and reflect on what the conditions have been like in past weeks at this time. This will allow Maria to approximate the probability that there will be heavy traffic today on the freeway. The decision facing her this morning is mapped out in Savage's model as follows:

	Traffic	No Traffic
Freeway	35 minutes	10 minutes
Normal Roads	20 minutes	20 minutes

Table 2: Traffic Decision Matrix

According to the standard model (SM), in these types of *normal form* decisions - static representations of one-shot choices - the method by which an agent ought to choose is simple. After assigning probabilities to all $s_k \in S$, she should look at the available outcomes and calculate the expected utility (EU) of each of her actions.[6] Since each o_m is a direct function of the corresponding s_k and a_n, this process entails weighting the outcomes (grouped by action) by their expected probabilities. The expected utility of any action then is equal to the sum of the utility values of each outcome weighted by the probability of the corresponding state. More formally:

$$exp_u(a_1) = \sum (p_{s_1}(u_{o[a_1,s_1]})) + (p_{s_2}(u_{o[a_1,s_2]})) + \cdots + (p_{s_n}(u_{o[a_1,s_n]}))$$

Repeating this expected utility assessment for each action in A will result in expected utility totals that can be ranked. It follows from the standard expected utility model that the only choiceworthy actions will be the ones ranked highest in terms of expected utility. Those actions in A which are best (or least detestable) in virtue of the fact that they lead to the highest possible utility payoffs constitute the choiceworthy set of actions, $C(A) \subset A$.[7] Choiceworthiness is taken to mean something like rational permissibility - the option or options that yield the highest utility are the only choiceworthy actions since it would be irrational to willingly settle for less. If, for example, I were to offer a friend the choice between either $5 or $10 (with no strings attached) you would certainly think it odd if he chose the former. In fact, it is a standard assumption that the utility functions of rational

[6] von Neumann and Morgenstern (1944).

[7] This is standard set theory notation and should be read as "the set of choiceworthy actions $C(A)$ as a subset of all available actions A." Note that this subset can be empty or can be a union with A.

agents are always monotonic or that more utility is always better than less and therefore always preferred.[8]

1.2 Some Assumptions about Preference

As an extended aside, it will perhaps be useful to examine some of the features of preference that are considered by modern normative decision theorists to be constitutive of rational behavior. The impetus for postulating these regulative assumptions about preference has to do with the Representation Theorem in rational choice theory. This theorem stipulates that a rational agent's preferences ought to be able to be represented as if she is maximizing her expected utility given her choice behavior.[9] Once again, this is an 'as if' stipulation in the sense that humans in their capacity as deliberative agents do not actively consider themselves to be maximizing their expected utility as they consider possible decisions. Rather they endeavor to satisfy their own preferences, and if their preferences obey certain general axioms then it turns out they will act as if they are expected utility maximizers automatically in virtue of their preferences satisfying those conditions. The fact that a representable set of preferences and subjective probabilities can be constructed from their behavior lends a kind of coherence to that behavior and validates it as rational, at least in a narrow sense. What exactly we mean when we talk about 'rationality' will be discussed further on in this work.

To begin, we assume for ease of discussion that an agent's preferences in all situations among her actions are complete, transitive, and reflexive. Completeness stipulates that the agent can always compare the value of two options vis-a-vis each other; that for every pair of prospects X and Y, either $X \preceq Y$, $X \succeq Y$, or $X \approx Y$.[10] For transitivity to hold among strict preferences, it must be that for all prospects X, Y, and Z, it cannot ever be the case that X$\succ$Y, Y$\succ$Z, and Z$\succ$X. In the case of transitivity of weak preferences, if X$\succeq$Y, Y$\succeq$Z, then it cannot be the case that Z$\succ$X. The key is that for both the strict and weak preferences to be transitive, if either of the first two relations are strict then the third is as well. Violations of this rule produce cyclic preferences that are irrational because they make an agent vulnerable to manipulation as a 'money pump' (but more to come on intransitive preferences later). And finally, reflexivity requires that each option is at least as good as itself, or that X$\preceq$X.

The dominance principle states that one possible action dominates another if it produces a better outcome in every possible state of the world - that if $o[a_1, s_k] \geq$ o[a_2, s_k] for all s_k $\in S$, then a_1 dominates a_2 and should always be chosen over it. Here too there is a strict version which holds for the $>$ relation and a weak version that holds for the $\geq$ relation between outcomes.

Decision theorists also postulate that an agent's preferences must obey an independence axiom. The value of an outcome must be exclusively the result of the features of that outcome and ought not depend on details such as the other options with which it is bundled. Formally, the strong axiom of independence states that if lottery ticket A_1 is (as good or)

[8] I make the assumption in this example that my friend is the type of person whose utility increases as their monetary payout increases but this need not be the case. The monotonic property applies only to utility, not necessarily to money or any other specific type of good.

[9] Savage (1972), Ramsey (1931), Anscombe and Aumann (diff)1963).

[10] This notation is standard for describing preferences - $\succeq$ is to be read as "is weakly preferred to," $\succ$ as "is strictly preferred to," and $\approx$ as "is indifferent to."

better than B_1 and lottery ticket A_2 is (as good or) better than B_2, then an even chance of getting A_1 or A_2 is (as good or) better than an even chance of getting B_1 or B_2 (and should therefore be preferred by the agent).[11] For example, suppose I prefer a new bike to a ham sandwich and I prefer a trip to Paris to a trip to Cleveland. It follows that I ought to prefer a gamble that gives me a bike with probability p and a trip to Paris with probability q to a gamble that gives me a ham sandwich with probability p and a trip to Cleveland with probability q.[12]

Another of the axioms concerned with enforcing independence is context independence, which states that when determining the choiceworthiness of actions, the only salient features of the situation are the agent's subjective beliefs about the likelihoods of the various states and her subjective desires among possible outcomes. All other background features cannot constitute rational criteria for choiceworthiness. This idea is contained within two complimentary principles:[13]

- α: If some act is choiceworthy among all options available in a decision, it must also be choiceworthy in any subset of the initial possible actions - if $x \in C(A)$, $x \in A^*$, and $A^* \subset A$ then $x \in C(A^*)$
- β: If some act is choiceworthy in an initial set of options in a decision as well as the superset of those options, then any other act similarly choiceworthy in the initial set must also be choiceworthy in the superset - for any x and y such that $x \in C(A)$ and $y \in C(A)$, if $A \subset A^*$ then $x \in C(A^*)$ if and only if $y \in C(A^*)$

The underlying notion is that the relative value of an act should never be subject to change with the addition or subtraction of options from the set unless the addition or subtraction of these options provides the agent with information that is relevant to what her actions will cause. The winner in a wide competition must still be the winner in a narrow competition and the winners in a narrow competition must either both remain winners or both be losers in a wider competition. For example, I cannot rationally think that France's national soccer team is the best in the world but not the best in Europe. And if I think that England and Mexico have equally good teams out of a list of ten countries, adding Ghana into the mix should not cause me to suddenly think Mexico is unequivocally best.[14]

Finally, there is the more difficult case of Savage's sure-thing principle (STP), which claims that for any choice between two lotteries, if both contain the same outcome weighted by the same probability, then that element should essentially be ignored because it is a 'sure thing.'[15] Additionally, the relative value of the lotteries should not change at all if that outcome is changed, as long is it is given identical probabilities in each lottery. So in practice, if you are deciding between lottery A_1 which pays out a new bicycle with probability .5 and an ice cream cone with probability .5 versus lottery A_2 which pays out a vacation to Miami with probability .5 and an ice cream cone with probability .5, you should choose identically if the ice cream cone in both lotteries is switched out with 50 cents, two ice cream cones, a Ferrari, or anything else. Since the agent is facing the same outcome

[11] Samuelson (1952), 672.
[12] For all values of p and q such that $p + q = 1$.
[13] Sen (1971).
[14] This particular example of football teams is borrowed from James Joyce.
[15] Savage (1972).

with probability .5 in either lottery, that aspect of the lotteries should not matter to her decision.

Unlike the other axioms included here, this idea has caused notable dissent in the philosophical and economic ranks. The infamous Allais paradox[16] serves as a supposed counterexample to the legitimacy of STP by pointing out that the opinions of the masses are about half of the time in opposition to its recommendation. Given the choice between lottery A, which pays \$1 million with certainty and lottery B, which pays \$5 million with probability .1, \$1 million with probability .89, and nothing with probability .01, a slim majority prefer A. However, if these lotteries are equivalently redescribed according to the sure thing principle as lottery C, which pays \$1 million with probability .11 and nothing with probability .89, and lottery D, which pays \$5 million with probability .1 and nothing with probability .9, about half of those surveyed think lottery D is the superior option. These lotteries are illustrated below:

	.89	.1	.01
A	\$1 million	\$1 million	\$1 million
B	\$1 million	\$5 million	\$0

	.89	.1	.01
C	\$0	\$1 million	\$1 million
D	\$0	\$5 million	\$0

Table 3: Allais Paradox Lotteries

According to expected utility theory, a rational agent ought to prefer B and D. So why do so many seemingly rational people pick A, despite the fact that it's expected payout is nearly 40% lower than option B? Some philosophers, notably Buchak, have postulated theories which validate this pattern of choice behavior by denying certain features of classic expected utility theory - these will be examined later. Savage's sure-thing principle has been defended by others and in my own opinion, the disconnect between the rational prescription and the intuitions of most people is simply a case of widespread irrationality and is neither remarkable nor troubling to the standard expected utility model. Therefore, I unhesitantly include the sure-thing principle as an axiom of rational preference.

1.3 Decision Trees

I now turn my attention to the dynamic variety of decisions. The features of dynamic choices are best visualized with decision trees, which are meant to represent a variety of possible outcomes that the agent may reach depending on how she chooses to navigate the tree.[17] Decision trees are composed of two basic features, nodes and the branches connecting them. In actuality, the entirety of a person's life can be conceived of as one gigantic tree with indefinitely many branches but for the purposes of clarity, it will be better to deal only with

[16] Allais (1953), 527.
[17] Raiffa (1968).

trees that are truncated - defined over a finite amount of time with a strict beginning and end.

These trees begin at an initial node, n_0 and branch outwards to all other nodes $(n_1, n_2, ... n_{\text{i}})$ such that all $n_{\text{i}} \in N$. The nodes of decision trees come in three types, each playing an important and unique role in dynamic choices. Choice nodes are designated by a square and indicate a time point where the path of continuation along the tree is up to the agent and the direction they move on the tree is determined by their actions.[18] Chance nodes, represented by circles, constitute time points at which the paths of continuation correspond to different probabilistic events that are not controlled by the agent, for example a coin flip.[19] The probabilities of each state that stem from that chance node are indicated on the tree. Finally, terminal nodes or endpoints are the nodes at the farthest extreme of the tree and act as the artificial conclusion of the tree imposed at some final time t_{f}.[20] Listed at each of the terminal nodes is a number (or sometimes a description) that indicates the total utility of the outcome corresponding to that particular path from the initial node to the terminus.

Below is an example of a simple decision tree which contains all of these features:

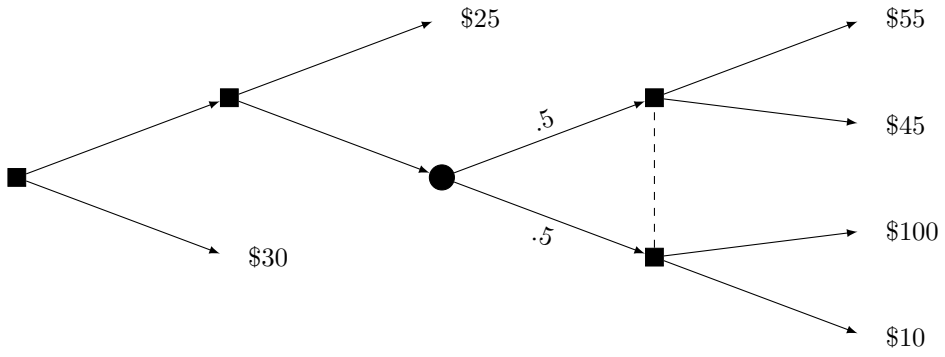

Figure 1: Sample Decision Tree

At the onset the agent must choose between receiving $30 or continuing along the tree. If she continues, she will then choose between a sure $25 or continuing on to the chance node. From there she will arrive at the information set indicated by the dashed line between the two square choice nodes. An information set is a grouping of acts such that the options at all points in the information set are the same and that when making the choice, the agent does not actually know which point she is at. Information sets as a concept are necessitated by the fact that the agent, when planning ahead, does not always know at which point of the tree she will find herself at some t_{i} thanks to the existence of chance nodes. These results

[18] As a side note, this work and rational choice theory generally remain noncommittal on issues relating to causal constraints on our free will - this statement not making a claim related to that family of issues.

[19] Where the events at a given chance node are typically complete and mutually exclusive.

[20] Time is mapped onto decision trees as a series of discrete steps, beginning with t_0 and proceeding up to t_{i} before terminating at the final time t_{f} where $t_1, \ldots, t_{\text{f}} \in T$. The different choice, chance, and end nodes all line up with these time points for the sake of simplicity. Obviously this is not the case in reality, in which time is much more fluid and messy.

will be determined by a dice roll, the flip of a coin, or the actions of other which are not predetermined.

From that point, depending on which outcome within the information set she finds herself in, she will either choose from ($55, $45) or ($100, $10). If she is an expected utility maximizer and a miser,[21] then she will compare the sure $30 to the sure $25 to the expected value of the chance option, which is (.5)$55+(.5)$100=$77.50. Because the chance node produces the highest probability-weighted expected payoff, that will be the option chosen.

As opposed to one-shot decisions, the agent faced with a dynamic choice has to choose not between individual actions but between different *strategies* or *plans*. A plan is a set of instructions that determines the agent's choice at every information set in the tree given her initial preferences among the available endpoints. The set of all plans in a tree is the collectively exhaustive list of possible paths that an agent could follow to reach the set of possible outcomes. At first glance, it seems like the right way to solve a decision tree is to look at the utilities on each terminal node and rank them, taking into account the probabilities associated with any chance nodes present. The rational option, then, would be to adopt the plan whose terminal node is the one with the maximum utility. If we solve the above decision tree in this way, it is possible to represent this extensive decision problem in the same form as Savage's model for static choices, swapping out the bundle of individual actions for all of the possible strategies in the decision tree.

	.5	.5
down	30	30
up, up	25	25
up, down	55	100

Table 4: Normal-Form Adaptation of Sample Decision Tree

This assumption - that the strategies of decision trees can be reduced to a static matrix and the choice can be made from there - is the principle of normal-form/extensive-form equivalence or a reduction from the extensive to the normal form.[22] The hypothesis is that a strategy in the extensive form tree is choiceworthy if and only if its corresponding action[23] in the normal form matrix is also choiceworthy. This equivalence was widely accepted in the past by members of the economic community.[24] It seemed intuitive to reduce dynamic decision trees in this way because it allowed them to be solved with a very straightforward one-step process. It was not until the mid 20th century that this view began to erode as new problems were identified. That development will be examined in the next section.

[21] Someone who values each additional dollar at the same amount of utility i.e., for whom there is no diminishing marginal utility of money.

[22] For a discussion of this principle, see McClennen (1990).

[23] Where that action is actually a single description of a series of steps along the tree.

[24] e.g., von Neuman and Morgenstern(1944).

1.4 Fickle Preferences and Dynamic Inconsistency

The normal-form/extensive form (NF/EF) equivalence was widely accepted until it was discovered that the static matrix model is reliably prone to failure when applied to dynamic situations in which the agent's preferences change and at some future time t_i run counter to her original intention at t_0.[25] This problem was first documented by Strotz, who illustrated this type of preference shift with the story of Odysseus and his encounter with the Sirens from Homer's Odyssey.[26] This literary example is one of the paradigmatic instances of temptation in western academic thought and is a useful depiction of dynamic inconsistency. Odysseus, on his way home to Ithaca, must decide whether or not to sail past the island of the Sirens, creatures whose enchanted song lures sailors to their death by convincing them to jump into the ocean or shipwreck on the rocky coast.[27] Curious to hear their song, but not wanting to commit suicide before reaching home, Odysseus faces a dynamic decision of the form:

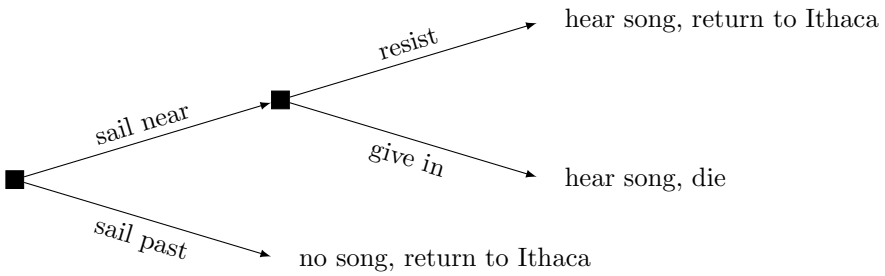

Figure 2: Odysseus and the Sirens Decision Tree

Odysseus ultimately decides to have his crew plug their ears with beeswax, rendering them effectively deaf, and have himself tied to the mast so that he is unable to act on his suicidal impulses when he hears the song. In thinking about the decision facing him, Odysseus realizes that a plan to hear the Sirens' song and resist it is not feasible in that despite his forming an intention to do so, no amount of willpower will overcome the temptation that he will encounter.

A feasible plan is one for which the agent will, as a matter of fact, find in her interest to execute at each choice point mandated by the plan.[28] There are two types of cases in which an agent can adopt infeasible plans. The first are situations where the agent authentically believes without pretense that she will carry out the plan but those beliefs are bad in that they are either false, unjustified, or both. This person wants to take her future self's beliefs and actions into account but is handicapped by her bad beliefs; we might call her a badly informed sophisticated agent. The second case is that of agents who do not consider their future self's willingness to adhere to the plan when making decisions; these agents are myopic.

[25] It is also worth noting that the NF/EF equivalence fails in two-player game theoretic situations in which agents make incredible threats and promises.

[26] Strotz (1995).

[27] Homer (1997).

[28] This particular phrasing suggested to me by Joyce.

The plan to sail past the Sirens but refrain from committing suicide is doomed to fail because the second he falls under their spell, Odysseus's desires for Ithaca will erode and he will reverse his prior intentions with certainty. Luckily, his clever solution works - he hears the song but makes it out alive. Odysseus's rationale here and his decision to tie himself to the mast constitutes what Strotz calls a strategy of precommitment to the initial plan.[29] Precommitment entails taking preventative measures in the present to prevent your future self from defecting from the original plan. But this has to be something stronger than just a resolution or promise to oneself. The act of precommitment has to make it possible for the agent to follow through with the plan as originally intended. There are two ways of accomplishing this: restricting action as in the case of Odysseus, or changing the incentives as in the case of the Biggest Loser competition which incentivizes contestants to keep the weight off not only for the sake of their health but because if they do not, pictures of their overweight bodies will be made public.

Of course the solution of precommitment is not always viable in the absence of effective mechanisms. So what ought Odysseus to have done if, for example, there had not been any rope aboard the ship to restrain him? In cases like this, Strotz advocates what he calls a strategy of consistent planning, and what has since been referred to primarily as sophisticated choice. A sophisticated chooser is one who adopts only plans which they reasonably believe to be feasible.[30] At the initial node in a decision tree, a sophisticated agent adopts a plan with a specific endpoint in mind. But sophisticated choosers mitigate their expectations of reaching that final outcome by considering the feasibility of each plan, and screening out all of those that are not dynamically consistent. In doing so, sophisticated choosers demonstrate an understanding of their own limitations when it comes to their commitment to a plan, a concept that is critical in understanding dynamic inconsistency.

Planning is an act of commitment and requires genuine intention to follow through on the agent's part. According to Bratman, intentions are not something to be taken lightly and "future-directed intentions play a central role in our psychology,"[31] As human beings we are not frictionless deliberators; there is always an opportunity cost to spending time reconsidering a plan.[32] Additionally, we are social creatures and have "pressing needs" for reliable coordination between us. Bratman writes, "if we were constantly to be reconsidering the merits of our prior plans they would be of little use in coordination and in helping us cope with our resource limitations. The nonreconsideration of one's prior intentions will typically be the default."[33] He argues that our psychologies actually dispose us to be fairly consistent and that our brains are adept at running computational algorithms in the background to determine when the potential benefits of reconsidering a past intention (perhaps due to the availability of new information or a change in circumstances) outweigh the costs of that reconsideration.[34]

[29] Strotz (1955),173.

[30] This is what McClennen calls reactive equilibrium with one's future self - McClennen (1990), 12.

[31] Bratman (1992), 2.

[32] I think there is an interesting hypothetical worry here about policies of reconsideration leading to an unproductive and possibly vicious regresses in which one first must deliberate whether to reconsider their current plan of action, then will be prompted to reconsider whether reconsideration is worth their time, etc. It is a fact of our psychological limitations that we cannot hope to account for and process the myriad of costs and benefits at each tier of analysis.

[33] Bratman (1992), 3.

[34] Bratman (1992), 6.

However, the presumption of nonreconsideration is not categorical, even in the absence of new information. Consider Kavka's toxin puzzle; you have been approached by an eccentric billionaire who wants to strike up a deal. At midnight tonight you will be asked to form a genuine intention to drink a toxin tomorrow afternoon which is nonlethal but causes quite a bit of discomfort. At midnight, you will be monitored by a mind reading machine which can with total accuracy detect if you are being honest or not. If you are found to be authentic in your intention, $1 million will be deposited in your bank account tomorrow morning *before* the time at which you will be presented the toxin. The money will not be taken back if you refuse the toxin, the payout depends only on your forming an intention tonight.[35] The aim of the puzzle is to show that our intentions are at best only partially volitional and are constrained by our reasons for the actions which are the object of the intention, mirroring the way our beliefs are rationally constrained by available evidence.[36]

My opinion of Kavka's toxin puzzle is that it would be impossible for any rational agent to pass the mind test. That would require the belief that she will execute the plan to drink the toxin if she chooses it and that in turn requires a reason for her to drink it. Does she have a reason at present to intend to drink it? Not really, because she knows that if she passes the test, the second the money is deposited in her account she is off the hook and will not drink it. And if she fails the test, there is certainly no reason for her to drink the toxin needlessly. This bears significantly on Bratman's point that one cannot intend to do something unless one authentically believes that one will carry through on one's intention and do it.[37]

In this case, the dominant strategy for her rational future self is to refrain from drinking the toxin. And her rational current self will know this and will therefore be unable to form the intention. An agent's inability to rationally believe that she will follow through with the plan makes it infeasible for her. I, for instance, cannot make a rational plan to sprout wings and fly from my house to the grocery store because the knowledge that it is physically impossible for me to do so. Because I cannot honestly believe in my ability to do so, I am prevented from forming a genuine intention to act on that plan. The existence of that intention is necessary to the integrity of a plan.

This issue of feasibility is why the NF/EF compatibility assumption fails to hold in cases of dynamic inconsistency. Barring the existence of an outside option to tie himself to the mast, Odysseus knows that the strategy which looks the best in the normal-form matrix (hearing the Sirens' song and then resisting it and continuing to sail to Ithaca) is not feasible.Feasibility is a criterion distinct from logical possibility and is in some ways more strict. Feasible outcomes are those which are not only hypothetically possible but which are probable and which a rational agent could reasonably expect given the circumstances at hand.

Odysseus realizes that his future self will inevitably succumb to the Sirens' song and therefore the seemingly optimal plan is one that he cannot rationally and in good faith intend for his future self to execute. This is what the sophisticated model of choice requires; that the rational agent take into account the fact that some possible plans are not feasible and therefore should be discounted as viable options.[38] Were the clever Odysseus not able

[35]Kavka (1983), 34.
[36]Kavka (1983), 36.
[37]See Bratman (1987) for further discussion.
[38]Hammond (1976), 167.

to tie himself to the mast, he would have circumvented the Sirens' island entirely, passing up the chance to hear their song knowing that it would mean certain death.

Below is the normal-form analog of Odysseus's decision tree. It demonstrates this problematic feature the NF/EF equivalence because it prescribes the plan which Odysseus, by correctly anticipating his future self's preference shift, knows is infeasible.

	1
down	no song, return to Ithaca
up, up	hear song, return to Ithaca
up, down	hear song, die

Table 5: Odysseus and the Sirens Normal-Form Variant

If Odysseus had not considered how his future self might behave and only consulted the normal-form abridgment of the choice problem, he would have erroneously chosen the option that would lead to certain death for himself and his entire crew.

In contrast with sophisticated choice, that type of behavior is characterized as myopic because the agent behaves near-sightedly and forms incorrect or incomplete beliefs. Myopic choice is a fundamental mischaracterization of certain circumstances. For example, Odysseus might have characterized the song as not particularly persuasive or characterized himself as possessing superhuman willpower. Neither of these are rationally grounded beliefs because they do not properly track the evidence that Odysseus has at his disposal. Beliefs are truth-apt in the sense that we can verify their correctness with reference to facts about the world. Myopia is leaving the house without an umbrella because it's sunny at the moment, despite adamant warnings from the weatherman that rain is on its way in a couple of hours. Myopic actions, as reflective of the underlying beliefs, are bad in that sense.

Dynamic choices involve deliberate coordination by the self across time. It is not enough to just choose a plan at the start and assume that one's self will reliably carry out those actions like an automaton. Frequently temptation, misinformation, or other exogenous factors will shift an agent's preferences at some future time. The problem with falsely equating the normal and extensive forms is that it prescribes plans without considering the future obstacles to implementing them, which often include obstacles imposed by features of one's own psychology. A person cannot rationally intend now for their future self to execute a plan that they have chosen in the present and that combined model frequently prescribes infeasible plans.

It is not the case that all dynamic inconsistency results from obviously irrational sources of temptation. One well-documented, though comparatively benign, source of dynamic inconsistency is hyperbolic discounting - the tendency to value an outcome increasingly more as the agent moves temporally closer to the time of actual attainment. The consequence of this discounting shape is that it creates temporary preferences for smaller, sooner rewards over larger, later ones. Dynamic inconsistency is prone to occur because hyperbolas distort the relative value of options with a fixed delay between the two in proportion to how far the choice-maker is from those options.[39] That is, most people have the preferences of (\$5 today) $\succ$ (\$6 tomorrow) but (\$5 in one year) $\prec$ (\$6 in one year and one day).

[39] Laibson, (1997).

Another way to visualize these preferences is to consider the case of an agent trying to stick with her diet. If she makes plans in the morning to get dinner with a friend in the evening, she initially prefers ordering a side salad to a chocolatey dessert. The dessert is a far away prospect and it does not appeal to her enough to cause her to defect from her diet. But when she is actually sitting at the restaurant table and the waiter asks if she would like to order dessert, faced with the prospect of that immediate gustatory gratification, her value for the chocolate cake will shoot up and she will eagerly order it.

Researchers in both behavioral economics and psychology have also empirically identified a myriad of anomalistic patterns of preference which contribute to the prevalence of dynamic inconsistency in our everyday choice behaviors. These include among others the common difference effect, absolute magnitude effect, gain-loss asymmetries, and delay-speedup asymmetries,[40] as well as a variety of other implicit cognitive biases and heuristics which bear on our choice behavior.[41] While hyperbolic discounting and these assorted phenomena deserve philosophical investigation, I will restrict my inquiry to the temptation class of examples of dynamic inconsistency. These cases are both the most philosophically rich and most pertinent to the potential rational axioms to be examined later.

1.5 Consequentialism, à la Hammond

Historically, the dominant theory of rational choice theory has been expected utility maximizing consequentialism, and one of the most influential proponents of this theory is Hammond. At its core, this theory stipulates that choices ought to be evaluated by their consequences alone.[42] Savage was also a consequentialist, and it is clear that his framework was designed with the consequentialist hypothesis in mind. The logical strength of consequentialism is perhaps most readily visible in the static cases already examined, but its prescriptions apply to decision trees and choices made in a dynamic context as well. It is vital to the purpose of this work that we examine what exactly the consequentialist has to say about dynamic inconsistency and why, since most of the significant disagreements over the issue are fundamentally disagreements about the validity of aspects of consequentialism.

Agents who employ consequential assessments of outcomes ought to do so at each subsequent information set and resulting subtree.[43] In essence, this means that the agent tasked with navigating a decision tree must at all times be forward-looking, irrespective of which node she occupies in the tree at the present moment. The reasons in virtue of which she makes her decisions must similarly be forward looking - sunk costs, regret, or things which otherwise might have been cannot rationally be factored into the agent's calculations. Something like a past obligation or promise can only rationally sway the agent to the extent that it has meaningfully altered the value of her future consequences. Unless it does this, it does not rationally impose restrictions on her present choice behavior (this is a distinction that I will return to later).

So at every choice point, the agent must exclusively consider what is causally downstream from the choice at hand. What has happened or failed to happen in the past by nature cannot rationally determine or influence how an agent will choose in the present *unless* those events

[40] Loewenstein and Prelec (1992), 575.
[41] Kahneman and Tversky (2013).
[42] Hammond (1988b), 25.
[43] Hammond (1988b), 26.

modified in some way her current preferences or the current value of her available outcomes. The consequentialist picture also insists that questions about "what might have been" are similarly irrelevant. At any given choice point, all parts of the tree not downstream are immaterial. This includes both the past and the possible but unrealized ways that things might have gone. These considerations lead Hammond to suggest three principles which purport to guide the consequentialist while navigating the decision tree.

The first principle is what some call the pruning lemma and what has also been referred to as the principle of dynamic separability (SEP) elsewhere in the literature.[44] This stipulates that any plan that can be considered choiceworthy in a truncated or pruned version of a decision tree must also be considered choiceworthy in all identical pruned trees.[45] A pruned or truncated tree is just a tree that has been snipped at the root from the larger tree from which it originated, but is otherwise structurally identical from that point onward.

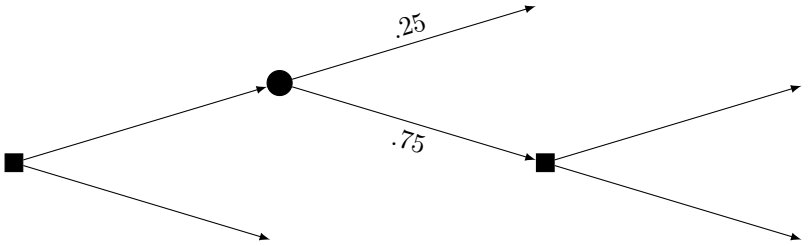

Figure 3: Example Tree

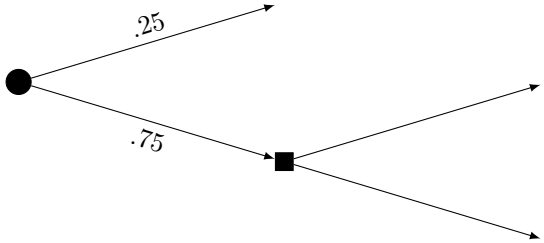

Figure 4: Pruned Version of Example Tree

Formally, we assume that at some node n_{i} in tree T there exists a nonempty set of possible plans $P(n_{\mathrm{i}})$ for the agent faced with that pruned tree $T(n_{\mathrm{i}})$. If we postulate another tree T^* in which n_{i} is the initial node and which is from that point onwards exactly identical to $T(n_{\mathrm{i}})$, then for any plan $p(n_{\mathrm{i}}) \in C(P(n_{\mathrm{i}}))$, it must also be the case that $p(n_{\mathrm{i}}) \in C(P(T^*))$. In other words, given some decision tree and resulting preference ordering of consequences, if the agent is faced with a subtree of a different tree which exhibits an identical structure

[44] e.g., McClennen (1990), 122.
[45] Hammond (1988b), 26.

of choices and consequences as in the original tree, their ranking over the consequences in the subtree must be consistent with their rankings in the original tree.[46]

For instance, Annie might strictly prefer going to the movies to doing the dishes in an abstract head-to-head matchup. In one tree she might face this decision on a Tuesday after commuting home from work. In another tree she might face this decision on a Wednesday after taking the dog for a walk. It is irrational, according to this theory, for the background context of the decision and the events and choices which have already transpired to alter Annie's relative preferences (unless those past events in the tree substantively change the attractiveness of one or both options). She should treat the options similarly in any case. Why would the day of the week change how attractive doing the dishes is?

The second principle suggested by Hammond is the dynamic consistency condition, which stipulates that any plan adopted by the agent at the initial node of the decision tree should be adhered to in all continuation trees.[47] It also implies that the restricted set of acceptable choices determined over the entire tree should identically restrict the agent's action at any information set within that tree.[48] So faced with some decision tree T, for any plan $p_1(n_0) \in C(P(n_0))$ the agent's preferences should remain unchanged so that $p_1(n_{\mathrm{i}}) \in C(P(n_{\mathrm{i}}))$ for all $n_{\mathrm{i}} \in T$. That is to say, the set of choiceworthy actions $C(A)$ determined at the initial node should designate options from the set A rigidly. Individual actions should not be added to or subtracted from $C(A)$ as the agent moves along the branches of the tree. A plan's completion at any choice node n_{i} relies crucially on the agent executing all remaining steps of the plan from that point onward. This indicates that for a plan to be rational, it must have a rational continuation optionat each choice node.

In practice, suppose that if an agent settles on a choiceworthy plan p_1 at n_0 and that plan demands that the agent choose a peanut butter and jelly sandwich over a ham and cheese sandwich at n_5. Assuming this instruction designates peanut butter and jelly as the only choiceworthy act at n_5, when she arrives at n_5, the agent must rationally choose the peanut butter and jelly as stipulated by the original plan because failing to do so amounts to a violation of the intended rigidity of $C(A)$ and makes it apparently temporally relative.

The important point is that, for dynamic consistency to hold, it must be the case that when an agent looks at a truncated tree, the act required by her plan has to be choiceworthy in that tree just as it was from the origin node in the larger tree. It need not necessarily be the only choiceworthy option though - it might be the case that both peanut butter and jelly and peanut butter and honey are choiceworthy but ham and cheese is not at the initial node. In that case, the agent may have an initial plan to choose peanut butter and jelly at n_0 but end up picking peanut butter and honey when they get to the counter of the deli at n_5.

While this does diminish the integrity of the agent's initial plan, it does not actually violate the dynamic consistency condition. We might view the dynamic consistency condition not as insurance of plan stability but as a prior consideration of plan feasibility. It is meant to address two concerns in dynamic choice behavior; (1) an agent should be able to choose a choiceworthy plan at each point and (2) the agent should never commit to a

[46] Hammond (1988b), 34.
[47] Hammond (1988b), 26.
[48] Hammond (1988b), 33.

plan which forces her to do something in the future which is not choiceworthy by her own lights.

As we have seen, the crux of the problem with dynamic choice is dynamic inconsistency - the phenomenon of preference shifts within the agent that causes him to dissent from her original strategy and adopt a different plan. Problematically, this frequently leads to suboptimal outcomes for the agent from a global perspective. While the consistency condition intends to address this issue by rationally prohibiting preference shifts, I do not think that this move is entirely successful. Even if one concedes that preference change is prima facie irrational, there could be other mitigating factors that in context make it reasonable.

Hammond and other economists tend to assume perfect rationality as a background assumption which automatically renders any preference deviation irrational. But such idealistic characterizations of agents are not particularly insightful. Consider that the entire life of an agent can be represented as one huge decision tree beginning at birth and terminating at death. It is a fact of maturation and growth that agents' preferences change, and even if it was possible to prevent that (which is in itself a preposterous assumption) it is misguided to assert that an agent ought to possess their 5-year-old preferences when they are 80 years old. In the real world there are cases in which we acquire new information or undergo morally beneficial character transformations and this principle is too rigid in its formulation to permit any distinction between good and bad cases of preference change.

The third principle is the consequentialist principle, which essentially stipulates that the shape of a tree ought to be irrelevant to the outcome realized by the agent. Hammond argues that a consequentialist behavior norm for dynamic decision making follows necessarily from and is characterized completely by the combination of three criteria.[49] First, that a revealed preference ordering exists for the agent given every non-empty set of decisions. According to standard economic theories of preference, a preference ranking is a pair of relations between two options. This just means that faced with two options, an agent will be able to pick one or the other (or might be indifferent between the two) and that this choice "reveals" her preferential rankings of the options.

Second, that this preference ordering satisfies Savage's STP extended to accommodate independent probabilities.[50] STP was examined previously in Section 1.2 and stipulates that an agent ought to ignore states of the world in which the outcomes are identical no matter what she chooses. (This does not imply other extensions of the sure-thing principle, like those proposed by Anscombe and Aumann.)[51]

And third, this ordering must also satisfy Samuelson's strong independence axiom. As explained in section 1.2, this axiom stipulates that if option A_1 is weakly preferred to option B_1 and A_2 is similarly weakly preferred to B_2, then an agent should prefer an even chance of getting A_1 or A_2 to an even chance at B_1 or B_2.[52] This is a combination of the dominance principle with the general features of independence - if every option in one lottery is at least weakly preferred to it's corresponding option in a second lottery then that first lottery must be preferred, irrespective of what rewards the particular bundles contain.

[49] Hammond (1988b), 28.
[50] Savage (1972).
[51] Anscombe and Aumann (1963).
[52] Samuelson (1952), 672.

The crucial bit about this principle is Hammond's argument that for every pair of decision trees T, T^* in which the set of feasible outcomes F are the same, the two trees are to be considered consequentially equivalent.[53] Therefore, the behavior in those two trees must also be consequentially equivalent and the structure of the decision tree must be irrelevant to the consequences of the acceptable or choiceworthy behavior. So if $F(T) = F(T^*)$ then $C(P(T)) = C(P(T^*))$ irrespective of the shapes of T and T^*.

For example, an agent may decide she prefers a cashmere sweater to getting either a Big Mac or a socket wrench. Then if she is first asked to choose between taking the wrench or moving on to the next choice, she should move onward in the tree and pick the cashmere sweater. And similarly if she is first given the option of choosing a Big Mac or moving on to the next choice, she should reject the Big Mac and choose the sweater at the subsequent node. The order in which the options are offered to her, the arrangement of the tree, ought not to sway her preferences. This assumes that only relevant factor is utility which designates final outcomes and rejects the idea that certain means to an end may be more valuable than others (insofar as the means or the path do not themselves change the value of the action - see the section 1.8).

Hammond refers to his own potential addict example[54] to further discuss the applications of this principle, comparing the original decision tree with the sophisticated version that includes an outside option to 'tie oneself to the mast' as follows:

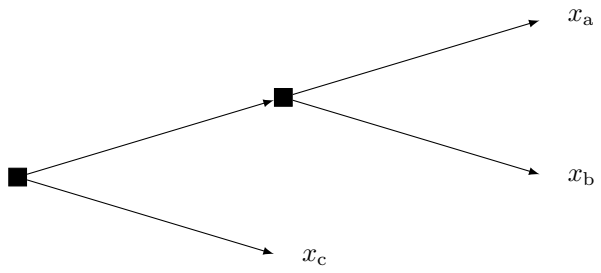

Figure 5: Potential Addict Tree

[53] Hammond (1988b), 34.
[54] Hammond 1976

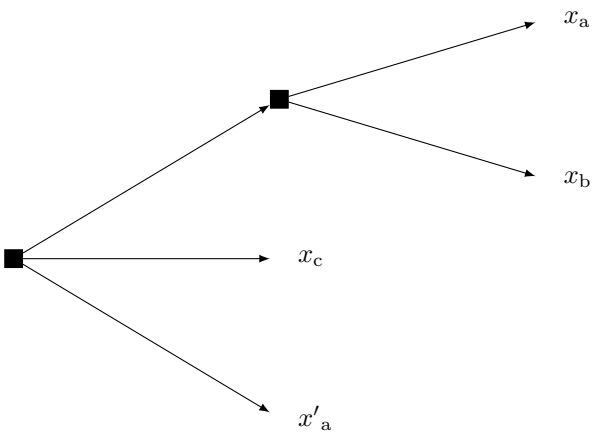

Figure 6: Potential Addict Tree with Precommitment Option

In this scenario, an agent is determining whether or not to use a pleasure-inducing drug that is known by her to be both addictive and deleterious to one's health when used to the point of addiction. In the tree, x_{a} refers to the option of using the drug up to the threshold of total addiction and then quitting 'cold turkey,' option x_{b} is to use the drug until irreversibly addicted, and option x_{c} is to refrain from ever using the drug in the first place. This tree generates the set of outcomes $\{a, b, c\}$. This case is in many ways analogous to that of Odysseus - an agent who understands that her future self will not follow through with x_{a} (rendering that option infeasible) will do the prudent thing and choose x_{c}. In the second tree, she also has option x'_{a} which would entail using the drug but taking some sort of preventative measure against her future self when she reaches that threshold such as destroying the contact information for her drug dealer or checking herself into a rehabilitation program. This option has the same consequence $\{a\}$ as x_{a}.

If the agent in question is a sophisticated chooser, she will opt for x_{c} in the first case and x'_{a} in the second since she will recognize that the only way to get outcome x_{a} is if she ties herself to the mast to prevent full addiction (otherwise she will choose x_{b}. She will therefore view that option in the first tree as infeasible. But according to Hammond, this choice behavior is in violation of consequentialism.[55] Because the consequences associated with x_{a} and x'_{a} are identical, the set of all feasible outcomes has not changed, only the shape of the tree has, and the principle stipulates that the structure of the path to the final outcomes should not meaningfully alter the choice behavior of a rational agent. This result is especially bizarre given Hammond's own admission that "sophisticated choice seems clearly the best,"[56] in both the potential addict case and this example. It certainly seems to undercut the strength of the very principle he is trying to defend as an axiom of rational deliberation.

It is difficult in this passage to ascertain what exactly Hammond thinks about this family of cases. He claims that rather than invalidating the consequentialist principle, the friction

[55] Hammond (1988b), 36.
[56] Ibid.

merely implies that cases like the potential addict should be treated as a game between two distinct persons each with their own preferences and agenda and says no more on the subject.[57] This is to assert that cases involving dynamic inconsistency are in some sense outside of the applicable domain of this principle. I am sympathetic to the intuitive appeal of this thought given the malleability of our personalities and preferences across time. But this maneuver on Hammond's part misunderstands the point of what it is to make intertemporal decisions rationally by construing them as antagonistic games. I will elaborate on a better alternative proposed by Velleman in Section 2.1.

Perhaps the best response on behalf of the sophisticated chooser is to push back on the notion that the set of feasible outcomes is equivalent in both trees. It seems clear that while consequence a is prima facie possible in both trees it is not an attainable consequence for the agent in the first tree because she will inevitably succumb to the addictive desires. One could argue that Hammond is conflating theoretical possibility with actual feasibility when the two are entirely distinct. This view implies that $F(T_1) = \{b, c\}$ and $F(T_2) = \{a, b, c\}$. Redescribing the situation in light of the qualified definition of feasibility alleviates the problem of violation of the consequentialist principle.

That this consequentialist principle is not prescriptively applicable to cases of dynamic inconsistency calls into question its viability as an axiom of rational behavior. In formulating the principle, Hammond seems to be committed to one of two accounts. One might interpret the principle as implicitly attributing total rationality (a set of idealized assumptions) to the agent such that she would never have any reason to stray from her perfectly rational desires regardless of the shape of the decision tree at hand. But this is almost unproductively tautological; this interpretation is reducible to the trivial statement that an unerringly rational agent will choose in accordance with their beliefs and desires. While true, this claim offers no insight into how an agent should choose when faced with the reasonable expectation of future inconsistency.

Alternatively, the principle might be interpreted as generating an unreasonably strong claim - that the rational agent who adheres to the principle must in all situations commit to the initial plan which conforms to their current, rational beliefs and desires, even in the face of clear evidence that their future self will dissent from that plan. That is, the agent should be unresponsive to her own beliefs about her future self. But this seems clearly irrational.

Rational agents should at all times apply Bayesian rules for updating their subjective probabilities for future outcomes and adjust their plans accordingly. Failing to update their plan at the initial node of a tree given the expectation of dynamic inconsistency on their part is akin to refusing to take into account weather predictions when dressing to leave the house for the day. If the weatherman says it will rain in the afternoon, a rational agent will consider this and their actions (e.g., packing an umbrella) will take this new information into account. The case of expectations of future irrationality should not differ procedurally from the case of expectations of rain - the agent should be sensitive to such new information and appropriately reactive to it.

Having now outlined Hammond's consequentialism, the theory to beat as it were, I next examine the different challenges to his variety of consequentialism. I will avoid the nuances between different consequentialist theories and instead focus on theoretical

[57] Ibid.

frameworks which reject expected utility maximization and have proposed alternatives to consequentialism as the prescriptive norm for rational choice.[58]

1.6 The Resolute Challenge

As previously mentioned, not everyone has been content with the response to dynamic inconsistency generally supported by the consequentialists. Some of the critics who stand directly opposed to the consequentialists advocate for resolute choice. While sophisticated choice dictates that only forward-looking features of the decision tree can rationally count as constraints on present choice, the theory of resolute choice disagrees and claims that backward-looking contextual features can also rationally sway an individual in the present. This section outlines the position taken by McClennen, who postulates resolute choice as a prescriptive alternative to the sophisticated model. Following that, I argue that features of the resolute framework make it ultimately unsatisfactory as a behavioral framework in dynamic decisions.

McClennen argues that rational choice itself fundamentally rests on two principles: the weak ordering principle (WO) and the independence principle (IND) - on this much, he and Hammond are in agreement. The former stipulates that an agents preference ordering among some set of options constitutes a weak ordering of those options just in case the ordering is connected, transitive, and context-free. This means the expansion of the initial set of options into a superset by adding new options should not change the relative preferences over the original options.

IND requires that the value of one gamble does not depend on what other gambles or risky options it is bundled with, that is, does not change from one bundle to another.[59] According to McClennen, the shifts in preference that precipitate moments of dynamic inconsistency are a result of letting go or a relaxing of either WO or IND.[60] Often, it seems the choices are especially not context-free, violating WO. Because of this characterization of preference changes, he shares Hammond's view that such shifts ought to be considered irrational.

There are four principles at play in dynamic decisions that McClennen examines. The first is simple reduction (SR) which stipulates that if there is only one initial choice node in a decision tree and the rest are all chance nodes, we can treat this problem as a simple function of the rankings of all final prospects.[61] This is distinct from NF/EF equivalencies because in this case the uniformity of chance nodes at every point past n_0 renders the decision tree meaningfully a one-shot decision followed by however many iterations of random selections.

The second is dynamic consistency (DC) - similar to Hammonds consistency condition, DC requires that choice in a decision tree be dynamically consistent at all time points in the tree.[62] This is the principle that is violated by Odyssean preference shifts. Third, the principle of normal form reduction (NFR)[63] requires that an agent's choice be invariant with

[58] See Levi (1997) and Seidenfeld (1988) for discussions of consequentialist theories which stand opposed to Hammond on several issues.

[59] McClennen (1990), 1.

[60] Ibid, 6.

[61] Ibid, 113.

[62] Ibid, 122.

[63] This is functionally identical to the NF/EF equivalence that I discussed previously, just under a different name.

respect to normalization and it should make no difference to choice whether one is presented with the normal or the extensive-form version of any plan.[64] Finally, dynamic separability (SEP) requires that in a dynamic choice, how one evaluates the tree must be independent of the context of that truncated tree - this is nearly equivalent to Hammond's pruning lemma. Importantly, no one of these conditions is logically implied by the conjunction of the other three.[65]

Having established this much, McClennen calls for a reinterpretation of dynamic choice. He notes that for the agent who accepts SR, he cannot retain NFR, DC, and SEP without violating either context-freedom (CF) or context-independence (CIND), which are well documented as irrational behaviors so this conclusion ought to be avoided.[66] McClennen correctly predicts that the sophisticated chooser will reject NFR as applicable to all logically possible plans at initial node because they will understand that NFR fails to account for the (in)feasibility of certain plans. For the sophisticated chooser "ex post choice behavior is to constrain ex ante choice with respect to a plan"[67].

At this point, he introduces the idea of resolute chooser who stands in opposition to her sophisticated counterpart by positing that constraint runs the other way too; that ex ante choice ought to constrain ex post choice behavior. This stance (supposedly) avoids violating DC like sophisticated choice, but at the expense of SEP rather than NFR.[68] The resolute chooser sees all nodes of choice points as valid continuation points. McClennen writes, "On this account, then, the sophisticated chooser can be said to regard certain plans as simply not feasible, given his projection of how he will on independent grounds evaluate certain future options; and in a parallel fashion, the resolute chooser can be said to regard certain ex post choices as simply not feasible, given the plan he adopted."[69] Whereas the sophisticated agent's choices are under constraints imposed by the nature of belief, the resolute agent's actions are constrained by past commitment.[70]

The intuition that McClennen's model of resolute choice aims to capture is that there is something amiss with a theory which prescribes a plan which is not always the one that results in the highest amount of utility. McClennen shows that an agent whose evaluative methods involve violating either CF or CIND and is sophisticated can be placed in a dynamic choice tree "in which he will be liable to choose a plan whose associated prospect is dominated with respect to sure outcomes by the prospect associated with some other plan defined by the decision tree."[71] Specifically, these are cases in which there is no external, physical constraint on the agent's behavior and the only thing stopping her from choosing the option with higher utility is herself. The resolute model claims that an agent can and should prevent her future self from changing course by looking backwards to the original plan that she adopted and "resolutely" sticking to it even when other options seem superior because her past commitment gives her a reason to do so.[72]

[64] McClennen (1990), 123.
[65] Ibid, 126.
[66] Ibid, 161.
[67] Ibid, 157.
[68] Ibid, 158.
[69] Ibid, 159.
[70] Ibid, 160.
[71] Ibid, 191.
[72] For an interesting analysis of this issue along different lines, see Sartre's story of the gambler.

There are legitimate concerns with this theory. The first is that it is not immediately certain that reasons can be backward looking in the way McClennen thinks they can be. That I committed to doing something in the past only counts as a reason for me if I am already the type of person who cares about regularity and consistency with respect to my past commitments. A related concern is pragmatic; there is no clear method for enforcing this resoluteness. Even if I decide in the present at t_0 to choose some plan p_1 and currently intend for my future self at time t_1 to continue with that plan, there exists no deterrent for my future self should they decide against p_1 and opt for another option. My current self cannot chastise, punish, or otherwise sanction my future self. So unless I am already sensitive to my past self's intentions in weighing the value of present decisions, nothing will stop me from abandoning the plan for a different one that better satisfies my new preferences.

Consider the fact that this narrow range of cases in which resolute choice works - when my future self is already disposed to care about my past self's opinions - are not really cases where there is a real risk of dynamic inconsistency. If my current self has some set of preferences and my future self's preferences will be to adhere to my past self's preferences (out of a desire to be resolute, irrespective of what those initial desires were), then for all intents and purposes I have the same preferences at t_0 and t_1. The resolute model is uninformative in this way because its success presupposes enough continuity of preference that dynamic inconsistency will not occur. But sophisticated choice is right in saying that in cases of legitimate preference shift, the only rational options are to treat certain options as nonexistent due to their infeasibility. One ought to effectively constrain the future self's available decisions, or take steps now to meaningfully change the values of the outcomes which the future self will consider.

1.7 Blatant Non-Consequentialism

McClennen is not the only voice to challenge the prescriptions of the standard model of expected utility maximization. Researchers in behavioral economics and psychology have demonstrated that there are a several examples of non-EU preference functions which fit the data of human behavior better than the standard model.[73]. The question of empirical fittingness aside, the normative goal of these alternatives must be to show that agents who adhere to a non-EU maximizing model of choice preference will not necessarily be dynamically inconsistent,[74] a common critique of non-EU theories.

Machina argues that non-consequentialist agents can in fact avoid automatically being Dutch-bookable and be dynamically consistent. Expected utility theories, because they are linear in the probabilities, mandate that preferences exhibit separability across mutually exclusive events.[75] Machina maintains that this property of separability is the most salient distinction between EU and non-EU, non-consequentialist models. Further, he divides separability into two component types: (1) replacement separability, which stipulates that if an agent would prefer to replace option x for option y in one lottery than she would similarly prefer switching in all lotteries of the same form, and (2) mixture separability, such that an

[73]See e.g., Edwards (1955), Kahneman and Tversky (2013), Hong (1983), Quiggin (1982), Hey (1984), and Loewenstein and Prelec (1992).

[74]Machina (1989), 1623.

[75]Ibid, 1627.

agent prefers option x over y in a mixture only if she would prefer x to y outright. That is, if $x \succ y$ then $(p(x) + (1-p)(z)) \succ (p(y) + (1-p)(z))$. These regulations of rational preference also extend to mutually exclusive sublotteries in a compound lottery.[76]

The Allais paradox preferences, which violate separability over sublotteries, are perhaps the most well-known case of common, intuitive, non-consequentialist preferences. Machina demonstrates that any agent who holds those Allais preferences can be presented with different decision trees in which they will act inconsistently by the standards of their own preferences.[77] It has also been demonstrated that when a non-EU maximizing agent violates mixture separability, replacement separability, or the independence axiom (i.e., the component aspects of the standard EU model) a clever agent can make book against her. If her preferences are $X \succ Y$ and $[pY, (1-p)Z] \succ [pX, (1-p)Z]$ for at least some values of Z and p, a bookie can alternately buy and sell options in an order such that the agent will assent to a sure loss.[78] (For an in-depth discussion of this type of diachronic Dutch-booking, refer to section 2.2.)[79]

Machina suggests that non-consequentialist agents take past uncertainties into account when choosing in the present, making their preference profile distinctly non-separable. Consequentialist standards are therefore inappropriate[80] because they include the "hidden assumption"[81] of separability. To motivate the reasonableness of non-separable (and therefore non-consequentialist) preferences, Machina invokes an example meant to prod intuitions, and then works to extend the non-separable, non-EU preferences to dynamic choice situations by introducing the concept of borne risk as a relevant factor to an agent's decision-making. I will assess both of these lines of argumentation and attempt to undermine the motivation for them by expanding the concept of a consequence to encompass all of the salient details about which Machina is concerned.

Machina provides the following example: consider a mother who is strictly indifferent between giving a treat to either her daughter Anna or her son Brendan and strictly prefers a fair coin toss to giving the treat to either of them outright. That is $A \approx B$ and $((.5)A \wedge (.5)B) \succ (A = B)$.[82] Now if the coin toss comes out in favor of Anna, Brendan might object that his mother is being inconsistent in light of her own preferences since she is now choosing to give Anna the treat, even though that choice is allegedly structurally identical to the initial one according to the separability constraint. Brendan will therefore urge her to flip again if she wishes to be considered rational in that light. These two interpretations of the situation are displayed below.

[76] Ibid, 1628.

[77] Ibid, 1637.

[78] Raiffa (1968), 83-85 and Seidenfeld (1988).

[79] Further, some have argued that a sophisticated non-EU maximizing agent can actually be made worse off by the acquisition of new information and are therefore averse to it in at least some situations (see Peter Wakker 1988 and Ronald Hilton 1989). This violates the value of information component of the standard model which stipulates that a rational agent should never pay money to avoid learning new information. For a complete characterization of this see Goode (1967). The introduction of new information ought to cause an agent to recalculate their forward looking plan of action. Consequentialism mandates that this recalculation be done in the right kind of way, by snipping the decision tree behind the agent and acting as if starting anew in the truncated tree according to Machina (page 1641)

[80] Machina (1989),1642.

[81] Ibid, 1639

[82] Ibid, 1643

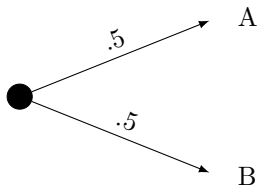

Figure 7: Mom's Decision Tree

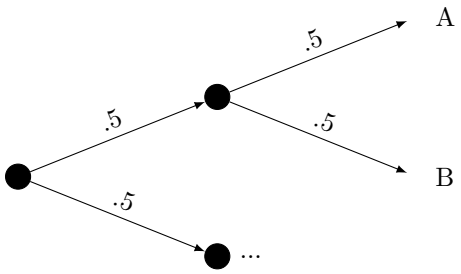

Figure 8: Brendan's Interpretation of Mom's Decision Tree

Applying the pruning lemma at the second node of Figure 8, we can see how it is (seemingly) reducible to the Figure 7. And hypothetically, Brendan could raise this objection ad infinitum until the flip comes out in his favor. Machina takes this to be a shortcoming of consequentialism insofar as it entails dynamic separability since we have the strong intuition (as does the mother in this case) that it would not be rational to flip the coin again.

My objection to Machina's suggestion is that in this case he has implicitly restricted what can count as a proper consequence. This issue is relevant to the paradox of voting that is often invoked to raise a challenge to consequentialism generally; for a rational and self-interested agent, the expected costs of voting will in almost all cases outweigh the expected benefits since the chances of their vote being the deciding one are almost nil and the inconveniences of going to vote are not insignificant. Given that fact, it is a wonder that any of us do go out to vote, and it must therefore be the result of an irrationality or a poorly calculated cost/benefit analysis.[83]

Against this critique, I and other consequentialists argue in favor of a more expansive conception of consequences or economic goods.[84] When an agent contemplates voting, she considers benefits other than the favorable outcome of the election weighted by the probability that her vote is the one which brings about that outcome. For instance, Renzi reasons that, additional to the marginal direct effect on the outcome of the election her vote will have, she will also receive a sense of gratification in having appropriately exercised her civic duty. She is the type of person for whom fulfilling her responsibilities as a virtuous citizen constitutes an important personal good. For another agent Nico, who does not value

[83]Downs (1957).

[84]See Broome (1990), Pettit (1991), and Joyce (1999), chapter 2 page 53-54.

civic participation quite as highly as Renzi, he may consider the additional benefit still too insignificant to outweigh the cost of the inconvenience to him.

Both of these agents are calculating their cost benefit analyses according to the subjective utility they derive from that additional benefit of voting. Their utility function is not $u_{\text{voting}} = f[(1e^{-10000})(\text{sway election})]$ but rather $u_{\text{voting}} = f[(1e^{-10000})(\text{sway election})+(1)(\text{fulfill civic duty})]$. Similarly, when mom gives the treat to the winner of the coin toss and rejects Brendan's sly maneouver, it is clear that mom is the type of person who values procedural fairness. That is, the presence of fairness in determining the recipient of the treat modifies the subjective value mom derives from giving it such that $u(\text{giving the treat to Anna via coin toss}) \succ u(\text{giving the treat to Anna arbitrarily})$. Fairness, civic duty, and other personal values like them, while not as easily quantifiable as monetary winnings or physical goods, still ought to be counted amongst the consequences of an action, otherwise the behavior of agents looks clearly irrational when really, the consequences are just under-described.

The under-description of consequences plagues quite a few of Machina's examples. Consider his attempt to extend the analogy of non-separable preferences to the single-agent, intertemporal case. Machina argues that consequentialism prohibits a rational agent from holding the following pairwise preferences: (Star Wars I) $\succ$ (Star Wars II) and (Star Wars I at 8 pm, Star Wars II at 10 pm) $\succ$ (Star Wars I at 8 pm, Star Wars I at 10 pm).[85] It is not the case that he is overlooking something trivial like the economic notion of complementarity and just misdescribing these goods. His critique is that there is no way to justify these preferences without violating the DS constraint.

Consider a more extreme (and admittedly morbid) example: let us suppose that I am an avid skier and I have the preferences (pair of skis)$\succ$(wheelchair). But if I was involved in a terrible ski accident in 2018 and lost mobility in my legs, it would be reasonable for me to now exhibit the preferences (wheelchair)$\succ$(pair of skis). This means we can specify a bundle of goods such that (skis in 2017, wheelchair in 2019)$\succ$(skis in 2017, skis in 2019), analogous to the Star Wars case.

In this case, I exhibit reasonable choice behavior because the context in which I have these preferences has changed drastically. Many preferences (all but our preferences over intrinsic goods) are conditional on certain background features. Past events bear directly on current consequences in many cases and one's evaluation can and ought to take those into account, just not when there exist no relevant current effects which modify the value of the available outcomes - this is what consequentialism stipulates. In the case of the ski accident, I am now a different person than I was before in virtue of having been disabled.

Similarly in the case of the Star Wars movies, I am now at 10 pm a person who has just seen the first Star Wars movie. In that way I am no longer evaluating the same good that I was two hours ago even though they look deceptively identical given Machina's shallow description of them. For the separability condition to hold correctly, we have to be examining the exact same good and in this case, although the two instances of Star Wars I are descriptively identical, they are importantly different given the fact that my memories ad experiences are different.

Relatedly, Machina argues that once an agent has begun traveling along a decision tree, she is rationally entitled to discard her original preference rankings because she has since

[85] Machina (1989), 1644.

t_0 consumed or borne some amount of risk in having previously chosen among uncertain options.[86]. Borne risk is a strange case. We might ask the agent who feels they have borne some risk - does that fact affect your ability to enjoy or appreciate the consequences in the future? It would be one thing if bearing certain risks in the past makes me jittery now and I get a bad feeling in my stomach at the prospect of bearing future risks. But Machina wants it to be the case that borne risk does not influence my current state in any meaningful way, yet has some explanative force in terms of why I choose certain irrational options. At the risk of sounding uncharitable, it seems as though Machina wants to have his cake and eat it too in this case.

Machina intends these examples to demonstrate that "if an individual informs you from the start that his preferences are non-separable (over time, over events, or over any other economic dimension), then it is inappropriate to impose separability ex post by explicitly or implicitly invoking consequentialism."[87] However, as I hope to have successfully illustrated, cases of apparent non-separability across events are remedied by correcting the problem of under-description of consequences and borne risk does not hold up well to scrutiny.

[86]Ibid, 1648
[87]Ibid, 1645.

2 A Closer Examination of Intertemporal Shifts in Desire

In the section prior, my aim was to trace the scholarly history of the problem of dynamic inconsistency and its development in the philosophical literature. As we have seen, the economists and philosophers who first researched this issue tended to assume that some sort of dynamic consistency condition (DC) was a reasonable constraint on rational choice behavior. I have previously outlined and defended against serious objections what I take to be the strongest normative position, that of consequentialism.

Having established that position, this section critically analyses the consequentialist conclusion in favor of a DC constraint and attempts to provide a slightly adjusted, alternative framework for thinking about our current self's posture towards our future self's preferences when we expect them to differ substantially from our current ones. We have already seen that it is difficult to enforce or incentivize that type of intertemporal coordination, but here I pose a more fundamental question about whether that type of combative attitude towards our future self is rational at all times.

Consider the case of a twelve-year-old boy Ryan who considers the prospect of being romantically intimate with a girl to be revolting, although he is aware of the fact that boys become suddenly infatuated with girls when they get a few years older. Ryan devises an ingenious plan to have his parents enroll him in an exclusively male military school until he is 18 years old so that he will be effectively prevented from flirting with or developing relationships with any girls.[88] Here, although he reasons in a sophisticated manner, Ryan's choice behavior is objectionable. The intuition this case frequently elicits is that there is something fitting or appropriate about this anticipated preference change and that Ryan should not be quite so antagonistic towards the expected behavior of his more mature future self.

Among the competing theories of rational choice that were examined in the first section, there is unanimous agreement that preference shifts which lead to dynamic inconsistency are irrational. Each of these theories prescribes axioms of rational choice which seek to avoid that type of preference defection. There are multiple possible positions one might take towards one's future self. The myopic chooser advocates a dismissive or neutral attitude towards her future preferences. In contrast, it is argued within the sophisticated choice paradigm that sometimes one ought to treat one's future selves paternalistically and take action to constrain them. This recognizes that one cannot, by a mere act of the will, determine what one's future self will want. And there is a third, unexplored option which is deference towards one's future self. Cases like Ryan's indicate that such an attitude might be warranted in some situations.

This section will first investigate what the point of rationality is and what exactly we mean when we use the word in this nuanced and philosophically technical way. The literature on dynamic choice problems is saturated with the word 'rationality,' but perhaps too little attention has been given to the project of constructing a unified, cohesive definitional framework of rationality. I will argue in favor of an account put forward by Velleman and will strengthen it with additional considerations.

[88] Gauthier (1997), 17.

Next, I examine Bas van Fraassen's arguments on what epistemically conscientious agents ought to do in cases where they learn that their future self will have different credences than they currently possess. He introduces a reflection principle for an agent's beliefs which stipulates that agents in the present moment ought to align their beliefs with those of their future self. I will sketch the most compelling version of this argument here, taking into account objections and suggested improvements to his principle.

Finally, in the sections following, I will examine whether or not each aspect of the epistemic reflection principle can be adapted to fit the case of changes in basic desires. Each of the component pieces will be examined, assessing the strengths and shortcomings of each in order to determine the success of this analogy holistically.

2.1 Rationality as an Intrapersonal Narrative

Velleman writes on the origins of the normative dimension of rational choice theory, specifically the axioms that are entailed by the representation theorem. As a reminder, the representation theorem states that if an agent has preferences which satisfy certain axioms, then a pair of functions assigning utilities and subjective probabilities to every outcome and state respectively, can be constructed. That is, the agent's unreflective choice behavior can be represented as if it deliberately conforms to those utility and probability mappings onto the decisions he is facing. This is sufficient to demonstrate that the agent will behave as if she is maximizing her expected utility with respect to those probabilities and utilities. Velleman acknowledges that this proof is sound but thinks the interpretation of the theorem is unsatisfactory.[89]

Among his concerns, the one most germane to this project is that it is not immediately intuitive how the rational choice theorist can find within the axioms of rational choice a prescriptive norm of rational preference. The imperative to maximize expected utility is more intuitively palatable when one presupposes that each available outcome already has a value determined by the deliberating agent, providing her with reasons to prefer the ones which are most likely to promote whatever she already values.[90]

Unfortunately, this is not the case. The numerical values that we assign to outcomes are merely "as if" values constructed out of the agent's preferences according to revealed preference theory. According to this view, there is a weak axiom of revealed preference which entails that if some good or bundle of goods X is chosen over another bundle Y when both are affordable and available to the agent, then that choice by the agent reveals that $X \succeq Y$. The strong axiom of revealed preference (SARP) is an expansion of WARP which prohibits indifference between the options. From the same choice behavior, WARP allows us to conclude that $X \succeq Y$ and by SARP, we conclude that $X \succ Y$ since it bars cases in which $X \approx Y$.[91]

On the subject of revealed preference, there is a divergence in interpretation between rational choice theorists. The standard economic interpretation is that the choice patterns directly reveal preference - this is the language used in the original formulations of WARP and SARP. From this behaviorist perspective, choices are thought of as observable properties

[89] Velleman (1993), 231.
[90] Ibid, 233.
[91] Samuelson (1948).

of agents and preferences are "operationally defined" in terms of the agent's overt choice behavior.[92] In contrast, the cognitivist account of WARP/SARP entails that $X \succeq Y$ or $X \succ Y$ is a fact about the agent's psychology that is used to explain why the agent chose X over Y. In this case, the preferences rationalize the overt choice behavior, although not all preferences are necessarily revealed by the choice behavior.

These comparative utility rankings of outcomes are not antecedently available to generate reasons for having the preferences which brought about that ordering in the first place. The standard model does not include explanations of the mechanism by which the preference ordering of different options comes to be available to the agent. On the cognitivist interpretation of revealed preference, that the agent does have certain preferences helps us as spectators make sense of their behavior but it does not provide a higher level justification of any kind for having those preferences (nor does the behaviorist account for that matter). Velleman's burning question is "why you ought to prefer things that promote what you value in this post facto sense."[93] That is, why is anyone correct to prefer what they prefer in the first place? The worry is that an expected utility value that emerges from probability and utility functions cannot have a justificatory role in a choice without circularity.

This criticism, framed as is, is only directly applicable to a behaviorist interpretation, which extracts probabilities and utility functions from the direct observances of behavior. On the cognitivist account, beliefs and desires causally explain why an agent has certain dispositions to make binary choices and claims that tendencies towards a particular behavior reflect the agent's underlying beliefs and values. The cognitivist might say that the agent is correct to prefer what she does because she takes herself to have causally-relevant reasons for the preferred action.

There is still, however, the further question for the cognitivist of why beliefs and desires ought to obey consistency conditions like transitivity, independence, etc. That is, we can ask in virtue of what an individual belief or desire rationalizes an action. But the deeper level of analysis is to understand why the cognitivist decision theorist claims that it is a fact of rational people that their beliefs and desires don't lead to preferences or choices which violate the previously stipulated list of rational axioms. If beliefs and desires are to have this causally explanatory role in rationalizing preference, we must uncover what it is about the axioms that they obey which makes them rational. Velleman's account of rationality answers these concerns for both the cognitivist and behaviorist families of interpretation.

Velleman first attempts to establish that an agent can indeed be (in)correct in her preferences in the first place by examining the phenomenon of intransitivity of preference, e.g., $A < B$, $B < C$, and $C < A$. This type of preference structure has been shown to be irrational through 'money pump' arguments which demonstrate the agent with intransitive preferences will willingly sacrifice money for no reason. They will pay some small amount to trade A for B, some small amount to trade B for C, and some small amount to trade C for A. At this point they have the exact same outcome they began with but have less money.

Velleman then demonstrates that such conclusions can successfully be avoided with a strategy of contextualization, which changes $A < B$ to A-not-B and this structure builds on itself in cases of sequential choices. Thus, the agent's actions in the

[92] Quoted from Jim Joyce's lectures in PHIL 443 in the Fall of 2017.
[93] Velleman (1993), 234.

original case become B-not-A, C-not-(B-not-A), and A-not-(C-not-(B-not-A)).[94] A better phrasing might be to say the agent preferences are A-when-A-and-B-were-options $\succ$ B-when-A-and-B-were-options and so on. Described in this way, Velleman argues that the agent does not technically return to their original state and does not violate the axioms of the representation theorem.

The intuition that he means to capture is that while this contextualization strategy does not violate the letter of the theory, it does violate the spirit of it.[95] I happen to disagree with the former half, not the latter. I believe that a consequentialist is committed to saying such broad strategies of contextualization still constitute a violation of the axiom of transitivity and therefore a violation of the letter of the representation theorem. Contextualization strategies insist that one can differentiate between otherwise identical goods based on the good's transactional hertory or how it was acquired, updating the value of the good in a backwards looking way. The consequentialist, in contrast, must claim that good A is good A in all situations, not just in the base case. Its value ought to to be rigid and resist fluctuation as the agent makes successive trades of goods.

Faced with a money pump style dynamic decision, this contextualization maneuver entails a violation of the principle of dynamic separability (i.e., the pruning lemma). The contextualizing agent has option A as her final outcome both at the initial node and at the terminal node. whether or not she came to possess A by trading it for C or (C-not-(B-not-A)) or ten gold coins or was gifted it by a five-headed dragon named Dave is irrelevant. Or, to reframe Velleman's somewhat clunky terminology, if I prefer (A-when-A-and-B-were-options) to (B-when-A-and-B-were-options), I should also prefer A to B in all head to head matchups. This would include against (B-when-B-and-C-were-options), (B-when-B-and-(A-when-A-and-C-were-options)-were-options), etc. This goes for any variation of semantic contextualization. How one comes to possess some good or experience is irrelevant in determining the value of that thing for the consequentialist because such features are fundamentally backwards looking (unless the attaining of it in some way transforms its value as in the case of sentimental heirlooms).

That being said, contextualization does also violate the spirit of the representation theory and I am inclined to endorse Velleman's account as to why that is. He claims that there must be some ulterior reason for having preferences which obey all the axioms of rational choice. In the original case, preferences are represented as three combinable values - in the contextualized case they become six disjoint, unrelated values.[96] Part of the purpose of adhering to the axioms is to ensure preferences are represented in as concise, economical, and powerful a way as possible.[97] This ensures that they will be "synoptically describable," which Velleman takes to be the "ultimate basis of the axioms' normative force."[98] By that he means that the preferences will make a certain kind of sense in terms of the agent's actions fitting together such that each one helps make sense of the others. Velleman thinks that kind of coherence of preferences is synonymous with their being rational.

[94] Ibid, 236.
[95] Ibid, 237.
[96] Ibid, 246.
[97] Ibid, 240.
[98] Ibid, 241.

This argument is partly an inference to the best explanation given the dissatisfaction with contextualization and partly an appeal to intuition; when people like Savage claim that intransitivity should strike us as "uncomfortable" it is because we should intuitively want to organize our preferences into some coherent posture towards the world. Using the analogy of narrative intelligibility in a fictional novel, Velleman argues that it does not matter to the comprehensibility of a story why some particular event happens or not, but rather how well all of the events can be "grasped together" as comprising a cohesive vector towards one type of outcome or another.[99] For example, the fact that Harry Potter has a lightning shaped scar on his forehead in itself does not change how intelligible the fantastical plot is, but that fact about him does make sense given the background context of J. K. Rowling's narrative.

There is an obvious benefit to this account of rationality which is that it explicitly attributes to rationality its motivating force. Synoptic describability provides an account not only of why we *ought* to behave rationally in a theoretical sense and why we frequently *do* in fact feel motivated to behave so. Savage's framework assumes that rational behavior is preferable merely in virtue of its being rational. This sort of circularity, while perhaps not vicious, asks us to consider rationality's underlying motivation to be a brute fact about the world. Velleman's account meshes better with our intuitions as agents and evolutionary facts about our psychologies.

It is not for the purpose of obeying some abstract imperative of rationality whose motivating force is sui generis that we want to make rational decisions. We desire to act rationally because it helps us and others make sense of our own internal preferences and choice history. That kind of coherence is attractive for the purposes of self understanding and understanding of others. Without perceiving our actions as cohesive in virtue of some narrative thread linking them together, our own conception of identity is undermined and it becomes difficult to view ourselves as a singular person that persists across time. Synoptic describability is also instrumentally valuable for the purpose of social cooperation. Ability count on others acting reliably in virtue of adherence to certain types of behavior enables the formation of trust and reliance, which are the precursors to efficient social cooperation and participation. Hence, our motivations for behaving rationally to the best of our abilities.

Next, Velleman makes an important distinction between two conceptions of rationality in rational choice theory, weak and strong. The weak version stipulates that an agent's actions must be intelligible in light of their preferences, but makes no judgements about the fittingness of those preferences in the first place. The thick or strong version makes more substantive demands on the agent, stipulating that the disposition of the agent's preferences must also be rational.[100] This distinction is necessitated by the question of whether or not masochistic, myopic, or otherwise misguided preferences can still count as rational. The weak conception answers in the affirmative, as long as those preferences are consistent and internally coherent. But according to the thick version, the having of those preferences is irrational in the first place.

Velleman thinks that the criterion of synoptic describability is more closely aligned with the thick version because it mandates a background explanation of why one ought to have

[99] Ibid, 245.
[100] Ibid, 247.

certain preferences in the first place.[101] I disagree here; in making this move he is conflating tiers of evaluations with types. The weak criterion is strictly rational and is concerned with consistency. The other conception is not more stringent but rather introduces a different quasi-moral criterion to assessing the rationality of preferences. The thick conception of rationality insists that we apply evaluative labels like 'good' and 'bad' to the dispositions underlying an agent's preferences. And it remains an open question whether that type of evaluation is within the appropriate purview of rational choice theory (although I will be taking up this issue in earnest further on in this thesis).

Ultimately, Velleman's proposed framework for thinking about rationality accurately captures many of our intuitions about what it means for us to talk about the rationality of an agent. However, it leaves me with one broad, lingering uncertainty. Contextualization is both in violation of the axioms of rational preference and in violation of our intuitions - Velleman's criterion for rational coherence addresses both concerns. But there are other theories which violate certain axioms of rational preference like Buchak's risk-weighted expected utility theory.[102] Cases like these are more difficult because they violate the axioms in such a way that gels with many people's intuitions, seeming to violate the letter but not the spirit of the theory. I am not sure how Velleman would respond to this nor am I sure how I should respond on behalf of his theory. It is a potential weakness of the Velleman approach that one can devise just as good a narrative with Buchak's REU theory as with the SM, even though the two rely on fundamentally different conceptions of what it means to make a rational choice. While this issue merits further philosophical treatment, for the purpose of this work I will be accepting a conception of rationality similar to Velleman's moving forward.

2.2 The Case of Changing Beliefs as a Proxy

To begin understanding how to address future changes in preference it will be useful to begin by examining the analogous case of intertemporal variance in belief. Van Fraassen has explored the issue of how an agent ought to respond to expectations that her future self will hold beliefs different from those of her current self. The main question van Fraassen sets out to address is whether or not a person can be rationally justified in believing something that is neither entailed by her past beliefs alone nor by her past beliefs in conjunction with current evidence available to her.

James, as cited by van Fraassen, defends a pragmatic and somewhat voluntarist conception of belief acquisition.[103] Van Fraassen is interested in the argument from James and points to the example of Darwin's theory of evolution which, while an especially strong theory, still has a nonzero probability of falsity. James says that there are pragmatic reasons for willing oneself to believe the theory despite its potential fallibility and those who refuse to do so are foolish. If we were to restrict the set of beliefs to which we assent to only those which can be deductively proven to be correct, we would be left with a very bare bones noetic structure. Therefore, it is in our best interests to also believe things in which we have less than perfect evidence.

[101] Ibid.
[102] Buchak (2013).
[103] James (1979).

I understand the idea behind this sort of pragmatism, but think that this Darwin example is an especially feeble since few philosophers hold the position that infallible certainty is a necessary condition for a belief to be justified. Bratman reminds us that we are social creatures whose behavior patterns frequently require careful coordination.[104] A better motivation for accepting fallible beliefs is the need to succeed at simple coordination games. For instance, I may wish to study in the Tanner Library later this evening with my friend Kelsey. Even if she has told me she will be there, it in no way follows deductively that she is, in fact, there. But even if I am not convinced, the available information and my hope that she is there together make it reasonable for me to choose to go to Tanner. In viewing action as in some sense reflective of underlying beliefs insofar as it is a commitment to them, it seems in this case I have willed myself to believe in something for which I have inconclusive evidence (Kelsey's presence in Tanner). Kelsey must go through a similar reasoning process. Only because we both assent to the assumption that the other person is in Tanner will we each show up, successfully coordinating.

Imagining a hypothetical case in which a rational agent is asked to gamble on the content of her future self's beliefs, van Fraassen demonstrates that anyone asked to wager on whether or not she will come to believe a certain proposition will certainly lose money as long as her current credence in a proposition differs from her expectation of her future credence in that proposition.[105] If the agent is certain that her future self's credence in X is p then her current credence for X should be p. If, on the other hand, she is uncertain what her future self's credence in X will be, her current credence should be her current expectation of her future credence for X.

To motivate this, van Fraassen demonstrates that she will find herself vulnerable to Dutch-booking as long as she violates these rules. The formal argument, which demonstrates the inevitability of the agent's vulnerability from asynchronous credences, is a form of a Dutch book (DB) argument - a gambling situation designed in such a way that the agent's own preferences and beliefs will lead her to accept a successive series of gambles, each of which she regards as fair (i.e., she is indifferent between having and not having the bet).[106] But cumulatively, the book of bets entails a certain net loss for the agent and a certain net gain for the bookie.

Suppose I learn that in a year from now my future self will believe in Darwin's theory of evolution with probability 0.9. Given that information, assume that my current credence for the theory is 0.78. This new information gives me evidence in support of Darwin's theory, but not conclusive evidence if at the same time I am not totally convinced my future self will be right. Incidentally, it does not matter whether the Darwin's account of evolution is actually true or false. I can now be trapped in a DB with certainty - a Dutch bookie can offer me a series of bets sequentially which will certainly diminish my monetary funds. In doing so the bookie deals a blow to my qualifications as a rational agent.

This situation leaves the agent with two options, according to van Fraassen; either she can refuse to form opinions about the reliability of her future judgements or she can have an exceptionally high opinion[107] of the judgements of her own future selves - in other words, have total confidence in her future selves. The first option has been defended by some who

[104] Bratman (1992), 2.
[105] van Fraassen (1984), 238.
[106] See de Finetti (1937).
[107] van Fraassen (1984), 243.

do not accept that higher order degrees of belief can constitute propositions. Proponents of this response insist that it is semantically vacuous to ask about the credence one has for one's own credences. This account is not plausible according to van Fraassen.[108]

The second approach, the one favored by van Fraassen, mandates total confidence in one's future self. This is expressed in her reflection principle (REF): $p_t(A \mid p_{t+i}(A) = r) = r$. This principle claims that an agent's "subjective probability for proposition A, on the supposition that her subjective probability for this proposition will equal r at some later time, must equal this same number r."[109] So if a rational agent knows she will assign probability r to an event A in the future, she should assign probability r to A right now. He suggests that this principle constitutes an additional criterion for rational belief and rational actions.

The motivation for accepting the principle and sticking to it reliably is that violations of REF necessarily make an agent vulnerable to Dutch-booking. Suppose that Josh has a rather low opinion of his future self's predictive capabilities when it comes to Michigan football; he thinks that the probability that Michigan beats Ohio State (W) is .25 on the condition that on the day of the game his optimistic future self will assign a .5 credence to the same proposition (W). Formally, for Josh (1) $p_t(W \mid p_{t+1}(W) = .5) = .25$. We must further stipulate that Josh assigns some probability now to the proposition that his future self will indeed hold that credence - (2) $p_t(p_{t+1}(W) = .5) = .2$.[110]

Now a Dutch bookie who knows those two facts about the agent's beliefs and nothing else can offer Josh two bets. The first bet (A) is won by Josh if his future self does indeed assign a .5 credence to (W) on the day of the big game. Since he thinks the likelihood of that happening is .2 and thus will be willing to take this gamble at 4-1 odds and put up \$2 to a bookie's \$8.[111]

The second bet (B) is conditional on the agent winning the first bet, that is, conditional on future Josh assigning probability .5 to (W). If this condition fails to obtain, the bet is rendered null and neitherparty wins anything. If the condition does obtain, then Josh wins the bet if Michigan football loses (¬W) and the bookie wins the bet if Michigan football wins (W). We know that Josh takes the probability of (W) given his future self's hypothetical credence is .25 and therefore, he views his chances of winning this bet to be .75. Based on this, he will agree to a gamble that puts his \$30 against \$10 from the bookie (1-3 odds).

Now that Josh has accepted the bets, the bookie will wait until gameday, when he will see whether $p_{t+x}(W) = .5$ or not. If his future self at time t+1 does assign that credence, $Josh_{t+1}$ wins the first bet and gets \$8. The second bet is now also fair game and will be determined by the outcome of the game. Now the bookie offers $Josh_{t+1}$ a third bet (C) in which he wins if Michigan football wins (W). Since $Josh_{t+1}$ thinks the odds are evenly split for that proposition, he will put up \$20 to the bookie's \$20.

At this point, the bookie has trapped Josh into a sure loss situation with no prior knowledge about him other than his beliefs (1) and (2). As the decision tree below illustrates, no matter the outcome of this situation, the bookie is guaranteed to come away with a \$2 profit and Josh is sure to give up \$2.[112] Below is a visual representation of this Dutch book in the extensive form.

[108] Ibid, 244.
[109] Ibid, 256.
[110] The numerical values used for these probabilities are identical to Christensen (1991), 232.
[111] Christensen (1991), 233.
[112] Ibid, 234.

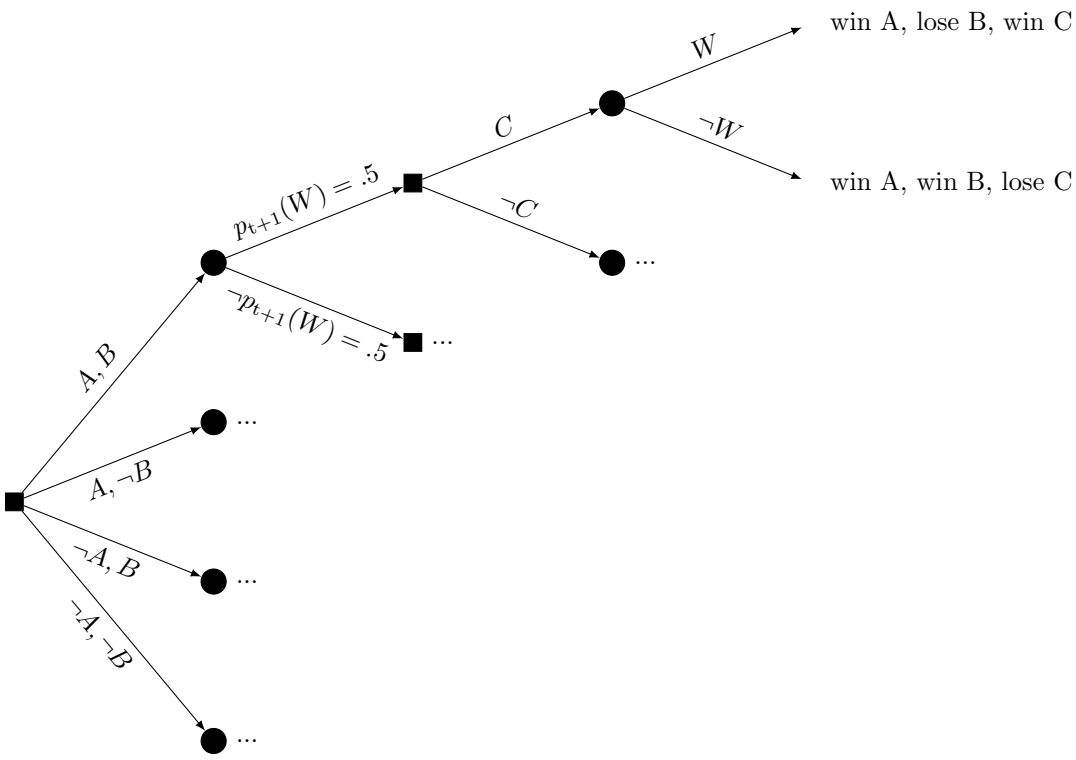

Figure 9: Extensive-Form Dutch Book

It is a mathematical fact that this sequential Dutch book strategy is possible in every single situation in which the agent's belief (1) disobeys REF. Unlike in the static case in which an agent's collectively exhaustive subjective probabilities across all outcomes total a value less than or greater than 1, in this case, the agent's beliefs need not be inherently irrational by themselves at any individual time point. Additionally, it is not necessary for the Dutch bookie to know anything at all about the agent other than (i) her preferences at the onset and (ii) the method in which she updates her beliefs to take into consideration new pieces of information (i.e., whether or not she adheres to REF).

The reflection principle is certainly not uncontentious but I will restrict my assessment to a few main doubts regarding van Fraassen's conclusion. First, what if our future self is epistemically impaired? This is the primary concern with his principle. Christensen uses the example of an agent who has ingested the drug 'LSQ' which, while not harmful to the body, does bring about in those who consume it an overwhelmingly strong belief that they have the ability to fly.[113] He claims that not only is reflection irrational in this case, but an agent qua rational is obligated to be irreflective.[114] Christensen further stipulates three things in this case. First, it is not necessarily irrational to opt for a temporary net loss in

[113]Ibid, 235.
[114]Ibid.

rationality if, for example, LSQ is a stimulant that causes euphoric sensations and one takes the proper precaution to stay away from the roofs of tall buildings. Second, the example does not necessitate that the agent suffers a net loss in rationality - it could be that LSQ actually heightens all of the brain's other faculties. And third, it is not necessary that the agent opted to take LSQ, it could have been that her normal Kool-Aid was accidentally switched with her friend's and this friend willingly laced her own Kool-Aid with LSQ. These stipulations are meant to show that the agent need not be irrational to be irreflective.[115]

The standard pushback against this type of objection to REF has been to argue that there are only certain conditions under which REF is intuitively plausible, even if it is at all times 'rational' inasmuch as it is necessary to avoid vulnerability to Dutch-booking. Skyrms for example, postulates that the rational force of the requirements for dynamic reflection depends on the kind of epistemic situation in which the agent finds herself.[116] Bayesian coherence does indeed require an agent to believe in the value of knowledge,[117] but according to Skyrms it only applies to cases in which a learning experience occurs.

In cases of generalized "black-box" learning - in which there is no external observational input, the agent just thinks about the matter at hand and updates based on her own thoughts - Skyrms thinks the value of learning theorem is still applicable. But he writes nevertheless we must find some way to distinguish between the kind of situation we expect to be a learning situation and kinds we expect to be "brainwashing, delusional, or otherwise epistemologically pathological."[118] Skyrms thinks that the agent does take herself to have undergone learning experience just in case her beliefs obey the reflection principle. This relationship is circular, but not viciously so.

In agreement with Skyrms's distinction, I argue that we ought to rationally defer to our future self's judgements only in cases when we believe that some sort of learning experience has occurred and that our future self is situated in an epistemically superior position relative to the present moment. Not all future selves are created equal and it is important that one is reflective only after some deliberation. That is, a rational agent must reasonably consider the background context in which her future self has decided to alter her credences. If it was the result of access to new information, she should certainly be more inclined to adjust her current beliefs. This actually fits well with what the standard model prescribes in terms of the value of information.

To motivate this distinction in the diachronic intrapersonal case, we might consider the synchronic interpersonal case by analogy. Samiur has been experiencing a persistent cough and sore throat for several days that he thinks is a common cold with probability .5. He tells his friend Yeager who suggests with .9 certainty that it is probably just a passing virus and all he needs to do is stay hydrated and get plenty of rest. Yeager happens to be an artist and Samiur decides to play it safe and consult another friend Julia, who is a doctor. She echoes Yeager's initial diagnosis with probability .9 and Samiur, contented that he has nothing serious to worry about, goes home, brews some green tea, and prepares to rest and recuperate.

As it happens, the opinions of Yeager and Julia were identical and they were both right. But Samiur would have been reckless to naively trust the diagnosis of someone without a

[115]Ibid, 236.
[116]Skyrms (1990), ix.
[117]Savage (1954 –chapter 7 and appendix 2), Good (1967) with proof
[118]Ibid, 97.

medical degree and he was right to seek out Julia's second opinion. Another way to frame this is to say that the value of the belief being offered to Samiur was contingent on who exactly was offering it up. The epistemic status of the person who is the source of a belief matters to the rationality of accepting that belief. Samiur should adhere to Julia's health recommendations and Yeager's aesthetic judgments because in those cases they can have expertise or have learned things which Samiur has not. Consider that this is analogous to the relationship an agent has with her future self. She can view that future person like Julia or like Yeager; she may view our future self as qualified or unqualified from her current perspective. And this fact ought to modulate her degree of cooperation with her other self.

The second problem with van Fraassen's model lies in his implicit endorsement of William James's voluntarist theory of belief. Even if I come into reliable information regarding my epistemically superior self's future beliefs and am rationally motivated to change my beliefs, it does not follow that I can change them at will, at least not easily. The willing of the belief is necessary because the mere learning of the fact that my future self assigns a different credal value to some proposition than I currently do does not automatically trigger a rationality module in my brain that readjusts that belief. This type of new informational input can certainly sway our beliefs but it does not, I think, in most cases bear upon on us as forcefully as other stimuli, such as witnessing something with our own eyes. There are psychological barriers to this process happening automatically and totally.

A full endorsement of van Fraassen's view implies commitment to the notion that selective belief changes are features of rational choice. Even though van Fraassen is willing to admit that one cannot always correct violations of REF, he maintains that irreflectiveness is an irrational feature of one's thinking. But it seems silly to suppose this when it is in violation of the 'ought implies can' imperative. This is an issue I will revisit later on with regards to my own suggestion.

2.3 What's the Point of Dutch Books Anyway?

In castigating cases of dynamic inconsistency, philosophers and economists are usually quick to invoke Dutch book arguments to make broad claims about the irrationality of dynamic inconsistency tout court. Ramsey, who first noticed this relationship, considered it to be symptomatic of a "deeper pathology."[119] It is demonstrable that in all cases of violating REF, an agent makes herself vulnerable to Dutch booking. But as Skyrms demonstrates, sometimes irreflective behaviors are necessary in order to prevent one's future self from deliberating to catastrophic conclusions.

If our future self is impaired, then we are faced with the choice of either violating REF or leaving ourselves open to DB. In these cases, one might reasonably ask whether we ought to obey REF regardless. One might press further and ask what exactly the rationally motivating force of DB arguments is. We must assess whether or not this kind of exploitability over time really does constitute an epistemic flaw.

One common misconception about DB arguments is to assume that it is a part of the rational agent's goal in the first place to avoid succumbing to a DB or other traps like it. This is not the case - the goal for the agent is actually to walk away from her decisions with as much money as possible or to maximize her wellbeing.

[119] Ramsey (1931), 109.

Given an indication of one's future self acting foolishly, the agent finds themself balancing the imperative to maximize expected utility and the competing, purported imperative to avoid DBs. It is of course reasonable to avoid a DB if one takes their future self to have benefitted from a learning process. But in cases of less-than-capable future selves, the right strategy seems to be to hedge your bets against the anticipated actions of your future self. It is not always rationally impermissible to accept a sure loss if it prevents a comparably larger expected loss. In those cases it is better for the agent to bite the bullet, so it is not obvious that REF is at all times a rational or reasonable principle to follow.

There are other cases in which the principle of Dutch-bookability generates claims of irrationality that are a bit odd. For example, a married couple who shares a bank account can easily be Dutch-booked if there is even the slightest bit of dissonance in their desires for how to spend their money.[120] That does not render their marriage fatally irrational, indeed it would be improbably odd if each of their desire profiles were mirror images of the others. Instead a healthy marriage ought to have some give and take, with each partner listening openly to each other and deferring appropriately to the better informed one on spending decisions. I argue we ought to think about the cultivation of healthy intertemporal cooperation in much the same way.

2.4 Extending the Analogy: The Case of Desire

Dynamic consistency is the phenomenon that occurs as a result of intertemporal preference shifts generally. The nature of an agent's preference orderings is a function of two variable inputs, unique to that agent; her subjective credences (beliefs) and her subjective utilities (desires). The kinds of preference shifts that are observed to cause dynamic inconsistency can plausibly be the result of changes to either one or both of those types of input. Having examined van Fraassen on the subject of shifts in belief, I now turn my attention to examining the case of shifts in desire. To my knowledge, this is not a subject that has been broached in the literature, and my hope is to give it as thorough a treatment as possible.

When imagining that an agent has undergone a change in desires or in their evaluations of the attractiveness of different actions, an important clarification must be made. I distinguish between two types of cases in which there is a change in desires or evaluations of actions between now and the future which can reasonably be regarded as the product of a learning experience or improvement similar to what Skryms identifies in the case of epistemic reflection. (1) There are cases in which a person's future preferences are simply better informed than their past self's with respect to certain relevant facts. These agents are epistemically better off and it is fairly easy to show that standard EU theory can accommodate these cases without additional theoretical maneuvering.[121] (2) Separately, there are cases in which the changes are to an agent's basic desires. In these situations, the person's most fundamental inclinations towards some prospect is different.[122]

[120] This case was suggested to me by David Manley in conversation and I find it an apt illustration of the limitations of Dutch book arguments.

[121] Skyrms (1990), Ramsey (1931), and Good (1967).

[122] Cases where they become more fitting as Elizabeth Anderson says or become more in tune with the good as a Platonist or Kantian might assert.

Of these cases, we might think that at least some of the time, these new preferences are in some respects 'better' than the old ones. These are cases in which Anderson argues that desires become more fitting insofar as they have become more in tune with, or are better approximations of, the good.[123] This is the type of case in which our 12-year-old protagonist Ryan finds himself. As he ages, his preferences are altered and he becomes more predisposed to feel sexual attraction towards others. But it is not the case that he reads books on sexual awakening or learns in school that sexual reproduction is healthy for adults and important for replenishing the human population. These are things he could have easily been told (and thus learned) when he was 12 years old. Rather, his attitude towards the same piece of information has by degrees morphed into the completely opposite stance.

Another, slightly less intuitive case could be that of developing a taste for fine wines. Newly-minted wine drinker Lyndsey may be told by an older family member that the epitome of good wine is an aromatic 2012 Syrah from the Côtes du Rhône region, yet she may initially prefer the taste of a cheap Rosé off the bottom shelf of the grocery store. But Lyndsey also reasonably believes that if she were to be exposed to a variety of high quality wines (selected by a competent sommelier, perhaps) her preferences would be altered and she would then prefer the taste of the finer wine.

It is tempting to say that this family of cases is really just another instance of learning new information and is not qualitatively different from the type of epistemic learning experiences that were discussed previously. I reject that idea that these shifts are reducible to acquisitions of new information. Lyndsey might believe at first that the cheap Rosé is the better of the two wines and after exposure will believe the opposite. But the modification of the belief structures is secondary to the underlying change in desires. That is, when asked to substantiate those beliefs, Lyndsey will not reference external facts about the world, but rather facts about her own internal tastes and predispositions which justify the beliefs.

It seems to me that the standard model just does not discuss this possibility that I might from my current perspective might recognize that there are certain changes in desires that I could undergo endogenously (i.e., not simply due to the acquisition of new information) which would leave me better off in terms of my preferences. Even if not totally endogenous, they might come from without yet still not be mediated by changes in belief.[124] Presumably this inclination on my part results from my imagining that some sort of process of maturation or positive realignment of desires will take place in that time.

In most cases, we as agents have an intrinsic desire for our own well-being and by extension a desire for the means to those ends. For that reason, I take my current desires to be the best possible approximation of what impulses are most conducive to that end. If we are to be affirmative advocates of reflective desires as well as reflective beliefs (at least in some cases) we must make sense of the idea that an agent can recognize that there are desires other than the ones I currently posses, the contents of which I need not necessarily know or be able to identify, which would better conduce to my well-being than my current ones.

It is clear that analogous reasoning takes place in the purely epistemic domain. Most of us recognize that there are all kinds of things we do not currently know, and if we did know them we would be better off. Take for instance, the contents of an upcoming exam

[123] Anderson (1995).

[124] This suggestion from Jim Joyce.

or the amount of traffic on the freeway I have to take to get to work. I accept from my current perspective the hypothetical claim that if I were to know these facts about the world then I would certainly be better off. This is an admission on my part that there are beliefs, additional to or wholly separate from the ones I currently possess, that would be more conducive to my well-being than the beliefs I have currently. Moreover, it is not necessary for me to know the contents of these beliefs for me to take this pro-stance towards acquiring them. The mere fact that some information will be on an upcoming exam gives me a reason to wish to acquire it and my lack of access to the contents of the beliefs I would acquire in no way dampens my interest in acquiring them.

The case of desires is slightly more complicated, and does not share the same level of tidy intuitive appeal. A reflective desires principle might look something like this: $u_{\mathrm{t}}(x \mid u_{\mathrm{t+i}}(x) = n) = n$. Let us call this principle $\mathrm{REF_d}$ (d for desires). This reformulation of van Fraassen's REF perhaps does not do justice to the robust idea we mean to convey by talking about shifts in desires. This axiom conveys only information about the actual number of utiles n that the agent assigns to a particular outcome. Because this formulation $\mathrm{REF_d}$ pegs specific utility values to individual basic goods, one might worry it does not capture the substantive internal process of value assessment and differentiation across classes of basic goods.

Take sneakers for instance - Matisse may value each model of sneakers differently depending on several variables. Some of these like brand and color are easily categorizable while others, like brute facts about Matisse's aesthetic tastes, are not. The unique combinations and relative weights of each input variable allows her to compute customized subjective utility comparisons. To mandate reflection of the final utility output is not to divorce it from the computational process which generates that output. It is to treat the utility quantities as indicative of all of the underlying variables such that, if we were able to compare a wide range of utilities across goods of different types, we would be able to construct an idea of her subjective value of a variety of goods and features. What we want is a model of Matisse as joined with her future self in terms of their evaluating processes.

In terms of an agent's judgements, they ought to result from the right kinds of processes. In the case of wine tastes for instance, those judgements cannot simply be read out of a book and internalized. These evaluative stances need to be justified in virtue of their integrity and the appropriateness of their formative processes. My hope is that this formulation is flexible enough to accommodate these concerns and also apply to both minor and severe shifts in desires. And it is also sufficiently specified that it would apply only to cases of basic desire shifts, not merely changes in information.

It is immediately clear that this case is harder to motivate than the original REF formulation, but I want to provide an example to suggest that it might at least be possible. Consider once again our intrepid explorer Odysseus. Suppose that in an alternate possible world, Odysseus is not debating whether to sail near the Sirens, but rather the Island of the Tasty Pies. The inhabitants of this island, instead of luring sailors to their untimely demise, are simply experts at baking the most delicious pies in the entirety of the Aegean Sea. The smell of these pies is so powerful that any sailor who breathes in their aroma immediately wants to veer off course and towards the island for a taste. The bakers happily provide these pies to any sailor who lands on their island at no cost and they taste so scrumptious

that no one ever regrets stopping by for an afternoon visit. The decision tree for Odysseus in this scenario looks like ther:

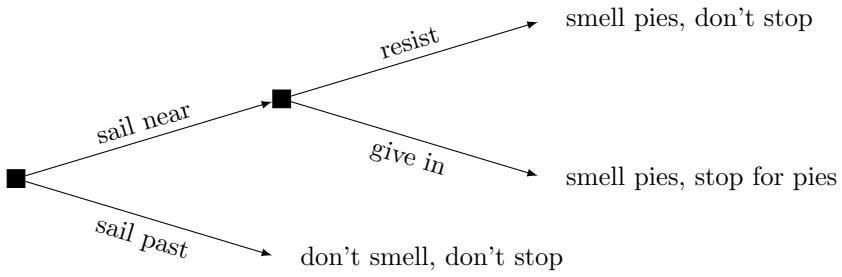

Figure 10: Odysseus and the Tasty Pies Tree

Odysseus of course, has been away from Ithaca for years and, as in the case of his historical counterpart, prioritizes returning home to his family and his kingdom over everything else. He knows that if he sails near enough to smell the pies cooking he will form a strong desire to land and spend an afternoon munching on the pies that will overwhelm his present desire to avoid unnecessary stops and keep sailing towards Ithaca. Importantly, this transformation that Odysseus expects will not be a case of a learning experience or gaining new information. Odysseus has never had the pies before and the mere act of smelling them does not actually confirm their tastiness or give him demonstrable proof of their gustatory worthiness.

As in the earlier case, Odysseus is here a sophisticated consequentialist and his preferences are (smell pie, don't stop)$\succ$(don't smell, don't stop)$\succ$(smell pies, stop for pies). Knowing that his future self will be overcome by a desire for pie upon smelling them, Odysseus will choose to either bind himself to the mast or dodge the island entirely if that option is unavailable to him. From a narrow perspective of rational choice, this choice seems praiseworthy in that he is taking steps to effectively exercise his present will over that of his pie-loving future self.

But his crew may be rightfully confused by his resistance to taking a detour of just a few hours to taste the most delicious pie he will every eat in his life. Those among them who have already tried the pie know that his future self will reflectively endorse his past choice to delay his journey home by just a few hours. Indeed, Odysseus himself recognizes that his future self will have no regrets about stopping to eat the pie and that this will increase his intertemporal wellbeing or utility on the whole.

This case suggests that there are times when being a sophisticated decision maker requires an attitude or posture other than resistance towards one's future self. Perhaps in this case Odysseus should be more deferential towards his future desires and allow himself to smell the fresh baked pies and take a minor detour, knowing that his future self will find contentment and be in a state of reflective equilibrium regarding that decision.

As an additional motivation, it is exceptionally easy for a clever bookie to take advantage of an agent with non-reflective desires for profit. Consider an agent who currently prefers going to the movies to see a Sci Fi film to going to the opera but whose future self will develop a taste for opera and will prefer the opera tickets. If she is initially endowed with

an opera ticket, the bookie can offer to give her a movie ticket in exchange for the opera ticket plus $\$\epsilon$. In the future, the bookie will offer to reverse the trade and give her the opera ticket in exchange for the movie ticket and an additional $\$\epsilon$ once again. Now the agent has willingly given up $\$\epsilon$ twice and possesses the exact same good that she had to begin with.

Before weighing in on the ultimate success of the analogous desire REF_d model I will critically assess each of the assumptions necessary to its success individually. First, we must examine the role of intra-personal trust and authority in the perception of these shifts in desire as either valid or invalid. Second, I will assess this reflection model in terms of the non-pragmatic, non-Dutch book arguments that lend support to van Fraassen's model. Third, I will turn to addressing the voluntarist implications for this model, similar to van Fraassen's REF, and briefly discuss the problem of motivation for the agent to meaningfully alter her desires.

2.5 Preference Authority and Improvement of Desires

When an agent takes her future self to have undergone a learning experience, she is inclined to trust the beliefs of that future self and absorb them into her own noetic structure. In the case of cognitive impairment, she treats her future self dubiously. Another way to frame these actions is to say that in the former case, she confers authority to her future self and in the latter case she withholds it. Abstracting away from the intrapersonal to the more familiar interpersonal, for us to accept someone else's claim that an act is rational or irrational, we must implicitly accept that there is a degree of authority backing up that judgement on their part. Here we must inquire as to what the normative grounds for this type of authority are. In this section, I will examine how authority plays a role in an agent's determination of whether or not to be reflective in their beliefs, and assess whether authority works in a similar way in the case of desire.

When we say a person accepts some fact or belief on authority, I take that to mean that she accepts someone else's endorsement of it as a reason of her own for accepting it.[125] Gibbard argues that there are two conditions under which this acceptance might take place. In the first case, one might attribute contextual authority to another. This kind of authority presupposes that the author of the judgement in question shares norms with the listeners and those shared norms lend credibility to the speaker so that the audience can use her reasoning as a proxy for their own.[126] Alternatively, there is what Gibbard calls Socratic authority, which involves the speaker guiding along the listener via strategic questioning, exactly as Socrates did in his dialogues.[127] By exercising this type of authority, the listener can be prodded towards a certain conclusion simply by realizing it is rationally implied by concepts or beliefs they already accept.[128]

A Socratic dialogue between present and future selves would never be able to get off the ground because the most fundamental premises will not be agreed on due to the divergence in evaluative utility functions. Claims of the form "act/good x should be valued equal to amount n as determined by $u(x)$" are some of the most basic that a person can make to convince another to do something but it only works if both interlocutors are relying

[125] Gibbard (1991), 174.
[126] Ibid.
[127] See e.g., Plato's Crito or Meno.
[128] Gibbard (1991), 174.

on the same utility calculus. The past and present self will value the same outcome in a fundamentally different way and without agreeing on the way to value one outcome relative to others, they will never agree on the choiceworthiness of that option.

Thus, the case of an agent's endorsement of REF_d must depend on the attribution of contextual authority to one's future self. The agent at present learns of her future self's desires and will accept them only if (a) she reasonably thinks her future self shares her own foundation of basic values and norms and (b) has been enhanced by the acquisition of some new information or experience that is unavailable to the agent in the present moment. The first constraint is a prerequisite of establishing authority while the second is the learning requirement insisted on by Skyrms. If either of these necessary conditions are violated then it becomes irrational, not to mention psychologically difficult, for an agent to accept the desires of her future self.

Now let us consider the case of an agent whose future self has desires which depart from her own at the present. This type of case entails a difference of utility functions. Thus, the agent will likely view her future self not as merely misinformed but as bizarre in some sense for sustaining alien preference rankings that are informed by entirely different utility functions. This is a type of disagreement that is more fundamental and presents a higher barrier to establishing credible authority. In this class of dynamically inconsistent cases, I argue that it is possible for a rational agent to view her future self as having a certain kind of contextual authority on matters of basic desires.

Under the guise of contextual authority, REF_d is plausible but only in some situations. Consider relatively innocuous cases of shifting desires like the desires for different varieties of wine. This change takes place in an agent's palette without disturbing any of the structure of underlying norms that Gibbard views as essential to establishing this authority. There is nothing about developing a taste for fine wine that mandates a person become more politically conservative, less religiously dogmatic, or morally compromised. It is simply the case that she now prefers a different, ostensibly superior flavor profile in wines than she did previously. It is only necessary that both she and her future self are looking for pleasurable gustatory experiences. If the two of them do not share this basic value then she would not defer to your future self.

There are other changes in desire that intuitively seem to challenge the contextual authority principle, but I argue that upon further investigation it is clear that the intuitive repulsion to these cases is a result of other mitigating factors. Consider Marlee who has been an ardent supporter of the Blue political party for her entire life thus far. Suppose she learns that in 20 years her future self, who has since become a lawyer and a recognized expert on politics, has switched her allegiance to the Red political party. Upon learning this, Marlee is not motivated to accept that change and refuses to try to cultivate a favorable disposition towards the Red party. Is this case a violation of REF_d?

The answer is no, for a combination of reasons. In the first case, political affiliation is not comparable to desiring delicious pies or fine wines. What is changing for Marlee is the preference relation among compound options. Political ideology is an amalgamation of basic preference orderings among values as well as the complementary beliefs that determine the valence of that ideology. For example, one might value promoting a feeling of security for herself over promoting a feeling of acceptance and comfort for someone else. But beliefs, not desires, transform that preference into political will. For example, if the agent believes that

policy X is the most effective means of providing that highly desired feeling of security and believes the Red party favors policy X, then insofar as she as a rational agent who desires the means to the ends that she values, she will wish to support the Red party. In that way political preference shifts are complex and not entirely attributable to the kinds of change in fundamental desire that are the focus of REF_{d}. A similar degree of clarification is also needed in the case of religious beliefs or other complex ideologies.

Furthermore, we might point out that there is not a clear case of improved desires here. Getting a J.D. does not qualitatively improve someone's desires because it is not the right kind of process. In fact, this is improvement more akin to the case of epistemic learning. In that case, we look for indications that our future self has come into contact with and appropriately synthesized previously unknown information in order to accept that they have undergone a learning process. In the case of desires, we look for appropriate exposure to the relevant types of desires in order to feel warranted in considering our future desires improved.

Learning about the chemical process of wine fermentation, the history of the techniques used for harvesting grapes, the properties of the wood used in the barrels, etc. is not what gives someone a nuanced appreciation of wine. Such ancillary information is perhaps interesting for an agent like Maria to know and will allow her to impress her friends at dinner parties but it does not qualify her as a wine expert. The process of epistemic improvement proceeds via the uptake and assimilation of new information that is relevant to the choice at hand. Conversely, the process of improvement of desires occurs via exposure to the object(s) of desire, not via learning information about that object.

In modern models of agent causality, it is typically theorized that there is a regulative feedback loop which begins at the point of acquisition of the action's goal and relays feelings of either satisfaction or disappointment back to the agent. This way, she can modify her expectations for that same outcome in the future to more accurately reflect her actual subjective appraisal of it once obtained.[129] This is the process by which desires evolve. It is important to stress that epistemic learning is an exogenous process that requires informational input not available to the agent, whereas desires evolve mostly endogenously, in a closed feedback loop. The agent does not receive information, she just experiences her own reactions to the consequence of a decision and adjusts her future attitudes towards that consequence appropriately to be in line with her experience at the present.

In this way future versions of agents can have contextual authority over their current selves in the form of hypothetical imperatives or claims about the state of their desires. In the case of Maria, she might think "If I were to continue tasting good wines in an intentional and attentive way, I would eventually take a liking to them and prefer them to the cheap stuff." Hence, she may find reason to continue to work at developing that mature taste for wine. All that she needs to know for this hypothetical claim about her desires to have motivational force for her is that her future self cares about maximizing her utility in the same way she does, but assigns utility in a superior way as a result of possessing a broader range of experience.

This conclusion ought to also hold in the case of Odysseus and the Island of the Tasty Pies. Here the rational thing to do is to be more receptive to the proposed preferences of his future self. In doing so, Odysseus is responding to the claim that "supposing I were to

[129] Railton (2014).

stop at the island for a quick bite of pie, I certainly would not regret it and would be happy with my decision in hindsight."

This type of authority is contextual because, crucially, it relies on the agent in the present moment recognizing her future self as fundamentally similar to her current self. When an agent experiences feelings of alienation from the identity of her future self, it is impossible for her to attribute authority to that future self. As Williams argues, an agent will only be moved by a reason for ϕ-ing if that reason appears in her subjective motivational set.[130] Our claim that Odysseus ought to listen to his future pie-loving self is correct only because his motivational set is such that he will after a period of deliberative reasoning find that he is motivated by that future self's desire for pie.

Here we ought to clarify that it is not the case that Odysseus ought to desire the pie in the present moment because his future self does, rather he should desire it *for* his future self. In the case of Lyndsey's tastes in wine, she ought to find within her motivational set a reason to develop a taste for fine wine now to enable her future self to be in a position to have superior desire profiles. With some desires (like the desire to refrain from using a harmful, addictive substance), the rational imperative is to have it immediately, in the present. In other cases there is not a reason to have the desire now but a reason to have the desire to act in such a way as to bring about the desired outcome for my future self. In these cases the agent is being appropriately responsive to her future self's anticipated wants and viewing that fact as generating a reason for her in the present.

In other examples, this contextual authority appears to be missing. Consider 12-year-old Ryan who feels clearly alienated from his future self's sexual attractions. His parent's claim that he ought to accept those desires does not constitute a misapprehension of her current motivational set, rather they are claiming to provide an external reason for him to act. But according to an existence internalist account of reasons, in order for a reason to ϕ to exist, the agent must be suitably capable of being motivated to ϕ. Existence internalism is not a claim about the meaning of the moral language that we use to explain why one ought to ϕ in terms of the reasons that support ϕ-ing. Rather, it is a claim about the truth conditions of attributions of reasons.[131] The claim that Ryan ought to be sympathetic to his future self's desires is therefore rendered false by the fact that attraction to girls appears nowhere in Ryan's subjective motivational set and therefore cannot count as a reason to act.

We might attempt to reform this problem by distinguishing between different types of desires that are differentiated by the fact of our persistence as agents across many time intervals. These are now-for-now and now-for-then desires.[132] In Ryan's case, it is not appropriate to have now-for-now desires to be romantically involved with girls at the present moment. But it is fitting to have now-for-then desires for it, desires for his future self to form those types of relationships. As a side effect effect, he should desire to refrain from taking preventative measures to restrict his future self's ability to act on those desires.

While this distinction appears a promising solution at first, it does not solve the internalist's dillemma. In this case, it is not just that Ryan does not want to be attracted to girls now, but that he does not ever want to be attracted to girls. He has now-for-now and now-for-then desires to avoid girls. For that reason there is nothing within his internal

[130] Williams (1997), 3.

[131] I owe this thought to points made by Peter Railton both in private conversation and in PHIL 429 and PHIL 640.

[132] This reframing attempt was suggested to me by Elizabeth Anderson in private conversation.

motivation set that can properly count as a reason for him in the present moment to be sympathetic towards his future self. His failure to recognize the fittingness of his future self's attitudes precludes a healthy acceptance of some changes in desire as fitting or appropriate for himself.

If we adhere to an internalist account of reasons and action justification,[133] it becomes clear that the scope of intertemporal contextual authority is limited to a narrow range of cases in which we perceive our future selves to be fundamentally the same person and comfortably identify with them. Essentially, this makes the applicability of REF_d limited to intertemporal cases of close proximity and likeness. While the extent to which identity can stretch across time will vary from one individual to the next, there is a longer list of identity prerequisites for REF_d to succeed than there are for REF since identification with beliefs requires minimal assumptions about rationality and lack of cognitive impairment, not more detailed and contextually-specific features of a future self.

2.6 Nonpragmatic Considerations

To motivate epistemic reflection as a criterion of rationality, philosophers have historically relied on two families of arguments, pragmatic and non-pragmatic. Dutch book arguments and their variants fall under the heading of the former. Their aim is to show that an agent, given certain irrational beliefs, will assent to gambles that guarantee a net loss of money to the bookie. They are pragmatic in that the argument identifies this wayward behavior[134] as the symptoms of an underlying pathological irrationality.[135] The upshot of these arguments is that given the irrational behavior, we can conclude that the underlying beliefs which motivate them are suspect.

Non-pragmatic arguments, on the other hand, aim to give a formal account of why the beliefs themselves are irrational from a purely epistemic perspective. Typically these approaches have endeavored to show that a person who violates axioms of rationality will have preferences that are suboptimal with respect to their accuracy. That is, in any situation there is an alternative belief structure which dominates their current beliefs with respect to producing accurate beliefs. (Where by accurate we mean as close an approximation to the truth as possible.) In this section, I will briefly describe certain of these non-pragmatic arguments for reflective belief structures and examine whether or not these models can work formally with REF_d.

There is some variability in the literature in terms of how philosophers argue we ought to measure inaccuracy. Van Fraassen, in introducing the reflection principle, argues that belief structures which obey REF will minimize inaccuracy as measured by a calibration index.[136] According to the calibration metric, inaccuracy is measured by calculating the magnitude of difference between an agents estimation of the frequency of an event and the actual frequency of the event.

Formally, imagine some proposition X which makes some claim about event ω. For example, in the case of the proposition that it will rain tomorrow, ω refers to the event that it rains tomorrow and X modifies that event to designate a position - for the sake of

[133] Which I take to be the correct account of reason justification.
[134] Or rather, the possibility of wayward behavior - as we know, there are no real-life Dutch bookies.
[135] Huttegar (2013).
[136] van Fraassen (1984), 245.

simplicity we will restrict it to either the affirmative or negative. We can postulate a scoring rule according to which we assign a numerical value to the designate the truth or falsity of a belief. This binary system would look like this:

$$X(\omega) = \begin{cases} 1 & \text{if X is true} \\ 0 & \text{else} \end{cases}$$

An agent then makes predictions or assigns credences to the propositions. In the intertemporal case, an agent will form different credences for $X(\omega)$ in different time periods, depending on whether the available evidence favors hypothesis X or its negation. Each individual case n such that $n \in N$ represents a point at which the agent forms a credence on X. The value or correctness of her credence for $X_{\mathrm{n}}(\omega)$ is calculated by weighting the score of a true belief by 1 and the score of an incorrect or false belief by 0 and summing the two values. That is to say that credences are expectations of truth-values:

$$C(X_{\mathrm{n}}(\omega)) = 1 * C(X_{\mathrm{n}}(\omega)) + 0 * (1 - C(X_{\mathrm{n}}(\omega)))$$

When these credence scores across all $n \in N$ for each individual credence are aggregated and then averaged, we find a value that represents the agent's implicit estimation of the frequency of $X(\omega)$.

$$\bar{C}(X(\omega)) = \frac{C(X_1(\omega)) + C(X_2(\omega)) + \cdots + C(X_{\mathrm{n}}(\omega))}{N}$$

The actual frequency of $X(\omega)$ can be represented similarly, by averaging the sum of all the binary scores for $X(\omega)$ in each case $n \in N$:[137]

$$\bar{X}(\omega) = \frac{X_1(\omega) + X_2(\omega) + \cdots + X_{\mathrm{n}}(\omega)}{N}$$

The calibration index proposed by van Fraassen measures the absolute value squared of the difference between $\bar{C}(X(\omega))$ and $\bar{X}(\omega)$. In the case that $\bar{C}(X(\omega)) > \bar{X}(\omega)$, the agent is overconfident in her estimation of the frequency of $X(\omega)$. In the case that $\bar{C}(X(\omega)) < \bar{X}(\omega)$, the agent has underestimated the frequency of $X(\omega)$. For a Bayesian, a perfectly calibrated agent who is neither epistemically conservative or overly bold will hold credences such that $\bar{C}(X(\omega)) = \bar{X}(\omega)$, that is, her estimation of the frequency of some event is identical to the observed frequency of the event and her hypothesis fits the data exactly. So if we selected some probability p and looked at the set of propositions an agent assigns p to, we can ask in that set of propositions what is frequency of truth? For the perfectly calibrated agent the answer is p (e.g., 90% of .9 propositions, 80% of .8 propositions, etc.).

$$\text{Calibration: } |\bar{C}(X(\omega)) - \bar{X}(\omega)|^2 = 0$$

In terms of the reflection principle, for van Fraassen it is imperative that for a rational agent who looks at a series of a propositions $X_{\mathrm{n}}(\omega)$ and assigns a probability of $C = .5$ to all of them, it is an axiom of rational probability that if she is asked what proportion

[137] This is an indicator function that represents the proportion of true propositions in the set of all propositions.

of her propositions is accurate, she is rationally obligated to say .5.[138] This would entail a minimization of the calibration index, indicating perfect calibration between the agent's beliefs and her confidence in her own beliefs.

However, the same fear of my future self's epistemic imparement seems to suggest this is not always optimally rational credal behavior. Imagine a meteorologist who relies on a forecasting model that supplies a prediction of rain of .8 frequently - but on the days it has made that prediction, it has only rained 70% of the time. Here there is a disparity between the actual frequency and the frequency determined by the model. Van Fraassen paints a picture in which accuracy for an agent, a model, or an algorithm (insofar as their purpose is predictive) is for it to rain 80% of the days that it predicts .8 chance of rain or that $\bar{C}(X(\omega)) = \bar{X}(\omega) = .8$. So in this case, the agent's estimates commit her to an established frequency that is not identical to the true frequency since her weather model itself is deficient.

We ought to have similar concerns about our future self committing to some $C(X(\omega))$ that is a bad estimate of the true frequency. In this case too it seems that we ought to trust the belief model only if there is good evidence that it is accurate. For example, if Conner's future self is somehow able to send him a text every day with a weather forecast, we might have reason to be a little skeptical. As he sees whether or not these beliefs are accurate (i.e., whether the predictions are well calibrated) then he develops reasons to either trust his self-texts from the future or disregard them. Van Fraassen thinks intertemporal solidarity is a motivating reason to be reflective in any case - that there is a need to stand behind oneself. This is a pragmatic concern about seeing oneself as the same person across time, though not a very good one. Remember, Skyrms instructs us to adhere to REF only in cases of learning. From the non-pragmatic perspective we might say you ought to conform to beliefs only if those beliefs are well calibrated over time. If they are, a rational agent has clear motivation for believing those beliefs.

In cases where the agent declines to follow REF, it is not the case that the agent fails to see herself as herself, it is that she recognizes the limitations of her future self. When a friend who is intoxicated says he should be allowed to drive home, a responsible friend inevitably will say no. In that moment she is not failing to see her friend as a person of value and an adult who is capable and deserving of autonomy. She merely recognizes that her friend is acting under a mental fog brought about by fatigue and alcohol. She accounts for that cognitive impairment and updates her preferences for her friend to drive home accordingly. I would argue against van Fraassen that an analogous condition applies here in terms of calibration. Neglecting to be calibrated does not constitute an inability to identify with one's future self.

Returning back tot he general discussion of non-pragmatic concerns however, calibration by itself is not sufficient as a guideline for proper scoring and the problem of relying on calibration exclusively is that it encourages uninformed choice behavior. Imagine a weatherperson Arooshee whose job is to make daily predictions about the chances of precipitation in the local area of Ann Arbor. Suppose that her annual salary is tied to her accuracy as a meteorologist, as determined by her calibration score. Suppose further that the actual frequency of rain in Ann Arbor annually is .45. That means in any given day across the year as a whole, the statistical probability of rain is .45.

[138] Joyce, Jeffrey, Di Finnetti

Arooshee has two options: she can try her best every day to read and interpret the models and calculate as accurate a probability as possible or she can just look at the historical data tables. Knowing that her salary depends on her accuracy on the job, Arooshee chooses to go on the air every morning and give the probability of rain as .45.[139] On the one hand, by van Fraassen's calibration metric she will be statistically perfectly accurate over the course of the next year. But on the other hand, she will be a horrible meteorologist. She will reliably predict rain with .45 probability both on days it pours all day and on days in which there is not a cloud in the sky. So from the perspective of her viewers she will usually be a more or less uninformative resource.

In the aggregate, her estimated frequency will very closely match the actual frequency of rain in Ann Arbor, but on the day-to-day basis, there will be high variability between her estimation of .45 and the actual outcome, which will be either 1 or 0 (since it will either rain or not rain on any given day). That is the reason that more sophisticated accuracy scoring metrics incorporate other accuracy conditions in addition to calibration. The widely cited Brier score, one of the earliest examples of this methodology, can according to the Murphy decomposition be conceived of as two principles - calibration and discrimination - the straight sum of which produces an accuracy score.[140]

The discrimination stipulation makes sense of the fact that Arooshee is apparently able to achieve such high accuracy as such a bad weather forecaster. Perfect discrimination mandates that for any prediction $C(X(\omega))$ it must be the case that in cases in which $C(X(\omega)) = p$ then $X(\omega)$ is uniform with respect to truth value (whether true or false). A perfectly discriminating credence function will be such that for any $X(\omega)$, all of the propositions that are assigned credence p will have the same truth value (be it either true or false). Considering a set of propositions assigned probability p by the agent, calibration dictates that the frequency of truths are equal to the associated probability. Discrimination stipulates that the best outcome is for those sets to be homogenous with respect to truth value and wants the ratio of truths to falsity to be as close to infinity or zero as possible, that is, as extreme as possible.

$$\text{Discrimination: } P(X(\omega) \mid C(X(\omega)) = p) = 1 \text{ and } P(X(\omega) \mid C(X(\omega)) = \neg p) = 0$$

Huttegar has shown that since REF is a required axiom, failure to adhere to it will necessarily produce inaccurate beliefs.[141] Therefore, irreflectiveness indicates irrationality for the agent. This additional argument nicely compliments the pragmatic Dutch book arguments to strengthen the case for REF. For any agent who violates the principle, she will surely have less accurate beliefs than she otherwise could have. It is therefore advisable for a rational agent to adhere to it in the interest of accuracy minimization.

However, reflective beliefs are rationally required *conditional on* the fact that the agent does not take her future self to be faulty in terms of her predictions. The perceived reliability of a source of estimation matters to rationality as Skyrms points out.[142] But these indexical measures of accuracy are important because inaccuracy reveals that an agent's beliefs are deficient in the sense that a failure to obey to the laws of probability produces beliefs which

[139] The wording of this example borrowed from Jim Joyce.
[140] Murphy (1973).
[141] Huttegar (2013).
[142] Skyrms (1990).

can be strictly dominated by the set of totally accurate beliefs. An agent is simply less epistemically accurate than she could have been and therefore ought to have been.

These non-pragmatic considerations add an additional dimension of cogency to the argument REF. But I argue that non-pragmatic arguments for REF_d cannot succeed due to fundamental differences in the logical structures of beliefs and desires. Beliefs have a "thetic" or mind-to-world direction of fit meaning that good beliefs are the ones which accurately represent the way the world is, rendering them true.[143] Desires on the other hand have a "telic" or world-to-mind direction of fit. Satisfaction of desires occurs then when the object of that telic attitude is fulfilled.[144]

As Humberstone points out, this distinction is not trivial but rather determines the structure of the conditional intention. In the case of beliefs, facts about the object of the attitude are held fixed and condition the propositional attitudes towards them. In the case of desires, facts about the possession of the desire is what is held fixed, and correctness is conditioned by the actions of the agent.[145] Humberstone considers thetic direction of fit to embody that direction as a matter of a constitutive rather than regulative principle and writes, "Thus the very concept of belief imports its own criterion of success, or [...] has its own "internal axiology."[146] It is in virtue of this internal axiology that beliefs can be measured in terms of accuracy.

Accuracy is a relational value that depends on the existence of external facts to which the beliefs can be referenced in order to determine their truth value. No such external marker exists in the case of desires to differentiate them based on regulative ("right" or "wrong") or evaluative ("good" or "bad") standards. For non-pragmatic scoring arguments to work in the case of REF_d, we would need to identify something in value theory analogous to truth. But unless one is a Platonist, a Moorean intuitionist, or an otherwise very strong realist about moral values, it is impossible us to identify this correctness condition, if it even exists.

And really, this is two separate problems. There is the metaphysical question of whether such a value could even exist and there is the epistemological question of how it would ever be possible to ascertain such a value. The likelihood of finding answers to either is, I think, rather unlikely. Therefore it is a strike against REF_d that it cannot be motivated by non-pragmatic arguments. This decreases the cogency of the argument as an imperative of rational choice, though the extent to which it should decrease our confidence might be ambiguous.

2.7 Feasibility Constraints

I have demonstrated that one's future self can, at least in certain cases, claim authority over one's present self and I have also illustrated the fact that REF_d cannot benefit from non-pragmatic considerations in the same way that REF does. Now I will turn my attention to a concern which precedes those relating to the motivational force of this supposed rational criterion - the question of whether the prescriptions of REF_d are even possible in the first place.

[143] Humberstone (1992) and Anscombe (1963).
[144] Humberstone (1992), 65.
[145] Ibid, 76.
[146] Ibid, 73.

As discussed previously, one worry with van Fraassen's original formulation of REF is that it strongly implies a voluntarist picture of belief acquisition or adjustment. Epistemic voluntarism is the view that the process of belief formation relies at least in part on an individual's will and is not simply the automatic registering of a degree of belief with respect to a given proposition.[147] That someone ought to align their current credences to those of her (epistemically capable) future self presupposes that such intentional manipulation of degrees of credence is possible in the first place. Were this not the case, REF would lose any claim it has to legitimate imperative force as an axiom of rationality.

At best, this voluntarist picture seems plausible in only a limited range of cases. Beliefs are directly responsive to outside information and evidence in a way that cuts out the conscious agent as a middleman. Data from the external world impinges on the mind and shapes the content of our beliefs somewhat forcibly. This is why 'Moorean' sentences such as "I went to the movies last Tuesday, but I don't believe that I did"[148] are absurd, despite their being logically consistent. It is epistemically incoherent that someone could be presented evidence of her having been to the movies on Tuesday night and accept that evidence as valid yet withhold belief in the proposition that she went to the movies on Tuesday.

Adjusting one's current credences to align them with future credences for the same proposition can only realistically take place under the conditions (i) that the agent takes her future self's credence as better evidence than her own, that is, more indicative of the actual probability or truth value and (ii) that the agent is not in the present moment faced with stronger evidence that contradicts those future credences.

Returning to the example of Josh and the Michigan-Ohio State football game, if he takes his future self to be better situated epistemically than he is at the present moment (perhaps he has seen the updated injury rosters or done some basic statistical analyses), then he will be motivated to readjust his credences appropriately. But this motivation is not a conscious act of the will - rather he has no choice but to be epistemically motivated by his future self's credences insofar as he considers them to be constitutive of evidence.

Similarly for desires, (i) the agent must take her future self's desires to be better informed than her own and (ii) the agent must not at the present moment have to grapple with stronger, countervailing desires. For this type of reasoning to proceed in the first place, it must pass the 'ought implies can' criterion. This is not always the case - addictions for example constitute a persistent sort of desire over which the agent feels (and in many cases is) powerless. But in general it does seem this assumption is more viable with desires than beliefs at least because most models of action take desires to have more endogenous origins than beliefs. That means tactics like self-deception might plausibly be more successful in terms of regulating desires than beliefs.

Ultimately, this is not a particularly worrisome issue for the case of $\mathrm{REF}_{\mathrm{d}}$. Even if it is not possible to force an about-face of desires immediately (thereby sparing oneself from susceptibility to Dutch books), in many cases that is not what is required in order to maximize expected utility globally. In cases like Odysseus and the Pies, he merely needs to passively allow his future self's desires to move him at the appropriate time. Resolving to have that attitude of passivity or actively working towards a more mature set of desires (as in the case of wine) neither requires an extraordinary act of the will nor accepting a

[147] James (1979).
[148] Moore (1942), 543.

totally voluntarist theory of desire formation. These feasibility considerations then are not an automatic defeater for $\mathrm{REF}_{\mathrm{d}}$.

3 Concluding Remarks

I will now briefly summarize some of the arguments that I made and reiterate where I stand on the strength and plausibility of the case for REF_d that I have advanced. Subject to the same background stipulations and limitations as REF (as proposed by Skyrms, Christensen, and others), it seems to me that yes, a rational agent ought to have reflective desires. Reflective desires offer an agent the opportunity to make her "rational narrative" a bit more coherent and sticky. It helps us avoid a certain class of Dutch book arguments. And it presumably makes us better off from a global utility perspective since cases in which REF_d is plausible to begin with are only those in which our future selves have a richer set of experiences that we take to have positively informed the posture of their desires.

Remember, there are broadly three families of response one might have towards expectations of dynamic inconsistency: hostility, indifference, and deference. As I have examined, the standard consequentialist theory (which I take to be the most compelling) calls for hostility - the sophisticated consequentialist ought to tie herself to the mast and force her future self to stay the course or else take preemptive action to ensure her future self does not get the chance to defect. As van Fraassen demonstrates, REF can be formally shown to rationally require the opposite. Given the knowledge that her future self will have beliefs (and therefore preferences too) which are out-of-sync with her current ones, she ought to alter her current beliefs out of pragmatic concerns about susceptibility to Dutch books.

I do not think it wise to pivot completely towards the side of deference. Hence I have tried to stake out a middle space that urges reflection for an agent only in cases where certain prerequisite conditions about her future self are met (e.g., epistemic competence, integrity of desires, valid claim to contextual authority). And I have further argued that this same qualified deference applies to the case of desires as well as beliefs (insofar as it is possible to modulate one's desires by an act of the will).

The case of REF_d is motivated by the same class of pragmatic concerns as REF, although it does not benefit from the added support of non-pragmatic arguments from accuracy. The perspective that I call cooperative paternalism vis-a-vis one's future self entails a delicate balancing act of determining when to be hostile and when to adopt now-for-then or now-for-now desires. It is cooperative in that one recognizes that one's future self may be better off with respect to additional information or experiences and one ought to make present adjustments or sacrifices so that one's preference attitudes will be consistent in a way that makes one coherent. And it is paternalistic in the sense that one ought not cooperate blindly, but ought to insist on there being a level of authority and trust in the reliability of one's future self as a prerequisite. One should view claims to authority by one's future self with a bit of reluctance or scepticism in that sense.

I have now, I hope, successfully argued for the reasonability of REF_d as a rational axiom in at least a narrow range of cases. But aside from investigating its merits in isolation, we must also ask ourselves whether this type of a principle could possibly be compatible with the larger consequentialism theory that I have endorsed previously. I am inclined to think that this principle does not mesh well with the rest of the theoretical apparatus. While REF can be showed to be an instantiation of the value of learning theorem and benefits from the formal proofs found in Good's theorem, REF_d has no analogous source of support in the literature.

There is a larger issue at play here, which is the historical insistence in the literature on a time-slice picture of rationality among economists and philosophers working on these issues. When we speak of an agent, we tend to speak of them exclusively as they are at that specific moment and view their preferences, beliefs, and desires all as fixed and applicable only to that particular person at t_i. As Strotz says, we are each an "infinity of individuals"[149] and it is tempting to look only at one agent at a time, making clean temporal distinctions between each version of the agent's self as if they were each their own person.

The issue with this type of thinking is that it dodges more complex questions about the nature of identity and its persistence across time, as well as meta-level questions about the nature of rationality itself. It is because of these considerations that I think Velleman's view bears heavily on the issues examined in this thesis. More attention ought to be paid to his and other alternative views of the unit of evaluation in rational choice theory as a space-time worm agent, not a time-slice agent.

This problem is substantive, and I am certainly not equipped to address it here. But I do think it is worth mentioning and I think it is essential that consequentialism reckon with this alternative picture and, if possible, synthesis it into the existing theoretical framework because there are certainly limitations and irregularities with the time-slice model. In some cases it seems obvious the agent should restrict her calculations to her immediate credences and utilities. But in other cases rationality mandates intertemporal cooperation which in turn requires a coherent explanation of who or what the agent is, not just locally but globally across all of her time-slices. And I think that consequentialism owes us a more adequate explanation of that.

[149] Strotz (1955), 179.

References

Allais, M. "Le Comportement de l'Homme Rationnel devant le Risque: Critique des Postulats et Axiomes de l'École Americaine." *Econometrica* 21, no. 4 (October 1953): 503–46.

Anderson, Elizabeth. *Value in Ethics and Economics.* Harvard University Press, 1995.

Andreou, Chrisoula. "Dynamic Choice." In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta, Spring 2017. Metaphysics Research Lab, Stanford University, 2017.

Anscombe, F. J., and R. J. Aumann. "A Definition of Subjective Probability." *The Annals of Mathematical Statistics* 34, no. 1 (March 1963): 199–205.

Bratman, Michael. *Intention, Plans, and Practical Reason.* Harvard University Press, 1987.

———. "Planning and the Stability of Intention." Minds and Machines 2, no. 1 (February 1992).

Broome, John. *Weighing Goods: Equality, Uncertainty and Time.* John Wiley & Sons, 2017.

Böhm-Bawerk, Eugen von. "Une Nouvelle Théorie sur le Capital." *Revue d'Économie Politique* 3, no. 2 (1889): 97-124.

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

Christensen, David. "Clever Bookies and Coherent Beliefs." *The Philosophical Review* 100, no. 2 (1991): 229-247.

De Finetti, Bruno. "La Prévision: Ses Lois Logiques, Ses Sources Subjectives." In *Annales de l'Institut Henri Poincaré*, vol. 7, no. 1, pp. 1-68. 1937.

De Finetti, Bruno. "Sul Significato Soggettivo della Probabilita." *Fundamenta Mathematicae* 17, no. 1 (1931): 298-329.

Downs, Anthony. "An Economic Theory of Democracy." (1957): 260-276.

Edwards, Ward. "The Prediction of Decisions Among Bets." *Journal of Experimental Psychology* 50, no. 3 (1955): 201.

Fisher, Irving. *Theory of Interest: As Determined by Impatience to Spend Income and Opportunity to Invest It.* Augustusm Kelly Publishers, Clifton, 1930.

Gauthier, David. "Resolute Choice and Rational Deliberation: A Critique and a Defense." *Noûs* 31, no. 1 (1997): 1–25.

Gibbard, Allan. "Wise Choices, Apt Feelings." *Cambridge, MA: Harvard University* (1991).

Good, Irving John. "On the Principle of Total Evidence." *The British Journal for the Philosophy of Science* 17, no. 4 (1967): 319-321.

Hammond, Peter J. "Changing Tastes and Coherent Dynamic Choice." *The Review of Economic Studies* 43, no. 1 (1976): 159–73.

———. "Consequentialism and the Independence Axiom." In *Risk, Decision and Rationality*, edited by Bertrand R. Munier, 503–16. Theory and Decision Library. Dordrecht: Springer Netherlands, 1988a.

———. "Consequentialist Foundations for Expected Utility." *Theory and Decision*i 25, no. 1 (1988b): 25–78.

Hey, John D. "The Economics of Optimism and Pessimism: A Definition and Some Applications." *Kyklos* 37, no. 2 (1984): 181-205.

Hilton, Ronald W. "Failure of Blackwell's Theorem under Machina's generalization of expected-utility analysis without the independence axiom." *Journal of Economic Behavior & Organization* 13, no. 2 (1990): 233-244.

Homer. *The Odyssey.* Penguin, 1997.
Hong, Chew Soo. "A Generalization of the Quasilinear Mean with Applications to the Measurement of Income Inequality and Decision Theory Resolving the Allais Paradox." *Econometrica: Journal of the Econometric Society* (1983): 1065-1092.
Humberstone, I. L. "Direction of Fit." *Mind* 101, no. 401 (1992): 59-83.
Hume, David. *Dialogues Concerning Natural Religion.* William Blackwood, 1907.
Huttegger, Simon M. "In Defense of Reflection." *Philosophy of Science* 80, no. 3 (July 1, 2013): 413–33.
James, William. *The Will to Believe and Other Essays in Popular Philosophy.* Vol. 6. Harvard University Press, 1979.
Jeffrey, Richard C. *The Logic of Decision.* University of Chicago Press, 1990.
Joyce, James M. *The Foundations of Causal Decision Theory.* Cambridge University Press, 1999.
Kahneman, Daniel. *Thinking, Fast and Slow.* New York: Farrar, Straus and Giroux, 2011.
Kahneman, Daniel, and Amos Tversky. "Prospect Theory: An Analysis of Decision Under Risk." In *Handbook of the Fundamentals of Financial Decision Making: Part I*, pp. 99-127. 2013.
Kavka, Gregory S. "The Toxin Puzzle." Analysis 43, no. 1 (1983): 33–36.
Laibson, David. "Golden Eggs and Hyperbolic Discounting." *The Quarterly Journal of Economics* 112, no. 2 (1997): 443-478.
Levi, Isaac. *The Covenant of Reason: Rationality and the Commitments of Thought.* Cambridge University Press, 1997.
Loewenstein, G., and D. Prelec. "Anomalies in Intertemporal Choice: Evidence and an Interpretation." *The Quarterly Journal of Economics* 107, no. 2 (May 1, 1992): 573–97.
Machina, Mark J. "Dynamic Consistency and Non-Expected Utility Models of Choice Under Uncertainty." *Journal of Economic Literature* 27, no. 4 (1989): 1622–68.
McClennen, Edward F. *Rationality and Dynamic Choice: Foundational Explorations.* Cambridge University Press, 1990.
Moore, G. E., 1942, "A Reply to My Critics", The Philosophy of G. E. Moore, edited by P. A. Schilpp. Evanston, IL: Northwestern University.
Murphy, Allan H. "A New Vector Partition of the Probability Score." *Journal of Applied Meteorology* 12, no. 4 (1973): 595-600.
Pettit, Philip. "Decision theory and folk psychology." In *Foundations of Decision Theory*, edited by Michael Bacharach and Susan Hurley, pp 147-175. 1991.
Quiggin, John. "A theory of anticipated utility." *Journal of Economic Behavior & Organization* 3, no. 4 (1982): 323-343.
Rae, John. *The Sociological Theory of Capital: Being a Complete Reprint of the New Principles of Political Economy, 1834.* Macmillan, 1905.
Raiffa, Howard. *Decision Analysis: Introductory Lectures on Choices and Uncertainty.* Reading (Mass.): Addison-Wesley, 1968.
Railton, Peter. "Reliance, Trust, and Belief." *Inquiry* 57, no. 1 (2014): 122-150.
Ramsey, Frank P. *The Foundations of Mathematics and Other Essays.* Ed. R. B. Braithwaite. New York: Harcourt Brace, 1931.
Samuelson, Paul A. "Consumption Theory in Terms of Revealed Preference." *Economica* 15, no. 60 (1948): 243-253.

Samuelson, Paul A. "Probability, Utility, and the Independence Axiom." *Econometrica* 20, no. 4 (1952): 670–78.

Sartre, Jean-Paul. *Being and Nothingness*. Open Road Media, 2012.

Savage, Leonard J. *The Foundations of Statistics*. Courier Corporation, 1972.

Seidenfeld, Teddy. "Decision Theory Without 'Independence' or Without 'Ordering': What Is the Difference?" Economics & Philosophy 4, no. 2 (October 1988): 267–90.

Sen, Amartya K. "Choice Functions and Revealed Preference." *The Review of Economic Studies* 38, no. 3 (1971): 307–17.

Skyrms, Brian. *The Dynamics of Rational Deliberation*. Harvard University Press, 1990.

Strotz, R. H. "Myopia and Inconsistency in Dynamic Utility Maximization." *The Review of Economic Studies* 23, no. 3 (1955): 165–80.

van Fraassen, Bas C. "Belief and the Problem of Ulysses and the Sirens." *Philosophical Studies* 77, no. 1 (January 1995): 7–37.

van Fraassen, Bas C. "Belief and the Will." *The Journal of Philosophy* 81, no. 5 (1984): 235–56.

Velleman, J. David. "The Story of Rational Action." *Philosophical Topics* 21, no. 1 (1993): 229-54.

von Neumann, John, and Oskar Morgenstern. *Theory of Games and Economic Behavior*. Princeton University Press, 1944.

Wakker, Peter. "Nonexpected Utility as Aversion of Information." *Journal of Behavioral Decision Making* 1, no. 3 (1988): 169-175.

Williams, Bernard. "Internal and External Reasons." In *Moral Discourse and Practice*, pp 363-373, 1997.