Author:
Axten, Nick G

Title:
The problem of hyperbolic discounting

The Problem of Hyperbolic Discounting

Nick Axten

A dissertation submitted to the University of Bristol
in accordance with the requirements for award of the degree of
Doctor of Philosophy in the Faculty of Arts.

March 2022

Abstract

In the standard theory of human decision making a rational agent faced with a set of options chooses an option that maximizes the total value of expected consequences. This theory introduces a number of characteristic variables: a set of options, a set of possible consequences of each option, and a measure of value, probability, and futurity for each possible consequence given that a particular option is chosen. A widely discussed question is whether the value attributed to any consequence does, or should, depend on its futurity. Observed behaviour appears to show that attributed value commonly decreases with futurity. There is an argument that this discounting is rational provided that it leaves the relative preference order among options unchanged as alternative consequences approach. But research in behavioural economics provides apparently compelling evidence that observed human decision making frequently exhibits a form of discounting – hyperbolic discounting – that violates this condition. If the inference is correct it appears to deprive the standard theory of its usually assumed explanatory merit.

On examination, this conclusion rests on a number of questionable assumptions. In this thesis I examine these assumptions and provide an alternative analysis of the process of rational decision making in terms of its adaptive basis and procedural constraints. I investigate, in particular, the causal structure of agency and associated issues of valuation, option definition, and probability assignment including, inter alia, the role of predictive accuracy, the combinatorial structure of valuation, proxies, the variety of alternatives, achievability, and the reference class problem. I derive a principle theory of rational decision making that generates a number of characteristic dynamical value profiles in response to various typically prevailing conditions. They exhibit exponential, hyperbolic, and other characteristic second order effects. I discuss issues of rational evaluation, amendment, testing, and wider philosophical implications.

Acknowledgements

This thesis is in part an attempt to repay a debt that I have long felt as a burden. In 1970 I joined the PhD programme in mathematical sociology at Pitt with generous support from the Fulbright Commission and the Andrew Mellon Foundation, with the aim of creating a quasi-algorithmic theory of institutional social action. In this I was enormously supported by the staff and graduate students at Pitt, especially Tom Fararo and John Skvoretz, and – over the road – by Herbert Simon at Carnegie Mellon. To my regret, I never completed the programme. One reason is that I became increasingly convinced that the theory I envisaged could not be explanatory unless it was underpinned by a theory of human motivation, and this proved considerably more difficult.

Anyone who has ventured into this field will soon find themselves descending into a labyrinth of unsolved philosophical problems, in which attempts to solve one problem quickly reveal others, equally puzzling. In due course, although I never gave up entirely, I backed off, comforting myself with the thought that since the problem is reasonably well defined someone else would surely sort it out. But as the years passed, during which other life projects took over, I became increasingly surprised to find few signs of this happening.

Eventually, feeling the years slipping away, I thought I should try again. To have any hope of success I would need some serious up-to-date academic support. So in 2016 I joined the MA programme in philosophy at Bristol, which I could get to on the 376 bus. It was a wonderful intense year, for which I must thank the staff and my fellow graduate students, who accepted me as one of them despite the almost 50 year age gap, and the management and staff at the Highbury Vaults, without whom our conversation would never have been half so productive.

Perversely, I spent much of the year studying physics, especially the structure of proper time, for which I must thank James Ladyman and Karim Thébault, but I continued to worry about the old problem. When I joined the PhD programme, Samir Okasha had the unenviable task of trying to get me to do something sensible. Eventually, although I had only a passing acquaintance with contemporary behavioural economics, we agreed that I should try to see what sense I could make of hyperbolic discounting.

The choice proved inspired. The topic turns out to provide a much clearer and better signposted route into the labyrinth and, whilst I had originally envisaged a self-contained formal 'fix', as I worked the pieces started to fall into place. Even the physics proved useful. But, for keeping me focused on the project in hand, I must first and foremost thank Samir. If I have strayed too far it is not for his failing to advise me.

Most importantly I must thank Geoff Keeling. By chance we started together on the MA programme and we have talked incessantly ever since, even though he is now in Palo Alto. He is a better philosopher than me, widely read, a great source of inspiration, and an astute critic. I am very fortunate to have him as a friend. Nearer home, I must thank Claire and the children and grandchildren, Mark and Vicky, Rachel and Chris, and Freya, Matilda, Oscar, and Phoebe, for tolerating my monomania. Further back, I should thank the long-gone staff at John Lyon School and the writers of the then-current physics, chemistry, and mathematics syllabus for instilling some fundamental competence despite my opposition, and the staff at Leeds, especially Bob Towler and 'Grebby' Grebenik, for indulging me as an obstreperous undergraduate and, I know not how, for recommending me to Pitt. I must also thank my parents. My mother lived to be 100, which is a good sign, but my father died at the ridiculous age of 65 without completing the historical research that occupied his final years. Tempus fugit. Finally, I must thank those I rely on most: librarians, publishers, and previous authors. If what follows is any good, it is largely down to them – although, of course, the faults are my own.

Author’s Declaration

I declare that the work in this dissertation was carried out in accordance with the requirements of the University's *Regulations and Code of Practice for Research Degree Programmes* and that it has not been submitted for any other academic award. Except where indicated by specific reference in the text, the work is the candidate's own work. Work done in collaboration with, or with the assistance of, others, is indicated as such. Any views expressed in the dissertation are those of the author.

SIGNED: NICHOLAS GEORGE AXTEN DATE: 10/03/2022

Contents

List of Figures

Chapter 1 Introduction

1.1 Intentional Action in Real Time

There is a standard theory of the aetiology of human action described, for example, by Davidson as follows:

> The choice of one course of action over another, or the preference that one state of affairs obtain rather than another, is generally the product of two considerations: we value a course of action, or a state of affairs because of the value we set on its possible consequences, and how likely we believe those consequences are, given that we perform the action or the state of affairs comes to obtain. In choosing among courses of action or states of affairs, therefore, we choose one the relative value of whose consequences, when tempered by the likelihood of those consequences, is greatest (Davidson 2004: 153).

A key feature of this theory, built into the standard notion of consequences, is that choices are made by reference to possible effects variously located in the relative future. This immediately raises an interesting question. To what extent if any does, or should, the relative value of the possible consequences of a course of action depend on their relative futurity? For example, should temporally more remote consequences count for less than temporally more proximate ones, other things being equal? If so, why? This is the question that I will seek to answer in the current work.

Although the theory, as stated, seems straightforwardly descriptive, the thought behind it is explanatory. It is that if we act consistently in the way described we will maximize the total value of the probable outcomes of our actions as we judge them, that this characterizes what it is to act rationally, and that humans are generally rational. Hence to the extent that we act as described, our choices can be explained as satisfying a general constraint, or set of constraints, implicit in our being rational. On this basis the theory is often described as normative in that it presumably characterizes what it means to choose rationally, or ideally rationally, whether or not humans generally do so and that it can be used prescriptively where they might not otherwise do so (Broome 1991: 90-5, Okasha 2016). For historical reasons, a formalized version of the theory in which it is assumed that value is aggregated linearly is called expected utility theory.

Whilst the theory may look simple, applying it analytically is not. The theory assumes, implicitly, that in any relevant situation an agent has certain options available, that each option, if chosen, entails various possible consequences each of which can justifiably be expected to occur, either immediately or after some delay with some relevant probability, and that the agent attaches value to each possible consequence on some reliable basis. Hence in order to apply the theory in a given case it is necessary to answer, from the agent's point of view, the following three questions:

1. What is an option?

2. What outcomes, or features of outcomes, are valued?

3. How are probabilities fixed?

Provided these questions are answered and it is assumed that the value an agent attaches to each relevant consequence remains constant or varies only predictably during any relevant period and that value is aggregated linearly, values can be determined from observed choices by what is called revealed preference methodology and choices can be predicted or explained in terms of these values.

Historically, most explicit analysis in terms of the theory has been in the fields of economics and experimental psychology, and researchers have generally answered the three questions listed above on the basis of what we may call a narrow definition of available options. In answer to the first question they assume that the set of relevant options consists of whatever is explicitly identified as an available opportunity in the given context, as conventionally classified. In commerce this is usually whatever is actually advertised as available to be bought or sold or whatever plan is on the table. In an experiment it is whatever the experimenter stipulates to be an option. In answer to the second question they assume that only outcomes or features of outcomes that are specified either directly or by direct inference within an option, as described, are relevantly valued, more or less in proportion to their stated monetary value or plausibly assumed desirability. And in answer to the third question they assume that probabilities are as stipulated by the researcher or as standardly or actuarially determined based on the narrowest appropriate reference class. Taken together, these assumptions prejudge the attitude that an agent may adopt in reaching a decision. They rule out, in particular, envisaged options, outcomes, and probabilities not considered relevant or appropriate by the theorist or experimenter. Experimental results are not infrequently rejected on the grounds that subjects appear not to be interpreting the task correctly.

There can be no a priori objection to these assumptions. But one consequence of adopting them has been, especially since the 1970s, the emergence of an increasing series of ‘anomalies’ – cases in which no consistent value assignments can be derived. The story has been told many times, for example by Starmer (2000). Whenever a new class of anomalies is discovered, theorists attempt to ‘patch’ the standard theory, generally by modifying the assumption that value is aggregated linearly. The effect has been to create a veritable menagerie of alternative versions of what Starmer calls non-expected utility theory. The effort is in many ways admirable, and it cannot be assumed that it will not end in success. But the results so far are equivocal. Anomalies continue to proliferate and the theoretical justification for the proposed patches, even where they are successful, is obscure and controversial.

The problem of anomalies is particularly acute in the topic I shall be examining, namely the relationship between value and futurity, because standard expected utility theory in its original form admits no significant relationship between value and futurity and hence the common observation that humans, and animals, typically prefer immediate rewards to delayed rewards is, prima facie, an anomaly. For this reason the topic is usually included in textbooks as an added extra, often in an entirely distinct section, as in Wilkinson (2008, Part 3) and Dhami (2016, also Part 3).

From a philosophical perspective the problem is of unusual interest. As noted, the original theory is usually assumed to define an intuitively justified normative standard of rational decision making rather than being merely a falsifiable descriptive theory of conventionally rational choice. But non-expected utility theory, as standardly presented, makes no such claim. Modifications are introduced purely on the grounds that they are justified by behavioural data (Dhami 2016: 5-10). Hence the ubiquity of anomalies appears to present a dilemma. Either the most straightforwardly justifiable and time-honoured principle of rational decision making embodied in standard expected utility theory must be abandoned with no obvious replacement available, or it must be admitted that humans rarely if ever satisfy this principle even when deciding carefully and hence that its explanatory value is void and its normative characterization is without significant effect. An extreme form of this dilemma relating to preference reversal will be described in §1.2.

There is, however, another possibility. It is that the identification of anomalies is, at least in part, an artefact of the methodology conventionally used in modelling observed decision making behaviour in its actual context. In other words, it is an artefact of the way the three question listed on page 10 are conventionally answered and of the way relevant standards of rational acceptability are defined. It is this possibility, particularly as it affects the inferred relationship between value and futurity, that I will be examining in the present work. The envisaged task is unavoidably complex. It involves investigating the way anomalies depend on and influence current modelling practice, and, equally importantly, developing a principled alternative response to each of the three questions. I shall proceed on this basis.

In the remainder of this chapter I will first give a brief overview of current research concerning the relationship between value and futurity in economics and philosophy, especially in relation to the issue of preference reversal. I will then describe and justify my proposed methodology. Next I will provide a more comprehensive review of current literature and of the main topics of current debate, and I will finish with a synopsis of the entire thesis and a summary of later chapters.

1.2 Discounted Utility, Preference Reversal

The question I wish to address concerning the relationship between value and futurity in decision making has been discussed most extensively within economics. It is standardly characterized as a question of time discounting (e.g. Frederick et al. 2002, Dhami 2016 Part 3). Given obvious facts about economic behaviour such as the usual expectation of a positive interest rate and typically observed priorities in investment and consumption it is generally taken for granted that short-term consequences are, ceteris paribus, given more weight in decision making than long-term consequences.

The currently standard economic analysis was first formalized by Samuelson (1937) in terms of discounted utility. His approach, following Bernoulli's response to the St. Petersburg paradox, described in §2.1, is to assume that the effective value, or utility, to an agent, of an expected outcome is not generally a linear function of its actual or monetary value. Samuelson's innovation was to propose, paradigmatically, that the utility of an outcome is an exponentially decreasing function of its expected futurity. He explicitly denied that this constituted a rationally justified evaluative principle but, given that economic theory is usually defended as a theory of rational behaviour (Hilton 2008: 10, Gintis 2009: 6), the denial has been frequently ignored. His formalization has remained influential within mainstream economics to this day (Frederick et al. 2002: 164, Dhami 2016: 583) but has been increasingly challenged. Evidence seems to support the assumption of a more convex, or hyperbolic, discounting relationship (Loewenstein and Prelec 1992).

The issue of time discounting has been discussed from a philosophical perspective by several significant authors. The usual assumption is that such an effect is ubiquitous in human decision making and the key issue is whether or not it is irrational. Opinions differ. Parfit (1984) proposes that it is justified by an agent's diminished psychological connection to their future self. Elster (2015: 103) asserts that in his view it is not irrational. Sullivan (2018) argues that all time discounting, not only of the future but also of the past, is irrational. But often the issue is ignored or set to one side as not of central importance. Broome, for example, says:

> I do not favour discounting myself. But I recognize there are arguments for it as well as against it, and I do not want to use up space in this book debating it (Broome 2004: 71).

Recently, however, the topic has received increasing attention owing chiefly to the discovery, attributed to Ainslie (1974) (Ainslie and Haslam 1992a: 65), that hyperbolic discounting entails preference reversal.

Preference reversal is a pattern of response frequently observed in both humans and animals in which expressed or predictable preferences reverse as constituent outcomes approach. A paradigmatic case is as follows. An agent *S* faces a choice between two alternative future outcomes *A* and *B* such that *A* is more remote than *B* and *S* initially prefers *A* to *B*. In due course, as both *A* and *B* approach, a moment is reached at which *S* prefers, and then continues to prefer, *B* to *A*. The smaller-sooner outcome comes to be preferred over the larger-later one as it becomes more immediately anticipated.

Several phenomena, including akrasia, impulsiveness, procrastination, temptation, addiction, and regret, that are of long established interest in philosophy and psychology are, as a result of Ainslie's discovery, now usually interpreted as involving preference reversal. As a result, interest in preference reversal is growing. A detailed investigation occurs, notably, in Bermúdez (2018).

Preference reversal is of particular interest in philosophy because, although it is a commonly observed pattern of response in human agents, respondents do not generally admit that they are acting unreasonably. This poses a version of the dilemma described on page 12, that either it must be denied that these respondents understand what it is to be reasonable or it must be denied that practical reason provides a diachronically stable criterion for rationally justified action. Neither of these is attractive. The former is, at least, condescending, and it places on theorists the onus of showing that they themselves, in their presumably rational conduct, avoid the implied error. The latter appears to undermine the almost universal philosophical assumption, implicit in the Davidson quote on page 9, that practical choice can be consistently optimizing.

This dilemma, if unresolved, adds to a long list of problematic conclusions within decision theory, evolutionary biology, and metaethics – including Sidgwick’s inability to reconcile duty and self-interest (1907: 508), Flood and Dresher’s discovery of prisoner’s dilemma (Flood 1958), Sen’s proofs of the impossibility of an ideal system of social choice (1970: 200), Gould and Lewontin’s critique of adaptationism (1979), and Parfit’s Repugnant Conclusion (1984: 388) – all of which tend, in their various ways, to undermine the assumption that a single principle of evaluative optimization exists. It is self-evidently a matter of some importance.

1.3 Methodology

In attempting to answer a question such as the one I am considering here it is usual practice in contemporary philosophy to pose the question, and then to develop a series of arguments based on a combination of prior theoretical claims and apparently plausible intuitive assumptions advanced by the author, leading to various conclusions concerning possible answers to the question. In such an approach, both the theoretical claims and the intuitive assumptions function as explanantia and the conclusions as explananda. The weakness of this methodology is that even if the argument is valid, it is sound only if the premises are true, and apparently plausible intuitive assumptions cannot be relied on to be true. On the contrary, there is hardly a single significant advance in the history of science that does not involve some apparently plausible intuitive assumption being false, and there is no reason to believe that the phenomena we classify as instances of human decision making and action are an exception to this rule. For example, there is evidence from data on the comparative timing of intuitive and neurological events that our intuitions about the nomological structure of decision making processes are systematically misleading (Chambon and Haggard 2013).

A possible response in theorizing about decision making and action is to seek to remove all reliance on intuitive data, as was proposed by the advocates of behaviourism (Thorndike 1911, Watson 1913) and, more recently, eliminative materialism (Churchland 1981, 2007). This is, in principle, misguided. Ultimately, all we have are intuitive data of one kind or another. All observation is ultimately intuitive. The key is not to discard intuitive data but to make a distinction between intuitive impressions *as data* and intuitive impressions *as truths* and, correspondingly, to treat occurrent intuitive impressions, whatever they are, as explananda rather than as explanantia in any process of theory construction.

This is, indeed, what we normally do. Suppose I see two trees *A* and *B*, and *A* appears taller than *B*. I am not obliged to assume that *A* is taller than *B*. I may do so, but I may instead construct an alternative account, provided that it gives an explanation of why *A appears* taller. For example, I may construct an account in which *A*, although shorter, is closer, and hence, by simple geometry, subtends a larger angle at the observer’s eye, and in which, in the relevant context, light travels approximately in straight lines, and that we judge the size of distant objects by the angle between incoming light rays at the eye as inferred from the size of the image projected on the retina, etc., etc. – and hence that tree *A* appears taller.

The crucial thing is that the account I accept will have all sorts of other implications in this and other cases and, in accepting it, I accept the obligation either to obtain relevant impressions as appropriate in those cases or to explain why I do not. I am not at liberty to pick and choose which impressions I attend to – to ignore ones that do not happen to suit my preferences. This obligation is surprisingly exacting and, as a matter of fact, we seldom if ever fully satisfy it. The best we can do is either to accept as veridical, or to explain as systematically non-veridical, as many impressions as possible, including whether or not a proposed account is an acceptable explanation, and, if we are serious about the whole project, to seek out new impressions that challenge our assumptions and, where necessary, reconstruct our assumptions accordingly.

Applying this rubric to the analysis of human decision making, the implication is that a proposed theory should either allow that choice depends on an evaluation of envisaged future effects more or less as it appears intuitively to do, or provide an account, based on other nomologically justified assumptions, of why it *appears* to do so. It is not sufficient to say that it operates in some other way without explaining why, nevertheless, it appears as it does. The intuitive impressions are an essential part of the data to be explained.

This methodology embodies a principle of reflective equilibrium (Little 1984) on the basis that epistemic justification depends ultimately on seemings (Tucker 2013) but it sets a more stringent standard of justification of claims of the non-veridicality of particular seemings beyond only the elimination of inconsistency – namely that in each such case the appearance of veridicality is explicable, notwithstanding its denial, in terms of other admitted assumptions. It does not self-evidently entail that a single best system of admitted assumptions either exists or is discoverable. Rather, it supports a process of marginal epistemic development in which apparent inconsistencies are discovered, investigated, and, so far as possible, partially resolved.

A key merit of this methodology is that it admits a rich supply of intuitive evidence to supplement or replace what is admitted under the conventional assumptions described in §1.1 without requiring that all such evidence be accepted as equally veridical. The former is important because if the conventional assumptions are relaxed without being supplemented or replaced there is a risk of triviality – that an analysis can always be contrived for any set of actual choices that trivially satisfies expected utility theory merely by specifying options, values, and probabilities to fit the data. The latter is important because, manifestly, intuitive evidence taken at face value does not form a consistent whole. By admitting intuitive evidence in this way, the proposed methodology rules out many theoretical proposals that, although perhaps achieving a superficial fit over a selected body of observations, are never plausibly explanatory (Okasha 2016: 421-5). Some more general philosophical implications are described in Chapter 7.

1.4 Literature

The pre-2002 literature on intertemporal choice, largely within economics, is very well surveyed in Frederick, Loewenstein, and O'Donoghue (2002) and the more recent literature in Ericson and Laibson (2018). It would be hard to improve on either, and I will merely summarize their content.

Frederick et al. introduces Samuelson's (1937) discounted utility (DU) model as the baseline for discussion and outlines its historical origins in the psychological theories of Rea, Böhm-Bawerk, and Irving Fisher and the emerging concept of time preference. It describes the DU model as one in which, "… all the psychological concerns discussed over the previous century were compressed into a single parameter, the discount rate," and as one that, as shown in Koopmans' "superficially plausible" axiomatization, involves assumptions of utility independence, consumption independence, stationarity, independence of discounting from consumption, time consistency, diminishing marginal utility, and positive time preference. With respect to the latter, Parfit's argument premised on diminished psychological connection to a future self is mentioned.

The paper then notes that research since the late 1970s has produced an increasing series of observations that appear to be incompatible with the DU model. The most prominent of these are more compatible with hyperbolic than exponential discounting. Many seem to demonstrate preference reversal. The paper includes an important meta-analysis showing that measured discount rates generally decrease with the assumed time horizon but notes that this effect can be explained by subadditive discounting, in which discounting is greater over more finely divided intervals. It describes other anomalies under a series of now-standard headings – the sign effect, the magnitude effect, delay-speedup asymmetry, preference for improving sequences, and other violations of independence – and discusses whether these are 'mistakes'. The discovery of these various anomalies has led to the formulation of a series of alternative models, many of which incorporate psychological rather than merely formal principles. Formal models are typically either hyperbolic or quasi-hyperbolic. Other models involve assumptions about self-awareness, habit-formation, reference-point based preferences, utility from anticipation, visceral influences, projection bias, mental accounting, choice bracketing, multiple-selves, and temptation. Examples of each are described and possible combinations discussed.

The paper then discusses the measurement of discount rates and presents a table summarizing the results of all known studies – 42 in total. It provides an interesting meta-analysis of reported discount rates against date of publication. The latter shows that reported rates are extremely variable, that the passage of time has not reduced variability, and that high or very high rates are common. It describes possible confounding factors that may contribute to high variability, including intertemporal arbitrage, non-linear utility, uncertainty, anticipated inflation, anticipation of variable future utility, and other psychological effects mentioned previously. It discusses the methodology, merits, and problems of field studies versus experiments, the methodological confounding of pure time preference with other factors, and the conceptual and semantic ambiguity associated with this. It proposes that the concept of time preference needs unpacking and advocates an explicitly psychological scheme involving three constituent motives – impulsivity, compulsivity, and inhibition.

The paper lists 219 references in total. It currently receives approximately 240 citations per year, many in papers in applied economics. It concludes as follows:

> … we believe that economists' understanding of intertemporal choices will progress most rapidly by continuing to import insights from psychology, by relinquishing the assumption that the key to understanding intertemporal choices is finding the right discount rate (or even the right discount function), and by readopting the view that intertemporal choices reflect many distinct considerations and often involve the interplay of several competing motives.

Ericson and Laibson (2018) is an NBER working paper written with advice from Loewenstein and Rabin that deliberately updates the "highly influential" paper just described. It aims to "review the latest research on intertemporal choice and identify important open questions". It takes the problematic character of time discounting for granted and introduces a meta-category of present-focused preferences defined in terms of the probability that an agent chooses, preferentially, an action that generates immediate experienced utility. It outlines methodological issues in the investigation of this effect and surveys currently proposed explanatory models and associated literature under a series of headings: quasi-hyperbolic models, unitary-self models with temptation, multiple-self models with simultaneous selves, objective risks that reduce future value, models with psychometric distortions, models of myopia, and models that do not generate present-focused preferences. It describes, in each case, a substantial theoretical and experimental literature showing both development and variation, mostly post-2002. It classifies all but the last in two dimensions: commitment versus no commitment, and dynamically consistent versus dynamically inconsistent preferences.

In order, so far as possible, to distinguish between substantive effects that need to be explained and effects that are current modelling anomalies the paper lists a series of problematical observations and associated questions. The questions include the following: why do individuals require such a high rate of return for money, why do people underestimate their own tendency to procrastination, why do small transaction costs affect choice disproportionately, why do households have such low levels of liquid wealth, and why do people prefer improving sequences. In each case relevant published evidence is described and discussed. The paper then similarly discusses a series of unresolved theoretical questions. These include: how soon is ‘now’, what is the role of temptation in decision making, how do psychological factors such as the perception of value and risk interact, how stable and consistent are time preferences across various domains, and how effective are self-management strategies. It concludes by arguing in favour of experimental paradigms evoking real effort rather than expressed monetary preferences. The paper lists 378 references of which 264 post-date the period surveyed by Frederick et al.

Beyond these survey papers, one classic text stands out as especially influential, namely the collection edited by Loewenstein and Elster (1992), *Choice over Time*. It provides a history of the development of the theory of hyperbolic discounting, a summary of its formal character, and an informative discussion of a wide range of supporting evidence and possible implications and applications. Dhami’s monumental work *The Foundations of Behavioral Economic Analysis* (2016) also deserves special mention.

The general trend in the economics literature cited, and since, is towards increasingly subtle testing of ever more diverse models under varying conditions. The set of currently proposed models embodies a variety of formal assumptions and envisaged mechanisms, most with several constituent parameters open to specification. Models are typically constructed in response to particular findings, tested using some plausibly relevant methodology, and revised to accommodate found discrepancies. The overall result is one of increasing conceptual complexity in which most developments are underdetermined by available evidence, given the unconstrained state of the currently accepted theoretical paradigm. For a more radical critique see Binmore and Shaked (2010).

There is however one distinct trend that appears to be emerging, namely to investigate the relationship between futurity and uncertainty. It is an issue that affects the interpretation of apparent time discounting and, in particular, the choice of the variable that is adopted as the mathematical argument of any proposed discounting function. As a matter of history, Bernoulli, in his response to the St. Petersburg paradox as described in §2.1, chose to discount utility relative to quoted monetary value, and Samuelson copied this in discounting utility relative to standardly measured futurity. But, as has been observed repeatedly, one might alternatively explain at least part of the St. Petersburg effect in terms of increasing uncertainty about the future. The infinite sum shown in (2.2.2) constitutes an appropriate analysis only to the extent that the casino can be relied on to pay the calculated amount, without limit, for any theoretically possible run of tails. This is not merely implausible but impossible. Unlimited funds cannot be available. The envisaged pay-out is, therefore, increasingly uncertain. It is possible that a similar correlation between futurity and justified uncertainty may occur much more widely.

Some early observations on the formal parallels between futurity and uncertainty are found in Prelec and Loewenstein (1991). More recent papers include Keren and Roelofsma (1995), Hong and Sagi (2003), Ng (2005), Weber and Chapman (2005), Epper et al. (2011), Rao and Li (2011), Takahashi (2011), Andreoni and Sprenger (2012), Schmidt (2014), Andreoni and Sprenger (2015), Miao and Zhong (2015), Hardisty and Pfeffer (2017), Luckman et al. (2017), Konstantinidis et al. (2018), Abdellaouia et al. (2019), and Chakraborty et al. (2019). They include both theoretical and experimental investigations, particularly of whether uncertainty and delay have qualitatively distinguishable effects. This issue is of considerable philosophical interest. It will be discussed at greater length in later chapters.

Non-dismissive consideration of time discounting in philosophy is less frequent but increasing. A major contribution occurs in the eleven essays included in Bermúdez (2018). As is usual in academic philosophy, these essays focus primarily on problems arising in an attempt to give a revised interpretation of historically sanctioned philosophical concepts in the light of contemporary extra-philosophical research. Mele (2018: 204) gives a helpful characterization as follows:

> In a project description that accompanied his invitation to contribute to this volume, José Bermúdez offered a definition of "a paradigm case for discussions of self-control" and asked two questions about the paradigm case. He wrote:
>
> > The paradigm case occurs when an agent makes at time t_1 a commitment or resolution to pursue a large, long-term benefit (henceforth: LL) at a later time t_3. At a time t_2, later than t_1 and earlier than t_3, the agent has the opportunity of a small, short-term reward (henceforth: SS). Although at the time of making the resolution the (discounted) value of LL is more powerfully motivating than the (discounted) value of SS, by t_2 the agent's preferences have temporarily reversed and now SS motivationally outweighs LL.
>
> His questions were these:
>
> > 1. How is it possible to exercise self-control in the paradigm case?
> > 2. When, how, and why is it rational to exercise self-control in the paradigm case?

The contributing authors provide a very diverse range of responses to this challenge. There is, however, no consensus as to how to eliminate the apparent explanatory redundancy of the concept of self-control within the standard theory of decision making as described by Davidson – namely that if choice maximizes expected utility then no separate faculty of self-control is required, and if it does not then any separate faculty of self-control is ipso facto ineffective. Likewise there is an unresolved tension between treating rationality as non-erroneous realization of de facto current preferences and treating it as imposing an additional – perhaps an 'all things considered' – external standard. These issues will be discussed more fully later.

Another significant philosophical discussion occurs in Sullivan (2018) – following Greene and Sullivan (2015) – in which it is argued that any kind of temporal bias in individual preference, not only with respect to relative futurity but also of the past versus the future, is irrational. The arguments are convincing, but they depend on a particular interpretation of the notion of preference in which *A* prefers *X* iff *A* judges that it would be better if *X* (see, for example, the definition of preference and regret on page 93). Given the associated assumption that rationality involves a principle that things should be best, a rejection of temporal bias in preference emerges more or less as a tautology. The argument does not, however, extend to actual judgements about what would be better in situ, since they depend also on nomology and prevailing boundary conditions, which are not temporally neutral. For example, it is often better to wait for more information before making a decision or to wait for conditions to be favourable before acting. The conclusion, then, applies to only a very restricted class of preferences. Suhler and Callender (2012), starting from less narrow assumptions, reach a diametrically opposite conclusion.

Several recent papers, whilst not directly concerned with the specific issue of intertemporal discounting, are nevertheless very relevant to the present enquiry. Schwarz (2021) identifies what he calls the problem of options, describing it as "an important (but largely ignored) gap in the foundations of microeconomics". He argues that the options among which an agent can choose are subject to significant constraints besides their association with valued outcomes – constraints that he calls ability, cover, and maximality. He develops, in outline, a revised theory to accommodate these constraints that makes reference to an agent's available methodology as well as to beliefs and preferences. The development has significant consequences, particularly concerning the propositional representation of alternatives.

Fumagalli (2020b) investigates a parallel problem arising in the characterization of options, which he describes as a "lacuna in the ... literature", namely the problem of specifying which features of an expected outcome are to be treated as significant for the purposes of computing its expected value. He considers, for example, what features of a food item, in context, might make it more or less desirable. The problem is that since the set of possible descriptive features is generally unbounded, unless some theoretical or methodological constraints are imposed there is a risk that any resulting analysis is trivial or unfalsifiable and/or that the supporting theory is overdemanding or incoherent. He argues that the problem is not insuperable since suitable constraints can be envisaged. He does not, however, describe any particular methodology for identifying relevant features beyond referring to "plausible reasons and/or evidence". He concludes that since some constraints can in principle be specified the problem "does not license general skepticism" about the possibility of justified decision theoretic analysis nor, consequently, about the explanatory power of rational decision theory. Although he does not explicitly discuss the issue, it is plausible that the features he admits as possibly relevant may include ones that vary with assumed futurity. Hence his analysis opens a parallel route into the investigation of apparent discounting.

1.5 The Current Debate

It is evident from the literature cited and other associated references that different authors hold very diverse views as to how decision making is to be conceptualized. This is perhaps not surprising since the status of intentional action in the physical world, of which decision making is a key constituent, is itself widely considered problematical. Hence there is generally no sound basis on which to build a robust response. Attempts to resolve any single problem often have the effect of exposing other problems elsewhere. For example, terms such as 'probability' and 'value', although apparently standard, have no self-evident interpretation. Verbal definition will not generally solve this problem since its usual effect is, in part, to transfer the burden of uncertainty from one term or concept to another.

One effect of this diversity is that attempts to treat the role of futurity in decision making as a self-contained subfield within expected utility theory typically fail. This is clearly seen in the commentaries provided by both Frederick et al. and Ericson and Laibson. It is my contention that this is inevitable so long as a number of widely discussed issues or dilemmas remain unresolved. In this section, therefore, I will briefly describe this set of problematical issues as they appear in the current literature. My implicit claim is that unless the theorist adopts a clear position with respect to each of them any resulting system will be analytically opaque. Much of the current work constitutes an attempt to vindicate this claim.

Firstly, a contrast is often made between thin and thick versions of decision theory, for example by Fumagalli (2020a). This corresponds, approximately, to an older contrast, dating back at least to Einstein, between principle theory and constructive theory (Brown 2005: 71) and perhaps to an even older notion of 'saving the phenomena'. The idea is that a principle or thin theory derives its explanatory power from the satisfaction in relevant phenomena of specified generic constraints whereas a constructive or thick theory derives its explanatory power from modelling the predictable operation of constituent processes. The latter type of theory may also be described as mechanistic, but this may appear to imply a greater degree of commitment to a unique ontic realization than is warranted by available evidence, as is illustrated in the critique of the search for 'true' utility outlined by Fumagalli (2013, 2019), and so is perhaps better avoided.

Usually in science the development of principle theory precedes the development of constructive theory. For example, Mendelian genetics preceded genomics and the theory of chemical valency preceded the theory of electron exchange. A principle theory such as general relativity can exist without any established constructive equivalent but the inverse is more problematical. For example, geology, although full of descriptions of constituent processes, has an overarching structures given, at least, by uniformitarianism, but geography, lacking such a principle, is doubtfully a science. A principle theory without a corresponding constructive theory must nevertheless be constrained by assumptions about constructive possibility. For example, Mendelian genetics must admit some biochemical mechanism. Research leading to the discovery of the DNA mechanism proceeded on this assumption (Watson 1968: 23). Ultimately, a developed constructive theory often reveals the corresponding principle theory as an idealization and allows constructive explanation of both ideal outcomes and marginal exceptions. Nevertheless, the principle theory, if generally validated, tends to retain its broad explanatory or didactic value. For example, chemical valency, as such, is assumed in standard diagrams of molecular structure and in associated nomenclature.

Applying this analysis to the issue discussed by Fumagalli, the implication is that provided that the notion of rational decision making admits a class of phenomena identified in terms of their approximately satisfying some relevant generic constraints and that the possibility of an eventual constructive account is not excluded by other compelling factive considerations there is no reason why a principle, or thin, theory of decision making should not be explanatory in much the same way as other principle theories are. No particular constructive account needs to be advanced, but the proposed theory must be compatible with constraints arising from non-negotiable features of the cognitive, computational, neurophysiological, biological, social, and ultimately physical context that might eventually be the subject matter of a constructive account.

There is considerable debate, however, as to how to characterize the relevant principles. Given that the envisaged class of phenomena comprises instances of rational decision making, the debate reduces largely to one concerning the definition and role of rationality. At one extreme is a view advocated by De Finetti (1937) which admits only the barest constraints of subjective consistency. At the other is a view advocated by Cosmides and Tooby (1997) that rationality is, in effect, an optimal adaptive response to mostly prehistoric environmental contingencies. In between are various formal proposals based on ideas developed by von Neumann and Morgenstern (1944), Savage (1954), and Jeffrey (1965) which characterize rationality in terms of a set of axioms to be jointly satisfied by any admissible set of expressed preferences (e.g. Joyce 2010, Bradley 2017) and, responding to a different line of thought, that intentional action must satisfy a global principle of free energy or prediction error minimization (e.g. Friston 2010, Hohwy 2013, Clark 2016).

Behind this divergence of views is a debate as to whether the rationality of preferences should be defined primarily in terms of their consistency or their aetiology. The issue can be traced back to Hume's (1739) claim that reason alone can never give rise to volition and hence, by implication, that preferences can be rationally criticized only in virtue of their mutual inconsistency. An alternative post-Darwinian view is that rationality is a form of evolved adaptive fitness and extends, therefore, far beyond mere consistency. The contrast is well illustrated in alternative accounts of the grounding of human cooperation – that it is either a tactical device to maximize each individual's payoff or an evolved trait or character virtue tending, in general, to promote adaptive survival or long-term success (Gintis 2011). An intermediate position is suggested in Fumagalli's (2020b) proposal that preferentially significant features of available options are open to identification on the basis of other evidence, since this does not limit them, a priori, to justification in terms of only either consistency or adaptive fitness alone. It opens the aetiology of preference to significantly wider theoretical investigation.

Cross-cutting this debate is another concerning whether decision theory should be viewed as descriptive or normative. If what is envisaged is a principle theory governed by a principle of rationality, the underlying assumption may be either that it represents a rational ideal to be advocated but not always achieved or that it models normal competent performance – or perhaps both, depending on the circumstances. A complication is that, as illustrated, standard axiomatic theory does not readily satisfy either characterization. It is often descriptively false and, viewed normatively, it imposes a standard of superficial consistency that often seems insufficiently responsive to the relativity of prevailing conditions. Hence the debate is easily diverted into attempts to solve either of these two problems. Among cognitive psychologists there is a parallel dispute in what is called the great rationality debate (Dhami 2016: 47). The issue is most acute in the many projects in behavioural economics to build predictively accurate models of decision making in response to available experimental data. As described above, the pursuit of descriptive accuracy has tended to produce models that have little obvious normative merit. This is widely thought to threaten the very foundations of rational choice theory and hence the fundamental status of economics (Dhami 2016: 29-32). But at a more philosophical level of debate it is not uncommon for theorists to adopt a relatively non-committal position, allowing that decision theory may be interpreted either normatively or descriptively depending on current explanatory aims (e.g. Okasha 2018).

On a more technical theme, considerable debate surrounds the widespread use of revealed preference methodology. This is a standard method, based ultimately on von Neumann and Morgenstern’s (1944) representation theorem, of constructing an interval measure of value for a given agent over available options from expressed preferences over probabilistic mixtures of described alternatives (cf. Wakker and Deneffe 1996). The key difficulty is that its effective use depends on a number of assumptions about the individuation and classification of options, the specification of probabilities, the consistency and invariability of an agent’s preferences, and the relation between preferences and expressions of preference. If, as is often the case, it fails to generate consistent results, it may be unclear whether the problem is that one or more of these assumptions is unsatisfied or that the assumption that an interval measure of value ought to be definable is false. If, conversely, assumptions are adjusted to render computed results consistent, the effect may be to make the entire derivation trivial. Both Fumagalli (2020b) and Schwarz (2021) conclude that some additional methodology is needed to justify claims about, in particular, the individuation and classification of options.

Underlying this is a wider dispute about the relationship between economics and psychology, often dated to the influential proposal by Pareto in the late 19th century that the scope of economics should be limited to the analysis of observed patterns of choice characterized without reference to psychological evidence. As widely noted, this leaves a problem of how to classify choices (Bruni and Sugden 2007, Fumagalli 2016, Dietrich and List 2016, Guala 2019). For example, if a driver sometimes turns left and sometimes right, are these choices to be classified thus or on some other basis? Pareto implies that there is a “pure naked fact” of the matter (Bruni and Sugden 2007: 155); but it is difficult to see how this is to be established except via unacknowledged psychological evidence. This obvious gap may account for the failure of later theorists to eliminate psychological elements from economics, as noted by Fumagalli (2016: 111-3).

An associated problem of specifying probabilities is also very widely debated, usually within a Bayesian framework in which it is assumed that the key task is to develop a calculus of subjective probability over a set of admissible propositions (e.g. Carnap 1950, Roeper and Leblanc 1999, Bradley 2017). However, this project only indirectly assists in solving a more immediate empirical problem, namely to specify how an agent can or should construct particular predictive probability assignments over outcomes associated with apparently available options given currently available evidence. Discussion of the latter typically reduces to a debate over the intractability of the reference class problem (e.g. Eagle 2004, Hájek 2007, Hájek and Hitchcock 2016). The problem is currently unsolved.

A number of other issues are implicit in the ongoing debate about methodology. Two were alluded to above in reference to possibly problematic theoretical assumptions, namely the intertemporal variability of individual preferences and the contingent relationship between preferences and expressions of preference. Another is that preferred conditions may be unobservable and hence that agents must often rely on proxies. And on a different dimension, the distinct motivational status of roles and duties is discussed to a limited extent in applications of game theory (Gintis 2009: 75) and in metaethics (Brink 1997), and somewhat more widely in the investigation of normatively coordinated action (Bicchieri 2010).

Finally, there is considerable debate in the wider philosophical literature concerning the causal status of choice and action (McLaughlin 1991, Gillett and Loewer 2001, McLaughlin and Cohen 2007, Price and Corry 2007). The issue is alluded to in discussion of causal decision theory but remains unsolved (Schwarz 2021). There is an obvious argument that it urgently requires a solution since unless the manner in which choice can be causally effective is explicated, any theory of its operation is seriously defective. The problem is standardly finessed by treating decision theory as a theory of preferences over described outcomes rather than of actions, but this significantly restricts the scope for explanatory analysis since it obscures the role of causal feedback from prior consequences. Indeed, it comes close to undermining the notion of consequences, on which the entire theory of rational choice is presumably based.

It is clear from this brief overview of current topics of debate that there are a number of foundational problems that inhibit the development of a coherent explanatory theory of rational choice and hence, in particular, of the relationship between differential futurity and expected value within that theory. An investigation of each of these topics will, to a large extent, form the subject matter of subsequent chapters. In consequence the scope of the enquiry is unusually wide, but it must be emphasized that this is a result of the state of the current debate not an unmotivated analytical preference. I will outline its structure in greater detail in the following section.

1.6 Synopsis

The present work may best be viewed as an attempt to follow through on the recommendation of Frederick et al. quoted on page 20 to “relinquish … the assumption that the key to understanding intertemporal choices is finding the right discount rate (or even the right discount function), … readopting the view that intertemporal choices reflect many distinct considerations and often involve the interplay of several competing motives”, and of Ericson and Laibson, quoted on page 21, to investigate “how … psychological factors such as the perception of value and risk interact, how stable and consistent are time preferences across various domains, and how effective are self-management strategies”. It involves, especially, investigating the relationship between futurity and uncertainty described on page 22.

Its style might, perhaps, be described as naturalistic, but this is rather uninformative in that there is little agreement as to what naturalism consists of beyond that it involves taking well established scientific theory seriously and not relying only on armchair speculation (Ladyman 2002: 4, Ladyman and Ross 2007: 5-7). The difficulty in the present case is that the various bodies of well established scientific theory that bear on the observed structure of human decision making are relevant only as background constraints. For example, neurophysiology, evolutionary biology, and information processing theory are all evidently relevant but they do not in themselves provide an unambiguous guide as to how explanatory analysis should proceed. Indeed, their implications are much disputed. The implications of behavioural economics are, self-evidently, uncertain.

The work’s style might be better described as Galilean. What this implies is illustrated in the contrast between Galileo’s abandoned proposal of 1590 that a body falls at a constant speed proportional to the difference between its density and the density of the surrounding medium, and his later proposal described in a letter to Paolo Sarpi dated 16 October 1604, that “the spaces passed over [by a falling body] in natural motion are as the squares of the time [since its release]”. The difference lies not in an abandonment of abstract modelling, nor in a specific reference to measurable variables, nor in any assumed universality, but in detailed attention to the dynamical form of the phenomena and a rejection of ‘hand waving’ accounts of approximations and disparities. As it happens, the dynamical form of the trajectory of an unrestrained falling body is mathematically simple – although the mathematics was a challenge at the time (Hanson 1958: 37-49) – whereas that of rational choice is evidently complex even in relatively simple cases and appears to depend on a considerable variety of contextual factors. But the analytical assumption is similar. It is that marginal adjustment to historically sanctioned formulas is unlikely to be sufficient and that what is required is a dynamical analysis that pays very careful attention to the precise role, if any, played by various factors conceivably involved.

So the approach adopted involves taking the description of decision making given by Davidson seriously and enquiring, in detail, about each aetiological component implicit in its formulation. These are approximately the topics of current debate outlined in §1.5. The assumption is that only if all these ‘moving parts’ are delineated to at least a first approximation can the system as a whole be properly articulated. It is assumed that the more usual policy in academic philosophy of examining one conceptual aspect at a time cannot readily succeed since it is the system as a whole that is functionally effective in virtue of the relationships among its parts, not any of its parts separately.

A striking parallel is Schwarz (2021). His investigation of what he calls the problem of options finds frequent echoes in Chapter 5, with some reservations that will be noted later. His general attitude to the task of devising an adequate theory of human decision making, and especially the inadequacy of current modelling, is very similar to my own. The chief differences are that he maintains a conventional type of analysis that omits time as a significant variable and that he aims only to provide a theory that covers “situations in which the agent has full control over her decisions”

(2021: 189), which he admits leaves significant residual problems. I believe that no such cases exist, or very few, and hence that a considerably broader body of theory needs to be devised.

The implied task is unavoidably complex. I hope therefore that the reader will forgive me if, firstly, I cover topics that may at times seem only loosely related and, secondly, I do not cover every topic in precisely the degree of scholarly detail that might otherwise be expected. The work is, in effect, an experiment in analysis. I hope that in the end it speaks for itself. In order to aid understanding I include a map of the principal dependence relations among key topics on page 41.

The task as it currently presents itself is unintelligible except by reference to the history of formal modelling in behavioural economics and decision theory. I begin, therefore, in Chapter 2, by providing a brief survey of this modelling. As mentioned on page 12, the topic is usually divided into two parts – models of preference as dependent on probability and models of preference as dependent on futurity. However, in view of the emerging recognition of interdependence between probability and futurity as described on page 22, this division is unhelpful. Its origin lies in the way probabilities have been conventionally quantified, which generally excludes futurity effects by fiat. Hence, in surveying existing models I will include both types without prejudice. A number of the issues outlined in Chapter 2 will be of considerable significance later in the analysis, particularly in Chapter 7. To assist the reader, a brief survey of experimental results generally classified as anomalies is given in the Appendix.

Insofar as probability varies with futurity it is conceivable that all observed futurity effects might be explained entirely in terms of variations in probability. On the other hand, there may be other explanatory factors involved. If so, they may or may not be rationally justified. I do not, a priori, either assume or deny any of these possibilities. My aim is to examine all apparent alternatives and, in due course, to model their implicit consequences. Ultimately it transpires that there are several factors, both subjective and objective, that vary more or less regularly with futurity. It is these that are the principal topic of investigation.

The chief focus is on the three issues described on page 10 – values, options, and probabilities. Before reaching this point, however, a number of preliminary issues need to be disposed of. The first is the question of rationality. As described, the standard theory is taken to be explanatory in virtue of its characterizing what it means to choose rationally, but there is little agreement in the literature as to what rationality in this context consists of. Requiring only consistency raises the difficulty that unless an account is given as to why prejudice should not outweigh evidence – which in turn requires a substantive theory of evidence – it appears to offer no clear defence against solipsism. Some additional criterion of epistemic or practical success, or adaptation, appears to be warranted. Sullivan (2018), for example, assumes that it involves preferring the best.

There is a convenient route into this issue via what is called the great rationality debate. This is the topic of §3.1. It raises three subordinate issues – the analytical status of behaviourism, the contrast between principle theory and constructive theory, and the relationship between rationality and adaptation. The underlying issue is the extent to which rationality can be characterized in terms of the satisfaction of generic criteria – and, if so, what those criteria are – versus in terms of the agent's deployment of relevant, presumably evolved, heuristic methods. The analysis suggests the possibility of a principle theory of decision making in which rationality, constrained by prevailing conditions, is the fundamental explanatory principle. This is the route I will eventually pursue.

The usual intuitive notion of rationality admits that it is constrained, at least in part, by the demands of hereditary survival that drive biological adaptation. I argue, however, that not all constraints are biological. At least two other recognizable evolutionary processes constrain rationally justified outcomes: cultural and cognitive. These, having much higher rates of evolutionary change, cannot be driven by the same demands of hereditary survival. I propose that cognitive evolution, in particular, is driven by the demands not of hereditary survival but of predictive success. Much of the remaining analysis hinges on this claim. It requires, inter alia, that rational agents have an evolved ability to formulate and test predictions. This in turn demands analysis.

Such an analysis reproduces a problem much discussed in evolutionary biology: adaptationism. This is a key topic of §3.2. The problem is that claimed explanations of particular phenotypic outcomes on the basis of assumed evolutionary demands frequently involve what are called 'just so stories'. A solution is proposed in §3.3, namely that evolutionary demands can be characterized in terms of a partially ordered set of existential problems that jointly characterize an ecological niche and admit, by cumulative analysis, increasingly precise aetiological explanation of particular phenotypic outcomes. This mode of analysis will be of key importance in Chapter 7.

It is proposed, then, that rationality depends on an evolved ability to formulate and to test predictions. This is explicable on the assumption that human agents possess an evolved computational capacity that answers, at least approximately, to the logical processing of conventional propositional content. The implications of this assumption, particularly concerning the relative status of connectionist versus algorithmic modelling, are discussed in §3.4. Again, this has significant later implications, especially in the analysis of probability assignment in Chapter 6.

There remains a fundamental issue mentioned in §1.5 – namely, on what basis can choice be causally effective. It is often treated as a metaphysical problem to be discussed in terms of physicalism and the status of meaning, but it is in fact a straightforward explanatory problem. Only insofar as choice, as such, is admitted to be causally effective does decision theory have any explanatory significance or, conversely, is the emergence of a capacity to choose explicable in terms of its adaptive consequences. It raises two fundamental issues – of causal determinism and temporal asymmetry. They are discussed in §3.5. Working from a suggestion made by Reichenbach, a theory is proposed, based on thermodynamic asymmetry, that characterizes records as recognizably improbable patterns. This provides a theory of evidence which in turn supports an explanatory account of the development of temporally asymmetric nomology and nomological inference, and hence a category of testable predictions, as required in the proposed account of rationality. It perhaps goes without saying that a substantive theory of evidence is of independent epistemological interest.

These various considerations are combined, firstly in an account of the causal role of choice in the determination of otherwise underdetermined outcomes in §3.6, and then in §3.7 in an investigation of the possibility of creating a principle theory of rational choice based on an analysis of constraints arising from adaptive problems defined in terms of not only biological but especially cognitive success. This will re-emerge as crucial in Chapter 7. Chapter 3 finishes with a summary of interim conclusions.

Chapters 4, 5, and 6 together contain an investigation of the three issues described on page 10, concerning the characterization, quantification, and explanatory role of values, options, and probabilities. The analysis rests on the assumption that to have any significant hope of ultimate explanatory adequacy it must take account of a wide variety of psychological and other evidence, especially concerning the functional constraints under which actual human decision making operates. The fundamental point is that agents cannot generally choose among consequences. As emphasized by both Schwarz and Fumagalli, they must choose among options. They must, then, have some way of identifying available options, of associating probable consequences with available options and, in anticipation, of assigning value to possible consequences, all based on imperfect data. Furthermore, both options and consequences may be arbitrarily complex, and options may come with all sorts of preconditions and probable additional consequences and admit collective, vicarious, or accidental realization. Even a choice to express a preference among described conditions with no expectation of their being realized involves some associated preconditions and probable consequences that are quite separate from the conditions described. A realistic theory of choice must acknowledge all these considerations. Manifestly, the usual modelling methodology does not do so.

Chapter 4 begins with question 2 – What is valued? It examines the characteristics that envisaged conditions must have in virtue of which they have, or may have, motivational value. It concludes that they must exist as quasi-representational features within a partially algorithmic computational system and that they acquire value either a priori or by statistical association. On this basis, issues of temporal structure, variety, intensity, separability, and functionality within the value system are examined in detail. Key conclusions are that inference to probable consequences requires that an agent has access to something that answers to a causal model of the world, that targets of valuation can be analysed into object-quality pairings evaluated combinatorially, in which objects and qualities are assigned positive or negative value independently by type, that interval scales of magnitude can generally be constructed from ordinal data by ranking, and that double counting of valued conditions is both ubiquitous and unavoidable. The result is a model of valuation considerably at variance with the usual model assumed in revealed preference methodology.

Chapter 5 continues with question 1 – What is an option? In particular, it examines the way options are defined or constrained by presumably achievable means. It defines three typical scenarios: S1, in which an option is a possible action, S2, in which an option is an externally specified and routinely achievable condition, and S3, in which an option is a desired outcome. It examines the process of planning and realization involved in each case, the evolution and evaluation of alternative methods, the role of fluency in action, epistemic activity, commitment, and the significance of roles and duties in constraining choice. Key conclusions are that achievability is recursively defined, usually with respect to an evolving system of partially envisaged methods, that planning terminates in either null or fluent constituent units of activity, that planning frequently involves intervening epistemic activity and hence nested choices, which admits a wider range of options than is typically assumed in standard modelling, that a valued condition of commitment is needed to prevent continual re-evaluation, and that a social context constrained by defined roles and duties substantially alters the range and reliability of available options, especially in relation to remote or delayed effects.

Chapter 6 considers question 3 – How are probabilities fixed? It first examines a number of issues associated with the concept of probability, particularly the reference class problem. In response it outlines a theory of probability according to which probability judgements are shaped by the adaptive value of successful prediction and in which, in the absence of determinate information, the best predictive policy is to assume that a currently envisaged case is typical of a known class of similar cases for which statistical data exist and to assign probabilities that reflect such data. It investigates sources of uncertainty admitted by this theory, including vagueness, limited data, limited attention, bias, deception, and physical unpredictability, discusses the basis on which nomological assumptions are derived, and identifies two opposing secular trends, adaptation and decay, that justify contradictory long-term predictive attitudes.

Chapter 7 combines the conclusions brought forwards in the previous chapters by investigating, in various typical cases, the predictable relationship between futurity and rational choice in the context of an established system of valuation, available methods, and partly resolvable uncertainty. The outcome is a body of theory, and an associated scheme of analysis, of the dynamics of rational preferability – and hence of rational choice – over apparently available options. It is a principle theory rather than a constructive theory as distinguished in §1.5 in that it rests on the analysis of rational and contextual constraints rather than on the modelling of heuristic or other methods developed to satisfy these constraints. Hence it satisfies the criterion of admitting normative explanation as described on page 9 whilst departing very significantly from the usual analysis in which it is assumed that agents are, in Bradley's words, "logically omniscient and maximally opinionated" (2017: 1).

The analysis identifies several distinguishable issues that constrain decision making. On this basis it models the typical dynamical profiles of six paradigmatic modes of futurity-dependent decision making, identified as involving promptness, deliberation, second thoughts, transitional uncertainty, forced choice, and long-term projection. These six have significantly different diachronic forms that variously exhibit hyperbolic, exponential, and other second order features. They together account for a wide variety of observed effects.

It is argued in §7.6 that these effects, although generally predictable, are open to higher level rational evaluation and possible amendment, realizing a principle of dynamic normativity. §7.7 considers issues of testing, particularly by reference to results described in the Appendix, and §7.8 discusses the significance of the foregoing analysis in epistemology and in philosophy more generally.

The Appendix contains brief descriptions of various research results that appear as anomalies within the current class of economic models. These are referred to at various points throughout the text in the form (A*n*). Not all involve variable futurity but all are relevant to the question of what type of modelling of intertemporal effects is empirically acceptable and, in particular, what assumptions underlie the quantification of value, futurity, and probability. The cases described include results that appear to violate generally accepted axioms of rational choice, results that exhibit unexpected effects that depend on differences in magnitude, probability, or futurity, and various more complex or persistent anomalies.

To aid understanding a map of the principal dependence relations among key topics appears in Figure 1 overleaf.

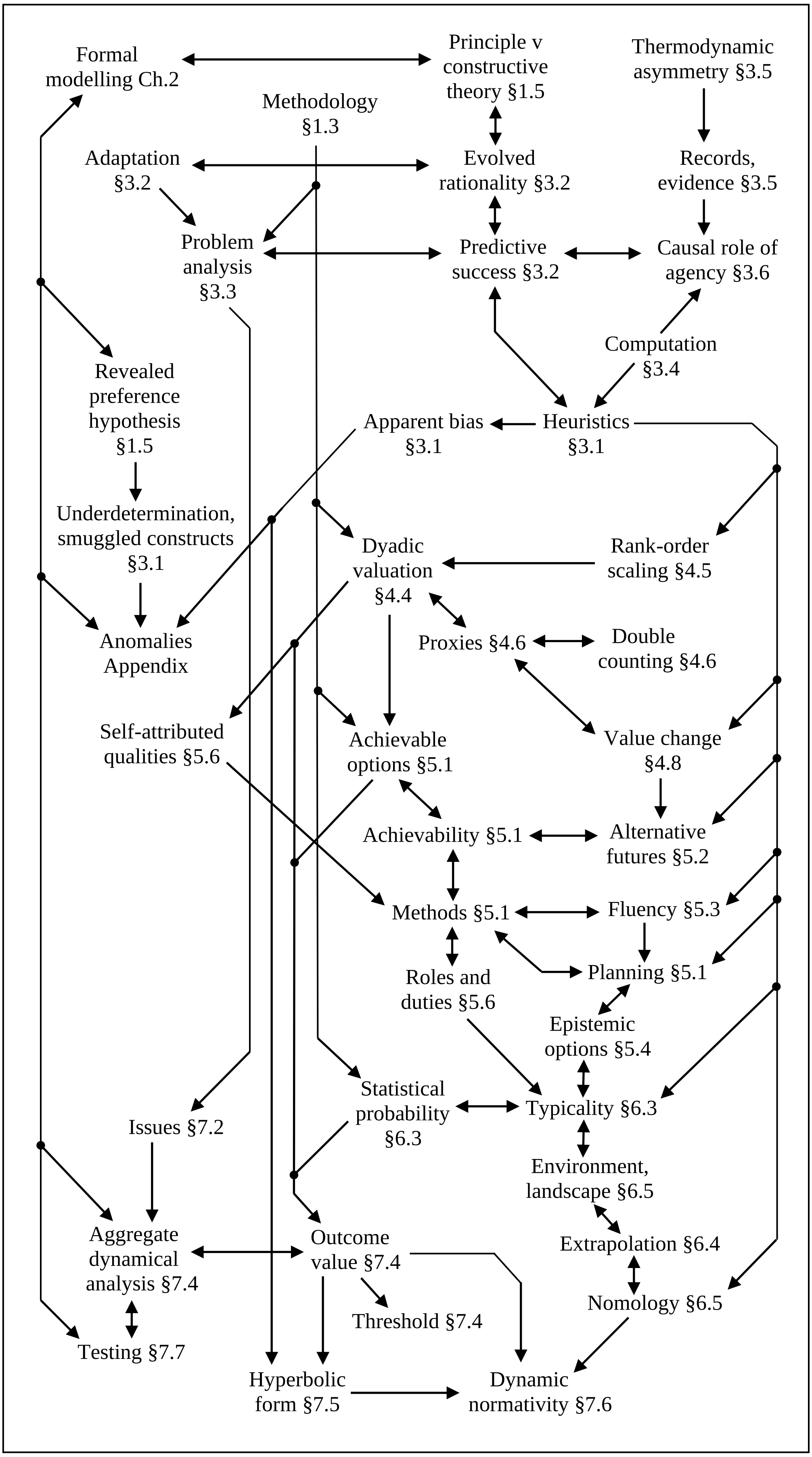

Figure 1: Conceptual structure

Chapter 2 Decision Making Models

2.1 Expected Utility

The origins of the modelling of decision making lie very largely in the analysis of gambling. A key practical problem was how, in a prematurely terminated game of chance, to divide the remaining stakes fairly between players. In the 1650s Pascal, Fermat, and Huygens all studied this problem and collectively propounded the principle that the current value of a future gain is proportional to the probability of obtaining it. This – the product of the notional value of a possible outcome and the probability of obtaining it – is its expected value. More generally, if an option (usually called, for historical reasons, a lottery) is a proposal reliably promising exactly one outcome from a set of mutually exclusive possible outcomes $\{x_1, \dots , x_n\}$ having, respectively, notional values $\{v_1, \dots , v_n\}$ and probabilities $\{p_1, \dots , p_n\}$, the expected value of the option is

$$\sum_{i=1}^{n} p_i \, v_i \qquad (2.2.1)$$

and the optimal policy of a player who wishes to maximize their winnings when offered a set of options is, ceteris paribus, to choose an option with the greatest expected value. In particular, a rational player offered the opportunity to purchase such an option ought to be willing to pay any sum up to its expected value.

This model is most obviously applicable to reliable finite games of chance in which stakes and outcomes have fully commensurable monetary values. To employ it more widely requires making a number of assumptions, as described in §1.1. The first is that in the target situation a well defined set of options can be identified, each involving a set of mutually exclusive possible outcomes each with a well defined probability within that option. This may be problematical even in games of chance. The options and odds offered or implied in conventional play may not accurately represent what may actually occur. Deviations may arise from, for example, a failure to observe relevant social or contractual norms or from various kinds of external disruption. To accommodate this a distinction is standardly made between risk and uncertainty. A risk is a possibility that has a well defined non-integer probability in the relevant context. Beyond this lies uncertainty (Knight 1921, Keynes 1921). Risk assumes, among other things, a unique reference class over which the relevant probabilities are quantified (Eagle 2004, Hájek 2007).

A second assumption is that all relevant outcomes have fixed and commensurable quantitative values regardless of their probability. In some cases a current market price will suffice. But for many valued outcomes no well defined market price exists, prices fluctuate, and, in any case, it is axiomatic in economics that agents' valuations may differ from the current market price for otherwise there would be no economic justification for exchange. Variously delayed versions of the same outcome may not be treated as equivalent. Outcomes with very low probability may be ignored.

A third assumption is that, in the target case, no other significantly valued possible condition, anywhere, at any time, has a probability that depends differentially on the option chosen. There is, in the modern phrase, no possibility of collateral damage or, conversely, collateral benefit. These assumptions are, no doubt, epistemically if not metaphysically onerous. The third is especially so, since it requires that it is possible in principle to estimate all significant long-term probabilistic consequences of any current choice. This is far beyond anything that can be justified on the basis of current scientific knowledge. Nevertheless, it does not prevent the expected value formula being used as either a rational evaluative principle or a practical rule of thumb in cases where the necessary assumptions are at least approximately satisfied.

Throughout the early 18th century it was commonly assumed that a rational player in a game of chance ought to be willing to pay any sum up to the expected value of an offer, as described (Starmer 2000: 333). This changed following Bernoulli’s investigation in 1738 of the St. Petersburg paradox, which can be described as follows. A player at a casino is offered a gamble. A fair coin is tossed. If it lands heads she receives \$2. If it lands tails play continues until it lands heads. If this occurs on the n^{th} throw she receives \$$2^n$. How much should a rational player be willing to pay to participate? Assuming that the chance of landing heads is exactly ½, the expected value of the offer is:

$$\sum_{n=1}^{\infty} ½^n.2^n = 1+1+1+1+ \dots = \infty. \qquad (2.2.2)$$

Hence a rational player should be willing to pay any amount, with no upper limit. Clearly no human player would do so. Research suggests that in practice the average acceptable price is less than \$5. Bernoulli’s solution was radical. Rather than invoking the possible violation of any of the assumptions described above as others had done (Dehling 1997), he proposed that the marginal value of a monetary gain to a player is not constant but is inversely proportional to that player’s current wealth. Hence, for any player, the relation between the effective value of a monetary gain and its notional monetary value is not linear but logarithmic. Thus, provided that the player’s initial wealth is greater than zero, the sum of the infinite series is finite and the paradox is eliminated.

This proposal assumes a person-specific interval-scale quantification of effective value, later termed utility. Adapting the expected value principle, if an option promises exactly one outcome from a set of mutually exclusive outcomes $\{x_1, \dots , x_n\}$ having probabilities $\{p_1, \dots , p_n\}$ and, for player j, utilities $\{u_j(x_1), \dots , u_j(x_n)\}$, the expected utility of the option for player j is

$$\sum_{i=1}^{n} p_i \, u_j(x_i) \qquad (2.2.3)$$

and the optimal policy for player j when offered a set of options is, ceteris paribus, to choose an option with the greatest expected utility.

The introduction of the concept of utility has the enormous advantage of extending the possible quantification of value from cases with a definable monetary value to other motivationally significant outcomes such as the experience of happiness or pain. It thus provides the basis for a general theory of value, and of morality, that came to be called utilitarianism – subject to a normatively motivated assumption of commensurability among person-specific measures of utility.

Bernoulli’s proposal entails that utility is an interval-scale measure, for which there appears to be no direct evidence. For this reason it remained unpopular with economists until the 1950s (Starmer 2000: 334). This changed after the publication by von Neumann and Morgenstern (1944) of a proof showing that an interval scale of utility could be justified entirely on the basis of assumptions about pairwise choice behaviour.

Their proof rests on the assumption that an agent exhibits strictly consistent choice behaviour over all possible pairs of options, where consistency is defined as satisfying a set of well defined and apparently plausible formal conditions – namely completeness, transitivity, continuity, independence, and monotonicity (see e.g. Wilkinson 2008: 87-9, Dhami 2016: 84-7). The proof works, essentially, by showing that the ratio scale exhibited by probability as standardly modelled transfers via the linearity of the expected utility function (2.2.3) to marginal utility. The listed conditions can be specified in various ways as axioms, and the expected utility model derived accordingly. This development has two key features. It establishes axiomatization based on rationally plausible principles governing choice behaviour as a method of theory construction (e.g. Koopmans 1960). It thus opens a developed theory to the objection that one or more of its underlying axioms is not in fact satisfied (e.g. Loewenstein and Prelec 1992: 120-1). And it creates a theory of valuation that assumes that probabilities are objectively given but that utilities can be, and perhaps must be, inferred from choice behaviour rather than being otherwise measurable (Dhami 2016: 92-3).

This situation was transformed by what Dhami calls "one of the most brilliant results in all of social science" (2016: 91), Savage's (1954) axiomatic construction of a model of choice behaviour that derives not only a measure of utility but also a measure answering the usual concept of probability from an agent's choice behaviour over options. Since the new measure is revealed in an individual's pattern of choice rather than being inferred from objective analysis of the situation it is generally interpreted as a measure of subjective probability, or credence, and the theory is called subjective expected utility theory.

The model assumes that the relevant decision space is partitioned into a set of mutually exclusive possible events $\{E_1, \ldots, E_n\}$. A decision maker j assigns a subjective probability $\mu_j(E_i)$ to each E_i such that for all i, $\mu_j(E_i) \geq 0$ and $\sum_{i=1}^{n} \mu_j(E_i) = 1$. An option f is an assignment of an outcome x_i to each E_i. If $f(E_i)$ is the outcome assigned to E_i in f and $u_j(f(E_i))$ is the utility of $f(E_i)$ for j, the subjective expected utility of option f for j is

$$\sum_{i=1}^{n} \mu_j(E_i)\, u_j(f(E_i)). \qquad (2.2.4)$$

If it can be assumed that, when offered a set of options $\{f_1, \ldots, f_m\}$, j always chooses an option with the greatest subjective expected utility and that the set of such choices always satisfies certain consistency conditions similar to those proposed by von Neumann and Morgenstern, then it can be shown that both subjective utilities and subjective probabilities satisfying normal scalar constraints can be constructed from suitable choice data. Working in reverse, if such quantities are defined, an optimal choice is one that maximizes subjective expected utility over current options.

Since both utilities and subjective probabilities are derived exclusively from observed choice behaviour and predictions of behaviour depend on these derived quantities, applications of the model have, in practice, rather limited predictive power. Even where a prediction is made, apparent errors can often be accommodated by repartitioning the event space or modifying the set of presumably relevant outcomes. For example, if j's preference for tea rather than coffee turns out to depend on the time of day or whether it is accompanied by cake or a biscuit or who makes it, outcomes can be differently classified accordingly. Ultimately the specificity of analysis can be increased so as to accommodate any conceivable pattern of choices. Only if this kind of ad hoc adjustment is resisted in at least some paradigmatic cases or if other theoretical constraints are added is significant testing possible.

Any effect commonly observed in choice behaviour might give reason to propose such a constraint. One such constraint is termed probabilistic sophistication (Dhami 2016: 91-2). It requires that agents treat outcomes associated with events, as in (2.2.4), and outcomes with associated probabilities, as in (2.2.3), equivalently. Another is that subjective probabilities mirror corresponding objective probabilities, as in the Principal Principle (Lewis 1986). A third is that, in general, equal increments in added objective value have uniformly decreasing added subjective value, as proposed by Bernoulli. A fourth, of particular interest in the current context, is the commonly observed effect known as time preference – that, other things being equal, agents generally prefer to receive gains earlier and losses later.

2.2 Discounted Utility

Interest in time preference dates back at least to Rae (1834) (Frederick et al. 2002: 164). Over the following century various introspectively grounded psychological explanations were canvassed until, in 1937, Samuelson proposed a formal model that reset the discussion. It gained rapid acceptance. He proposed that the utility of any outcome increases over the period of time during which it is pending by a constant ratio $(1+\rho)^{-1}$ per unit time, where $\rho \geq 0$ is the discount rate. This implies a discount function that is exponentially decreasing with expected delay. Formally, if an outcome x_i occurring at a time t_0+t, where $t \geq 0$, has a utility $u_j(x_i)$ for j at t_0+t, then x_i has a utility $e^{-kt}u_j(x_i)$ for j at t_0, where $k = \ln(1+\rho) \geq 0$.

Time discounting introduces a significant theoretical issue. Provided that the utility associated with an event or condition is invariant with respect to time there is no need to distinguish whether or not an evaluated event or condition is temporally extended. Utility depends on either its simple aggregate magnitude or its average magnitude and duration. But if utility varies with time this is not so. On the plausible assumption that each infinitesimal time slice of an extended condition ought to be accounted for proportionately, the utility associated with an extended condition ought always to be computed as an integral with respect to time. Dhami (2016: 589 (9.8)) gives an example. However, it is then difficult to assign utility to momentary events such as discrete economic transactions since the relevant time interval in such a case is zero.

Since, for reasons of history and data, the standard economic paradigm quantifies many outcomes as discrete events, modelling utility in continuous time is highly inconvenient. The alternative is to model continuously extended conditions as sequences of finite time slices, each with an associated utility – that is, to construct models in discrete time. This is the method conventionally used. Assuming a set of equal time intervals $t \in \{0, \ldots , T\}$ and that k is independent of i, in Samuelson's model the expected utility for j at $t = 0$ of an option in which outcome x_i occurs with probability $p_i(t)$ in interval t is

$$\sum_{i=1}^{n} \sum_{t=0}^{T} p_i(t)\, e^{-kt}\, u_j(x_i) \tag{2.3.1}$$

and, again, the optimal policy for player j when offered a set of such options is to choose an option with the greatest expected utility. The alternative in continuous time, assuming that apparently momentary events can be treated as nominally extended, is

$$\sum_{i=1}^{n} \int_{0}^{\infty} p_i(t)\, e^{-kt}\, u_j(x_i)\, dt. \tag{2.3.2}$$

To incorporate time discounting into the subjective expected utility function (2.2.4) requires a function $f(E, t)$ that maps event E to outcomes at t. In continuous time the discounted subjective expected utility for j at $t = 0$ of option f is

$$\sum_{i=1}^{n} \int_{0}^{\infty} \mu_j(E_i)\, e^{-kt}\, u_j(f(E_i, t))\, dt. \tag{2.3.3}$$

Whilst modelling in discrete time is convenient in economics it is difficult to justify in the analysis of decision making more generally. Broome (2004), for example, equivocates. The issue will be discussed more fully later.

Like Bernoulli, Samuelson assumes interval-scale utility. This soon ceased to be thought problematical, given von Neumann and Morgenstern's result. Although Samuelson set out his proposal in a short paper without claiming that it has deep rational justification it was in due course shown to be derivable from a set of rationally plausible axioms (Koopmans 1960). This gave it "a scarcely needed further boost to its dominance as the standard model of intertemporal choice" (Frederick et al. 2002: 167).

Among the key features of Samuelson's model is that it exhibits stationarity. This means that if two options are differently located in the future but are otherwise unvarying, the ratio of their expected utilities remains constant as they approach. Formally, for any pair of options $\{f, g\}$ with expected utilities $\{U_t(f), U_t(g)\}$ for j at t, if for all i and all $0 \leq t \leq b < T$, $p_i(t)\, u_j(x_i) = 0$, then for all $0 \leq t \leq b < T$, $U_t(f)/U_t(g)$ is constant. The effect is that while two options both remain pending their preference order is fixed. It is a logical consequence of having a constant discount rate.

Although the assumption of generally exponential time discounting came to be accepted almost universally in economics, doubts started to arise in the 1970s as to its justification. Its key advantage – stationarity – came, in view of the evidence of the frequent apparent inconstancy of human choice, to be viewed as a disadvantage. An acrimonious debate has followed, revolving in part around whether economics ought to be a descriptive rather than a normative discipline (Dhami 2016: 44). This has, perhaps, hindered clarification of the issues involved.

The usual way of formulating the question involves generalizing (2.3.1) or its equivalent as follows (e.g. Dhami 2016: 587 (9.1))

$$\sum_{i=1}^{n} \sum_{t=0}^{T} p_i(t)\, D(t)\, u_j(x_i) \qquad (2.3.4)$$

and enquiring as to the possible form of the discount function $D(t)$. In defining $D(t)$, a number of considerations have been advanced. One is evidence that human discounting per unit of delay is much greater in the short term than in the long term and that a similar effect has been observed in animal foraging behaviour. Another is that human preferences often reverse as competing options approach. A third is the ubiquity of an inverse law of perceptual intensity, including of memory and reinforcement (Chung and Herrnstein 1967, Mazur 1987, Thaler 1981, Rachlin and Raineri 1992, Ainslie and Haslam 1992a, Camerer and Lowenstein 2004, Webley and Nyhus 2008). The simplest formula that satisfies these constraints whilst remaining finite at $t = 0$ is $D(t) = (1+kt)^{-1}$ where $k > 0$. On this basis the expected utility for j at $t = 0$ of an option in which outcome x_i occurs with probability p_i in interval t is

$$\sum_{i=1}^{n} \sum_{t=0}^{T} p_i(t)\, (1+kt)^{-1}\, u_j(x_i) \qquad (2.3.5)$$

or, in continuous time,

$$\sum_{i=1}^{n} \int_{0}^{\infty} p_i(t)\, (1+kt)^{-1}\, u_j(x_i)\, dt. \qquad (2.3.6)$$

This entails a hyperbolic rather than an exponential relationship. The contrast is illustrated in Figure 2. Key differences are that for small delays the current discount rate under hyperbolic discounting is greater than under exponential discounting but for large delays it is smaller. At large delays the expected value under hyperbolic discounting approaches zero less rapidly. Hence long-term effects do not decrease in significance as rapidly. And for any permanent condition, for which the total expected utility is proportional to the area under the curve, the hyperbolic total approaches an infinite value whereas the exponential total approaches a finite value.

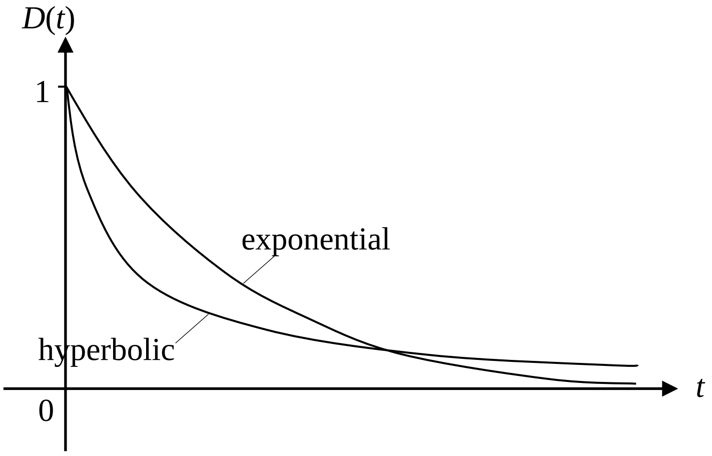

Figure 2: Exponential versus hyperbolic discounting

But the most frequently discussed effect is preference reversal. It is illustrated in Figure 3.

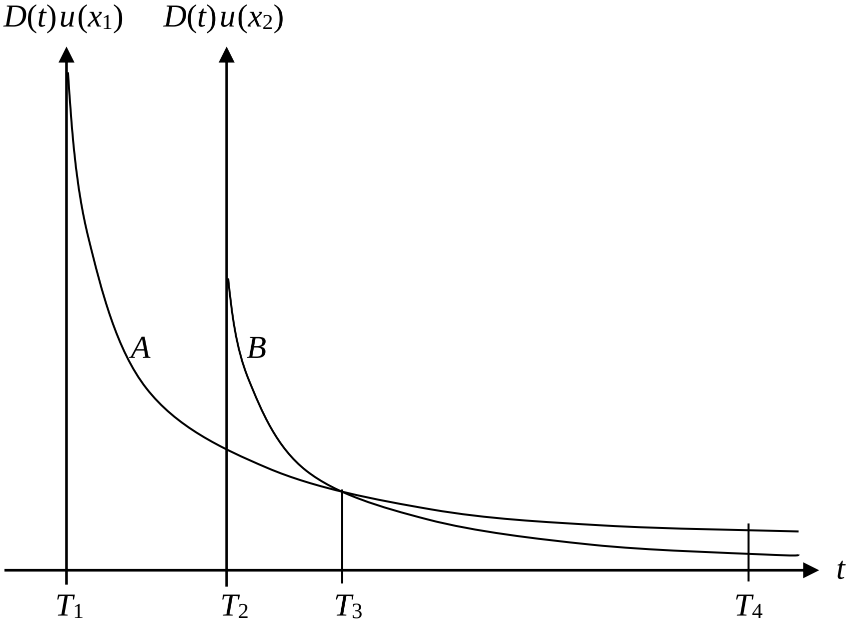

Figure 3: Preference reversal

This figure superimposes two graphs, *A* and *B*, plotting the expected utility against delay of outcomes x_1 and x_2 with undiscounted utilities $u(x_1)$ and $u(x_2)$ occurring at T_1 and T_2 respectively, where $u(x_1) > u(x_2)$. Since t measures delay, the conventional temporal order of events is right-to-left. Viewed from T_4, T_1 is more remote than T_2 and the expected utility of x_1, represented by *A*, is greater than the expected utility of x_2. But at T_3 the relationship reverses. Viewed from any point between T_2 and T_3 the expected utility of x_2, represented by *B*, is greater. Beyond T_2, x_2 is in the past.

Although the hyperbolic discounting model described above has been widely influential it has competitors. All assume the general hyperbolic form shown in Figure 2 but achieve it in other ways. A significant motivation is that hyperbolic discounting is most naturally formulated in continuous time, as in (2.3.6), but, for reasons of analytical tractability, discrete time is usually preferred in economics (Wilkinson 2008: 229).

A quasi-hyperbolic model in discrete time was first proposed by Phelps and Pollak (1968) and revived by Laibson (1994). Assuming a set of equal time intervals $t \in \{0, \ldots, T\}$ and that k and β are independent of i and $0 \leq \beta \leq 1$, the expected utility for j at $t = 0$ of an option in which outcome x_i occurs with probability $p_i(t)$ in interval t is

$$\sum_{i=1}^{n} (\, p_i(0)\, u_j(x_i) + \beta \sum_{t=1}^{T} p_i(t)\, e^{-kt}\, u_j(x_i)\,) \qquad (2.3.7)$$

and, again, the optimal policy for player j when offered a set of such options is to choose an option with the greatest expected utility.

This model constructs the expected utility of an outcome from two elements. The first is an undiscounted component that is included if the outcome occurs or is expected immediately. The other is an exponentially discounted sum over all subsequent time periods, additionally discounted by a constant factor β. If β is small, immediate effects dominate. If k is small, long-term effects remain relatively significant. The model implies a sharply discontinuous decrease in expected utility between the first and second time periods. It is conventional to express (2.3.7) in the form

$$\sum_{i=1}^{n} (p_i(0)\, u_j(x_i) + \beta \sum_{t=1}^{T} p_i(t)\ \delta^t\, u_j(x_i)) \qquad (2.3.8)$$

where $\delta = e^{-k}$ and hence $0 \leq \delta \leq 1$. Hence it is popularly called the (β, δ) form of hyperbolic discounting (Dhami 2016: 614).

2.3 Prospect Theory

In parallel with the development of the hyperbolic discounting model, other significant modifications to the expected utility model have been proposed, primarily to accommodate an increasing range of discovered anomalies. The approach that has achieved greatest prominence, due principally to Tversky and Kahneman, is called prospect theory.

Prospect theory can be understood most perspicuously as adding two novel formal constraints to the generalized subjective expected utility function that, in continuous time, has the form

$$\sum_{i=1}^{n} \int_{0}^{\infty} \mu_j(E_i)\ D(t)\ u_j(f(E_i, t))\ dt. \qquad (2.4.1)$$

In place of the assumption that subjective probability satisfies the Principal Principle, prospect theory assumes a non-linear relation between subjective probability and objective probability as standardly quantified. It typically has a reverse S-shaped form defined by the Prelec function. And in place of either a linear or a logarithmic relation between utility and objective value prospect theory assumes a monotonic-increasing S-shaped relation in which gains and losses are evaluated asymmetrically, usually defined as a power function. It also usually assumes some version of hyperbolic discounting.

The Prelec function w is a continuous monotonic-increasing function that maps [0, 1] to itself. In the present notation, if p is the objective probability of E_i as standardly quantified, $\mu_j(E_i) = w(p) = e^{-\beta(-\ln p)^\alpha}$, where $\alpha > 0$ and $\beta > 0$ (Prelec 1998). If $w(p)$ is plotted against p, the shape of the curve depends on α and β. A typical example where $\alpha = 0.5$ and $\beta = 1$ is illustrated in Figure 4.

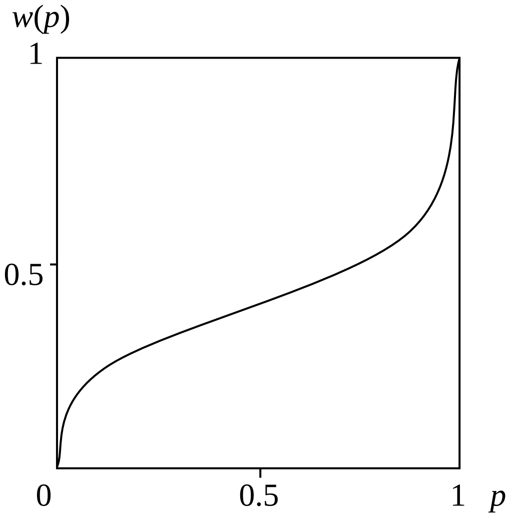

Figure 4: Plot of the Prelec function (after Dhami 2016: 28)

On this basis, subjective probability overestimates small objective probabilities and underestimates large ones. A consequence is that the total subjective probability of a set of disjoint alternatives may exceed one.

The utility function v is a continuous monotonic-increasing function that maps $\mathbb{R}$ to itself. It is concave in the domain of gains and convex in the domain of losses. A power form of the function defined in Tversky and Kahneman (1992) can be expressed as follows. If $y = f'(E_i, j, t)$ is the current value for j of the outcome assigned to E_i in f at t standardized to a current zero reference point and a unit scale of sensitivity for j,

$$u_j(f(E_i, t)) = v(y) = \begin{cases} y^{\gamma^+} & \text{if } y \geq 0 \\ -\lambda(-y)^{\gamma^-} & \text{if } y < 0 \end{cases} \quad \text{where } 0<\gamma^+,\gamma^-<1,\ \lambda>1. \quad (2.4.2)$$

In this formula, λ represents the degree of loss aversion and $^1/_\gamma$ the degree of declining sensitivity to gains and/or losses. A simpler version of the formula has $\gamma^+ = \gamma^-$. If $v(y)$ is plotted against y, the shape of the curve depends on λ and γ. Tversky and Kahneman estimate from available data that $\lambda \approx 2.25$ and $\gamma^+ \approx \gamma^- \approx 0.88$. A typical example where $\lambda = 2.5$ and $\gamma^+ = \gamma^- = 0.5$ is illustrated in Figure 5.

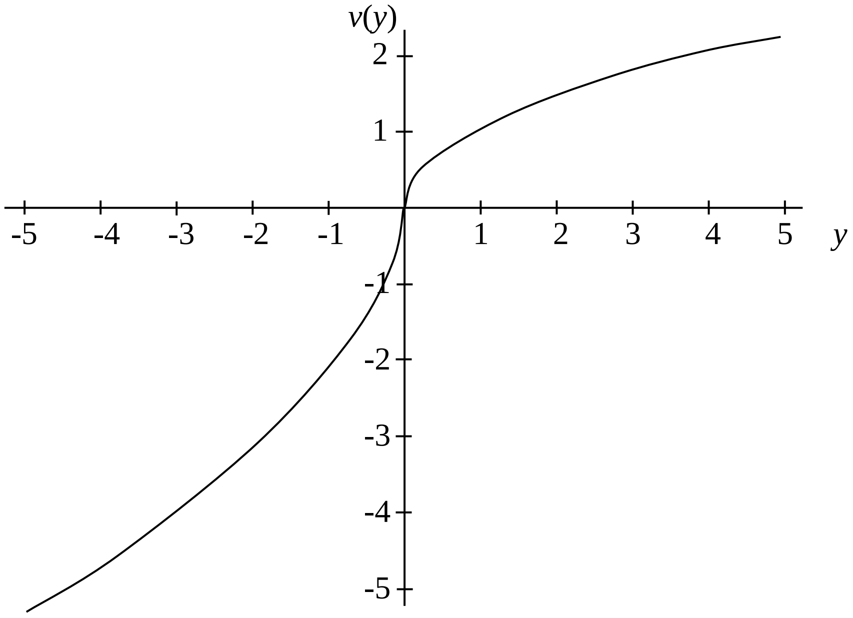

Figure 5: Plot of the power form utility function (after Dhami 2016: 132)

This function, unlike a normalized version of Bernoulli's logarithmic function, is inverse-parabolic in form. This has two significant consequences, that the gradient at $y = 0$ is infinite and that the gradient at extreme magnitudes reduces less rapidly. The contrast is illustrated in Figure 6. From the fact that the gradient at $y = 0$ is infinite it follows that very small gains and losses have disproportionate significance. For example, a few pence saving on even an expensive item may predictably swing a purchasing decision. Similarly, a series of small gains accounted for sequentially usually has greater weight than their sum accounted for as a single event, and a gain followed by an equal loss, if accounted for separately, is usually treated as a net loss. From the fact that the gradient at extreme magnitudes is steeper it follows that large gains and losses have a less disproportionately reduced significance.

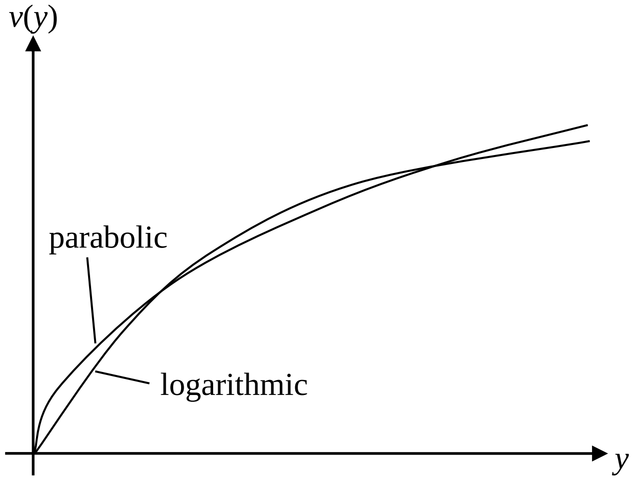

Figure 6: Logarithmic versus inverse parabolic relationship

Other formulas defining relationships that are structurally similar to those illustrated in Figures 3, 4, and 5 but have subtly different implications, especially at extreme magnitudes, can easily be devised (Loewenstein and Prelec 1992: 127-33, Starmer 2000). Dhami (2016: 146, 156) lists some alternatives. Unless there are theoretical reasons to prefer a particular formula, the choice in each case depends on the fit of the resulting models to available data.

One line of development aims to correct the non-additivity of subjective probabilities exemplified in prospect theory. A possible solution is rank-dependent utility theory (Quiggin 1982). It involves transforming subjective probabilities into decision weights via a probability weighting function under the constraint that if outcomes are rank-ordered by value the decision weight of each outcome is equal to the difference between the weighted magnitude of the cumulative probability of all outcomes of equal or higher rank and the weighted magnitude of the cumulative probability of all outcomes of higher rank, and that the total decision weight is one. This works, and is adopted in a revised version – cumulative prospect theory – but it has the effect that decision weights depend critically on the set of outcomes actually considered, regardless of their low probability, and on their precise rank order (Starmer 2000: 348). It depends, in other words, on the criterion, or procedure, for admitting outcomes as relevant.

Prospect theory is usually described as a descriptive theory (Dhami 2016: 26) – that is, it is expected to satisfy normal scientific standards of testable fit over a relevant class of observed cases. The role of axiomatization in such a theory must generally be to elucidate its structural features rather than to define independently justified foundations since, if testing refutes an assumption implicit in the theory, independent justification cannot save it. It follows that alternative axiomatizations with different explanatory implications may be equally valid.

Furthermore, many current models, including those of prospect theory, may involve a significant equivocation, as follows. Such a model typically consists of a formal system *S* together with an explicit or implicit rule that maps conditions in *S* to implied decisions, but this rule may take either of two subtly different forms. It may require either that if condition *C* is satisfied then decision *D* is made, or that if condition *C* is satisfied *and a relevant decision is made* then decision *D* is made. In reported research this distinction may be concealed insofar as field data report only observed choices and experimental data report only forced choices and hence, in both cases, uncompleted or non-decisions are omitted. The distinction becomes especially significant if condition *C* is time-varying since in that case the implied decision *D* may depend on when it is made. The issue is not resolved by stipulating that decisions depend on preferences and preferences exist only when a choice is made since it then recurs in the determination of preferences.

2.4 Other Models

Whilst the line of development described above has gained considerable influence in behavioural economics there are alternatives. Starmer (2000: 332), now over twenty years ago, puts the total number "well into double figures". I will describe the most prominent of these briefly. I will not include game theory because, although it is extremely significant, it is not an alternative but an extension of standard decision modelling into multi-player situations. Thus insofar as it models decision making rather than systemic collective behaviour it assumes rather than replaces one or other of the agent-focused schemes described here, adding, at most, the possibility of an equilibrating heuristic (Dhami 2016: 40-4).

A major line of development consists in the creation of a range of procedural theories (Starmer 2000: 350). The fundamental insight is that decision making is a process that takes time and has a cost, both in effort and, often, in benefit forgone or loss incurred in the interim. This line of thought, and supporting research, led Simon (1956) to propose that economic decision making does not in general aim at an optimal outcome but only at one that is good enough – or ‘satisficing’. Models developed on this basis typically assume either an a priori criterion of acceptability or a process of option definition and valuation plus a stopping rule such that an immediately apparent benefit prompts a quick decision but more a complex or evenly balanced choice takes longer. Typically, a decision is made if and only if a context-dependent threshold of apparent marginal benefit, or some heuristically equivalent procedural criterion, is satisfied. It is a theory of bounded rationality.

Models of this type are often constructed to accommodate observed effects that appear as anomalies in standard expected utility theory. A characteristic example is the cognitive hierarchy model (Camerer et al. 2004). It proposes that in some tasks in which induction implies an extreme choice, players typically restrict inference to a characteristic number of steps. Hence aggregate data display an otherwise unexpected series of peaks corresponding to integral inference patterns. A more extreme version allows that an agent may have available, or may construct by imitation, deliberation, or experiment, a variety of decision making methods and may select by trial and error, in context, a method and a set of preferences so as to generate a suitably consistent pattern of choices. This is the discovered preference hypothesis (Plott 1996). The same principle may apply to the evolution of an equilibrating or otherwise acceptable strategy in game theory, perhaps satisfying a Nash equilibrium (Binmore et al. 2001). Turan

(2019) develops a model in which participants may adopt a strategy that exploits the inconsistent time preferences of others.

Since utility-maximizing decision making, even if possible, is computationally expensive, many authors develop models that incorporate heuristic short-cuts. Evidence for the use of heuristics, of which agents are generally unaware, can be derived from observation of systematic biases in decision making (Tversky and Kahneman 1974). The resulting discoveries have prompted a significant, continuing, and controversial research program (Dhami 2016: 44-8, 1339-47).

Several projects investigate issues of implicit statistical sampling – that agents must extrapolate from limited data given in prior experience – and on the efficacy of various sample-based heuristics. An interesting selection of models is presented in Chater and Oaksford (2008). Brighton and Gigerenzer (2008) describe a simple decision heuristic called ‘take the best’. Hertwig and Pleskac (2008) investigate the efficacy of relying on small samples. Hansson et al. (2008) investigate the bias introduced by a naïve sampling method. Stewart and Simpson (2008) describe a simple heuristic called decision by sampling and compare its implications with those of prospect theory. Usher et al. (2008) describe and compare two connectionist models that generate decisions by implicit learned statistical inference. Elsewhere, Ericson et al. (2015) describe and evaluate a model that generates a decision from a weighted sum of four easy-to-compute features based on available value and time data. Dhami (2016 Part 7) gives a fuller account of this research.

It is usual, as in prospect theory, to quantify outcomes as gains or losses rather than by absolute magnitude – that is, in relation to a current agent-specific neutral magnitude, or reference point. Similarly, the way options are presented in context influences what is chosen. This is referred to as framing. There is research into and modelling of both. For example, Hart and Moore (2008) describe a model of contract-based reference point setting and Fehr et al. (2009) a fairness-based model. Shefrin and Thaler

(1992) describe a life-cycle-based model of framing. Other models concern the subjective perception of time or compensating responses to anticipated outcomes. For example, Prelec (2004) describes a model that has impatience as the core variable. Ebert and Prelec (2007) describe a model that admits variable sensitivity to future time. Loomes and Sugden (1987) describe a model of regret, and Gul (1991) one of disappointment aversion.

Observed choice often appears inexplicably variable. A possible response is to build a decision making model containing a random process. Starmer (2000: 374) discusses several competing proposals. Goeree and Holt (2004) describe a model of noisy introspection. A model may explicitly incorporate neurological assumptions and aim to represent neurological processes for which there is independent evidence given by, for example, PET or fMRI scans. This approach is in its infancy (Dhami 2016: 1644).

Peters (2019) observes that all standard economic modelling of accumulated utility in sequential processes assumes ergodicity, in which an integral over time is assumed to be equal to a sum over probabilistically weighted possible states, and that this condition is almost never satisfied. He traces the problem back to an error made by Bernoulli 1738, never previously noticed. The effect is that almost all models of sequentially accumulated value under risk are wrong. Evidence suggests that some apparent risk aversion standardly interpreted as anomalous is explicable as a rational response to observed non-ergodicity.

Finally, there are several theoretical approaches that reject the principle that decision making is psychologically unitary. Among the most well known is Loewenstein’s analysis of visceral factors (1996). This assumes that, within an agent, rational and emotional processes work in opposition to each other and that in cases where visceral factors dominate, short-term consequences have disproportionate weight. This accounts for, or replaces, hyperbolic discounting. Thaler and Shefrin (1981) propose a planner-doer model, in which the planner evaluates long-term consequences and the more myopic doer reacts to immediate desires and may override the planner. These proposals typically assume that an individual should be modelled as a competitive coalition of multiple selves. It creates a technical problem of how to model a process of self-control – a process in which a self guarding a particular set of interests either does or does not prevail – as discussed in Bermúdez (2018).

2.5 Conclusion

In this chapter I have briefly surveyed the principal lines of development in the more or less recent literature on the modelling of individual human decision making. The central theme is the agent's assumed evaluation of expected utility, but theoretical development and empirical observation has led this in several sometimes controversial directions. A key issue, which becomes increasingly apparent, is that the desire to accommodate and respond to an increasing variety of apparently anomalous observations, particularly as given in experimental results, has led to the development of an almost bewildering variety of competing models among which there is no obvious basis for disinterested selection or ultimate unification.

Prospect theory has the best claim to providing a coherent and empirically justified extension of original expected utility theory but there is still some not inconsiderable contradictory evidence (Dhami 2016: 172-81) and it contains a number of features that appear puzzlingly ad hoc. Conspicuously, it does not readily integrate time preferences into the rest of the theory. This is made strikingly clear in Dhami's 'agenda for the future' (2016: 206-7), in which all five points concern the status of time preferences and, in particular, their relation to risk preferences.

A number of lessons emerge from this survey. One is that although expected utility theory faces many difficulties it continues to provide the basic conceptual framework without which not much in the analysis of decision making, including the description of anomalies, makes sense. It is the sine qua non of interpretation. Recent proposals in behavioural economics, including prospect theory, do not alter this. They are, in Dhami's words, "an enhancement … not its antithesis" (2016: 2).

Nevertheless, the multiplicity of current proposals devised in an attempt to accommodate apparent anomalies, is, if anything, becoming increasingly diverse. The usual modelling methodology, which is to create a mathematical formalization of superficial regularities based on what I have called a narrow definition of available options, shows little sign of achieving broad explanatory success. A plausible explanation, advocated by, for example, Kahneman and Tversky, Camerer and Loewenstein, Ainslie and Haslam, Thaler, and many others, is that, as outlined in §1.4, it pays insufficient attention to psychological and causal constraints on human action and to the structural role of decision making within human action more generally. Even prospect theory, which began by emphasizing psychological processes, has, under pressure to achieve technical consistency – and like much recent modelling – become increasingly formalistic, at the cost of apparent normative or explanatory relevance (Dhami 2016: 192-3).

Moreover, the theoretical options that are available in response to the demand to give due weight to psychological and causal constraints on decision making and action are a matter of considerable dispute. This dispute, centring on the role of heuristics, finds a focus in what is called the great rationality debate. I will examine this issue and related issues of adaptation, computation, and causation in the next chapter.

Chapter 3 Rationality and Causal Structure

3.1 The Great Rationality Debate

In §2.4 a number of proposals that involve modelling cases of human decision making as realizations of particular heuristic processes were described briefly. Whilst to the outsider these proposals might be thought relatively uncontroversial they have in fact generated very considerable dispute within psychology. The dispute standardly goes by the name of the great rationality debate (Dhami 2016: 47).

The great rationality debate is an acrimonious metatheoretical dispute, principally over the status of the work of Kahneman, Tversky, and others investigating the role of heuristics and biases in human decision making (Stanovich and West 2000, Stanovich 2012, Dhami 2016: 1426-31). The dispute is not about whether humans use heuristics and hence exhibit biases – this is indisputable – but whether this constitutes a deviation from or an approximation to rationality. Kahneman and Tversky (KT), often termed Meliorists, emphasize the former. Their critics, especially Gigerenzer and his colleagues (G&C), often termed Panglossians, emphasize the latter. Dhami describes the dispute as "muddled" (2016: 1427). He says:

> Given the very different nature of the core issues dealt with in the KT and G&C frameworks, it is staggering that so many of the leading researchers could take such strong positions and fish in very muddy waters (2016: 1428).

The debate ranges over a variety of issues, including what is the normatively correct decision making model, whether particular choices are correctly interpreted as violating relevant norms, whether intuitive inference is based on frequency or probability, how the probability of singular events is to be quantified, whether analysis allows for effects of context, framing, and base-rate variation, whether apparent bias is a kind of error, whether proposed heuristics are sufficiently well defined for claims to be tested, and in what way neurological modelling is significant. But, as Dhami implies, the strength of the dispute suggests some deeper unresolved issue or issues.

Two such issues are evident in the literature: the legacy of behaviourism and the definition and assumed role of rationality (Okasha 2016). I will examine each of these briefly, and how they interact, in order to orient subsequent analysis. Much more could be said but I hope this will be sufficient in the present context.

Behaviourism is a term much used, by both advocates and opponents, but one with very little shared content. As Greenwood remarks:

> All behaviorists were committed to the view that observable behavior (as opposed to conscious experience) is the subject matter of scientific psychology, but that was about all they agreed upon. They disagreed on almost every other substantive issue (Greenwood 2015: 359).

The confusion is compounded by the fact that in philosophy the term is often associated with the eliminativist thesis, championed by Ryle (1949), that mental states are nothing but dispositions to behave in observably distinct ways (Antony 2007: 151). Nevertheless, despite being described as "widely discredited" (Okasha 2016: 411), behaviourism remains influential. This is puzzling.

A plausible explanation is that the broad concept of behaviourism draws upon several underlying ideas not all equally discredited. One is a kind of radical physicalism – the idea that behaviour is just matter in motion and ought to be analysed as such. This does not rule out the analysis of intervening mechanisms but allows that they can be characterized only as physical or neurophysiological, not representational. In practice, whatever arguments are proposed in its defence, as a research project – except, at most, in relation to the behaviour of very simple organisms and simple reflex responses – it is a non-starter. No one has ever tried it, and it is unclear how an attempt could be coherently characterized in its own terms.

A second idea is a kind of radical functionalism – that behaviour is to be understood as a pattern of relationships between prior and subsequent effects in the environment and, moreover, that these relationships can be appropriately characterized without reference to intervening mechanisms. It is assumed that the classification of relevant effects is functional not merely physical – that, for example, 'food' and 'eating' are functional not physical concepts – but that this classification depends only on the patterns of observable relationships via which they are related, not in virtue of any mechanisms via which these relationships are effected. Ultimately, mechanisms are classified entirely by their effects, not vice versa. The resulting type of theory parallels what in physics is termed principle rather than constructive theory, as is exemplified by, for example, Newton's theory of gravitation and Einstein's theory of relativity, neither of which propose any mechanism via which the specified relationships are realized (Brown 2005: 71).

This idea is much less obviously objectionable. Nevertheless, there is a problem implicit in the assumption that the classification of relevant effects does not rest on information about intervening mechanisms. It is a general problem implicit in all radical functionalism, as mentioned on page 30 in relation to Pareto's definition of economics, namely that in identifying and classifying functionally related features in the environment theorists may smuggle in parts of their own intuitive classification based on introspectively derived information without noticing and that if the classification smuggled in is inappropriate the resulting analysis will typically fail. Consider, for example, a dog running about in a wood, sniffing in the undergrowth. We intuitively assume that the dog is engaged in some kind of complex olfactory discrimination but, being much less well equipped, we can have very little idea what pattern of classification of environmental features is involved. It is reasonably easy to test whether a dog can discriminate features that we discriminate, such as the presence or absence of covid-19 in a patient, but this is not the same as mapping out the dog's own classificatory system. Almost certainly, the latter would require a long programme of research into the biochemistry and neurophysiology of canine olfactory processing (e.g. Jenkins et al. 2018). This observation mirrors Wittgenstein's widely quoted remark that "If a lion could speak, we could not understand him" (1953: 223).

This leaves a third idea, a non-radical functionalism. It assumes that behaviour is to be understood as a pattern of relationships between prior and subsequent effects in the environment but that its characterization is to be informed by data about intervening mechanisms. Analysis may focus more strongly on the elucidation of either relationships or mechanisms but cannot ignore the other. One that focuses on relationships is usually described as behaviourist. One that focuses on mechanisms – a constructive analysis – tends towards either cognitivism or neurophysiology. There is, however, considerable overlap between these approaches. There is no reason to claim that a generally behaviourist approach as so characterized is discredited except in the sense that it has not yet been unequivocally successful.

Both Meliorists and Panglossians adopt a generally constructivist approach but they place a differing emphasis on the status of intervening mechanisms and, especially, on rationality as a covering principle. Concerning the latter – the definition and assumed role of rationality – there are again several competing ideas. The first is characteristic of standard decision theory. It is that rationality is defined by a set of formal constraints, or axioms, that must jointly be satisfied by any admissible system of preferences. It sets no limits on what is an admissible preference but only on the pattern of relationships among them, requiring, more or less, that they be maximally consistent.

However, a significant problem arises. It is that the data via which consistency is evaluated consists not of preferences but of expressions of preference, that every expression of preference is, in context, a unique case, that every case can be characterized in arbitrarily many different ways and hence that any partition of cases determined by a finite set of expressions of preference can in general be characterized in many different ways, and hence that there is in general no assumption-free way of deciding which features of a case are relevant to the observed expression of preference. Put differently, every expression of preference is a token of many different types, data specifies a relation over tokens, but the proposed definition of rationality specifies a relation over types.

This is the gap that is standardly filled by what I call on page 10 a narrow definition of available options. Manifestly, however, adopting this definition results in many patterns of preference that might ordinarily be considered acceptable being classified as irrational. For example, cats are generally carnivorous – that is, they prefer to eat meat rather than vegetable matter. But it is commonly observed that domestic cats sometimes eat grass. This is, prima facie, a clear reversal of preference and hence, assuming that cats are possibly rational, ought to be classified as irrational. But this is not generally conceded. On the contrary, it is usually assumed that there must be a good reason why cats eat grass – which is to say that there must be a justifiable account in terms of which it is not irrational. In fact there is a lively debate about it (Shultz 2019). It shows that inconsistency among expressions of preference under a superficially justified classification is not itself generally considered evidence of irrationality. Indeed it cannot be, since there will often be inconsistent superficially justified classifications – a drink of tea or coffee in the morning or afternoon, alone or with friends, at home or out, with cake or a biscuit, after a meal or separately, and so on. It is a typical problem of underdetermination by data, not unlike the reference class problem that I will discuss in §6.1. The implication is that, despite its undoubted mathematical brilliance, without some supporting theory or methodology to decide what features of a case are evaluatively significant the usual axiomatic definition of rationality is unworkable.

This conclusion is implicitly accepted by both Meliorists and Panglossians. Moreover, both sides assume that rationality is to be defined in terms of the satisfaction of some standard of effectiveness or adaptation, as is broadly admitted in what I have termed non-radical functionalism. But they differ in both the standard itself and the explanatory role assigned to it. For Meliorists, rationality is a priori, as in formal logic and probability theory, and it is not invoked as an explanatory principle. It is an ideal standard, seldom or imperfectly achieved by humans. Hence whilst it may provide an important motive for analysis, most or all the explanatory work in actual cases is done by the elucidation of cognitive processes that may or may not achieve it. Indeed, the gap is the motive for possible meliorism. For Panglossians, rationality is not just a possible outcome of the operation of common cognitive processes but is the core principle explaining their existence. As in evolutionary biology it is assumed that such processes are an evolved product, shaped by adaptive success. Hence it is assumed that they typically satisfy a principle of rationality even if not in ways that are standardly recognized as ideal engineering solutions – rather as birds fly, but differently from aircraft.

It may be helpful at this point to summarize the available alternatives, in a rather simplified form, as follows:

Definition of rationality:	P principle (behaviourist)
	C constructive (heuristic)
Role of rationality:	D descriptive
	E explanatory

PD = behavioural economics
CD = Meliorist
PE = neoclassical economics
CE = Panglossian

Figure 7: The theoretical role of rationality

It should be noted that normative status, in the sense of admitting prescriptive use, rests more on the theorist’s attitude to deviations or alternatives than on the structure of the theory itself – provided that deviations or alternatives are admitted methodologically. Meliorist theory, although descriptive in the above sense, can be interpreted normatively, as its name implies. Conversely, Panglossian theory, although involving rationality as an ideal standard, might be thought unsuitable for normative use in much the same way that biological adaptation is not generally thought to admit a prescriptive interpretation.

Returning to the point of dispute in the great rationality debate, there are grounds to argue that both the main positions are open to significant objections. The Meliorist approach, which seeks to develop only a descriptive theory of heuristics and biases, is open to an objection that Dhami (2016: 1428) quotes Gilovich et al. (2002) as calling the "*we cannot be that dumb critique*". It is that the vast range of human achievement, especially in science, is inconsistent with a theory that is, ultimately, a catalogue of biased heuristics. Dhami objects that the success of science does not exclude the idea that scientists employ heuristics and are subject to biases in their thinking and decision making. This is correct. But the reverse problem remains – to show that a theory that consists of a catalogue of heuristics and biases, with no overarching principle of rational effectiveness, does not exclude successful science. This is a problem not of whether heuristics and biases are admissible but whether, on this basis, scientific success is explicable, or even describable. A catalogue of mechanisms does not resolve this question. Some implicit principle of relative success is needed.

Conversely, a Panglossian approach, in assuming that a principle of rationality is explanatory, requires an implicit mechanism via which the explanatory connection is realized. The general assumption is that the relevant mechanism is one of biological adaptation. More particularly, it is assumed that ubiquitous features of intuitive human decision making, including intuitive classification and heuristic procedures, are evolved phenotypes shaped by natural selection and, therefore, are optimally adapted to the relevant causal structure of the historically experienced environment.

This is a version of evolutionary psychology as advocated by, for example, Wright (1994), Cosmides and Tooby (1997), and Pinker (1997).

This account fills the explanatory gap as required, in that it defines rationality in terms of long-term instrumental effectiveness and accounts for its ubiquitous realization in terms of hereditary adaptation. However, the account of evolution it relies on has been subject to two significant lines of criticism. It is criticized both for misunderstanding the mechanism of biological adaptation and for ignoring non-biological adaptation (e.g. Kitcher 1985, Dupré 2012 part IV). Since these undermine an apparently plausible account of the causal basis of rational decision making I will examine them next.

3.2 Adaptation

A significant criticism of a Panglossian account is that it is an example of adaptationism. The issues involved are complex (Sterelny and Griffiths 1999: 43-8) and only partly relevant in the present context. Briefly stated however, what is at issue is whether the ubiquity or distribution of an identified phenotype within a population can properly be explained as being an optimal means of realizing some relevant type of effect within the prevailing environment. In this the word ‘optimal’ is crucial for it is this that offers the possibility that precise details – of, say, wing shape or camouflage pattern – can be explained, rather than merely ‘something of the kind’. In the case of rationality, the offered prospect is that particular intuitive or heuristic methods might be explained as being similarly optimal rather than merely ‘good enough’.

The adaptationist position is to endorse this possibility. Indeed, the usual assumption is that every identifiable phenotypic feature can be explained in this way (Gould and Lewontin 1979). The assumption is guaranteed, it is claimed, by the content of Darwinian evolutionary theory which entails the survival of the fittest and is illustrated in formal modelling in which the adaptation of a trait in a population is represented as a hill-climbing process towards a condition of maximum fitness (Okasha 2018).

Consider, for example, the human hand. The human hand is evidently adapted to grasping. This is of obvious advantage in the struggle for hereditary survival, in enabling tree-climbing, wielding large implements, and so on. The usual adaptationist claim is that it is maximally adapted to these tasks as they presented themselves during the relevant evolutionary period, sometimes interpreted, in the case of humans, as comprising approximately the Stone Age. But the human hand is also reasonably well adapted to many other tasks: removing skin parasites, picking soft fruit, evaluating surface texture, punching, signalling, forming a cup to hold water, and so on. It is not plausible that it is maximally adapted to each of these separately since they involve competing priorities.

The adaptationist claim is attractive in that it provides an apparently plausible mode of explanation of various particular phenomena that are otherwise seemingly extremely improbable, such an eye or wing. But even in these extreme cases optimality seem too strong. The human retina is functionally inside out, the posterior position of the human spine is poorly adapted to a vertical stance, the immune system is subject to auto-immune disease. And many other cases, such as facial hair, have no characteristic in relation to which they are even approximately optimal. A frequent criticism is that explanations constructed on this basis are, if not false, then ‘just so stories’ that rest, to an uncomfortable degree, on superficial plausibility rather than detailed evidence (Gould and Lewontin 1979, Sterelny and Griffiths 1999, Dupré 2012, Green 2014). Even where they are presented as sophisticated mathematical models the underdetermination of parameters by independent evidence invites worries about circularity (e.g. Ross 2013, Froese and Ikegami 2013, Ransom and Fazelpour 2015, Kogo and Trengove 2015, Wiese 2016, Colombo and Wright 2017, cf. Friston et al. 2016: 876).

Nevertheless, the inverse mode of explanation – that either the entire organism or some particular feature is maximally adapted to the totality of actual instrumental constraints as they apply to actual phenotypic resources distributed within the relevant population – seems scarcely better. Indeed, it is often criticized as being tautological. Insofar as this criticism is directed at evolutionary theory as a whole it is clearly unwarranted in that it fails to acknowledge the power of the theory to explain the developmental structure of speciation, which was Darwin's original project – encapsulated in the title, "*On the Origin of Species* …". But as a criticism of an envisaged explanation of actual outcomes it is justified. It highlights the typical impossibility of itemizing and relevantly weighting all the factors that jointly, directly and by interaction, contribute to a relevant outcome being precisely as it is. It is a significant problem in the analysis and explanation not only of physical traits but also of behavioural traits, including features of evolved rationality, whether under a Meliorist or Panglossian interpretation or some middle way. I will consider it more fully in §3.3.

A second possible criticism of a Panglossian approach is that it ignores non-biological adaptation. It inherits this attitude from evolutionary psychology, or sociobiology, in which most or all significant human traits are assumed to be biologically determined. This ignores two other discernibly different modes of evolving development: cultural and cognitive. I will discuss how these differ in the remainder of this section.

That many if not most human traits, both physical and behavioural, have an evolved biological basis is undeniable. Despite a remarkable expansion of practical and intellectual competence in the geologically recent past, humans are biologically very similar to other animals and share many metabolic and behavioural characteristics that persist, with variation, through successive generations. On this basis it is possible in principle to construct a purely biological account of rational competence – namely that its evolved content is that which has tended over the relevant evolutionary period to generate actions that serve optimally to promote the survival, in the prevailing environment, of relevant hereditable units (Godfrey-Smith 2009). For example, Cosmides and Tooby (1997) argue that contemporary human behaviour is determined by a legacy of adaptation to environmental contingencies prevailing during, loosely speaking, the Stone Age, and Friston (e.g. 2010) argues that human behaviour is such as to minimize a quantity measuring the agent's overall adaptive relation to its environment characterized in terms of free energy.

But although accounts of this kind are attractive in principle, the almost bewildering variety of patterns of choice and action observed in humans in various regions and eras invites scepticism. The doubts are similar to those surrounding other forms of adaptationism – that there is a very considerable gap between known preconditions and the diversity of observed outcomes – but it is exacerbated in this case by obvious differences in the velocity of development and the characteristic modes of transmission. On this basis two additional forms of evolutionary development are commonly recognized: cultural and cognitive. Whilst there is no general theoretical agreement as to how these two operate, both are clearly distinguished from biological evolution by their very much faster rates of change. Hence a purely biological account of rational competence, and hence of decision making, can be confidently rejected.

Although there is no general agreement as to precisely how either cultural or cognitive evolution operates it is possible to sketch a plausible outline as follows (Heyes 2019). Cultural evolution appears to involve the diffusion of objects and information by physical transmission or copying, the diffusion of evaluative attitudes by imitation, and selection among available evaluative attitudes mostly by authoritative or peer influence. For example, in many communities hat wearing is strongly conventional. Hats are standardly made or acquired, individuals copy their peers' or mentors' hat wearing, and, to the extent that deviation from accepted standards prompts disapproval, modifications are preferentially eliminated. In this process, immediate pragmatic considerations are not the main constraint except in the sense that agents act in a way that avoids socially relevant disapproval. Whilst the adaptive merit of evolved cultural forms may not be immediately apparent and may sometimes be non-existent, it is not in this respect wholly unlike biological evolution. A biological analogy occurs in sexual selection – namely that, provided that no greater selective disadvantage arises, in the competition for favourable attention an adaptive advantage may accrue from satisfying apparently arbitrary reciprocated preferences, such as for a certain pattern of behavioural display or plumage. This is capable of accounting for extremely diverse shared outcomes in both biological and cultural cases.

By contrast, cognitive evolution is driven by predictive success, usually but not always in the achievement of positively valued outcomes or the avoidance of negatively valued outcomes. Novel epistemic resources and evaluative attitudes arise by reasoned or accidental modification or extension of existing models and those that, in practice, yield more successful predictions, particularly of desirable effects as currently judged by those involved, tend to be maintained and copied and others tend to be discarded. For example, a cognitive methodology that more successfully predicts the location of a food source or the probability of stormy weather will tend to be retained and practised and may, if accessible, be imitated by others. Less successful alternatives are lost, often almost completely without trace. Since marginally less competitive alternatives are continually eliminated and therefore may escape notice, the large rate of failure is easily underestimated. Nothing in these accounts supports a principle of absolute or optimal rather than relative adaptation.

The implication is that a purely biological account of the aetiology of decision making processes cannot succeed (Simon 1990, Blackmore 1999, Fracchia and Lewontin 2005, Lewens 2015). Rather, several different strands of evolutionary development operate together, on different timescales. This conclusion multiplies the complexity of the analytical problem but leaves the problem posed by the rejection of adaptationism untouched – namely that if phenotypes or other evolved forms cannot be paired one-to-one with modes of adaptation, the alternative appears to be an uninformative holistic or tautological account that asserts merely that the totality of features of an evolved system is adapted to the totality of the environmental constraints applying over a relevant period. The prima facie attractiveness of adaptationism suggests that this omits something important. It is to this issue, which bears on the characterization of rationality, that I now turn.

3.3 Problem Analysis

The crucial issue is this. There is a fundamental assumption, both intuitively and in evolutionary biology, that each evolved form satisfies a principle of adaptation, and that to be adapted, or well adapted, is a relation to some kind of need or demand or other generic effect. This is the relation that presumably exists between the hand and grasping. It is only via this type of relation that the idea of adaptation is comprehensible. And yet the adaptationist programme, that tries to tell the story of a direct explanatory link from demand to outcome generally fails. Usually, many competing demands interact and the outcome is only explicable on balance.

Conversely, however, a holistic account asserting, implicitly, that the totality of outcomes is attributable to the totality of demands or other constraints, even if correct, is completely uninformative. No useful analysis can proceed on this basis. Among researchers the problem is avoided by concentrating on the analysis of particular processes or relationships – sex selection in fruit flies, the musculature of the eye, and so on – in which some relevant functional demands are assumed to be particularly relevant, perhaps without being specifically itemized as such. But in a more general philosophical analysis this is rather unsatisfactory.

Underlying both the intuitive and scientific approaches a significant principle can be discerned. It is that every relevant entity – which may be an individual or a colony or some other suitably structured unit or conglomerate – faces, in sustaining itself in its environment, a set of what I will call ‘problems’. These need to be jointly satisfied, in virtue of the entity’s exhibiting what may be called a unity of purpose or fate (Okasha 2018). Obvious examples include nutrition, excretion, escaping predation, infection control, sexual contact, dispersal of progeny, and communicating with allies. These problems are, to various degrees, standard and generic. Each set, in effect, defines a class of beings that inhabit a common, broadly defined niche. Indeed, in allowing parametric specification this vindicates the concept of a niche. Various derived combinations define other classes and hence alternative niches. For example, not all organisms reproduce sexually. Viruses lack nutrition but require a host. Additional problems such as locomotion, threat detection, and camouflage characterize more specific classes. Problems can be partially ordered in terms of generality on this basis.

For each relevant entity to maintain a state of adaptation within such a niche, each relevant problem must be satisfied more or less efficiently in situ by some one or more methods or combination of methods. The availability of such methods generally depends on the hereditable and other resources available. More specific methods are defined as responses to, and in turn create, increasingly specific or derivative problems.

Crucially, hereditable resources, as used, are shaped by evolutionary selection depending on the pattern of overall problem satisfaction, but with higher level problems typically having greater selective force. Consequently an analysis of evolutionary effects can start moderately effectively with higher level problems and proceed, usually with increasing precision, by incorporating the analysis of successively more specific problems, with the proviso that no complete or holistic analysis is ever generally possible. So, for example, an analysis of skeletal structure might proceed via physical integrity, stability under gravitation, accessing food, soft tissue support and protection, escaping predation, general locomotion, functional manipulation, parturition, nurturing, sexual display, grooming, and so on. Each adds additional and partly competing constraints. The idea is that a sufficient analysis of overall fitness can be approached but not achieved by successive approximation. It is this pattern of incrementally effective analysis that makes adaptationism both apparently but also imperfectly successful.

This proposal bears a strong similarity to conventional functionalism (Godfrey-Smith 1994, Sterelny and Griffiths 1999: 220-4). There is, however, a crucial difference between the notion of a problem as used here and the notion of a function. A problem resides in a putative relation of a relevant entity to its environment prior to any means of solution. A function is typically understood to reside in the means, as an evolved solution. But the relation between means, and hence functions, and problems in the sense defined is not one-to-one. It is usually many-to-many. Grasping is in this sense a problem. The hand, with associated musculature, neurology, etc, is a means of solution. But there are other means – using the teeth, knees, pliers, etc. – and the hand is party to the solution of many other problems – picking fruit, removing parasites, communicating, etc. Since evolutionary selection depends on the solution of relevant problems by any available means, an analysis in terms of functions misfires. The invention of cooking, for example, radically changed the way humans solve the problem of nutrition (Wrangham 2009). No analysis of anatomical or behavioural functions, as already evolved effects, gets a grip of this. In other words, what is attempted in functional analysis is, misleadingly, trait-centred rather than target-centred.

The proposal resolves a puzzle thought to favour propensity theory – that a trait can apparently have a function even if it has no relevant evolutionary history, if it solves a relevant problem. Thus a pacemaker has a biological function in the relevant sense of solving the derivative problem of heartbeat regulation even though it is both novel and non-biological. In this way the proposal satisfies the description given by Sterelny and Griffiths (1999: 223) of the typical work of anatomists and physiologists, also identified as functional analysis, except that the latter is described as concerning “the activities that an organism can perform: flying, digesting food, detecting viruses …” rather than the problems that these address: locomotion, nutrition, infection control – which begs the question, why fly?

Since solutions depend on the means available and on current environmental resources and constraints they are not guaranteed to work. A dodo, for example, is well equipped to solve the problems of hereditary survival on a remote island without significant predators but is ill equipped to solve the same set of problems on an island visited by hungry humans. Similarly the problem of infection control in humans is radically affected by the availability of vaccines and antibiotics. The key point is that core survival problems endure despite all but the most extreme environmental changes and, furthermore, that solutions need not be wholly biological. Hence a diachronic or counterfactual analysis is possible, as is an analysis comprehending a not wholly biological class of target phenomena. Both these possibilities will prove highly significant in the analysis of presumably rational decision making continued in §3.7.

Before this, however, several issues concerning the causal structure of decision making and action warrant analysis. They are matters of long established concern in the philosophy of mind and the philosophy of causation. They are, for example, implicit in the discussion of self-control in Bermúdez (2018) referred to in §1.4. In the following three sections I will make a number of proposals in response to these issues. The proposals are, like several others, frankly hypothetical, but they appear consistent with known constraints and illustrate the form a non-trivial response may take. Some such response is ultimately required.

3.4 Computation

A key feature of human decision making, as noted by Davidson and as assumed in economic modelling and intuitive description, is that choices are made by reference to possible effects variously located in the relative future. In this section I will examine a crucial underlying question the answer to which will significantly affect later analysis. It is this. What is the relationship between such an effect as envisaged and the corresponding effect eventually realized? I will start by describing and rejecting a number of initially plausible proposals, working towards a version that apparently has greatest current plausibility – namely that envisaged effects are potentially descriptive constructs, within an adaptively evolved partially algorithmic computational system, that purport to describe actual or hypothetical conditions. I will, in due course, pull out these proposed features and examine their implications and justification.

Preliminary intuition suggests that envisaged effects are actual states of the world as perceived or to be perceived via the senses or as subsequently recalled. So, for example, if I put out the waste to be collected for recycling tomorrow, it is the actual collection of the waste tomorrow that I envisage. More careful consideration shows, however, that this cannot be generally correct. There are many familiar counterexamples. For example, I may have forgotten that it is a bank holiday week and so the collection is a day later. If so, the event I envisage will not occur. Nor is this just a matter of futurity. I may come home fully confident that the waste has been taken and be surprised to find it has not. Similarly I may envisage an event that never occurs – perhaps even one that can never occur, such as meeting a ghost. Moreover, some effects exist only in virtue of cultural classification, such as getting a job or paying a compliment, and many are known only by report or as acknowledged fictions and so cannot be perceived as such.

All this is well known, and the folk notion of relevant effects has adapted to it although not in the first instance so as to generate a fully coherent theory. The usual intuitive accommodation is to assume that relevant effects are real but occur in an alternative version of the world that is accessible by one or more senses analogous to perception such as memory and imagination. The standard philosophical account of possible worlds might be interpreted as a sophisticated version of this view. However, although it accommodates cases that are suitably similar to ones that are, by common consent, actually perceived, it has considerably more difficulty with culturally defined effects and fiction.

Fiction is a particular puzzle, involving a characteristic epistemic equivocation sometimes described as the suspension of disbelief. Cultural classification ought to produce a similar equivocation but instead it is commonly resolved by fiat – by assuming that culturally defined features are straightforwardly real and that people who classify things differently are simply wrong. So political, religious, cultural, and even philosophical disagreements typically descend into disputes about what is real. But, given that systems of cultural classification evidently evolve, this cannot be coherently maintained. The problem is exacerbated by the fact that scientific classification is also culturally non-standard and evolving so it creates one or more additional candidate versions of reality, leading to yet further disputes about what is real.

An alternative account is suggested by the observation that many relevant effects can be described in the natural language used by the relevant agents and that some, especially fiction, appear to exist primarily in this form. Furthermore, the evaluative attitude to a condition may depend on how it is described – on whether, for example, an action is described as bold or rash, wise or cowardly, merciful or lenient. This suggests the possibility that every relevant effects is exactly as described by some available linguistic formula. The proposal has the extra advantage that it offers an account of the productivity of possible conditions in terms of the combinatorial structure of language.

However, it too encounters objections. Two stand out. One is that there are discernible effects for which there is no available and sufficient specifying description. Smells are an obvious example. I may detect a smell that I can easily recognize but cannot adequately describe. The other is that perception often appears intuitively prior to language. I may see something, and recognize it, but struggle to describe it. There is a considerable industry that exploits the fact that people assign value on the basis of qualities that they perceive only subliminally.

As is widely recognized, a possible solution is suggested by advances in information technology since about 1950 (Cummins and Cummins 2000). It is now entirely usual for devices to exist that reliably instantiate arbitrarily complex systems of dynamically interrelated internal states in a way that answers effectively to the notion of encoding and processing potentially descriptive information. Each typically has an interface for the input and output of externally interpretable language-like expressions and hence must admit internal states that answer to the categories picked out in such expressions, but they are typically embedded in a much richer internal system of implicit categories and procedural units for which no adequate linguistic description in the interface language need be available.

Such a device typically has an interface the operation of which answers to the notion of sensory input. Again, this need not operate strictly on the basis of standard linguistic categories or processes. On the contrary, it has recently become usual for input processing to operate via statistical pattern recognition embodied in neural network technology (Churchland 2000). For example, photographs uploaded to social media are often subject to face recognition processing by means of which images of known friends have identifying labels automatically attached. Generated or stored contents can be sequentially indexed, vindicating a notion of linear temporal order. In such a system, the categorization process is often conceptually opaque even where the categories themselves are conceptually transparent. Processing may be a mixture of the probabilistic and the algorithmic (cf. Fodor and McLaughlin 2000, Smolensky 2000).

A theory that assumes that human agents instantiate a system of approximately this kind – a partially algorithmic computational theory of mind – can account for a number of otherwise problematical observations. The recognition of distinct smells can be explained as involving a process much like face recognition and the fact that there is no shared naming system for smells by the assumption that in humans smell recognition is rather variable and difficult to calibrate or olfactory classification is of relatively limited inferential value. The productivity of imagination and planning, including fiction, is as easy to accommodate as under the linguistic hypothesis. Linear temporal order and apparent concurrence can be accommodated by sequential indexing. Prediction can be accommodated by comparison of differently indexed contents. Intentional attitudes, logical inference, and heuristics and biases can be accommodated by appropriate assumptions about internal processing regularities. Innate competence can be accounted for in terms of a priori procedural structure. Learning can be accounted for in terms of admitted functional modification. And mental dysfunction can be accounted for in terms of storage and processing errors. There remains, however, a serious problem in showing how this can provide either a causally or an intuitively acceptable account of decision making as a process realizing rational choice. I will consider this next.

3.5 Causal Asymmetry, Records

It is implicit in all the previous discussion, all the way back to the original quote from Davidson on page 9, that choice can be causally effective. The assumption, which reflects intuition, is that agents make choices that depend only partly on past conditions and that some effects in the relative future can be more accurately predicted from information that includes data about these choices than from information that excludes such data. Without this assumption the standard theory is at least intuitively misleading since what are interpreted as chosen effects could always be equally well predicted without reference to any choices being made. Some theorists are willing to bite this bullet (e.g., Lockwood 2005: 254) but it is at least perplexing.

Within behavioural economics the assumption that choice is causally effective is usually taken for granted without further discussion. But in the philosophy of mind and the philosophy of causation it is not treated as obvious and is widely discussed (e.g. Gillett and Loewer 2001, McLaughlin and Cohen 2007, Price and Corry 2007). Since it is possible that some of the difficulties encountered in behavioural economics arise from a failure to examine the relevant issues carefully I will do so in this and the following section. The resulting discussion may at first appear tangential. This is incorrect. The issues are fundamental and the resulting investigation yields valuable results. It provides, inter alia, a theory of evidence that is of considerable significance in the explanatory analysis of choice and action.

The problem can be divided into two parts. The first is determinism. If, in general, every choice is fully determined by its prior context then whatever follows from a particular choice follows equally from its prior context. This issue may be disposed of immediately. Whatever philosophical tradition may dictate – especially in the form of Leibnitz's principle of sufficient reason – nothing in contemporary physical science provides compelling evidence in favour of the sort of determinism envisaged. The only extant attempt to devise such a theory – superdeterminism – is more often disputed than advocated (Earman 1986, Bell 1987, Larsson 2014, Hossenfelder 2020, Baas and Bihan 2020). Norton (2007), in particular, argues convincingly to the contrary.

To deny that every choice is fully determined by features of its prior context is not to assert that some other transcendent mode of causation prevails. The claim is merely that adding data about choices actually made enhances predictive power, in much the same way that adding data about which way up a tossed coin falls at the beginning of a football match adds power to a predictive account of what occurs next. Nor does it require Dennett's "intentional stance" (1987). It is just a general observation about underdetermination. The implication is that, whatever the precise aetiology of choice, information about choices made can be irreducibly significant in enhancing the accuracy of prediction of subsequent effects.

This assumption of underdetermination accords with the intuition that many trivial details of observed events – the precise shape of a cloud, the trajectory a dead leaf blown on the wind through a wood, the glint of light on ruffled water as seen momentarily by a particular observer, and so on – cannot plausibly be explained, at that level of detail, without reference to transient local contingencies. The converse claim that all such details are in principle predictable at any earlier time given sufficiently accurate data on the state of the universe at that earlier time is unsupported by evidence. The observation that having more information often allows marginally improved prediction does not, by induction, constitute such evidence.

The second part of the problem is temporal asymmetry – that choices are assumed to reflect past effects and to constrain future effects, and not vice versa. On this the evidence from physics is more equivocal. On the one hand there is a sound basis on which to define a universal temporal direction, conventionally measured forwards from the big bang. But on the other hand, despite the impression given by intuitive experience, the principle that causation is fundamentally asymmetric – that earlier effects cause later effects but not vice versa – is, with one conspicuous exception, not widely supported in theoretical physics. Temporally symmetric explanatory analysis, or even partial retrocausality, is at least equally supported (Feynman 1965, Cramer 1986, Stehle 1993, Penrose 2004, Laughlin 2005, Aharonov and Vaidman 2008, Zee 2013, Evans 2015). The exception is thermodynamics.

The second law of thermodynamics expresses the universally observed principle that in all closed systems improbable macroscopic structure tends, in a well defined sense, to decay with advancing time. If change occurs, it is invariably in the direction of increasing aggregate disorder as quantified by a statistically defined measure of entropy. This statistical measure of disorder is not precisely correlated with the intuitive notion of disorder. For example, the separation of oil and water in an undisturbed container involves, for reasons relating to intermolecular bonding, increasing entropy (Silverstein 1998). But in general the relationship is strong. As Vallino et al. remark (2013: 340), the words written on a page, although contributing to its intuitive orderliness, contribute very little to its relatively low entropy. Nevertheless, if the page is burned, the intuitively recognized orderliness of the words is lost along with the structure of the page. The intuitive impression that significant structure tends to decay is thus well justified by known physics.

The significance of this was noticed by Reichenbach. He describes a footprint in sand, which thermodynamic processes will eventually erase, as inherently improbable and hence, while it survives, as "[a] 'record' that at some earlier time a man (sic) walked over the sand, thus causing footprints"

(1956: 150). The crucial observable feature of the footprint in this account is not its history, which is not observable – the footprint being, like all other observables, merely *present* on the occasion of its observation – but the wealth of distinctive structural detail it exhibits. Such detail is inevitably destroyed by random physical interaction. Hence, while it survives, it is at every moment inherently improbable. Its continuing existence is, therefore, indicative, at least, of its surviving continuously within a window of relative stability, perhaps from or to some characteristic condition or event.

In this way, instances of inherently improbable structure can serve as indicators, or records, of remote conditions or events. Evolving relations between records can tracked, and observed instances of their origination and/or destruction compared either directly or by comparison with other associated records. This, in total, supports an evolved system in biological agents of correlated, sequentially indexed internal record keeping that we call memory – memories being themselves internally preserved records.

We are, on this basis, as capable of extrapolating patterns forwards in time as backwards. We often do so. I have no good reason to be significantly less confident that the room I am in will be here tomorrow than that it was here yesterday. Nevertheless, given the ubiquity of thermodynamic decay there is a persistent asymmetry that we regularly observe. It is that, in relation to precise structural details, describable initial conditions associated with similar records are often strongly correlated whereas describable final conditions are varied and often variously combined. For example, particular features of a footprint, while it survives, are strongly correlated with independently recorded features of the foot that made it but not with features of the wind or waves involved in its destruction. Hence the observation of a particular pattern of distinctive structural detail is generally much more distinctly informative about conditions in its relative past than about conditions in its relative future. No prior notion of asymmetric causation is required. This conclusion applies very widely to, for example, natural and manufactured objects, photographs, documents, and memories. It is a direct consequence of thermodynamic asymmetry.

Reichenbach fails to follow up on his original insight in this way. Instead he continues immediately with the development of a theory of asymmetric causation due to thermodynamic branching based on the then common assumption that the universe is of infinite age and hence that the second law of thermodynamics is only a temporary fluctuation. This is no longer tenable. Regrettably, his proposal was published posthumously so he had no opportunity to revise it. Subsequent theorists have tended to follow up on ideas arising from the latter, particularly the notion of a branching structure of macroscopic conditions (Albert 2000, 2014, Loewer 2007, 2012a, 2012b, Lockwood 2005) or an associated proposal that asymmetric causation is inferable from statistical data identifying a common cause (Salmon 1984, Pearl 2009, Hofer-Szabó et al. 2013). Frisch criticizes the former as involving "an equivocation on the notion of macro-state" (2010: 25). Pearl eventually rejects the latter, describing it as a "blunder" (Pearl and Mackenzie 2018: 50). The original insight seems to have been lost.

The error involving the forward branching structure of macroscopic conditions need not concern us. But the error involving the common cause principle is, despite Pearl's recent denial, sufficiently well entrenched that a more explicit analysis is warranted. The problem is, essentially, that its plausibility rests on an unrepresentative choice of cases. A commonly cited example is as follows (Reichenbach 1956: 157, Salmon 1984: 158-167). Some members of a group of actors share a meal containing poisonous mushrooms and simultaneously fall ill. It is then claimed that the statistical association between mushroom eating and falling ill indicates asymmetric causation because the single common event of mushroom eating precedes the correlated multiple events of falling ill. Furthermore, even if inverse cases are possible – such as that several actors fall ill simultaneously and the show is cancelled – it is claimed that the coincidence is invariably or almost invariably already explicable by citing a prior common cause such as mushroom eating, and hence, by Reichenbach's exclusionary principle (1956: 159), no inverse causal inference is warranted. But this argument relies on a failure to notice inverse cases that lack a prior common cause. In fact they occur quite frequently. For example, there are punters at a horse race. Some win, some lose, at random. After the race a queue forms at the bookmaker's booth. Joining the queue predicts winning, just as mushroom eating predicts falling ill. And there is in this case, by definition, no prior common cause of winning, hence Reichenbach's exclusionary principle does not apply. Hence the common cause argument, if valid, entails retrocausality. Examples like this are quite common: people arriving at a theatre, rainwater collecting in a gully, wind blowing a tree down, salt crystallizing from solution, and so on. As Pearl concludes, and as discussed again in §6.1 below, it is an error to suppose that asymmetric causation can be inferred from observed statistical relationships alone.

Returning to Reichenbach's original insight it should be emphasized that, in contrast to the theories proposed by Albert, Loewer, and Lockwood, it implicitly identifies records in terms of their exhibiting a high level of persistent and distinctive structural detail rather than a particular type of causal or intertemporal connection. This persistence is not evidence of thermodynamic equilibration but, on the contrary, of temporary local resistance to thermodynamic equilibration. This is what makes the pattern special and recognizable. Destruction is ubiquitous. Here is something that has not yet succumbed.

This constitutes, in effect, a theory of evidence. Pace Carnap (1950), Williamson (2000), and many others, evidence is not everything you know. It is a small subset of currently surveyable material exhibiting a high level of persistent and distinctive structural detail. Detail is everything. It is for this reason that courts question witnesses at length, that physical evidence is subject to forensic analysis, that original documents are preserved and examined, that information concerning the provenance of paintings and other works of art is important, and that it is otherwise trivial circumstantial detail that makes recalled memory compelling. It is why forgery is the enemy of knowledge – for forgery creates detail falsely. Interpreted in terms of information theory (Shannon 1948), it is redundancy that validates a record. This is why a checksum is generally added to a transmitted message, in order to confirm its status as information rather than noise. And it explains in terms of the greater wealth of still-detectable detail and the more limited time for probable corruption why recently generated evidence is usually the most informative.

It is import to emphasize how small the subset of evidence is. The vast majority of actual conditions and events pass without contributing distinguishably to any discernible structural pattern that would serve, after any significant period, as a reliable record of past conditions or events. But because accurate prediction is of adaptive benefit, and accurate prediction depends on recognition of intertemporal patterns, and recognition of intertemporal patterns depends on reliable evidence, a huge investment – biological, cultural, and cognitive – is continually made to contrive the creation, preservation, and recognition of detailed records, in memory and other media. This involves a continual dedicated sheltering of material patterns, including the content of memory, from thermodynamic decay.

3.6 Agency

This theory is capable of providing an account of the causal role of choice as follows. Agency is, primarily, a mechanism that increases the reliability of prediction. Consider, for example, two complex macroscopic conditions such as *A*, a pile of various pebbles and several empty jars, and *B*, the same pebbles variously placed in the jars, in the temporal order (*A*, *B*). Given underdetermination – or just imperfect knowledge – there is no generally reliable one-to-one mapping from possible versions of *A* to possible versions of *B*. However, it is possible to insert a partly cognitive process *C*, involving a characteristic component recognized intuitively as choosing, between *A* and *B*, so as to concentrate the underdetermination in the relation (*A*, *C*), leaving (*C*, *B*) as approximately determinate, so that *B* can, given relevant cognitive resources, be more or less reliably predicted from *C*. On this basis, given that *A* is past relative to *C*, the underdetermination of the relation (*A*, *C*) can be eliminated to the extent that there exist relevant records *R* persisting from *A* to *C* and hence that *C* can be adjusted to compensate for variation in *A* as inferable via *R* at *C*. The existence of such records does not depend on any reliable causal relation between *A* and *C* but on there being, detectably, in this case, no relevant intervening disruption. Hence reliable prediction from *A* to *B* is possible via *R* and *C*, provided that *C* depends in some computationally predictable way on *R*. No equivalent inverse effect is possible because of the asymmetry of records. The effect is typically most reliable if *A* and *C* are relatively close. Ensuring that (*R*, *C*) and (*C*, *B*) are, in each relevant case, approximately determinate requires a grasp of prevailing nomology, reliable cognitive and motor processes, and stable boundary conditions. Humans evidently pay considerable attention to these matters.

The characteristic effect of this mechanism is revealed most clearly in cases intuitively recognized as involving repetitive choosing and arbitrary or contrarian choice. Suppose, as described, an agent is faced with a pile of pebbles and several jars. The agent picks up pebbles successively and places each in one or other of the jars. Whilst there is normally no predicable correlation between the initial state of the pebbles and their final locations the agent can, approximately, ensure such a correlation by a suitable pattern of choice. Such a pattern of choice can be organized to satisfy almost any arbitrary decision principle, however trivial. It might be varied to reflect the type, or size, or shape, or colour of the pebble, or its numerical position in the sequence, or a ritual instruction, or so as to maintain a balance or imbalance between the jars, or – so far as is humanly possible – a merely random assignment. Delaying or avoiding a choice is similarly open to adjustment. The decision principle itself may be chosen on a similarly arbitrary basis, recursively. Observing such patterns, the agent may become aware of their openness to deliberative or even perverse modification. This plausibly grounds the intuition of free will and justifies the attribution of responsibility insofar as underdetermination is otherwise assumed.

This proposal vindicates Dretske’s account of the causal structure of intentional action (Dretske 1988, McLaughlin 1991) with one crucial modification. In the current notation, Dretske's illustration of his proposed analytical structure is as follows (1988: 84 Fig. 4.1):

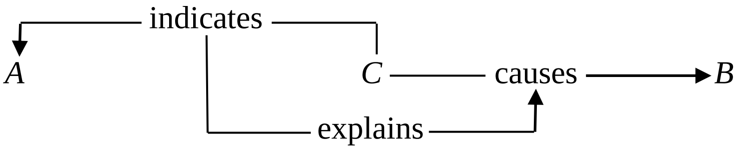

Figure 8: Dretske's intentional structure

His problem is to define the relation between *A* and *C* in a way that does not assimilate it to the usually assumed causal relationship between *A* and *B*. His proposal is to characterize it in terms of the meaning of *A* as indicated at *C* based on historically observed correlation. As several commentators in McLaughlin (1991) observe – notably Dennett, Cummins, Horgan, and Kim – this cannot do what is required. It provides, inter alia, no way of distinguishing correlations of type (*A*, *C*) that are appropriately characterized as indicator relations from other correlations. The current proposal provides a response. The connection is via records. They are distinguished by the unusual and distinctive structural detail preserved from *A* to *C*, not by any otherwise distinguishable intertemporal relation. It is thermodynamics that guarantees the inferential significance of this detail. Indeed, without this special type of connection it is not clear that any non-adjacent prior element is distinctly identifiable as of type *A* at all.

This proposal provides an account of the intuition in the example of punters queuing at the bookmaker's booth described in §3.5 that queuing does not cause winning. The intuition derives not from a local statistical relationship but from the experience of active intervention in similar cases – namely that arbitrarily adding people to or removing them from the queue does not alter whether or not they win, for which there is independent evidence. Hence the intuition depends on an assumption about the asymmetric causal effectiveness of action. As an externally observed effect, action is subject to the usual non-asymmetry of statistical inference, but as a subjectively experienced or reported effect it has a unique temporal asymmetry arising from the intuition of its involving choice relative to evidence. It is this that underpins the assumed directionality of Pearl's *do*(·) operator, in that the experimenter is aware of choosing what to do.

It may be remarked that most bearers of records, including individual humans, are uniquely identifiable persistent macroscopic physical objects. This may account for the prominent role of macroscopic physical objects in folk ontology despite the conceptual and explanatory difficulties that this is known to involve, as evidenced in the ancient problem of conceptualizing change, especially death, and despite the overwhelming dominance of process analysis in modern science. It may also be remarked that the theory plausibly accounts for the historical difficulty in categorizing the products of biological evolution. Living things bear all the hallmarks of reliable records. It is therefore an obvious step to infer their initial creation by analogy with that of other approximately similar things such as works of art or craft, as inferred from observation. The same applies to prominent landscape and cosmological features. It is a considerable intellectual step to escape this pattern of inference, given that we have only very limited local evidence of their aetiology, such as of pigeon breeding and coastal erosion.

Concerning the usual assumption that the past is fixed, it is commonly observed that distinct records may survive independently and that where several records of a single condition or event exist, modifying one does not modify the others. Hence there is good reason to treat detailed records as presumably veridical and the conditions they indicate as non-malleable, and to assume by default that if no record is currently known an undiscovered record may yet be found that would resolve an envisaged uncertainty.

There is a residual question posed by the methodology outlined in §1.3 as to why it is commonly thought that scientific observation supports an assumption of determinism. A possible answer is that scientists, in choosing what to investigate and model, choose processes that are, so far as possible, deterministic. Experimenters go to considerable trouble to isolate their systems from extraneous inputs and they usually suppress results that are assumed to be significantly contaminated (e.g. Shankland 1964). The frequently impressive technological application of the resulting findings depends on the fact that suitably similar conditions can be reproduced more widely. For example, machinery is designed, sited, maintained, and used precisely to achieve the necessary stability. The reported results are not false, but they are a misleading guide to the complexity of the universe as a whole. This issue is discussed at greater length in Toulmin (1953) and more recently in Cartwright (1999), Horst (2011), and associated literature.

If this analysis is correct it follows that the apparent asymmetry of effective action and of the past versus the future is based ultimately on thermodynamic asymmetry via the asymmetry of records. This in turn grounds the folk theory of asymmetric causation by analogy or extrapolation and Pearl's theory of statistical causation (Pearl 2009, Pearl and Mackenzie 2018) via the admission of a *do*(·) operator that implicitly assumes both the efficacy and the intuitive asymmetry of agency.

3.7 Metatheoretical Alternatives

Having established a series of crucial background assumptions concerning the rational and causal grounding of agency it is now possible to reconsider the metatheoretical alternatives discussed in §3.1, particularly as set out in Figure 7 on page 68. In this section I will summarize the conclusions so far and survey possible lines of development. The aim is to map out a way forward through subsequent chapters leading, eventually, to the analytical framework to be developed in Chapter 7.

It is clear that if the usual theory of intentional action as described by Davidson is to be vindicated it must be on the basis that rationality is admitted as an explanatory principle rather than as only a descriptive ideal and that it ought to be at least approximately satisfied in most admittedly competent human decision making. Furthermore, since any approximately adequate constructive theory of human decision making must, on the model of genomics for example, almost certainly be of very great complexity and be embedded in a much wider body of cognitive, neurophysiological, and biochemical theory as yet imperfectly developed, it is to be expected that an initial theory of rational decision making, if any is possible, must be a principle theory rather than a constructive theory. A subsequent constructive theory would then be justified as such in virtue of its constituent parts demonstrably satisfying constraints implicit in this already specified a principle theory. For example, if a specifically neurological theory of rational decision making is possible it must be on the basis that it provides an analysis of processes that are, in context, demonstrably characteristic of cognitive and/or behavioural effects already recognized as instances of rational decision making. Otherwise it is just neurology. In other words, an initial theory ought to be located in category PE in Figure 7.

This, of course, is the category containing neoclassical economics – which perhaps explains the marked reluctance of adherents of neoclassical theory, as described by Dhami (2016), to accept any weakening of its key features, either by abandoning rationality as an explanatory principle (PD), or in moving to a constructive theory (CE), or both (CD). And yet it is clear from the research described that neoclassical theory, in the usual axiomatic form originally established by von Neumann and Morgenstern, does not provide a generally satisfactory empirical account of individual human decision making. Either it is insufficiently constrained to entail any particular consequences or it is subject to frequent falsification.

Three main responses to this discovery can be observed in the literature. One is to abandon the attempt to satisfy a well defined standard of rationality. Another is to pursue a constructive analysis, with or without an assumed standard of rationality. A third is to ignore the empirical evidence and continue the investigation of axiomatic systems as a purely mathematical or intuitively justified – typically Bayesian – enterprise. All, surely, concede too much. All assume that the battle to devise an empirically adequate principle theory of agency that has rationality, suitably defined, as its fundamental explanatory principle is already lost without sufficiently considering possible alternatives.

The analysis above suggests an alternative based on four key assumptions. Firstly, rationality is more than just consistency over preferences. The notion of consistency necessarily requires a distinction between types and tokens – which is to say, a system of classification. But there is no natural classification of preferable features of the world absent some notion of function or purpose. Nor is there any universal notion of function or purpose that, it is generally agreed, rationality must satisfy. Hence the system of classification relevant to particular agents must be discovered by analysis of the demands or priorities under which they themselves operate. It cannot be assumed to be well defined without further analysis or be imposed by stipulation. It follows, for example, that in the expected utility formula (2.2.3) it is the specification of a set of outcomes $\{x_1, \dots, x_n\}$ that does much or most of the ultimately significant explanatory work, and in the subjective expected utility formula (2.2.4) it is the specification of a function f mapping events to outcomes. Hence whether either formula forms the basis of an empirically significant analysis depends on whether or not a well-justified methodology exists for their specification. Supplying such a methodology is a significant problem, as is implicitly recognized by, for example, Fumagalli (2020b) and Schwarz (2021).

Secondly, the rationally justified priorities under which a human agent operates cannot be only those arising from the demands of hereditary survival that drive biological evolution. Most selective effects in agency operate on a much faster time scale than those of biological evolution, and a vast and astonishingly varied collection of derivative human priorities has developed within the envelope afforded by biological survival, many parts of which are only indirectly connected to discernible biological imperatives. Both effects can be accounted for by assuming that the key principle that drives short-term selection in cognitive methodology is not biological survival, or even social acceptance, but predictive success, on the basis that predictive success, although empirically determinate, operates on a comparatively short time scale and is otherwise relatively unconstrained as to its content.

Thirdly, a principle theory of human agency, although not explicitly directed at the elucidation of constructive mechanisms, must respect applicable constructive limits. It makes no sense to develop a theory that requires the realization of effects that could not be realized, even approximately, by any available neurophysiological, psychological, or other means. The theory must, in other words, satisfy a notion of bounded rationality. It cannot, as Bradley puts it, assume that human agents are logically omniscient and maximally opinionated. Almost all the interesting problems of human agency arise because, on the contrary, people are not logically omniscient and maximally opinionated.

Finally, a theory needs to answer the 'when' question described on page 57. A theory of preferences with no rule stipulating when a preference is converted into a decision is not, strictly speaking, a theory of decision making at all. Hence, since various features both of relative preference and of the context to which it applies vary over time, it follows that any relevant theory must engage with the dynamics of relative preference in context and with its implications for the timing of consequent decisions. This is not, as might perhaps be assumed, a constraint that applies only to constructive theory – only to, for example, a theory of realized heuristic methods or neurophysiological processes. It applies equally to any non-trivial principle theory. If this is correct, the lack of attention to dynamics in the usual proposition-based form of decision theory is not a convenient idealization. It is a serious flaw.

To provide an explanatory analysis of rationally justified demands or priorities in decision making involves the same adaptationist dilemma as arises in evolutionary biology, as discussed in §3.3 – namely that assuming a one-to-one explanatory relationship between particular demands or priorities and associated modes of solution is typically an error, but assuming only a holistic relationship between all demands or priorities and any or all established solutions is vacuous. The same solution is available, namely problem analysis.

Consider, for example, decision making associated with the provision and wearing of contact lenses. The primary constraint in the design and manufacture of contact lenses is that they are part of an employed solution to the problem of gathering accurate visual information, which might be classified as primarily an epistemic or cognitive problem. But contact lenses also have to be consistently wearable and hence not bearers of infection – which might be classified as primarily biological – and they must be predictably available and socially acceptable – which might be classified as primarily cultural. These problems jointly constrain relevant agential methodology, but with varying degrees of comparative priority.

Crucially however, responses generated by rational decision making to specific problems such as these are not characteristically shaped, as in biological evolution, by the non-survival of unsuccessful respondents. Rather, each response is a cognitively derived product shaped by cumulatively successful prediction, in which the relevant demands and priorities are represented as envisaged conditions having assigned value, modes of intervention are represented as available options, and the predictability of consequential relationships is represented via assigned probabilities. In other words, decision making, as an evolved activity or competence, is itself shaped by three higher level meta-problems – namely justified valuation, option definition, and probability assignment. These three meta-problems are jointly characteristic of rational decision making as a generic phenomenon. It is presumably not an accident that they correspond, one-to-one, with the three key questions listed on page 10. Their recognition as such plausibly reflects the fact that we have a reasonably well developed insight into the problematic character of decision making, just as we have into the problematic character of life, and hence of nutrition, infection control, and so on.

In each case of rational decision making the primary test of a solution methodology is that it is predictively successful. On this basis the analysis of an emergent solution methodology can proceed as described in §3.3 by examining how the various problems arising at each successive problem level are, jointly, more or less successfully responded to. Hence the type of analysis appropriate to rational decision making *in general* is one that focuses on the emergent methodology associated, in various cases, with the three generic meta-problems just described – valuation, option definition, and probability assignment. In other words, a principle theory of rational choice should be concerned in the first instance with investigating the structure, and especially the dynamics, of the response to these three characteristic meta-problems in various standard circumstances, rather than with the minutiae of particular derivative preferences or choices.

In this analysis, the policy adopted by Kahneman and Tversky of investigating heuristics and biases is likely to be useful in identifying the constraints under which decision making operates, but it does not generally follow that apparent bias indicates a failure of rationality. If a pattern of response wholly disregards any admittedly relevant problem it is in that respect rationally defective. But the response to a set of problems almost always involves some compromise. Hence the response to each particular problem may be, and usually must be, to some degree imperfect. Whether this is judged to be a failure of rationality depends on what is thought ordinarily feasible at the time. For example, human memory is not expected to satisfy a standard of reliability appropriate to commercial IT systems, intuitive computational precision may reasonably be sacrificed in favour of speed, flexibility, or resilience, and apparent preference reversal may reflect varying information. Moreover, performance can often be improved by supplementary means. For example, limitations of human eyesight can be alleviated by the use of spectacles, binoculars, or night vision goggles, and memory by note-taking. Whether a failure to use available means is judged irrational depends similarly on the balance of competing considerations. The judgement is likely to be often a matter of debate.

The key task is, then, to map out how each of the three meta-problems described above is approximately solved within the dynamic methodology of human decision making, to examine what consequences or biases follow, specifically with respect to variable futurity, and to investigate possible amendment. This is the task I shall address in the remaining four chapters.

3.8 Conclusion

In this chapter I have examined the relationship between rationality and adaptation implicit in four contemporary approaches to the analysis of decision making – neoclassical economics, mainstream behavioural economics, the Meliorist approach of Kahneman and Tversky, and the Panglossian approach of Gigerenzer and his colleagues. I have identified three patterns of evolution, biological, cultural, and cognitive, that have significant adaptive consequences but are distinguishable in having different rates of change. I have examined the role of adaptationism and anti-adaptationism in evolutionary biology and proposed a quasi-functionalist type of explanatory analysis that assumes that a relevant entity, given its established adaptive response to its environment, faces a partially ordered set of describable problems and that this allows increasingly precise explanatory analysis by successive approximation over a expanding set of derivative problems. It formalizes common intuitive practice.

Decision making requires some mechanism for achieving rational effects. The only theory currently available that is capable of accounting for this involves the idea that cognition is a form of computation. But neither a fully algorithmic type of processing as envisaged by Turing nor a full connectionist type as envisaged by Churchland appears capable of explaining both the statistical grounding and the systematicity of observed cognitive activity. It is probable, therefore, that cognitive processing is partially algorithmic – that feature-detection is statistical but that arbitrary patterns of features can be stored and processed algorithmically. This admits the type of comparative combinatorial functionality usually assumed.

A theory of decision making as computation assumes a causal structure in which information is available from the past and is effective, via a mechanism that detects and responds to structurally complex patters, in generating or varying significant contingent consequences exclusively in the future. This is generally assumed, often almost without comment, together with an associated assumption of asymmetric causal determinism. But given that fundamental physics apparently involves neither asymmetric causation nor causal determinism it is in fact a major analytical problem with uncertain conceptual implications.

I have examined various aspects of this problem and developed a proposal in which the apparent asymmetry of macroscopic processes, including human decision making and action, arises from the ubiquitous asymmetry of thermodynamics. From this is derived a theory of records, and hence of evidence, based on recognizable specificity of retained detail in an environment marked by ubiquitous thermodynamic decay, and a theory of agency based on adaptive selection of mechanisms reliably instantiating computational processes that exploit the availability of records to enable increasingly reliable prediction in otherwise underdetermined diachronic scenarios.

On the basis of this account I have examined, in outline, the possibility of constructing a principle theory of rational choice based on an analysis of the constraints placed on evolved heuristic methods by the need to solve a variety of adaptive problems defined in terms of not only biological and cultural but especially predictive success. It establishes a task, in the proposed analysis of the impact of variable futurity in decision making, of investigating the way in which human decision making methodology provides an approximately effective intertemporal response to each of three key problems – valuation, option definition, and probability assignment. It is to this task that I now turn.

Chapter 4 Valuation

4.1 Preliminaries

The key question as posed on page 9 is this. In human decision making, to what extent if any does, or should, the relative value of any envisaged consequence depend on its relative futurity? Having laid out some necessary background it may be useful at this point to look forward to the proposed response. Without pre-empting the conclusions in detail I will sketch the structure of the analysis as follows.

Decision making is a generic behavioural methodology, evolved in humans by biological, cultural, and cognitive adaptation over the more or less recent past. It has a certain general methodological or heuristic structure, approximately as sketched by Davidson, developed in response to three distinctive problems – valuation, option selection, and probability assignment – grounded in the predominantly thermodynamic causal structure of the environment. Each of these problems, as responded to by some method in some relevant context, has implications for resulting decision making that may vary with the assumed futurity of envisaged consequences. As a result, decisions may display common patterns of variation, perhaps interpreted as a kind of bias, that may appear as instances of hyperbolic or other discounting. The analytical task is to identify these cases, to assess their ubiquity, and to examine methods of detecting and if possible compensating for avoidable bias.

The analysis is complicated by the fact that it has been common practice among researchers to identify options, and usually also probabilities, via what I have called a narrow definition of available options and then to infer values from expressed choices via what is called revealed preference or trade-off consistency methodology. The effect is that any effect that might result from intertemporal variation of options or probabilities is interpreted as resulting from intertemporal variation in valuation. This approach, although sanctioned by tradition, has been called into question recently by the observation, as noted on page 22, of a possibly systematic relationship between probability and futurity. This renders its implications moot even where they are not straightforwardly inconsistent.

It is this methodology that accounts for instances of intertemporal variation in expressed choices being standardly interpreted as evidence of varying value and for apparent preference reversals being interpreted as evidence of inconsistent valuation, and hence irrational. It is essential to eliminate this complication. It requires, first of all, establishing a theory of valuation that does not depend on the assumed validity of revealed preference methodology. This is my next task.

4.2 Relevant Conditions

The aetiology and form of human valuation was a topic widely discussed in the 18th and 19th centuries, as outlined in Camerer and Lowenstein (2004), but development was disrupted in the 20th century by the introduction of a behaviourist methodology that deprecated all explicit reference to introspectively generated data (Wilkinson 2008: 10-11). The justification for this was in part that as a result of its reliance on introspectively generated data psychology had "failed signally … to make its place in the world as an undisputed natural science" (Watson 1913: 163). Unfortunately the behaviourist project also failed (Dietrich and List 2016). Nevertheless, more recent developments have tended to confirm Watson's criticism. Starting in the 1950s, a computational project initiated by Turing (1950), exemplified by Newell and Simon (1972), and immortalized in the acronym GOFAI – 'good old-fashioned artificial intelligence' – attempted to develop a theory of decision making based on the strategy of formalizing a set of algorithmic inferential processes operating on data structures representing current goals and other cognitive constituents identifiable, for the most part, as introspectively available mental contents. It too has proved less that ideally successful (Greenwood 2015: 466, 482-5, Hinton 2017).

Evidently it is difficult to develop a successful theory of human decision making that either ignores or reproduces the structure of introspectively available data. Given what we know about the complex relationship between thought and action, this is perhaps not surprising.

A key implication is that, in the project to develop a naturalistic analysis of human valuation, introspectively generated data are both essential and unreliable. I propose therefore to adopt the hybrid methodology described in §1.3. Its key principle is that we treat intuitive data as explananda rather than as explanantia. Thus the analytical scheme and associated theory should be designed, inter alia, to explain why intuitive evaluative impressions are as they are rather than merely to formalize their superficial structure and, in particular, insofar as it fails to reproduce their superficial structure the theorist is under an obligation to account for the discrepancy, if not immediately then in due course. To minimize the additional explanatory burden, a reasonable policy is to accept intuitive impressions at face value insofar as consistency allows. As well as admitting obviously relevant data, this allows the implied explanation of otherwise unexpected or perplexing effects to count in favour of a proposed analysis and it provides a principled adjudication in the ongoing dispute as to whether, or to what extent, 'as if' modelling is either necessary or acceptable (Dhami 2016: 138, Okasha 2016).

The central question is as follows. What characteristics do envisaged conditions have in common in virtue of which they have, or may have, motivational value? As outlined in §3.4, it is now generally assumed that, notwithstanding the inadequacy of GOFAI, decision making is a form of computation and that envisaged conditions exist as potentially descriptive computational constructs. But whilst this appears well justified it only partly answers the above question. The problem is that, as stated, it admits too much. It requires only that all relevant conditions be representable within a partially algorithmic computational system instantiated in the valuing agent. Intuitively, this omits two key characteristics. One is that a relevant condition is classified as possible. The other is that it is classified as positively or negatively desirable. Both add significant constraints. I will discuss the former next and the latter in §4.4.

4.3 Temporal Structure

The intuition that some conditions are possible and others are not can be accommodated most straightforwardly by assuming that an agent tacitly or explicitly traces out what are assumed to be alternative partial regularities among actual or envisaged conditions in the context of assumptions about what is actually the case. This accords with the intuition that evaluating whether or not an envisaged condition is possible often involves envisaging a way in which it might be realized. By implication, what needs to be added to the generic computational theory is that the system contains resources that, in effect, constitute a causal model of the world. It embodies assumptions about both what has been and is the case, and about the – usually diachronic – regularities that connect actual and possible conditions as so represented. Evidently, in humans, such a model need not be entirely coherent nor, by current scientific standards, accurate. It may admit fiction.

The fact that conventional AI systems do not generally operate in this way – via a system of global valuation of envisaged conditions within a broadly based causal model of the world – may perhaps account for their frequently noted failure to exhibit anything like human situational awareness (e.g. Penrose 1989, Fodor 2000, Piccinini and Bahar 2013, Hinton 2017).

Among the central assumptions embodied in this account is that relevant conditions are envisaged, from the agent's current perspective, as occurring within or throughout specific time intervals. Considerable uncertainty surrounds how this is to be treated theoretically some of which has been described in §2.2 and §3.5. In the rest of this section I will survey the issue more systematically. In the background are two notable facts about the conceptualization and quantification of time. The first is that estimates of relative temporal position and duration, although open to subjective variation, are not readily conflated with any other dimensional estimates, and alternative estimates are always locally commensurable by reference to described events. For example, there is normally no conceptual confusion in the idea that one momentary event occurs before another, although there may be dispute about the fact of the matter in a given case. No other dimensional intuition – for example, of size or probability or monetary value – is so generally unambiguous. The second is that elapsed time has a unique status in the calibration of natural phenomena in virtue of which it is uniquely measurable. The resulting standard quantification of time as an independent variable occurs almost everywhere in science. It is fundamental within the SI system and in a vast range of time-series data across many fields.

From the point of view of modelling decision making these facts pose a dilemma. It is apparent, both from intuition and experiment and from consideration of issues of data storage, that the conceptualization of both temporal location and duration, whether of actual or possible conditions, must exhibit a degree of granularity. Perceptual assimilation consists in the conversion of a flow of real-time sensory data into a granular format, as famously described by Miller (1956). A computational account of processing of this kind – of, for example, speech recognition – is readily available (e.g. Lewandowsky and Farrell 2011). Furthermore, there is reason to believe that the motivationally relevant human conceptualization of time is more variable than the standard quantification (e.g. Allan 1979, Ebert and Prelec 2007). But insofar as the systematic structure of human decision making is assumed to be either a rational or an evolved response to actual environmental contingencies and these contingencies are natural phenomena governed by physical law, the standard quantification of time ought to be preferred. Departures ought to be interpreted as accidental or heuristic processing effects, or as deviations from or approximations to standard quantification – that is, as a kind of error or simplification. Moreover, since typical rates of change of observed conditions vary enormously, the optimal degree of temporal granularity needed to support predictively successful inference must depend significantly on the type of effect involved. Tracking the movement of a fly, for example, requires a much finer level of temporal granularity than tracking the movement of a commercial aircraft. Hence modelling that assumes a single scale of granularity over all conditions, as occurs in most discrete-time models of discounted utility, is problematical. On the other hand, effective predictive inference based on knowledge of recurrent processes imposes a condition of commensurability between the past and the future. Hence, if granularity is admitted, its level for each relevant type of effect must be approximately uniform through time.

For all these reasons, the standard continuous-variable quantification of elapsed time must in general be assumed to be analytically prior to any cognitively justified granular quantification. Normative theory in particular ought in the first instance to involve standard quantification.

Just as predictive inference imposes a condition of commensurability between the past and the future, evaluative comparability imposes a condition of commensurability between alternative conditions, whether currently realized or not. This conclusion is in some ways psychologically surprising. Suppose I must choose whether to stay or go. The conditions associated with my staying are, for the most part, actually, observably, present. The alternative conditions associated with my going are, by contrast, hypothetical. But this contrast is misleading. The available choice is not really between the present and the future but between alternative futures, and the persistence of currently realized conditions into the future, when viewed from the present, is no less hypothetical that the emergence of alternative conditions. At most, persistence is more straightforwardly predictable than change. If current conditions are unstable – if the building is on fire – even this cannot be assumed.

Within the granular structure of conditions individuated by occasion and duration, conditions themselves typically have a characteristic internal temporal structure. This may be merely of qualitative persistence or of the initiation or termination of an otherwise persistent condition, but it often involves considerably more internal variation. Cases of the latter type are intuitively identified as events or processes. They may be characterized by highly complex patterns of internal structural contingency and development – explosions, conversations, holidays, football tournaments, epidemics. On this basis, granular structure is recursively nested.

An envisaged occurrence of an extended or complex condition is typically characterized by a nominal occasion and duration. There is no a priori reason why the value attached to such a unit should equal the sum of the values attached independently to its parts. On the contrary, value typically accrues to a unit as a whole – as in the case of, for example, a rational argument or a musical performance – and depends on various characteristic structural and functional features. Evaluatively, a whole may be more, or less, than the sum of its parts. This generally bypasses the problem of ergodicity described by Peters (2019).

It is notable that significantly valued conditions in the envisaged future are typically associated with an intuitive feeling of hope or fear whereas valued conditions in the present or past are typically associated with a feeling of satisfaction or regret. Two issues arise – whether the class of valued conditions is, ceteris paribus, the same in both cases, and what is the functional significance of the dichotomy.

Concerning the former, whilst there seems to be no fundamental reason why we should not attribute value very differently to conditions in the present or past as compared to otherwise identical conditions in the future, there seems little evidence that we do. What we fear we also, generally speaking, regret, insofar as it appears to have occurred. What we hope for we also, generally speaking, take pleasure in, and vice versa. At most, past ills are excused and past good fortune tempered. Furthermore, despite the fact that prevailing values are known to vary, people typically judge envisaged conditions in both the past and the future by contemporary standards, anachronistically. Past evils are not excused on the grounds that they were thought justified at the time, and people generally have no intuition that what they now accept may, at some point in the future, be judged abhorrent, or vice versa.

Concerning the second issue – namely, the functional significance of our differing attitude to the future versus the past – the short answer is that the former is motivating but the latter is not. We strive to change the future. We do not, generally speaking, strive to change the past. Since we value both, this is a puzzle – one that I will discuss in the following section.

Finally, the standard method of analysing the role of futurity in decision making is, as discussed above, to assume that value is subject to intertemporal discounting. But the evidence cited in its support is almost all of expressed choices and, as has been repeatedly observed, the claim that this indicates a direct relationship between value and futurity is undermined by the fact that factors other than value contribute to expressed choices. If this is admitted, the evidence in favour of universal value discounting seems rather weak. Samuelson explicitly denies that it is rationally necessary, Broome doubts it, Sullivan denies it on analytic grounds, other philosophers generally ignore it or explain presumed effects indirectly, for example by citing a weak epistemic connection with ones future self, its adaptive benefit is obscure, introspective evidence is at best equivocal, and behavioural evidence is open to alternative explanation.

4.4 Targets

Amongst the class of envisaged conditions, not all are valued. Many are viewed with indifference. I will refer to those conditions or features of conditions that are significantly valued, either generally or by a particular agent, as targets.

An early attempt to identify and classify targets is implicit in the development of classical utilitarianism in the 18th and 19th centuries, in which it is assumed that value accrues hedonically from actual or expected happiness grounded, paradigmatically, in the personal experience of pleasure and pain, the primary evidence for which is necessarily introspective. As a theory it has been controversial almost from its inception, as is illustrated in the famous dispute between Bentham and Mill concerning the comparative value of pushpin and poetry, and it continues to face a number of widely rehearsed objections (Feldman 2004). One objection is that, on its most straightforward interpretation, it appears to imply that people only ever act selfishly. But this conclusion is grossly at variance with a second line of intuitive evidence – that, in our various ways, we care about our friends, families, and communities, even about the future of the planet, that we are often angry on hearing about damage done to valued objects or harm done to others, and to animals, and that, within our groups, we value conformity and propriety. It is possible, for example, for a decision maker to remark, without self-contradiction, "OK, but I am not happy about it." The burden of this evidence is that the class of possible targets is wider than is allowed by hedonic intuition alone. This raises a perennially troublesome question, namely how are inconsistent intuitions to be reconciled.

The methodology described in §1.3 provides a solution. It allows that we may discount a problematic intuition provided that its occurrence can be explained. The intuition that, uniquely, happiness is what is valued is problematic in just this sense. So the question is, can it be explained? If so it can be discounted, and apparently conflicting intuitive evidence such as that we care about our friends, families, and communities becomes clearly admissible.

It can be explained as follows. Just as a sensation of physical pain is approximately but not invariably correlated with tissue damage – the link is disrupted, for example, by anaesthesia – a sensation of happiness is approximately correlated with a high or rising level of expected value computed on current or currently anticipated conditions. It thus gives prima facie evidence of what is valued. But in any particular case it may be misleading. For example, a commonly observed phenomenon is of what is called referred pain – pain experienced as if from a false location such as, in extreme cases, an amputated limb. Similarly, a father might positively value his daughter's happiness on her graduation but be distressed that he will miss it. The distress is not evidence that the intuition of positive value is false for it has a different origin, namely that he will not be present. Furthermore, an evolutionary explanation of the propensity to feel pain or unhappiness can be given, namely that it flags up a currently relevant issue for prompt attention. On this account, the felt pain or unhappiness is not itself, or not normally, a target condition but only a typical but possibly erroneous correlate of such a condition – that is, a proxy.

This explanation provides a possible solution to another problem associated with the hedonic theory, that of temporal continuity. Happiness, like pleasure and pain, is conventionally assumed to be a fluctuating state in real time. It is therefore natural to quantify its total over any given time interval by integration. This is at variance with the apparent granularity of temporal quantification described in §4.3 and with the evidence of subadditive discounting described in (A18). Direct intuitive evidence of value attribution normally has no such cumulative implication. For example, if I value the experience of eating a meal in a particular restaurant there is no obvious intuition that the attributed value is a linear function of its duration. There is, more likely, an optimal duration, just as there are optimal levels of many other variables, such as the quality and quantity of the food, the noise level, the attentiveness of service, the companionship, and so on.

Evidence of the complex interaction of variables in valuation is suggested by an example discussed by Wilkinson (2008: 50). In a series of studies of pain in patients undergoing colostomy examination, patients were asked to give a rating of current pain at one-minute intervals during examination and, subsequently, of aggregate discomfort. The latter is best explained as depending on the maximum during the entire examination and the mean during the last three minutes. Patients whose examination is extended by an unnecessary but relatively painless final interval report lower aggregate discomfort and lower resistance to repeated treatment (Katz et al. 1997). These matters will be discussed more fully in §4.6.

Rejection of the hedonic theory frees up other intuitive evidence of what we value for direct examination. In doing so it allows several other generally problematical issues to be clarified. One is the observed absence of loss aversion in trading activity (Dhami 2016: 232). Generally, as described in (A6), agents demand a significantly higher monetary payment to relinquish items they own than they are willing to pay to acquire them. But there are, after allowing for transaction costs, significant exceptions to this rule. They include the exchange of cash, generic commodities, and goods intended for resale. And then there are counter-exceptions. An antiques dealer may acquire an item for resale and then be unwilling to part with it. After receiving a coin in payment, a person may think of it as a memento of a happy occasion and choose to keep it. And so on. Standard economic methodology deprecates the sort of introspective evidence that would distinguishes these cases. Hence they appear as anomalies. Introspective evidence distinguishes them.

There is a similar issue with quantification. Experimentalists tend to adopt a method that reproduces conventionally expressed promises of cash payment, as illustrated repeatedly in the Appendix. But theorists admit a variety of less well defined constructs including, for example, wealth (Dhami 2016: 1463-9), utility flow (Halevy 2014), reward (Lohrenz and Montague 2008: 459), happiness (Benjamin et al. 2012), welfare (Sen 1970), and wellbeing (Broome 2004). Furthermore, whilst evidence indicates that valuation is, in general, quantitatively dependent on a current reference level, various different methods are used to quantify this level. It may be assumed to depend on, for example, current assets, or an established goal or contract, or an expected outcome, or a notion of fairness or obligation (Dhami 2016: 228-9, 250-5, 264-72, 439-44). Once again, the admission of additional intuitive evidence makes it possible to identify an appropriate basis of quantification in each case.

Intuitive evidence of what is valued, either positively or negatively, offers a surprisingly wide spectrum of candidates. It includes, at least, a wide range of presumably possible, individualizable, spatiotemporally instantiated current and future conditions. In the rest of this section I will discuss some key analytical issues that arise within this characterization.

A typical class of possible targets exhibits the kind of combinatorial productivity characteristic of natural language (Fodor 1998: 94-108). On this basis, alternative conditions, actual or imagined, are open to systematic valuation. For example, as mentioned, I may attribute value to any of a range of conditions, actual or imagined, each characterized as possibly arising in the personal experience of eating a meal in a restaurant. Assuming a computational theory of mind, this combinatorial productivity is accounted for on the basis that for each envisaged condition there exists, in some physical form, an internal representation of its relevant features and of their arrangement in the prevailing context, as in a descriptive algebra, and that the attributed value is a function computed on this representation. This creates a distinction between conditions that are conceptually impossible in that they are not describable in the relevant algebra and conditions that are classified as practically impossible in that they are assigned zero probability given the assumptions embodied in the agent's current model of the world. If motivational force depends on the product of value and probability these alternatives are not readily distinguishable in observed outcomes since neither entails any consequent action.

The implied distinction between conceptual and actual impossibility provides a possible account of the effect, already described, that whilst conditions in the past are generally valued on more or less the same basis as conditions in the future, this valuation is motivationally ineffective. Whatever nomological relations are admitted by their causal model of the world, human agents generally assume, as discussed in §3.6, that conditions in the past unlike those in the future cannot be modified and hence that all envisaged alternatives to assumed past conditions have a probability of zero. Hence no motivational effect results. The same limitation applies to alternatives to presumably fixed or inevitable future conditions. It is, perhaps, amongst the strongest reasons for distinguishing sharply, as in expected utility theory, between value and probability.

A striking and distinctive feature of the human value system is that it appears to target not only actual or envisaged conditions but also the actual or envisaged objects and qualities involved in the realization of those conditions. According to our intuitive model of the world there are objects classified by type, including ourselves, that we cherish or detest – distinguished, generically, as friends and enemies – and qualities classified by type, often admitting relative quantification, that when instantiated in such objects we enjoy or admire or that engender disgust or aversion (cf. Gintis 2009: 49). The two are frequently linked inversely in that, for example, agents positively value good fortune in friends but ill fortune in enemies and proximity in things they like but remoteness in things they dislike. A relevant object may be composite – a couple or a collection, for example – or it may be may be global, or abstract. I may, for example, value a just world or the works of Shakespeare. In this system, relations among objects are conceptualizable as qualities instantiated in collections – for example, the distance or direction from *A* to *B* can be conceptualized as a quality instantiated in the ordered pair (*A*, *B*). The system as a whole may have a distinctly Manichaean quality.

Combined with other evidence, including linguistic structure and the established status of predicate logic and of vector representation in science, this suggests that at least a part of the human value system involves the combinatorial valuation of dyadic pairs of classifiers – of an object-type and a quality-type – each of which makes a distinct quantitative contribution to the resulting valuation. I will term these contributions ‘affinity’ and ‘salience’. On this basis, if I register that my favourite mug is broken, what happens is that I implicitly assign to the envisaged condition a value that is, in this case, a multiplicative function of the degree to which I like the mug – its affinity – and the degree to which I assume that, in general, being broken, when exhibited in an object that I like, is desirable – its salience. Similarly, if I register that a child is crying, the value I assign depends on the affinity associated with its being, in this case, my child, and the salience associated with that type of crying. Within this system, existence is generally treated as a quality having positive salience. Thus the mere existence of an object having positive affinity confers positive value – and conversely in negative cases.

I will not pursue this proposal in greater detail at present except to say that there is reason to suppose that pairings involving qualities felt as threatening are treated additively not multiplicatively, and hence that being threatening is relatively undesirable, but to a differing degree, in both friends and enemies. I introduce it here principally in order to expand the possible interpretation of intuitive evidence, particularly in cases of hedonic versus non-hedonic valuation. Hedonic valuation is unusual in that the relevant object is invariably oneself. Hence according to the proposed theory, hedonic valuation depends only on salience; the significance of affinity is concealed. This accounts for its distinctively one-dimensional structure (cf. Bruni and Sugden 2007: 160). Most other cases of valuation, including cases normally classified as involving altruism, reciprocity, justice, competition, and conflict, are two-dimensional. In such cases human valuation seldom treats all other people, or objects, equally – Nagel, Rawls, Parfit, Broome, and many others notwithstanding. Affinity intervenes. Its effect is seen repeatedly in social relations, in politics, and in behavioural economics. For example, many findings in game theoretic research demonstrate an effect interpreted as indicating variable levels of reciprocity. This is readily interpretable in terms of varying affinity (Dhami 2016: 357-86). Its aetiological justification is discussed in §6.6.

A number of other points may be mentioned more briefly. Where an object or quality is classified under more than one type its affinity or salience is that associated with the most specific type for which a distinct value is defined – perhaps a singleton. This requires a partial order of valued types. The experience of valuation may entail an effect that is also valued. Fear is unpleasant, and may itself be feared. This is an instance of a more general possibility, of double counting, that will be discussed in §4.7. And finally, a condition may be desired instrumentally, in virtue of its being a currently relevant means to some otherwise valued effect. In such a case the intuition of value is misleading. The key distinguishing evidence is that the intuition of value evaporates if the associated effect is achieved by other means.

4.5 Valence, Scale

The value attributed to a condition usually depends on its comparative magnitude relative to a current value-neutral reference condition. In economics this is illustrated by the endowment effect, described in (A6). If an ordinal metric is defined over a class of conditions it typically creates a distinction between gains and losses. Gains and losses are often valued inversely. Positive differential or incremental value, or valence, is, ceteris paribus, associated with desire.

For a given ordinal metric there sometimes exists a preferred interval metric. For example, wealth is standardly quantified in terms of cash value and sound intensity in terms of energy density. The former is exploited in economic modelling, the latter in physics. But even where an interval metric exists, evaluative discrimination is typically not proportionate. Recognition of this is the basis of the nonlinear utility functions of Bernoulli and of prospect theory (§2.3 Figures 5 and 6). A similar nonlinearity is observed in many psychophysical processes. It is formalized in the Weber-Fechner law, which holds that in general the relative differential sensitivity of a subject to a stimulus condition – the just-noticeable difference during fluctuation – is proportional to the absolute magnitude of the stimulus. This generates a logarithmic relation between effective sensitivity and absolute magnitude of the type assumed by Bernoulli and incorporated in, for example, the decibel scale of sound intensity. Its relevance to economics is discussed in Ainslie and Haslam (1992a: 71-3).

It seems possible, prima facie, that this might provide the basis for the development of general evaluative metric in terms of which all relevant conditions can be appropriately quantified. But unfortunately it works only in cases where an interval scale of absolute magnitude already exists, which is seldom the case for intuitively valued qualities. Qualities such as pleasure or pain, honesty, social status, administrative authority, or artistic merit, although admitting ordinal judgement, have no natural interval scale of measurement. Even physical size is problematical.

The lack of a natural interval scale for many apparently valued qualities led, until about 1950, to scepticism among economists as to the merit of analysis based on an assumption of cardinal rather than ordinal preference relations (Starmer 2000). As noted in §2.1, this changed with the publication by von Neumann and Morgenstern (1944) of a representation theorem showing that, assuming only some apparently plausible consistency conditions, an interval-scale of utility can be constructed from data on pairwise choice behaviour over probabilistic mixtures of alternatives. The result was to establish, almost universally, revealed preference or trade-off consistency methodology as the standard method of generating interval measures of relative value (Wakker and Deneffe 1996, Dhami 2016: 92-5, 140-8, 631-3, Wilkinson 2008: 54-7). Its central assumption is that, for any agent, a unique numerical decision utility can be assigned to each of a set of hypothetically available outcomes by observation of the agent's expressed choices over a set of appropriately specified options.

The difficulty is that, as amply demonstrated in subsequent research some of which is described in the Appendix, the consistency conditions upon which the theorem depends are seldom satisfied. This makes the methodology not just inaccurate but fundamentally unsound. Furthermore, as illustrated in §3.1, even where the results in a given case are not actually inconsistent there is no principled way, without considering broader psychological evidence, of deciding which qualities of each specified option the computed valuation applies to, since the various members of any finite set of options can be distinguished from each other in arbitrarily many different ways. Hence even if a quantification is appropriately derived it may be assigned to the wrong conditions – and, conversely, apparent inconsistencies may be the result of misidentification.

Part of the difficulty is that revealed preference methodology misinterprets the analytical task. The initial problem is not to construct a scale of relative valuation over alternatives but to construct a scale of judged magnitude over potentially valued alternatives – something that answers to, for example, weight, or price, or population size. Whether discriminated alternatives are valued on this basis, and to what degree, is another issue. If valuation is to satisfy the additivity assumption of expected utility theory, judged magnitude must support interval-scale quantification even in cases where no natural interval scale exists.

The issue is, therefore, whether, in general, for any relevant class of conditions an interval quantification of magnitude can be constructed from data available to the agent. The proposed method must work in all relevant cases not only those for which an absolute interval scale already exists or where there is an obvious basis for consistent choice. In effect, it must be able to generate an interval scale from any relevant set of ordinal data.

There is such a method. It depends on the statistical distribution of observed object-quality pairings. Provided that various individual members of a specified class of objects can be identified and observed and that, cumulatively over some relevant period, observed cases can be rank-ordered in terms of the degree to which each apparently exhibits a specified quality, an interval scale for that quality over that class of objects can be constructed based on the position of each in the resultant ranking. The simplest scale is linear with rank, but others are possible. Linearity combined with preferential sampling of extreme or unusual cases produces a scale that is more sensitive to significant differences. If gains and losses are sampled asymmetrically an asymmetric scale is produced. The reliability of the scale created in this way depends on the representativeness of the recorded observations with respect to the implicit class of cases – that is, on sampling. In the limit, it is the sampling methodology that defines the class. The method adjusts to evolving data.

Rank order methodology of this kind is widely used in science (Krabbe et al. 2007) and market research (Law 2016), and is the usual basis for quantifying psychometric scores (Urbina 2011). It is also used in internet search engines and has been suggested as a basis of human memory and information retrieval (Steyvers and Griffiths 2008).

There is ample intuitive evidence of the epistemic and cultural significance attached to rank order and sampling. It is seen in the popular interest in pop charts, sporting records, and 'lists' of all kinds. There is also ample evidence of the dependence of qualitative judgements on sampling. The answer to the question 'How large is this, as an elephant?' versus 'How large is this, as a mouse?' depends ultimately on sample results not on any a priori definition of what is large. What is admitted as a valid sample of, for example, millionaires, or rock-climbing routes, or cases of sexual assault, is likely to change over time. Hence, since scaling depends on what is currently admitted, assigned magnitudes are likely to change over time. Alternatively, sampling rules may be adjusted to maintain scale values.

There are several reasons to think that after allowing for preferential sampling of extreme or unusual cases intuitive magnitude is generally linear with rank. It is the simplest fully generalizable non-parametric method available. It can explain, on the basis of sampling by availability, a commonly observed effect, namely that people discriminate more finely among familiar than among unfamiliar cases – for example, in judging the homogeneity of in-group versus out-group members (e.g. Quattrone and Jones 1980, Park and Rothbart 1982). And it can explain in cases where a scale of causal magnitude exists the typically curvilinear relationship between intuitive magnitude and causal magnitude, as follows.

For reasons related to the ubiquitous combinatorial interaction of independent causal factors, two types of frequency distribution of causal magnitude are commonly observed. They are the two-tailed normal or Gaussian distribution (Patel and Read 1996) and the one-tailed power law or Zipf-Mandelbrot distribution (Li 2002). For example, the height of adult humans has a normal distribution whereas the population size of cities has a power-law distribution. For any such distribution, the mapping from causal magnitude to linear rank has the same form as the corresponding cumulative frequency distribution. Typical cumulative frequency distributions are illustrated in Figures 9 and 10. These are at least approximately similar to the distributions represented in Figures 5 and 6. Discrepancies can be accounted for or accommodated in various ways.

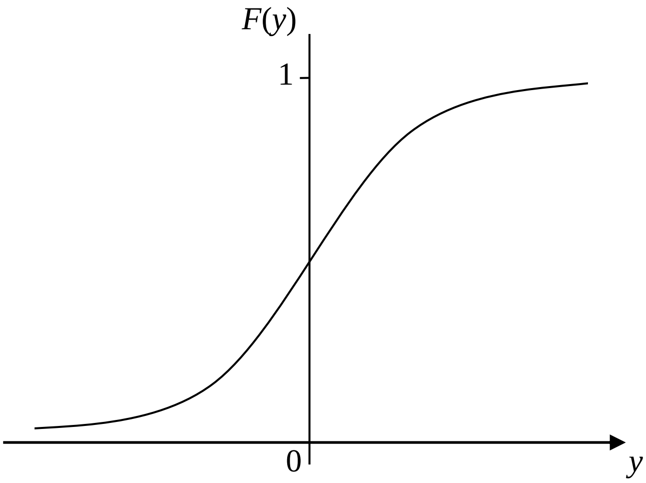

Figure 9: Cumulative frequency of a Gaussian distribution

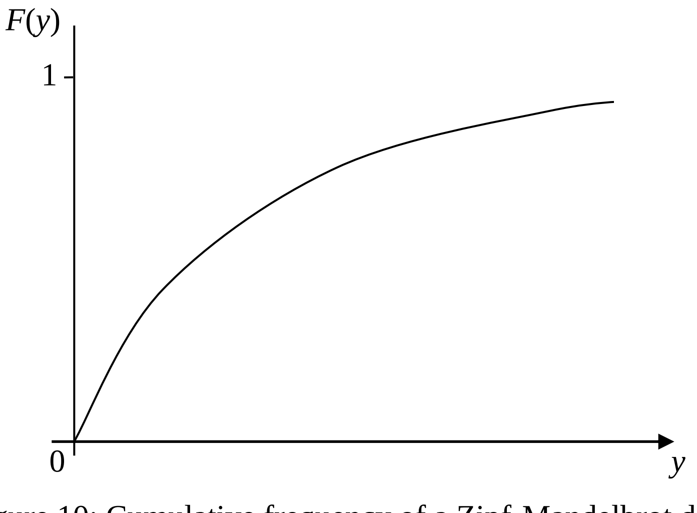

Figure 10: Cumulative frequency of a Zipf-Mandelbrot distribution

Ultimately, the form of the relationship between causal magnitude and motivational value via intuitive magnitude is an empirical issue subject to the resolution of various associated theoretical and methodological questions. But it may be noted that the two-tailed data reported in Abdellaoui et al. (2007) and reproduced in Dhami (2016: 143-4, Figures 2.8 and 2.9), although clearly asymmetric between gains and losses, appear to show a somewhat more nearly linear relationship near zero than is shown in Figure 5. The issue will be discussed further in §4.7.

Regardless of how any value scale is derived, evidence of intuitive valuation suggests that each must have a functionally adjusted zero value and unit range. The zero value may be set either a priori or dynamically by reference to a value neutral or current average reference condition. Varying experience may involve scale adjustment to maintain unit range. Within these constraints, generic differences in motivational weight reflect differences in affinity and salience. For example, physical injury has a natural zero and, usually, a larger salience than, for example, tardiness. The latter reflects the fact that, for humans, moderate personal physical injury is typically associated with many more value-negative outcomes than moderate tardiness. Being late may annoy me, but cutting my finger with a sharp knife easily overrides this concern. How this overriding is effected is discussed in §7.5.

4.6 Double Counting

Useful evidence concerning what conditions are valued is provided by introspective reports of respondents when presented with various actual or imagined scenes. Common experience, however, suggests that it is not entirely reliable. A person may, for example, be happy or annoyed yet not know what exactly has prompted their happiness or annoyance. Significant objects and qualities may be involved in a scene indirectly, by nomological or statistical association rather than by being immediately present, and observation and valuation may be at least partly subliminal. Nevertheless, if the usual theory of decision making is to be vindicated it must be the case that for each agent on each occasion there is, over a class of relevant scenes, a set of conceivably realized conditions to which value is attached. On this basis, to evaluate one scene relative to another is to identify a subset of such conditions, namely ones that are apparently differentially realized in the pair, and to compute a total value difference accordingly.

We know that, in general, scenes can be described in many different ways. The analytical problem is, then, to partition the set of possible partial descriptions into valued and non-valued members such that, in every relevant case, the total computed value difference between two comparable scenes given by the best justified set of valued partial descriptions properly reflects their relative desirability. This raises a general problem of statistical separability.

To clarify what is at stake, consider the following. Suppose that I like having house plants at home. Presumably I value their being alive. But do I also value their being green? Normally, for house plants, being green is a sufficient condition, and an effective proxy, for being alive, so if I value both my valuation is subject to double counting. But I may also like deciduous plants, or artificial plants. Hence neither being alive nor being green can be generally discounted as a relevantly valuable condition. The problem is that usually-correlated conditions, including proxies, may come apart. The entertainment industry depends very largely on the value commonly attached to disconnected proxies.

A general issue of separability is widely discussed, especially by Broome (1991a, 2004). He introduces it, not as a problem but as an assumption – namely that different varieties of good "can be evaluated independently from" each other (2004: 43). But this fails to distinguish conceptual separability from statistical separability. Obviously, in house plants, being green is conceptually distinct from being alive, but the two are strongly correlated and the former often serves as a proxy for the latter. For epistemic reasons valuation must often apply to proxies, and if proxies are partially correlated this creates a problem of inconsistent valuation and, in particular, of double counting.

A particularly interesting case is of pain versus physical injury. What makes this unusual is that pain is directly observed only by the injured party. Hence if value is attached only to pain, injuries to others are insignificant. Since there are good reasons for being concerned about injuries to others – at least to close family members and to sustain reciprocity – this is unsatisfactory. Conversely, if value is attached only to overt injury the information supplied to an injured party by the sensation of pain is ignored. This is especially problematic if there are no other currently observable signs of injury. Alternatively, if value is attached not to pain itself but to overt effects caused by pain – crying out, etc. – this has the bizarre consequence that an injured party judges the severity of their own injuries by observing their own behaviour. It is reminiscent of Williams's comment, apropos Fried's discussion of whether a man should rescue his wife, that such a thought is "one thought too many" (1981: 18).

A plausible explanation for the occurrence and persistence of double counting is that the primary adaptive function of the value system is to enable a rapid response to immediate threats and opportunities. Generally speaking, the true significance of an apparent threat or opportunity, even if this notion is well defined, is not an observable. It must be estimated from other partially correlated observables that serve as proxies. Since a rapid response is required, over-reporting of proxies is adaptively preferable to under-reporting. This often results in double counting.

In the extreme it may lead to evaluative overload and pragmatic paralysis. This is a well known engineering problem. For example, in the HSE report on the explosion and fire at the Milford Haven Texaco refinery in 1994 it is recorded that for several minutes at the height of the crisis alarms were sounding in the control room at a rate of one every two or three seconds – 275 in 10.7 minutes (HSE 1997). The two controllers on duty were completely overwhelmed. In ordinary human activity the equivalent response is panic.

Hence despite its short-term adaptive value, some mechanism for suppressing double counting, or its cumulative effect, is to be expected. One possibility is that, by some evolved mechanism, the value system compensates for correlated data. The most obvious possibility involves a kind of factor analysis of historic data that generates a canonical set of approximately orthogonal intuitive indicators to which value is exclusively attached. This can work, especially to suppress double counting of strongly correlated factors, provided that the pattern of correlation in the environment remains stable. But the technical challenges of producing a fully orthogonal system, especially in response to relatively recent data, are formidable. Statisticians struggle with this problem in the real world. It is most likely to work best in the control of very long-evolved patterns of response such as eating and locomotion. Even then, in atypical conditions it may produce dysfunctional consequences such as obesity or seasickness.

A second possibility is that suppression of duplicated responses to correlated data is done dynamically, on-the-fly. If pain and physical injury are known to be correlated, when both occur together the response to one can suppress the other. For example, observation of physical injury may suppress sensitivity to pain and so preferentially motivate action to remedy injury rather than pain. Resetting reference conditions may achieve a similar effect. However, both are likely to take a significant time to kick in.

A third possibility is the emergence of a system of deliberation and action based on a set of rationally, socially, or culturally developed and inculcated values that, in the event, suppress the effectiveness of and/or supersede others already extant. One can take painkillers, visit the doctor, and do what the doctor orders. How this can work within the assumed motivational system. The extent to which it can ultimately reduce or eliminate double counting, will be discussed more fully in Chapter 7.

Double counting is of particular relevance in the interpretation of risk aversion. The standard analytical response to observed risk aversion is to construct a single asymmetric utility function, as illustrated in Figure 5. This assumes implicitly that, throughout the relevant class of cases, the resultant valuation *v*(*y*) depends on a single quality instantiated in each current choice situation, standardly quantified as *y*. But if statistical separability cannot be assumed this assumption is unjustified. Rather, it is possible that several imperfectly correlated and independently valued qualities are involved, variously instantiated within the assumed class of cases. A plausible second candidate is the binary quality 'win/lose'. Its independent relevance is apparent in many other choice situations such as sporting tournaments, simple games such as noughts and crosses in which no other quantification exists, and in situations in which even a positive outcome may constitute a loss, such as a qualified election or plebiscite.

Admitting a second factor and double counting allows an alternative account both for the general asymmetry of gains and losses and, more particularly, for the disproportionate sensitivity of agents to small gains and losses. The latter is reflected in the extreme gradient near zero of the power-form utility function illustrated in Figure 5. Thus, for example, buying a can of beans that has been accidentally priced 1p cheaper than those nearby may feel like a win despite the trivial monetary saving. This restores the possibility that the weight attributable to conventional scalar quantification, which is otherwise the usually dominant factor, is more approximately linear near zero. The latter appears to be supported by data in Abdellaoui et al. (2007) and reproduced in Dhami (2016: 143 Figure 2.8), which appears to show a less extreme gradient at the origin, as mentioned in §4.5. Disaggregation of relevant data might clarify the issue.

Further evidence of the variability of value-relevant factors in decision making and the weak justification of the assumption, common in economics, that all can be interpreted as having a measurable equivalent cash value is provided by Ebert (2010). It describes the development of motivation-based rather than money-based measures of value – measures correlated with expended effort rather than cash value. It reports that subjects appear to use different valuation strategies in the two cases. In general, the effective value attributed to a future reward was found to be more reduced using a motivation-based rather than a money-based measure.

4.7 Functionality

Valuation is not an abstract logical process. It is an embodied process that takes time, consumes resources, and is inherently open-ended in the sense that in the evaluation of an envisaged future there is no obvious limit to the set of possibly relevant conditions that may contribute non-trivially to its overall desirability. This is a version of the frame problem (Pylyshyn 1987) exacerbated by the openness of the future. Hence two key issues arise: what to consider and when to discontinue exploration – attention and stopping.

It is usually assumed that attention is an effect associated with agential prediction of expected effects. This is a development of ideas first proposed by Helmholtz (1878). The assumption is that an agent's epistemic relation to the environment is mediated by a process involving the continual generation, from a continually updated model of the world, of predictions of effects variously located in the immediate and/or more distantly envisaged future. In one modern version, usually referred to as a theory of predictive processing, it is assumed that the prediction of immediate effects extends all the way down to the level of real-time peripheral sense data. Predicted and actual data are compared dynamically and reports of discrepancies are fed back, modifying the current sensory-motor configuration and, if necessary, the generative model of the world, so as to minimize ongoing discrepancies. Hence it is often referred to as the theory of prediction error minimization (Frith 2007, Friston 2010, Hohwy 2013, 2016, Clark 2013, 2016, Kiefer and Hohwy 2018). On this basis, attention can be characterized in terms of the specificity and continuity of ongoing prediction. For example, the action of catching a ball involves the continual prediction of a variety of sensory data of which the visual size and apparent location of the ball relative to the hands within the agent's current field of view is, ceteris paribus, the dominant part. This, intuitively, constitutes the focus of attention. The theoretical question is what determines this focus.

Supporters of the standard version of predictive processing commonly claim that the key factor is the predicted prediction error rate, or precision. Clark, for example, claims that attention is "*nothing other than* the process of optimizing the precision (inverse variance) of critical prediction error signals" (2017: 115, original emphasis). Others express doubt (Ransom and Fazelpour 2015, Ransom et al. 2017). A key problem is that if the predicted error rate is assumed to reflect actual unpredictability, the implication in the ball-catching case is that the relative location of the ball is the most unpredictable feature of the scene. This seems unlikely. There must be many other things that are at least equally unpredictable, such as passing cars, birds, the movements of other players, the play of sunshine and wind, etc. But if the significance of a predicted error rate reflects what is attended to the account is at least incomplete if not circular.

A more plausible proposal is that, in addition to unpredictability, there is value currently attached to the ball being caught. In short, agents attend preferentially to predicted but uncertain conditions which, if realized, entail significant marginal evaluative consequences. If, for example, in the ball-catching case the movement of another player is such as to appear to entail a risk of collision and hence to have potentially significant marginal evaluative consequences, it becomes a focus of attention. By implication, human attention, and indeed our epistemic relation to the world generally, is significantly constrained by the content of the value system. We attend mainly to what we care about. This is a major part of the standard human response to the frame problem. By implication, attention depends as much on valuation of possible medium and long-term consequences as on immediately predicted perceptual effects.

There must, however, be at least one exception. Events may occur that are entirely unpredicted but nevertheless value-relevant. If I trip over unexpectedly, the tripping motion becomes a focus of attention despite being unpredicted. It generates a mass of problematic signals, but not ones that are initially associated with any uniquely relevant prediction. Presumably a generic attention-redirecting response is triggered. Such a response may be triggered not only by external conditions but also by various internal conditions – conditions that are detectable, have attached value, and are routinely monitored and evaluated. They include conditions relating to various forms of bodily damage and malfunction and conditions relating to the current state of cognitive processing. The latter include sensory deprivation, cognitive overload, inferential or memory failure, repetitive processing, and a low current rate of prediction error – commonly recognized as boredom – and their inverses. On this basis, exploratory activity without any immediate pragmatic justification can be accounted for as standardly motivated by, inter alia, the amelioration of boredom. As with pain, medication may interfere with the normal operation of these effects.

As mentioned, open-ended valuation creates a stopping problem. In order to satisfy a principle of optimization, valuation should halt only when there is no possibility that its continuation will result in the currently preferred choice ceasing to be preferred. But this condition cannot generally be evaluated directly, on pain of infinite regress. Hence the observation that stopping is ubiquitous can be accounted for within expected utility theory only by assuming the operation of a second-order procedure – one that treats valuation non-recursively as an evaluated activity.

There is both intuitive and experimental evidence for such a procedure (e.g. Thura et al. 2012, Hawkins et al. 2015) operating as follows. Delay in reaching a conclusion is a valued condition that increases negatively with time. It is felt intuitively as urgency. Stopping discharges urgency. Data from prior processing supports the prediction, on a case-by-case basis, of probable pending value adjustment associated with further processing. Typically, then, when the expected benefit implied by the discharge of urgency exceeds the expected benefit of further processing, stopping becomes the preferred option. This mechanism achieves an effect that

Simon (1956) characterizes as satisficing, namely that evaluation is halted when the expected outcome satisfies a current aspiration level, but it does so without the usually problematic requirement of a fixed aspiration level. Rather, the aspiration level is set dynamically (Simon 1982: 417). As Simon presciently observes, this arrangement makes it “difficult to draw a formal distinction between optimizing and satisficing procedures” (1982: 418). It is important to emphasize that without the assumption of a stopping procedure the expected utility principle entails no real-time consequences. The consequences typically implied in standard modelling depend crucially on the temporal or inferential boundaries on processing that the researcher chooses to assume or impose. This is seldom explicit.

The third way in which functionality constrains the value system is that valuation has a characteristic feeling. In view of this, a distinction is often made between the visceral and the rational (Loewenstein 1996). Visceral factors are usually understood to have a direct hedonic basis, to be inflexibly correlated with local objective conditions, to vary rapidly in response to such conditions, and to have a unique neurological basis (1996: 273). A prevalent view is that they typically defeat rational deliberation in a way that generates negative outcomes despite the latter being anticipated. Models have been devised to account for this apparently paradoxical effect, such as the planner-doer model (Thaler and Shefrin 1981). They typically assume that, in effect, an agent has several ‘selves’ operating with different priorities that may set constraints for each other, much as a group of interacting individuals might do.

This presents a dilemma. If the theory admits a principle for uniquely aggregating the effects of the various selves it is unclear on what basis the selves are identified as distinct. This makes the empirical justification of any particular model problematical (Ambrus and Rozen 2015). But if it admits no such principle it will not yield coherent predictions. Neither option is attractive.

A further objection is that the theory is ill-motivated. The key observation justifying its development – namely that humans often do things despite being aware of negative consequences – is plainly not sufficient to vitiate a unitary form of decision theory, for the following reason. It is standardly admitted that choices may, and often do, involve both positively and negatively valued features. For example, people pay for goods despite, presumably, having a preference for keeping the cash. In order to justify the claim that the expectation of future loss ought to subvert a current preference it is necessary to show not merely that the loss is anticipated but that the associated gain is of insufficient compensating value. Evidence that an agent is aware of a prospective loss and experiences anticipatory regret does not demonstrate this. It merely shows that the prospect of loss is recognized as such. For example, that I choose to sell my favourite painting in order to pay for dental treatment knowing that I will miss the painting terribly is no evidence of inconsistency. Similarly, choosing to drink or smoke despite knowing that it will make me ill is not necessarily evidence of irrationality (Rachlin 2018). Even suicide is not unjustifiable. Hence the observation that expected losses are recognized, regretted, but not avoided, is no evidence that a unitary theory is false.

It is likely that most if not all significant valuation of perceived or anticipated conditions involves some characteristically associated feeling. Sometimes the feeling is very intense, presumably reflecting its evaluative magnitude. Being a detectable condition, it may itself be valued. Hence it is a possible source of double counting. One may fear fear or pleasurably anticipate pleasure. But this does not itself imply that a separate viscerally-based form of explanation is needed. Rather, it is more plausible that visceral effects are a characteristic feature of a particular type of unitary system that admits double counting. Loewenstein’s (2010) “soul-searching” admission that global problems may receive insufficient rational attention because they fail to induce a sufficiently strong emotional response implicitly recognizes this possibility. In short, emotion, even strong emotion, can be interpreted as a normal feature of a unitary value system, often indicative of urgency, not the inconsistent residue of another more primitive system.

4.8 Value Change

Evidently, values are not immutably fixed. Levels of affinity and salience may change, presumably in response to intervening experience. In order to admit this obvious possibility I shall, as in §4.4, do no more than sketch a minimally plausible account. The aim is not at present to develop a full-blown theory but merely to fill an analytical gap that current modelling generally ignores, namely the variability of individual valuation through time. In doing so I shall follow the usual scientific practice of assuming that, whilst explanatory variables such as affinity and salience are not permanently fixed, they change only via some relevant, usually incremental, process. Any coherent psychological theory of utility as conventionally quantified, if one were available, would require a similar extension.

A plausible theory of value change – that I will outline here merely to indicate what is possible – is that assigned levels of affinity and salience change incrementally, ultimately from innately determined levels, more or less as assumed in standard reinforcement learning theory but without any extraneous notion of reward or punishment. Effects normally attributed to extraneous reward or punishment are produced by perceived events that generate in the agent an abrupt variation in current net valuation, as standardly computed. Their effect is to modify the levels of affinity attributed to objects apparently involved as immediately relevant precursors of those events and of salience attributed to transient qualities exhibited by those objects, in such a way as to transmit value from currently valued conditions to statistically associated conditions, the direction of adjustment being such as would amplify the driving variation. Hence value, both positive and negative, diffuses through the system by statistical association. The adaptive effect of this is, in general, to improve anticipation of valued conditions.

For example, a child frightened by a proximate barking dog – that is, by an event that involves a sudden increase in the perceived self-exhibited negatively valued quality of felt alarm in the presence of a barking dog – attaches increased negative affinity to that dog and/or to dogs in general and increased negative salience to heard barking. Subsequently, both proximate dogs and heard barking are negatively, or more negatively valued. Since barking is a threatening quality, these are computed additively. If, in the then-current context, the child experiences an immediate positively valued interaction such as maternal reassurance, this will usually reduce the value-changing effect, provided that the counteracting effect is sufficiently large. If not, other concurrent conditions associated with barking dogs, such as a characteristic location, may in due course also become more negatively valued – and so on. The size of any such effect is likely to be subject to parametric variation among individuals and by age, prior experience, etc.

There seems to be no good reason to believe that other values, including those entailing economic preferences, are not similarly shaped. It is clear, for example, that consumer brand preferences are strongly dependent on imagery and learned association. In particular, the instrumental association of available resources, including money, with the achievement of positively valued effects, results in the attachment of value to the availability of those resources. On this basis, having money is valued because it is experientially associated with valued effects, not intrinsically.

4.9 Conclusion

The aim of this chapter has been to investigate, in the light of both intuitive evidence and the available record of prior theoretical conjecture, inference, and research, the main aetiological and structural characteristics of human valuation of envisaged conditions. Its most important conclusion is that valuation is both structurally and functionally complicated. The task, as it were, of assigning value to detectable or imaginable features of the world in a way that generally distinguishes prospective conditions that are relatively conducive to adaptive success, broadly construed, or its converse, is not straightforward. This conclusion is considerably at variance with the usual assumption, apparently implicit in much behavioural research, that actually evolved preferences must satisfy a simple functional relationship defined over superficial features of alternatives as standardly described.

Once this conclusion is allowed, effective research can continue only if a broader range of intuitive evidence is admitted. Since intuitive evidence is both necessary and unreliable, a selective methodology is needed. For this reason I have relied on the methodology described in §1.3, namely that, prima facie, an intuition should be treated as veridical unless its occurrence can be explained by other justified assumptions. This allows many common assumptions about what we value, both personally and collectively, to be taken at face value without irresolvable inconsistency.

To be a target, a cognitively represented condition must be both inferable and, in some way, valuable. The former requires something that answers to a causal model of the world and the latter to a grounding, ultimately, in evolutionary adaptation extended by statistical association. Temporal discrimination typically has a nested granular structure based on uniform quantification in a way that distinguishes the relative past, present, and future. Identified conditions may be diachronically complex.

Intuitive evidence speaks strongly in favour of there being targets that can analysed into object-quality pairings, evaluated combinatorially, in which objects and qualities are assigned values independently by type. This creates a system in which hedonic value is distinctively one-dimensional since, for any agent, it varies only with quality. Admitting object-quality pairings by type accounts for the productivity and systematicity of valuation. Consistency requires a partial order over valued types.

Conditions generally admit valuation distinguished in terms of both valence and scale, and differences as gains or losses. Such scales can be constructed by sample ranking relative to relevant value-neutral reference items. Given standard objective distributions, this typically generates a set of dynamically normalized approximately logarithmic value scales. Asymmetric sampling of gains and losses generates asymmetric scaling.

It is an irresolvable problem for uniform valuation that significant conditions are often partially correlated. If orthogonal factors cannot be extracted, this poses a dilemma between double counting and omitting some significant cases. Mitigated double counting appears ubiquitous. Double counting both gains and wins, and their inverses, can account for the extreme sensitivity to small gains and losses commonly observed.

The human value system rests fundamentally on constraints imposed by biological adaptation but also on cultural and cognitive evolution. These operate on different principles and at different rates. Their combination undermines any purely biological account of valuation. Valuation is subject to two underdetermined constraints, attention and stopping. Both require some form of second-order feedback. Visceral influences can be accounted for as a form of double counting. Value assignments are not generally fixed but may vary in response to experience, presumably so as to enhance anticipation of significant effects.

The main limitation of this account is that it fails to distinguish clearly between preferring and choosing. The distinction is explicit in the Davidson quote on page 9 in which he refers to both “the preference that one state of affairs obtain rather than another” and “the choice of one course of action over another”. These are not equivalent. One may prefer *A* over *B* but be prevented from choosing *A* because *A* is not an option. Intuitively, it is not ‘available’. What, then, is an option? It is to this question that I now turn.

Chapter 5 Options

5.1 Achievability

There is a clear intuitive distinction between preference and choice. Preference is an attitude whereas choice is an event. Choice initiates a process that may or may not satisfy a preference. An agent cannot normally choose what to prefer, although choices may indirectly modify preferences. An agent may choose to express a preference or to act to satisfy a preference or to do neither. Preference is defined over a set of targets, as described in Chapter 4. Choice is defined over a set of options that may or may not be targets. The aim of this chapter is to investigate the relationship between targets and options and, in particular, to examine how the characteristic effect of choosing an option is to modify the expected probability that a target will be realized.

A necessary condition for something to be admitted as an option is that it is assumed to be, in some relevant sense, achievable. If an envisaged effect is not assumed to be achievable then no process of choice can reasonably be assumed to lead to its realization and hence it is not, in the relevant sense, an option. This excludes not only impossible effects but also effects the assumed probability of which does not depend significantly on the agent's choice. For example, buying a lottery ticket is in many circumstances an option since, unless other constraints exclude it, it is an effect that can be achieved by a known and usually feasible course of action. But winning the jackpot, even if possible and desirable, is not generally an option insofar as it is not assumed to be reasonably achievable by any currently available means. The probability of winning contingent on buying a ticket, as implied in the agent's current causal model of the world, is not sufficient to transfer optionality from the latter to the former.

However, the notion of achievability raises a significant problem. It can be illustrated by describing three typical scenarios in which options are defined. The first (S1) is as follows. An agent thinks about their situation, envisages something they might do or not do and evaluates the likely consequences. Many decisions relating to social or leisure activity and many conventional or moral decisions are of this type. The second scenario (S2) is one in which an agent is in a situation in which two or more alterative possible effects are described and the agent expresses or is asked to express a preference between them, the implication being that the preferred effect will be realized by some prearranged or confidently expected means. Most experiments described in the Appendix are of this type. The third scenario (S3) is one in which an agent thinks about their situation, envisages something that might be better or worse, and considers a possible action or course of action to adjust what is likely to occur. Most ordinary practical, economic, and epistemic decisions are of this type.

The problem is as follows. In S1, what is initially envisaged as an option, and hence what is assumed to be achievable, is standardly classified as a possible action. It is achievable, presumably, in virtue of its being within the agent's current competence. In S2, what is initially envisaged as an option and hence what is assumed to be achievable is standardly classified as a possible condition. It is achievable, presumably, in virtue of the implicit guarantee embodied in the current context that its realization is an assured consequence of the otherwise trivial act of expressing a preference. The latter is assumed to be performable in virtue of the agent's current linguistic or other indicative competence. But in S3, although what is initially envisaged as an option is also classified as a possible condition, it is a condition for which there is typically no implicitly prearranged course of action guaranteeing its realization in the current context. Rather, a suitable action or course of action must, if possible, be constructed in situ, in a context-dependent developmental process, from component actions that are, in the event, only contingently performable. The question is, then, on what basis is the initially envisaged condition classified as presumably achievable. This raises a difficult issue concerning the causal relationship between agency and outcome that I have only partly addressed in Chapter 3. It is especially difficult in cases of collective action.

Suppose, for example, I am in an unfamiliar place at about lunchtime and I think I would like to get a sandwich. Introspective evidence speaks strongly in favour of the assumption that I can admit getting a sandwich as an option without having any fully worked-out plan in place for how to achieve this effect. It is sufficient that I am confident that, by some series of presently ill-defined choices made in response to discovered constraints and opportunities, I can in due course realize a course of action to achieve the desired effect.

This quasi-teleological structure is difficult to accommodate within a scheme of analysis in which it is assumed that the intentional structure of behaviour consists in the successive realization of a set of predetermined contingent relations connecting component acts and associated effects, as is usual in associationist, algorithmic, and causal accounts (e.g. Hull 1943, Miller et al. 1960, Dretske 1988). For even if the condition that constitutes my getting a sandwich is, in fact, programmatically or statistically linked to some prior or currently known set of actions, the idea that my getting a sandwich, as a possible option, is confined to or constituted of this set is not plausible. Among other objections, it does not adequately accommodate either the complex combinatorial structure or the discoverability or learnability of alternative or novel methods.

An account that better accords with available introspective evidence is as follows. Action planning is an inherently underdetermined recursive process involving the construction, in context, starting from an originally envisaged condition, of a system of interconnected methods, usually via a process of trial-and-error selection and assembly. Given, for example, the initial thought of getting a sandwich, I can readily envisage at least one and probably several possibly relevant methods each of which, if appropriately realized, has getting a sandwich as a possible consequence. Each method implicitly specifies a presumably performable course of action together with a set of related prerequisite and consequent conditions grounded in a relevant causal model of the world. It carries the implication that, ceteris paribus, if the prerequisite conditions are satisfied and the action is performed then the consequent conditions will, probably, be realized. On this basis, the claim that an envisaged condition is presumably achievable in a given context is vindicated, recursively, if a method exists in which, in that context, the specified action is routinely performable, the consequent conditions include the envisaged condition, and every prerequisite condition is either already satisfied or presumably achievable.

For example, one method that might typically be associated with getting a sandwich involves the action of buying a sandwich in a shop. This method involves a set of prerequisite conditions including, typically, that there is a shop selling sandwiches, that I am in the shop, and that I have sufficient money. It also involves a set of consequent conditions, including that I am still in the shop, that I have less money, and that I get a sandwich. Hence assuming that there is such a shop nearby and I have sufficient money, getting a sandwich is presumably achievable if buying a sandwich is, in the assumed context, a routinely performable action and being in the shop is either already realized or presumably achievable. If being in the shop is not already realized, the planning process sets this as a newly envisaged condition and begins the task of finding a method that has it as a probable consequence, and so on, until either the initial assumption that getting a sandwich is achievable is vindicated or the search is abandoned.

What is notable about this process is that it makes it possible to embark on a course of action without first vindicating the claim that each relevantly envisaged condition is presumably achievable. In this sense it is not algorithmic. For example, I can justifiably set out to look for a shop without knowing whether I will find one or, if I do, whether it will be open and sell sandwiches, and I can justifiably act to ensure that I have enough money before setting out. If one method is or appears to be failing I can switch to another. Moreover, an envisaged consequent condition may be arbitrarily remote. A person might envisage, for example, building a settlement on Mars or making an effective invisibility cloak. In such a case the trial-and-error process of planning and execution may be of unpredictable scale and complexity, notwithstanding any initially justified confidence that ‘it must be possible’. This account vindicates the intuitive notion of trying. The varying intuitive status of envisaged consequent conditions in action is partly characterized by Bratman in his distinction between intentions and settled objectives (2009a: 19).

There is considerable evidence that recursive trial-and-error processes are significantly more effective in solving complex problems than was once assumed. It is seen especially in current developments in AI, particularly in the success of the techniques of back propagation in neural networks and deep reinforcement learning (Churchland 2000, Lewandowsky and Farrell 2011: 293-8, Hafner et al. 2020, Hauptmann and Adler 2020). The emergence of this type of evolutionary processing in AI is reminiscent of the emergence of Darwinism in evolutionary biology. It involves a mechanism characterized by underdetermined internal variation and selective feedback depending on actual outcomes rather than the predictable working out of a pre-existing design. It does not standardly involve the emergence of intuitively intelligible intermediate propositional contents such as appear to characterize human action planning but no fundamental principle excludes it.

Returning to the three scenarios described on page 134, the question of what is an option now appears rather more nuanced. Firstly, whilst the distinction between a condition and an action is explicit in the notion of a method it is less clear in the notion of an option. Conceptually, the realization of an action can be characterized as a condition satisfied – namely the condition that the action is performed. Hence choosing an action in S1 can also be interpreted as choosing a condition. But choosing a method is not the same as choosing an action since the same action with different preconditions and/or consequences may be involved in several different methods. On this basis it is possible to act in a way that advances several methods simultaneously without immediately choosing between them. Secondly, it seems that the definiteness of options in S2 is largely an artefact of the assumed context. It relies on the agent trusting the implicit guarantees and failing to consider other alternatives. Both the existence of anomalous experimental results and introspection suggest that this cannot be relied on. And thirdly, what initially appears in S3 to be a choice among envisaged outcomes emerges as a nested series of choices of epistemically and pragmatically directed actions bound together by the search, in situ, for a feasible method to achieve a particular outcome. In this, the status of an envisaged outcome as an option is not formally established until the assumption of its achievability is vindicated, which may not be much before it is in fact achieved. It follows that what is achievable and hence what is an option is, in anticipation, often significantly conjectural. Furthermore, an option may be abandoned before or after its achievability is established, and a subordinate option may continue to be pursued after its original justification has been abandoned. For example, I may abandon trying to get a sandwich because I meet a friend who offers to buy me lunch, or I may decide on entering the shop to get a pie instead. Such a system of nested options may occur in any of the three scenarios described.

This analysis of achievability closely parallels the analysis of what Schwarz calls the Ability condition:

> Ability. A proposition *A* is an option (for an agent in a given choice situation) only if the agent can make *A* true (in the sense that there is an available decision that would render *A* true) (2021: 170),

but with three important differences. The first is that Schwarz categorizes the Ability condition as a property of an agent rather than as a property of an envisaged outcome relative to a method. This parallels the contrast between a function and a problem discussed in §3.3. Prima facie, it excludes collective or vicarious action. Secondly, as remarked on page 34, there are very few non-trivial outcomes that can be made true, outright, by a single non-proximate decision. As the saying goes, ‘There’s many a slip …’. And thirdly, Schwarz does not allow that the target condition can be recursively vindicated. This is in line with his general omission of dynamics. Some of the difficulties of his analysis which he recognizes, such as the marksman problem, flow from this omission.

These considerations raise a number of issues that will be discussed more fully in the following sections. The derivation and evaluation of alternatives will be discussed in §5.2. Fluency, feasibility, and competence will be discussed in §5.3. The status of epistemic options will be discussed in §5.4, commitment in §5.5, and effects related to social interaction in §5.6. More detailed discussion of probability will be deferred to chapter 6.

5.2 Alternatives

In order to be chooseable an option must be not only presumably achievable but also one of a set of alternatives that are, in some way, similar but distinct and so can be compared in terms of their expected value. In this section I will investigate the structure and development of alternatives, principally in relation to an envisaged future. As before, it will be convenient to proceed by considering the three scenarios described on page 134. In the absence of any plausible explanatory alternative I will continue to assume that that our characteristic intuitive impressions of the decision making process provide, inter alia, admissible evidence of its constitution.

S1, as described, involves exactly two alternatives: an action to be performed and the same action left unperformed. This immediately raises a question as to the content of the latter since, prima facie, it consists of something absent. It is an issue that arises quite widely.

An initially attractive proposal is that if an envisaged action is absent then evaluation must apply to whatever is assumed to comprise the set of current conditions. This, however, cannot be correct. Change is ubiquitous. An envisaged action is often in addition to or in variation of a course of action already in train, not a uniquely disruptive intervention, and extant environmental processes continually generate partly predictable contextual fluctuations irrespective of any envisaged action. Hence evaluation cannot assume background invariability. What is required is a notion of a default future – of the future as predicted on the basis that currently ongoing and justifiably expected processes and actions continue without the addition of any significant modifying choices made by any relevant agent. A positive action envisaged in S1 is presumably a defeater of this condition – one that supports the prediction of an alternative future. Its evaluation, in context, is implicitly of the difference between this and the current default future.

The notion of an alternative future raises a significant issue. On any occasion, any conceivable future is in principle of unlimited potential complexity. But there are several reasons, including arguments from evolutionary adaptation, computational tractability, and introspective, observational and experimental evidence, to suppose that evaluative processing cannot fully admit this complexity. Some significant simplification must occur.

Several modes of simplification are available. One is that only differences between alternative futures need to be computed. This rules out as irrelevant all effects that are assumed to be causally unrelated to any choice among current options since they can be assumed to be unaffected by that choice. This enormously reduces the burden of investigating approximately concurrent but spatially remote effects but it leaves considerable uncertainty regarding an ever-expanding class of long-term effects, especially ones that depend significantly on other agents' contingent responses. The issue will be discussed further in §5.6.

A second mode of simplification rests on the principle discussed in §4.7 that only effects that are judged to be of non-zero value by the agent are standardly attended to. This again reduces the burden of investigating remote effects. But there remains a problem, in that evaluatively significant effects may occur disconnectedly, possibly in the remote future. Hence it fails, in itself, to provide a clear duration-based cut-off.

A third mode of simplification is implicit in the dyadic value theory described in §4.5. If an agent assigns significant affinity only to members of a relatively small class of currently known objects, such as friends, family members, and personal possessions, the problem of open-ended valuation will be considerably reduced by the fact that long-term effects may rarely involve any of these objects.

A fourth mode of simplification arises from the increasing uncertainty of the future. This renders potentially valued effects, especially choice-dependent differences, increasingly insignificant. The effect is likely to be amplified by the stopping procedure described in §4.7. The issue will be examined more fully in Chapter 6. Whether any additional principle of duration-based simplification, as is perhaps implicit in discounted utility theory, is either usual or rationally justified will be considered in Chapter 7.

S2 introduces three new issues regarding alternatives. One is that more than two alternatives may be available. The second is that ‘do nothing’ is usually excluded. The third is that the relevant consequences of each alternative are, or appear to be, specified explicitly within an assumed quasi-contractual environment rather than being consequentially inferred.

However, the significance of these differences is not entirely clear. Any multiple choice can be construed as a nested series of binary choices, either pairwise or to accept or reject each single option, under the constraint that the final choice of an option terminates the choosing process. Observed effects related to the order of presentation of alternatives as described in (A24) may provide evidence of distinctly sequential processing.

Furthermore, whatever alternatives are specified there is always the possibility of an agent rejecting the entire package. There is an implicit choice, as in S1, of whether to act in conformity with the relevant situational norms or to deviate. In many social situations attempts are made via criminal or contract law or the threat of informal sanctions to render deviant alternatives non-preferable. But as the frequency of shoplifting shows, these efforts are not always successful.

And, crucially, the usual specification of an option as a described effect typically leaves its evaluative significance underdescribed. The way in which the specification functions in creating a choice is that it implicitly identifies a possible action that serves, as in S1, as the basis for the agent's inferential construction of a corresponding partially envisaged future. Evaluation is implicitly of this future, not merely of the condition as described. For example, for an option specified in the form, "You will receive $100 next Friday," a relevant valuation typically involves not only the prospective receipt itself as an event but also the prospective conditions of awaiting receipt and of subsequently having and eventually spending the money, presumably on some desirable object or, perhaps, to repay a debt. The receipt of the money, as an event, may be relatively trivial. A bank transfer may hardly be noticed.

Insofar as several options are specified, several alternative futures will need to be constructed in parallel. As the number increases, pairwise evaluation of differences becomes increasingly complex, especially to the extent that consequences are projected into the remote future. Again, some method of simplification is required.

One such method arises as a side effect of the stopping process discussed in §4.7. Provided that the potential value increment associated with further processing of an alternative has an upper limit, the early discovery of a sufficiently negative outcome may halt evaluation even if the alternative is still presumably achievable. I will describe an alternative that can be eliminated in this way as non-viable. If all alternatives appear to be non-viable processing must restart with a reduced threshold. Ultimately, an alternative having zero expected benefit may be the best available.

A problem of multiple alternatives also occurs where several logically independent choices arise in parallel. Consider, for example, choosing where to live and what job to apply for, approximately concurrently. In principle these choices may be logically independent. But intuitive evidence strongly suggests that they are usually considered combinatorially.

As the number of choices and/or alternatives increases this leads to an geometrically increasing set of possible combinations. Given the difficulty that this entails, the failure to treat such choices independently requires explanation.

A plausible explanation is that, for the purposes of choice, alternatives are distinguished by reference to valuation and valuation by reference to envisaged futures, and that alternative futures are, by assumption, mutually exclusive. It follows that if constituent conditions combine independently, each combination of conditions, and so each combination of associated options must be evaluated separately. The resulting combinatorial explosion, which Fumagalli (2020b) quoting Gilboa (2009: 116) calls the "explosion of cardinality", adds to the necessity of simplification, achieved principally by the rapid elimination of non-viable combinations. For example, a combination of options that entails a future in which one lives far from ones place of work might be quickly discarded.

S3 appears to be the most widely applicable of the described scenarios and is dynamically the most complex. It typically involves no stable set of currently envisaged alternatives. Alternatives are generated and eliminated successively during a planning process, the dynamical structure of which is approximately as follows. Its starts from the thought of an envisaged amelioration of a current or currently expected condition. Provided that this amelioration is admitted to be, in context, presumably achievable it creates an initial pair of alternatives – of the default future versus a non-specific amelioration. The problem is then to vindicate the presumed achievability of amelioration by a viable method. It involves a recursive search, as described in §5.1, for an interconnected series of methods grounded ultimately in satisfied conditions and routinely performable actions under a constraint defined in terms of the net valuation of required and expected effects over the implied future – often, presumably, subject to double counting. This tends to generate a depth-first search in which choice selects among remaining alternatives identified as both viable and presumably achievable. It may eventually eliminate all non-default alternatives.

Several effects may redirect or interrupt the analysis of alternatives. For example, an original ameliorative issue may be resolved in some other way or be displaced by another, or prevailing conditions may vary unexpectedly, disrupting inference. The longer the total duration of the search process, the more likely such an effect is to occur. Furthermore, at many points in the process there is likely to be a choice between choosing and delaying, with epistemic activity intervening. This is a version of the stopping problem described in §4.7. I will examine the issue further in §5.4. First, however, I will investigate the status of the units of performable action from which planned activity is constructed.

5.3 Fluency

It is implicit in the account proposed here, reflecting our intuitive understanding of decision making, that choices occur within, and serve to redirect, a system of action-production that is to a large extent procedurally autonomous (cf. Railton 2009). For example, a competent tennis player chooses, within limits, when, where, and how fast to serve the ball, but beyond this the serving action is executed fluently without intermediate choice-making. Indeed, if a player becomes explicitly aware of intervening features of their performance and attempts to vary the course of the action by choice during execution, its overall efficacy is likely to be considerably impaired. All human action – walking, driving, eating, speaking, manipulating objects, observing, calculating, planning, choosing – involves fluent components of this kind.

A key feature of fluent action is its apparent modularity – that it is composed of specific formative units each of which, unless disrupted, has narrowly predictable dynamics and is distinguished, typically, by a characteristic aetiology, neurophysiological instantiation – including dedicated or preferred input and output channels and data streams – prerequisite initial and boundary conditions, and limited parametric variation. Fluent episodes can run in parallel with minimal cognitive interference. Units thus typically exhibit what is standardly termed encapsulation (Pylyshyn 1999). Deliberative choice, which occurs between rather that within such episodes, generally has the effect of setting or modifying initial and boundary conditions and parameters so as to start, stop, interrupt, restart, or redirect fluent processing. It is not part of that processing. In particular, the internal dynamics of fluent action – other than the choosing process itself – are not value driven. Indeed, some trivial effects such as drumming ones fingers and repetitive actions as in OCD occur almost entirely independently of deliberative valuation.

Some types of fluent action are, presumably, biologically evolved phenotypes emerging predictably under relevant environmental conditions. Elementary language use is an obvious example. Many more, however, are learned by repetition of an initially deliberative pattern of activity – that is, by practice. During practice, fluency takes over from deliberation, presumably as relevant dynamical connections are established. It is inconceivable that people could play the piano, or write books, or drive, or play tennis as they do without this transitional capability. It frees up the relatively slow process of value-based decision making, allowing it to attend to higher level strategic problems and possibilities. It is especially in this respect that human action is significantly different from most animal behaviour and from the performance of most current AI systems.

The claim that fluent action is modular is quite distinct from the more frequently discussed claim of neural modularity – that, in general, cognitive processing occurs in distinctly modular neural units between which little concurrent information is exchanged, either locally or globally (Fodor 1983, Pinker 1997). The usual counterargument is that the latter is incompatible with evidence of both learned fluency and neuroplasticity (Carruthers 2006). The current claim that fluent human action is under the partial control of an overarching unitary value system adds weight to this counterargument.

Computationally, a fluent action unit is plausibly represented as a relatively self-contained algorithmic or partially algorithmic program unit of a kind that standardly occurs in many AI systems. The value system stands outside this – except in that valuation is itself a fluent process. It modulates processing, typically by adjusting operational parameters. The partial encapsulation of fluent processing that this entails may account for both the early attractiveness of GOFAI as a possible account of the structure of human action and its success at the level of routine performance but also for its failure at a more strategic level where dependence on an overarching value system predominates.

Considering once again the three scenarios described on page 134, the action referred to in S1 may often be assumed to be available as a fluent unit even if, when executed, its fluency may be disrupted by emerging events. This would account for it being conceptualized as a single option with typically predictable consequences. For example, the thought might be to make a cup of tea. This is a familiar action the entire course of which can be rehearsed introspectively with typical prerequisites and anticipated consequences, and hence one that can be chosen as a single option. The fact that it may not always proceed as envisaged – if, for example, there is no milk in the fridge – is for the moment ignored.

In S2, where options are standardly presented as possible conditions, the corresponding actions are at first sight almost trivially fluent, consisting only in the expressing of a preference either verbally of by some equivalent indicative behaviour. On consideration, however, rather more is involved, in that the effective expression of a preference is part of a complex socio-linguistic system of processing in which messages are encoded, transmitted, and decoded. The procedure in competent adults is paradigmatically fluent, having been learned in an innately grounded context-driven developmental process over which we, as agents, have almost no deliberate control. Only its inputs and outputs and some qualitative features of its performance are introspectively accessible and deliberatively adjustable.

In S3 the initial options are not generally defined in terms of fluent actions but in terms of presumably achievable consequences. The objective of the planning process is to vindicate this presumed achievability by envisaging a connected structure of methods, subject to apparent viability, in which all prerequisite conditions are either already satisfied or currently achievable and all actions are fluently performable given these preconditions. In this, fluently performable actions provide a conceptual bottom level at which planning can stop and at which, during execution, activity can be initiated and, by default, allowed to continue autonomously subject only to parametric modulation. Without such a bottom level, deliberative analysis in both planning and execution would be overwhelmed by the analysis of trivial tactical details to the exclusion of value-driven strategic objectives.

All or almost all fluent action includes an epistemic component via which practical intervention is adjusted to prevailing conditions. The two aspects, practical and epistemic, are intimately integrated, usually to such a degree that the intervening connections defy introspection. Thus a skilled pianist or juggler has no explicit awareness of how perception, including visual and proprioceptive perception, and bodily movement are mutually integrated during performance. Moreover, bodily movement is not exclusively a means of practical intervention. Some, especially eye movement, occurs primarily as an integral part of epistemic processing. Other types of fluent action, such as reading and calculating, are primarily epistemic or involve symbolic rather than practical manipulation. Even planning can be envisaged as constructed of fluent units.

I will not at present expand on the idea that the structure of fluent action is partially algorithmic. As mentioned previously, there are ample resources described in the computer science and AI literature and widely implemented, including classical programming, neural networks, predictive processing, and deep reinforcement learning, to justify the claim that insofar as fluent action is encapsulated it can generally be modelled in some such way. Accordingly, the most important outstanding analytical problem is not to model fluency but to integrate fluency into a value-driven strategic system in which both fluent and deliberative modes of action exist interdependently as parts of an evolved heuristic methodology.

5.4 Epistemic Options

Not only actions but also options may be predominantly epistemic. This possibility arises directly in S1, in which actions constitute options. For example, one may choose to read an article in a newspaper or on Wikipedia. This may, in context, be part of a distinct practical project, but it may be merely habitual or directed at the satisfaction of inquisitiveness. In S2 the available options may involve gaining access to various alternative bodies of information. Typical examples include the options offered on a telephone helpline and the links displayed on a webpage. But it is in S3 that the full complexity of epistemic activity is revealed.

Hitherto I have appeared to assume that during planning each prerequisite condition is either apparently true or apparently false and that, if false, the only relevant issue is whether it is presumably achievable and hence can be made true. But there is another possibility, namely that it is unknown or uncertain. Such a situation is often modelled under the rubric of a Bayesian epistemology. Since Bayesian analysis is ubiquitous I will briefly illustrate its usual methodology so as to contrast it with the more naturalistic scheme proposed here.

Bayesian epistemology standardly admits two ways in which processing can continue in response to an unknown or uncertain prerequisite condition. One is that the agent adopts an a priori degree of subjective probability on some presumptive basis and continues accordingly. The other is that the agent receives some probabilistically related but otherwise epistemically secure evidence and adopts a revised degree of subjective probability that satisfies a standard updating rule based on that evidence. There is a vast and mathematically sophisticated literature devoted to this issue (e.g. Ramsey 1931, Jeffrey 1965, Joyce 1999). It is of great analytical interest, but its relevance to actual human decision making is weakened by a lack of attention to wider epistemic options. Like almost everything described in the Appendix, most research implicitly adopts S2 as the only relevant scenario.

Joyce (2010) may serve as a characteristic example. Its aim is to show on rational grounds, by considering various cases, that degrees of subjective probability, or credence, must sometimes be imprecise. One case (2010: 286) is as follows. Two urns contain balls marked $1000, $500, and $0. The agent is told nothing about the proportions. A ball is drawn at random from the first urn but not revealed. The agent is offered a choice between receiving either the sum shown plus $500 or the sum shown plus the sum shown on a second ball drawn at random from the second urn. Which should they choose?

Before examining this further it should be noted that Joyce assumes without comment that the agent can and should confidently accept the information given – including that an expressed choice will be honoured – as, or as if, true. Since the fundamental motivation of the paper is that "belief is not all-or-nothing" (2010: 281) the justification for this is not obvious. Next, he implicitly assumes that the rationally preferable choice depends only on a single epistemic consideration, namely the subjective probability of each relevant monetary outcome. On this basis, although he does not describe it in quite these terms, he allows the agent just two epistemic options – to adopt a precise level of subjective probability that the sum shown on the second ball will, when drawn, be greater or less than $500, each to be represented by a real number in the range [0, 1], or to adopt an imprecise level of subjective probability that it will be greater or less than $500, each to be represented by a set of real numbers each in the range [0, 1], one for each possible sum shown on the first ball. The latter is justified by the agent's admitted, and separate, uncertainty concerning the correlation between the urns' contents. He then argues, based on the principle that an adopted level of subjective probability should not support any more precise conclusions than the data on which it is based, that the former is rationally untenable. This leaves only the latter.

Most of the rest of the paper is taken up with an attempt to solve various problems apparently generated by this notion of imprecise subjective probability. What is not investigated is whether the agent might have other options.

An obvious possibility is to seek further evidence, delaying the adoption of any specific level of subjective probability until more information is available. Joyce, in effect, stipulates that this is not an option. No doubt the artificially constrained context described renders most of the more obvious methods non-viable, but this is not a good reason to omit it as an analytical possibility. In many other contexts, for example in science or the law, it is exactly what rationality would commend. At least, any conclusions based on such a stipulation cannot be readily generalized to cases in which it is inapplicable.

If the validity of the stipulation is denied, a more intuitively plausible analysis of the case, as a version of S3, is as follows. The agent thinks about their situation and envisages that it might be made better in virtue of the possibility of winning a cash prize conditional on their performing one of two described actions. However, which of these actions has the greater expected value depends on a currently unknown precondition, namely the contents of the second urn. Given this consideration, planning establishes a reduction in uncertainty of this precondition as an instrumentally valuable possibility. Such an effect is evidently achievable in some contexts, for example by looking into the urn. The question is whether it is achievable in this context. Joyce’s description does not necessarily exclude it. The agent might, for example, scan through memories of previous cases to try to estimate the chances of various types of outcome, or try to read the supervisor’s body language, or resort to bribery or prayer or clairvoyance. All may prove unsuccessful but they are not, a priori, irrational.

These possibilities are not in conflict with the usual assumption, which Joyce evidently accepts, that rational choice depends on the agent’s assessment of the balance of probability of the various outcomes associated with each option weighted by the anticipated value of each. Rather, they justify an extended analysis that includes subordinate choices and associated activity directed towards the reduction of uncertainty. I will discuss further implications for the theory of subjective probability in Chapter 6.

For the present one additional issue should be noted, namely Joyce’s implicit assumption that, for the assumed agent, the only value-relevant consideration is an envisaged monetary payment, quantified as such. Viewed naturalistically, this is evidently false. There is also the possibility, as mentioned in §4.7, of the agent experiencing feelings of regret, equanimity, or fortuitous satisfaction depending on the outcome, assuming that the second option is chosen. It follows that even if, as Joyce stipulates, the agent has no data on which to distinguish outcomes in terms of their probability, there may well be a basis on which to distinguish them in terms of their value. Hence uncertainty of probability may not be the dominant issue. A more significant issue may be the agent’s current attitude to risk. This undermines Joyce’s argument that an adopted level of subjective probability should not support any more precise conclusions than the data on which it is based, since it is no longer clear that there is any “precise conclusion” that the data, as given, supports. On the contrary, for a risk averse subject the data tends preferentially to support the first option whereas for a gambler it tends to support the second.

5.5 Commitment

As mentioned at the start of this chapter there is a crucial difference between preferring and choosing. Choosing is an event that, whether or not it involves any immediately observable activity, sets an internal commitment to some current or prospective activity. The question is what this commitment consists of.

Commitment appears to have two aspects: to assign intentional status to relevant actions in virtue of which they are in due course contingently initiated, and to suppress further comparative valuation. Assigning intentional status has the effect of initiating or fixing to initiate a planned system of presumably performable actions usually involving, in part, the pursuit of presumably achievable conditions for which current planning remains incomplete. For example, if I intend to buy a sandwich I may initiate or fix to initiate, inter alia, a connected series of actions including walking to and entering a particular shop and picking and paying for a sandwich but leaving the selection of a sandwich to be a choice among discovered options to be realized when I get there. Having adopted this plan after due consideration, I typically resist repeatedly re-evaluating alternatives, such as ordering a pizza or going hungry. The rationale for the latter is straightforward – if plans are not pursued consistently even when conditions that render them temporarily less preferable are encountered, much effort will be wasted and few desirable objectives achieved.

In a conventionally programmed system, assigning intentional status to a prospective action might involve passing control to an algorithm that serves as its implementation. This would, unless explicitly interrupted, continue independently of any subsequent re-evaluation. A range of evidence suggests that, other than in cases of obviously fluent action, commitment in human action does not operate in this way. Firstly, a plan may admit arbitrary or contingent initial delay. And secondly, evaluatively significant conditions that radically disrupt current activity may arise at almost any point during its execution. For example, an agent may have second thoughts about a choice already made, perhaps as a result of evaluating some previously unnoticed precondition or consequence or receiving new information or an offer of assistance, or be interrupted by the discovery of an unconnected threat or opportunity, or by an accident or injury. The effect may be to expand the range of apparently available options or to revise the value status of previously considered options.

Intuitive evidence suggests that the probability of disruption generally depends on the revised balance of value over these existing and alternative options, as usual. Thus an ongoing project is likely to be interrupted by a non-incapacitating accident if and only if the loss expected to be incurred in failing to respond to the accident significantly exceeds the loss expected to be incurred in interrupting the project. There appears to be no additional interrupting mechanism other than via an effect approximating to actual physical incapacitation – injury, immediate pain, loss of balance, etc.

This assimilation of interruption to normal valuation leaves a problem of accounting for the commitment effect that is rather like the stopping problem discussed in §4.7. The problem is to describe, within the proposed unitary value system, a mechanism that inhibits continual re-evaluation and hence the frequent overturning of choices that have already been made.

A similar analytical response is available. It is that self-observation of repetitive re-evaluation or of the overturning of established choices confers on an agent a quality that, characteristically, has negative salience – a quality intuitively recognized as inconstancy or vacillation. Anticipation of such an attribution weighs against any envisaged advantage to be gained from re-evaluation but, as usual, it may be counterbalanced by a sufficiently large prospective advantage. It is, for example, easily outweighed by any really significant threat or opportunity. Its salience may vary from agent to agent, leading to varying degrees of habitual perseverance. This can readily account for at least part of the effect that Bermúdez (2018) identifies as self-control.

5.6 Roles and Duties

A large proportion of human activity involves social interaction, either face-to-face or indirect. As usually classified, interaction may be either competitive or cooperative or a mixture of the two.

Some competitive interaction is, prima facie, relatively easy to account for within a fundamentally individualistic value-based decision system as assumed in standard decision theory. Evaluation is of prospective conditions however they are assumed to arise, and choices are made by each agent to maximize expected marginal benefit from their own perspective. This can provide a coherent account even in cases in which value is attributed, either positively or negatively, to qualities currently or prospectively exhibited by others – including, for example, the pleasure or pain, or success or failure, experienced by others. Hence it admits, inter alia, simple altruism (Dhami 2016 Chapter 6).

There is, however, a problem in accounting for cooperative interaction that, on consideration, extends also into competitive interaction. It is that cooperation, and most non-trivial competition, depends on each relevant agent operating on the basis of an implicit causal model of the world in which the actions of others are treated not as ordinarily caused phenomena but as chosen projects of rational agents that are capable of supporting or frustrating the agent's own plans, and vice versa. Whilst there is considerable debate over the aetiology and content of this folk modelling of agency there is widespread agreement about its explanatory significance (Hutto and Ratcliffe 2007).

The problem arises as follows. In general, unless there are very significant cultural, institutional, or organizational constraints on admissible options and admittedly relevant consequences, the task faced by an agent in trying to make choices that optimize aggregate outcomes in complex interactive situations is generally intractable. Briefly, so great is the variety of actual and conceivable individual and shared concepts, beliefs, values, conditions, methods, and competences, and so great is the complexity of positive and negative feedback relations connecting individuals' choices, that the combinatorial possibilities in any realistic interactive situation rapidly exceed any plausible computational capacity. This conclusion is supported by both theoretical analysis and historical observation (Newell and Simon 1972: 108-12, Simon 1990). The implication is that, in general, prior to any assessment of what is either presumably achievable or viable, the range of options that agents typically consider, either for themselves or for others, must be severely restricted.

This conclusion considerably limits the explanatory power of the type of game theoretic modelling typically used in the analysis of interactive behaviour, contrary to the impression sometimes given (e.g. Gintis 2009: xiii). Consider, for example, the Centipede Game (A27). In order for a game theoretic analysis to get started it has to be assumed that each player has only two options on each turn – either to play on or to steal. Why this is so is not part of the analysis; it is simply an analytical stipulation. Since the stipulation may be violated, a more complete account must explain both where the stipulation comes from and why it is satisfied in the present case, insofar as it is.

From this perspective it is clear that the Centipede Game, like many other game-theoretic scenarios, is parasitic on a common feature of modern Western culture, namely the concept of a game that includes a distinctly constituted rule-maker and a set of participating players who observe the stipulations of the rule-maker, subject to two further options: to play or not play. This concept is not universal. It stands in marked contrast to the more traditional concept of a game in which both the normative principles of play and of legitimate participation are culturally evolved (Watson 2019).

In all cases involving any significant element of cooperation the standard game-theoretic analysis must assume that participants share a relevant concept of a pattern of activity with stipulated or evolved rules and choose to abide by its constraints. It is this assumption that does most of the explanatory work. Some such constraint is almost universal in interactive human behaviour. Participants hold shared concepts of possible or admissible patterns of conduct and choose, in various degrees, to act accordingly. Such patterns of conduct are usually, especially in sociology, characterized as systems of roles and duties. Their definition is part of a mechanism for reducing otherwise intractable computational complexity.

A system of roles and duties operates chiefly by restricting participants' available options. Consider, for example, a nurse in a busy hospital. The nurse must make a very large number of choices every day. The vast majority involve highly constrained sets of options most of which are unintelligible outside the organizational context in which they occur, in which the systemically related options of other agents are assumed to be similarly constrained. Moreover, the process of choice among alternatives is often more nearly one of approximate rule-following than of comparative valuation of aggregate instrumental consequences.

That participants comply with the constraints defined in a system of roles and duties, even approximately, presents an explanatory problem similar to the problem of commitment described in §5.5. It admits a similar solution, namely that participants recognize and positively value a self-attributed quality generally characterized as conscientiousness. This constitutes, approximately, what Gintis calls a meta-preference, or character virtue. As he says:

> One might be tempted to model … character virtues as self-constituted constraints on one's set of available actions … , but a more fruitful approach is to include the state of being virtuous … as an argument in one's preference function, to be traded off against other valuable objects of desire … (Gintis 2009: 73-4).

Roles and duties need not be explicitly specified in situ as in a kind of public job description. In many social and organizational settings there is an informal culture of admitted and excluded options that participants implicitly know of and subscribe to. This applies both to role-based duties and to duties assigned to individuals more generally by prevailing moral, religious, or legal doctrine, or by convention within a given culture, either universally or by social status. In all cases its operation depends on individuals accepting the definition of relevant duties as defining available options for relevant agents and, individually, assigning positive salience to conscientiousness. It does not depend on their attaching value directly to instrumental outcomes except insofar as such directly attached value may, in situ, add to or override the value attached to conscientiousness.

Within the dyadic value theory proposed in §4.4 it follows that the realization of such positively valued self-attributed qualities – commitment, conscientiousness, etc. – is, insofar as it is observable, desirable in friends and proportionately undesirable in enemies. Individuals typically differ in the salience they attach to each.

In addition to the discussion within sociology there is a very extensive literature concerning the definition of specific roles and duties in particular social and organizational settings – in medicine, the law, business, public administration, etc. There is much less discussion of the explanatory status of roles and duties within contemporary behavioural economics or decision theory (Akerlof and Kranton 2005). Dhami (2016), for example, contains no relevant index entry. It is replaced by a more limited discussion of social norms (Bicchieri 2010). Early proposals made by Simon in 1959 and 1963 (1983: 308-12, 344-50) appear to have faded from collective memory.

It may be noted that an account in terms of roles and duties provides a different, or an additional, account of prosocial effects than one invoking gene-culture coevolution (Gintis 2011, Dhami 2016: 1064-84). Whilst it admits that the specification of options and values must have some biological basis, it allows that the evolution of particular socially defined options within a population has its own cultural and cognitive dynamic. In particular, the set of recognized options and the associated value system of each individual is subject to rational and experiential modification partly realized via conventional education and other mechanisms of social influence and control. The effect can be partly accounted for by an evolved characteristic that Simon (1990) calls docility. On this basis, most of the work done in accounting for prosocial conduct in any significant detail is likely to rest on sociological and psychological rather than on biological foundations.

5.7 Timescale

The methods by which options are realized vary greatly in terms of the time interval over which each typically becomes effective – that is, in their characteristic timescale. Many have an approximately predictable duration or minimum duration. For example, getting a sandwich via the method of finding a shop selling sandwiches typically takes at least several minutes from start to finish but not hours. Serving in tennis takes at most a few seconds. Writing a dissertation takes considerably longer. Getting a sandwich next week or next year is evidently a different task from getting a sandwich today and admits significantly different options with different preconditions, consequences, and associated probabilities.

The implication is that there is in general a non-trivial relationship between the probability of a presumably achievable outcome being realized consequent on some choice being made and the expected time interval between choice and outcome, due merely to the time constraints imposed by available methods. Since, according to the standard theory, the expected value of an outcome depends on its probability, it follows that the aggregate expected value attributed to an option depends in part on the anticipated delay of its various expected consequences purely on account of this relationship between probability and time interval.

Moreover, the relationship between probability and time interval depends on the type of activity involved and the circumstances in which it occurs. For example, fluent actions typically have a different and much more abbreviated achievement profile than deliberate actions, epistemic activity depends strongly on available sources of information, commitment generally extends the achievement profile of a selected option, and exploitation of an available system of roles and duties frequently admits not only extended reliability but also collective or vicarious achievement. For this reason the relationship between probability and time interval needs to be analysed separately in each of these cases.

However, although the existence of a non-trivial relationship between the time interval from choice to outcome and the probability of such an outcome is in many cases obvious, the quantification of the relationship depends on the quantification of probability. This is not straightforward. I will examine the issue systematically in the next chapter.

5.8 Conclusion

In this chapter I have attempted to answer the first question on page 10, ‘What is an option?’ The underlying issue is that however agents assign value over possible conditions – that is, whatever their preferences – they cannot simply choose that a desirable condition be realized. At best, such a condition can be realized only by some means. The options available to an agent are therefore constrained by available means, not merely by desirable outcomes. This is what Schwarz (2021) calls the problem of options. It is concealed in the scenario typically envisaged in theoretical analysis, namely S2 as described on page 134, in which options are paired one-to-one with described outcomes and no collateral effects or other alternatives are admitted as relevant. Evidently this is an idealization even in experimental research, given that non-compliance is an available but unadmitted option. In the generality of human action it is usually very far from being satisfied. Indeed, in many choice situations available options are defined more readily in terms of means – that is, in terms of currently performable actions – than in terms of any specifically preferred consequences, as in S1. This is particularly the case where the issue is one of duty or obligation or where there is no distinctly preferred outcome in view, as in convivial social interaction. And in other cases, as in S3, the question of whether means are currently available may be uncertain. The issue is then whether suitable means can be discovered or devised.

To be an option a condition or action must be presumably achievable. Achievability is often initially uncertain. It is, where possible, vindicated recursively within a process of planning and constituent activity realizing an evolving system of partially envisaged methods. An option standardly characterized as involving only a verbal expression of preference typically has envisaged implications other than those literally described.

Choice assumes a comparison of alternative futures – usually a default future and at least one alternative. Since futures may be arbitrarily complex, simplification is necessary. It may be achieved in several ways, particularly by focusing on probable value differences and rapidly rejecting as non-viable options apparently involving significantly negative value consequences.

Planning terminates in null or fluent units of activity, the latter often having an apparently algorithmic form. Value driven choice selects among, initiates, sets parameters, and may interrupt or halt the operation of such units. Fluency may be established by practice. Many options involve epistemic activity. Epistemic options are typically under-represented in standard decision theoretic modelling.

To be effective, choice requires a commitment mechanism based on negative value being attached to a self-attributed quality of inconstancy or vacillation. Defined roles and duties constrain participants' options, usually improving predictability in otherwise intractably complex interactive situations. Their realization requires a similar mechanism based on positive value being attached to a self-attributed quality of conscientiousness.

Many options are predictably realized by methods having a characteristic timescale. Hence their probable realization typically varies with the futurity of envisaged outcomes in a way that depends on available methods. The analysis of this relationship, and of other futurity-dependent effects, depends however on the quantification of probability. It is to this question that I now turn.

Chapter 6 Probability

6.1 The Reference Class Problem

As is clear from the material described in Chapters 1 and 2, reference to probability is ubiquitous in decision theory. It does not follow, however, that there is any unanimity as to what feature of the world, if any, is being referred to. On the contrary, there is at present a very lively debate, especially within philosophy, as to the merits or otherwise of a variety of possible interpretations of the concept of probability (Hájek and Hitchcock 2016).

Probability is both an intuitive and a formal concept. Many authors give intuitive examples. For example, Hájek (2019) suggests the following:

> ‘The Democrats will probably win the next election.’
> ‘The coin is just as likely to land heads as tails.’
> ‘There is a 30% chance of rain tomorrow.’
> ‘The probability that a radium atom decays in one year is roughly 0.0004.’

On the other hand an axiomatic system commonly identified as defining probability was set out by Kolmogorov (1933) as follows:

Ω is a non-empty set, (6.1.1)
F is a set of subsets of Ω including Ω,
F is closed under complementation and union,
P is a function from **F** to $\mathbb{R}$ such that:

1. $P(A) \geq 0$ for all $A \in \mathbf{F}$,
2. $P(\Omega) = 1$,
3. $P(A \cup B) = P(A) + P(B)$ for all $A, B \in \mathbf{F}$ such that $(A \cap B) = \varnothing$,

and a conditional probability $P(A|B)$ is standardly defined as a ratio:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \quad \text{provided that } P(B) > 0. \qquad (6.1.2)$$

It is unclear, however, to what extent these intuitive and formal notions overlap except that insofar as intuitive probability is numerically quantified it is assumed that no probability is less that zero or greater than one and a more or less precise ordering of comparable probabilities is generally assumed. Philosophical desirability notwithstanding, it is unclear that there is a single coherent intuitive notion of probability in general use – any more than that there is a single notion of number that comprehends, for example, atomic number, mass, velocity, temperature, hardness, population size, money supply, page number, telephone number, and the numerosity of mountains or ideas.

Nevertheless, many attempts have been made to formulate an account of probability that both satisfies Kolmogorov's axioms and reflects intuitive assumptions about how degrees of probability are assigned and used. Hájek (2007) distinguishes ten current alternatives. He argues that all are subject to significant objections. Most importantly, he argues that all current theories containing a non-trivial principle of quantification are afflicted by a single overriding problem known as the reference class problem.

The reference class problem has been known about at least since Venn (1888) and was first so named by Reichenbach (1949). It is usually described, after Venn, as follows. Suppose that John Smith is a man aged 50. What is the probability that he will live to 61? This is evidently something that may be of interest to him. Perhaps he should consult relevant life tables. But now suppose that he is an Englishman, consumptive, living in a northern town, and so on. Each classification is likely to imply a different probability. Which is to be preferred? If the answer is that the most specific applicable classification is to be preferred, this is likely to leave him in a class containing only himself for which no non-integer probability is inferable, for ultimately he either lives or dies, so the probability is either zero or one. What, then, is the relevant probability and to what does it attach?

It has often been assumed that the reference class problem is specific to theories that quantify probability on the basis of an observed or assumed relative frequency. Hájek argues convincingly that it applies more generally. For example, it recurs in a theory based on an assumed equiprobable distribution over possibilities. If, for example, it is assumed on grounds of equiprobable distribution that the probability that a coin lands heads versus tails is 0.5, why should it not be argued similarly that the probability of living in the northern hemisphere is 0.5? And there is an analogous problem in interpretations that assume some form of prior randomization that we may call the admitted data problem. Consider, for example, a typical game of cards, such a bridge or poker. Normally the probability of various combinations of cards is assumed to depend on the number of cards of each relevant type in a standard pack or in the remainder of the current pack. But in fact, after shuffling, the sequence in which cards are dealt is entirely predetermined. This is explicit in the game of duplicate bridge in which a dealt sequence is presented repeatedly to different combinations of players. Hence, strictly speaking, the normally assumed probabilities never apply in any actual game. Their appearing to do so depends on the current state of the cards being concealed. Indeed, the complexity of both bridge and most versions of poker arises in part from the progressive revelation of potentially relevant data during play. A similar but more subtle effect also arises in games involving physical chance such as roulette or dice. Real-time processing of high-speed imagery of the motion of the ball or dice could, with increasing certainty, predict how they will land. The standard calculation of probability depends on such data not being admitted. Its use is classified as cheating.

Recognition of the reference class problem is a major motivation for the adoption of a theory that interprets probability as a subjective attitude or, more particularly, a degree of belief or credence. But, as Hájek argues, if such an attitude is to serve as an effective guide in rational decision making it must be constrained by objective features of the circumstances in which it applies, in which case the problem recurs in the identification of these constraints. Consistency alone is not sufficient to make probability relevant.

Both Hájek (2003, 2007) and Eagle (2004), argue that the root of the problem is located in the ratio formula defining conditional probability, (6.1.2). This formula requires, in the denominator, a non-zero unconditional probability that, they argue, generally does not exist. For example, in defining the probability that a coin lands heads given that it is tossed, the formula requires, in the denominator, a number representing the probability that a coin is tossed. Hájek objects that in most cases there is no such probability. What is needed, he argues, is a theory of probability based on a primitive notion of conditional rather than unconditional probability. Several attempts have been made to develop such a theory (Roeper and Leblanc 1999, Hájek 2011). None has achieved widespread acceptance.

Oddly however, there is little pressure from the main users of probability theory – those working in statistics and statistical science – to develop such a theory. Analysis is not generally inhibited by any obvious difficulty in defining denominators.

The explanation appears to be that the standard formalization of probability is already implicitly conditional. It is conditional on the admitted content, in any assumed instance, of a universe of reference identified in (6.1.1) as Ω. In any particular statistical application this is what characterizes the relevant population and its correspondingly relevant attributes. It is specifically chosen to admit only cases of certain relevant types occurring within a certain domain. Nowhere is it assumed that it admits, or might admit, the entire content of the actual universe. Within this system, conditional probabilities are insensitive to the arbitrary inclusion of additional non-cases since the same non-cases will be reflected equally in both the numerator and the denominator of (6.1.2) and hence the corresponding factors cancel out. For example, the proportion of heads to coin-tosses is unaffected by the spurious inclusion of, say, foot-stamping in the set of admitted events. The problem of uncountable alternatives is usually avoided by finite sampling – which is more or less unavoidable given that both experiential and experimental data are necessarily finite – and in more complex cases it is often possible to define a relevant probability density function. Hence, in practice, the mathematical problems identified by Hájek and Eagle seldom inhibit effective analysis.

Nevertheless, this does not resolve the reference class problem since no choice of Ω resolves the question of whether John Smith will live to 61. The problem is that a variety of estimates can be derived, from complete uncertainty to ultimate certainty, given that what will happen is initially unknown but that in the end he will, presumably, either live or die. So the problem is not one of finding an appropriate conditional probability but of quantifying imperfect data. I will investigate this problem shortly. First, however, there are several other issues that arise in the standard account of probability that will be relevant in later analysis.

It is sometimes claimed in philosophical discussion that exact unconditional probabilities exist as raw physical propensities. Quantum mechanical effects – especially of radioactive decay – are often cited as examples, as on page 161. On this basis, probability is assumed to be inherent in the physical constitution of the universe rather than being only a form epistemic generalization. This offers the attractive possibility that all sound probabilistic judgement may be traced back to elementary physical effects propagated by complex patterns of deterministic interaction.

One difficulty with this is that it is known that deterministic systems can behave unpredictably or chaotically without the involvement of any fundamental probabilistic propensities and hence that the ontic status of examples such as those mentioned is uncertain. Bohmian mechanics, for example, provides a different analysis of quantum probability from that of the standard theory (Maudlin 2002). On this basis Schwarz (2018) argues that that the search for an ontic interpretation of probability is misguided and that the point of probabilistic theory “is not to express facts about some probabilistic quantity, but rather to capture noisy relationships between ordinary, non-probabilistic quantities" (2018: 1210). This accords with the many arguments against propensity analysis advanced by Eagle (2004) and, in particular, his claim, quoting Clark (2001), that "the issue of determinism versus indeterminism really ought to be (is) irrelevant to an interpretation of probability theory" (2004: 387).

There has been increasing interest in recent years in a notion of imprecise or second-order probability, in which estimates of probability are treated as values of a variable having an associated probability distribution (Gaifman 1986, Williamson 2000, Klumpp and Hanebeck 2009, Caster and Ekenberg 2012). Joyce (2010), described in §5.4, presents a version of this approach. Several different notions of imprecision exist, represented by, for example, sets of possible probabilities, fuzzy sets, numerical intervals, or density distributions. Set-based and interval-based schemes, although often more tractable than density distributions, appear to violate the intuition of central tendency and leave a problem of establishing a best estimate, as needed to justify a unique preference in decision making (Caster and Ekenberg 2012). An alternative is that imprecision is itself imprecisely quantified.

A well understood source of uncertainty arises from sampling error (Fisher 1937, 1956). The key fact is that whilst it is possible to estimate the relative frequency of a trait in an actual or envisaged population by measuring the frequency of that trait in a representative sample of members of the population, the estimate is subject to a probable error that is, typically, inversely proportional to the square root of the sample size. Where a sample that is representative of a specified population contains a subset that is representative of a more homogeneous subpopulation, the observed frequency of a relevant quality in the subset may not provide a better estimate of the subpopulation frequency than one derived from the entire sample because, although the subset is more representative, it is smaller, and hence the estimate is subject to greater sampling error. This is what makes the reference class problem particularly intractable.

A final issue concerns inferential asymmetry. Probability is, in common usage, usually a measure of uncertainty of prediction – that is, of inference from evidence of given conditions to an expected outcome. It is seldom envisaged as merely describing a distribution over possible factive states. This commonly introduces an intuitive asymmetry that mirrors the assumed asymmetry of causation discussed in Chapter 3. Thus although statistical texts regularly caution that correlation does not imply causation, the urge to draw such an inference is strong. Unexplained correlation is intuitively a puzzle. It is not epistemically neutral.

An apparently plausible rule for inferring a causal relation is that we have evidence that *A* causes *B* if the probability of *B* conditional on *A* is greater that the probability of *B* in general – that is, if

$$P(B|A) > P(B). \tag{6.1.3}$$

But this generates a problem discussed by, for example, Eagle (2004: 402) and Ahmed (2007: 122). It arises as a simple consequence of the ratio formula (6.1.2). Given that,

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \text{ and } P(B|A) = \frac{P(B \cap A)}{P(A)} \text{ and } A \cap B = B \cap A,$$

it follows that,

$$\frac{P(A|B)}{P(A)} = \frac{P(B|A)}{P(B)}\;. \tag{6.1.4}$$

Hence $P(B|A) > P(B)$ iff $P(A|B) > P(A)$.

The implication is that whenever we have prima facie probabilistic evidence that *A* causes *B* we have equivalent prima facie probabilistic evidence that *B* causes *A*. This accords with the conclusion reached in §3.5 that causation cannot be inferred from probabilistic evidence alone and with Pearl’s (2018) renunciation of the opposite principle, as discussed on page 86.

In the light of issues discussed in this section and other background considerations it is possible to formulate a set of desiderata that, it appears, need to be satisfied in the construction of an empirically adequate theory of probability as required in effective human decision making. For clarity I will assemble these desiderata systematically in the following section.

6.2 Desiderata

The standard account of decision making based on valuation of probable consequences depends unavoidably on a vindication of the notion of a quantifiable and relevantly applicable measure of probability. As is apparent, this is not straightforward. The usual practice of merely stipulating a numerical probability, as frequently illustrated in the Appendix, does not resolve this but avoids it. It is not surprising, then, that the relevance of experimental results is often disputed (Dhami 2016: 4).

The required account need not be one that is appropriate in all fields in which a concept of probability is used – in, for example, statistical mechanics or quantum mechanics. Nor is the task merely to provide a conceptual analysis or interpretation of an available intuitive or formal concept. Rather, the problem is to explicate how, in Hájek's (2007) phrase, probability serves as a 'guide to life' – how judgements of probability, in context, are located, quantified, and employed in the effective evaluation of envisaged options and hence in actual decision making. On this basis I propose that the account should satisfy the following desiderata.

1 Agent-based Futurity

The notion of probability at issue is one that applies primarily to conditions or events envisaged as possibly occurring or prevailing in the future relative to a particular agent at a particular moment. It may by extension be applied to uncertain past or present conditions or events but only to the extent that they are envisaged as partial determinants of future conditions or events.

2 Subjective-Objective Duality

The required notion of probability must have both a subjective and an objective aspect. The subjective aspect is indicated, at least, by the fact that degrees of probability are assigned by agents to envisaged future conditions and events. There is a fact of the matter about these assignments that is partly independent of any facts of the matter about the conditions or events themselves. The objective aspect is indicated, at least, by the fact that the assigned degrees of probability are features of an action-generating system that has testable real-world outcomes.

3 Adaptive Origin

There is good reason the believe that the assignment of probability to envisaged conditions and events is part of an evolved biological, cultural, and cognitive system shaped by relative adaptive success. Since success is externally constrained, it follows that something in the world must answer at least approximately to assigned probabilities. It does not follow, however, that each assigned probability must have, or represent, a correspondingly quantifiable ontic counterpart. A more plausible hypothesis, consistent with the standardly assumed status of scientific theory, is that a degree of probability is a hypothetical construct that is part of an effectively employed model of an instrumentally accessible part of the world. The ontic status of its components is a matter of, at best, conjecture.

4 Alternatives

The key role of the assignment of probability to envisaged future conditions or events is to enable prediction of relative likelihood among a set of apparently possible alternative outcomes in some envisaged case, based on currently available data. The set of assumed alternatives must usually be inferred, either directly or by extrapolation, from recorded features of members of a class of cases from which the data relied on is derived.

5 Hybrid Quantification

The competing considerations partly summarized in §6.1 strongly suggest that there is no unique basis on which degrees of probability are assigned to envisaged conditions or events. There is reason to believe that frequency data are always involved somewhere, but this is by no means sufficient. Frequency involves some kind of classification. Arbitrary counting does not generate meaningful assignments of probability. Probability applies to cases not merely to aggregates, so some relevant similarity among class members must be assumed, and it is in virtue of this similarity that probability is assigned. Furthermore, frequency-based probability is often precisified via causal assumptions. This is reflected in a symmetry-based account, such as that a fair coin fairly tossed has a probability of exactly 0.5 of landing heads. Indeed, dependence on causal assumptions may occur more generally, as in an assessment of the probability that a person will die based on an consideration of various possible effects leading to their death – heart attack, tuberculosis, accidental drowning, assassination, etc. – rather than simply on their current class membership. It is further extended in statistical analysis by, for example, factor analysis, which may sometimes be interpreted as implicitly causal.

6 Token Specificity

The required notion of probability must be one that applies to tokens of relevant types, not only to types. It must, for example, speak to the question of whether *John Smith himself* will survive, not merely to whether *a person like John Smith* is, or would be, likely to survive. This necessarily exposes it to the reference class problem. Where this applies there will typically be no uniquely justified case-specific probability. Rather, there will be a range of differently justified probabilities based on different modes of classification whether or not these are associated with distinct causal effects. Furthermore, as classification becomes more exclusive, the body of implicitly associated statistical data generally becomes increasingly limited and hence is increasingly subject to sampling error. A best estimate for predictive purposes of case-specific probability must balance these competing effects.

7 Second-order Probability

The resulting quantification is most plausibly of a central estimate of probability based on current considerations together with a measure of uncertainty, or variance, arising from the uncertainty and/or incompleteness of those considerations. Extending the set of current considerations does not necessarily increase precision since it may reduce the range of relevant data. First-order and second-order probability cannot be straightforwardly combined into a single scalar quantity without loss of information. Their aggregation in logical compounds depends on the structure of the assumed possibility space as admitted in a current causal model of the world. Commensurability in decision making requires their transient collapse into a single effective quantity at each point of definitive choice.

8 Numerical Precision

Assignments of probability must admit of at least ordinal measure and transient commensurability. Intuitive and experimental evidence beyond this is difficult to assess. Neurological and computational considerations speak against a realization that fully satisfies the standard probability axioms. Approximate, stochastic, and heuristic alternatives can easily be envisaged. Accuracy and Dutch book arguments have no purchase if real number quantification is impossible. Nevertheless, as an idealization, real number quantification is at least convenient.

9 Classification

The envisaged system assumes routine collection of suitably classified statistical data embedded in an evolving causal model of the world. Classification is generally Boolean. It may be reconstructed from time to time so as to provide more efficient modelling of structures apparently exhibited in available or newly acquired probabilistic evidence and more successful prediction of outcomes, either by discrete conceptual adjustment or by probabilistic updating of detector processes.

10 Rationality

A probabilistic system is rationally justified to the extent that, in situ, it contributes to the successful prediction of outcomes.

6.3 Probability in Decision Making

Given these desiderata I propose to characterize the concept of probability involved in decision making as follows.

1) A probability is an imprecise scalar quality assigned to each of several alternative versions of a recognized type of object or condition or event in a recognized type of environment based, inter alia, on relative frequency data available to the agent.

2) A *typical case* of a given type in a given type of environment is a single object or condition or event of that type in which the realized version is not more accurately predicted from all data available to the agent than from its being a case of that type in that type of environment.

3) In such a case the most accurate available prediction comprises a set of probabilities assigned to alternative versions of that type in that type of environment.

4) Methods of testing accuracy of prediction exist.

5) Methods of selecting typical cases exist.

6) Accurate prediction is of adaptive benefit.

Several features of this characterization require comment. It treats probability as an attributed quality much like colour or size or weight. When effectively used, assignments of probability have an objective basis, but it does not follow that they are independent of human judgement. On the contrary, they depend on a prevailing system of classification very little of which can be explicated without at least implicit reference to the diversity of human cultural and cognitive evolution. Treating probability as an attributed quality allows that is can be attributed either intuitively or as an explicit measurement as in statistical analysis and stochastic modelling. The same computational methods need not apply equally in all cases.

Assignments of probability are inherently conditional. The key idea is that distinguishable objects or conditions or events of a given type can be realized in various mutually exclusive versions. Furthermore, various types of environment can be distinguished. A distinct probability is assigned to, at most, each version of a specified type within a particular type of environment. No concept of absolute probability is assumed.

Relativity with respect to a type of environment is significant in two respects. It is a common observation that contingent relations among objects, conditions, or events vary from one environment to another. But patterns of variation are often systematic. Hence information relevant in one environment can be relevant in another by systematic transformation. For example, probabilities applying in a game of cards played without replacement are systematically transformed as cards are removed from play. Systematically related environments may be envisaged as parts of a landscape.

Assignments of probability are assumed to be scalar but, in varying degrees, imprecise. In science a probability is usually represented as a real number in the range [0, 1] with, where shown, a similarly estimated standard error, or a conventional multiple thereof. The error term represents a degree of approximation, itself approximate. Intuitive assignments of probability involve a similar degree of implied approximation. Rationality cannot require absolute precision.

All assignments of probability are based partly on statistical data – namely, on the relative frequency of observed realizations of various versions of a given type within a characteristic set of environments. There is, however, an ‘inter alia’ proviso. It allows, in addition, considerations of symmetry, similarity, systematic variation, and assumed nomology. These must be appropriately reflected in some relevant statistical data but they often provide greater precision than statistical data alone would supply. This is particularly so across related environments within a landscape where the data in each environment is sparse but is supplemented by corresponding data in other parts of the landscape. For example, epidemiological data is usually assumed to be generalizable across subpopulations.

Furthermore, there is reason to believe that humans tend to oversample rare or unusual cases, as discussed in §4.6. This makes rare cases – whether successes or failures – appear more common than they are. Hence probabilities near zero or one tend, respectively, to be over or underestimated, as shown in Figure 4 on page 54. Conversely, intermediate cases are comparatively undersampled and hence the total probability over the set of admitted alternatives remains equal to one. No compensating mechanism is required.

The proposal introduces the concept of a typical case. This might alternatively be described as representative or, loosely, random. The key idea is that where a particular object or condition or event of a particular type in a particular environment is observed or envisaged, and only limited information is available, the best predictor of the realized version is simply that it is of that type in that type of environment, and the best prediction is the set of probabilities assigned on this basis. Where there is a choice of types, the optimal type is the most restrictive type for which relevant probabilities are defined.

Type-based prediction may work both in presently realized cases such as a card lying face down on a table and in prospective cases such as a coin about to be tossed. This equivalence is instructive. The fact that they are intuitively treated as probabilistically similar shows that it is the availability of data that is crucial, not futurity. Futurity is relevant only in that it limits available data. But it also shows why, so far as decision making is concerned, futurity is such an important matter. Futurity makes probabilistic analysis necessary because, as described in §3.5, records, as such, are rarely if ever better indicators of future conditions.

And finally, three factual claims are made, namely that it is possible to test the accuracy of prediction, that it is possible to select typical cases, and that successful prediction is of adaptive benefit. Testing the accuracy of prediction is common both in science and, implicitly, in ordinary life. The testing of relative probability is always subject to sampling error but gross disparities are detectable insofar as they greatly exceed typical divergence. Various recognizably random or quasi-random procedures designed to defeat or conceal possibly relevant discriminatory factors exist for selecting typical cases. Shuffling cards is such a procedure. And the generally adaptive benefit of successful prediction is evidenced in past technological development and is underwritten by evolutionary theory. This combination of factual claims plausibly justifies an assumption of the approximately objective status of emergent classification and of associated probability assignments. It allows claims of rationality to depend on broad predictive success rather than on either a specific aetiology or freedom from error.

The proposed account of what I will call, for clarity, ‘statistical probability’ plausibly explains why humans assign quantified probability to envisaged events and conditions despite the lack, before the emergence of quantum mechanics, of any compelling evidence of probabilistic causality. Its merit is that, on the often reasonable assumption of typicality, it provides a best estimate of likely outcomes in cases where limited information makes unique prediction impossible. This is of obvious adaptive benefit. Furthermore, insofar as envisaged possible outcomes partition an assumed possibility space, it explains both additivity and unit total probability in each relevant set of possible outcomes. It does so independently of any notion of partial belief. It makes it possible, for example by doubting perfect typicality, to assign a non-zero statistical probability to a possible outcome *and* to believe that it will not happen – which is otherwise a puzzle, highlighted by Dhami (2016: 193-6). I will discuss this further in the next section.

By extension, it also plausibly explains the attention paid in concept formation to partitioning. This is not universal, nor self-evidently implicit in the observed world, nor is it inherent in a minimal notion of description – which merely involves identifying apparent features. Indeed, many qualities such as colour, shape, texture, geographical location, and function defy ready partitioning. Nevertheless, constructing partitions has a long evolutionary history, extended into science. It is most explicit in the definition of scalar variables such as age or length or mass. Both statistical and nomological analysis depend significantly on this development.

The proposed account also plausibly explains why, in making a prediction, an agent assigns a probability distribution over envisaged alternatives rather than merely assuming that the most probable alternative will occur, despite the fact that, prima facie, the latter is less likely to be wrong. The answer is that the predictive system is part of a decision making system that also involves valuation and that in decision making the crucial comparison is of expected value not merely probability. Hence it is important to preserve probability distribution information at least until a determining set of aggregate values is computed. Even then some probability information may be preserved, perhaps to facilitate revaluation where necessary and to enable updating of stored data.

It should be noted that this series of proposed explanations satisfies the methodological requirement outlined in §1.3 that any theory that denies the veridicality of some relevant intuitive impressions must account for their occurrence, given this denial. The proposed theory denies that, in general, individual realizations of particular objects or conditions or events have any absolute probability other than zero or one. Intuitive impressions of absolute probability other than zero or one are, therefore, generally non-veridical. Hence their occurrence requires explanation. In summary, the explanation is as follows. Envisaged future objects, conditions, and events are, prior to the accumulation of more detailed descriptive information, commonly and perhaps necessarily assumed to be typical members of known types. Hence associated statistical data, if available, applies. The same is not generally the case for past or present objects, conditions, or events since for them more detailed descriptive information usually already exists, or may exist, via observation or available records. Hence an intuitive probability is inferred for them only if such information is severely or deliberately limited. In either case, intuitive probability typically changes as extra information justifying reclassification emerges until, perhaps, a point is reached at which no relevant type having sufficient associated statistical data is available, at which point an unstable condition of uncertainty, often marked by hesitation and equivocation, takes over. Only in rare cases in which an assumption of typicality is unvarying, such as for specific types of radioactive decay, is an assumption of absolute probability well founded.

Notwithstanding its justification in terms of adaptive benefit, the dependence of the predictive system on classified data introduces various kinds of bias, uncertainty, and possible error. Intuitive data are collected, stored, and recalled by an individual during their lifetime. Hence the system is subject to constraints and biases arising from first-person experience. They include inappropriate, idiosyncratic, or inconsistent classification, imperfect detection, unrepresentative experience, value-based selective attention, limited storage, and imperfect recall. Stored data may fail to reflect environmental variation either by domain or over time. Most obviously, records are accumulated historically and so may fail to reflect secular trends or local fluctuations. Cultural variation is unlikely to be fully recognized. All data are subject to vagaries of past and future social interaction and, especially when acquired by report, of in-group selection, mutual reinforcement, and possible fraud. And finally, classification and sampling must necessarily fail to reflect the complexity of the environment especially in respect of currently undetectable, unrecognized, or misclassified features and relationships. I will examine possible effects of and compensating responses to these sources of bias, uncertainty, and error, especially as they relate to intertemporal choice, in the rest of this chapter.

6.4 Predictive Reliability, Vagueness

The evolved action-generating mechanism relies, in the absence of more secure information, on estimated probabilities of envisaged conditions and events. These estimates are subject in various degrees to error, uncertainly, and unreliability.

Imperfect reliability takes a number of distinguishable forms. Firstly, the entire system is agent-relative. Culturally given classification, shared data, and explicit statistical analysis establish a degree of commonality, but idiosyncratic features cannot be eliminated. In ordinary decision making these may be very significant, especially the dependence on personal experience. Conversely, information received from others usually lacks adequate experiential confirmation and so depends not only on the trustworthiness of the source but also on shared classification. Shared language does not necessarily entail shared classification since the latter depends significantly on personally instantiated recognition procedures that may differ in detail. It is especially problematic across cultural boundaries and in interdisciplinary or intergenerational comparisons. The degree of variability may be concealed in generally superficial communication.

Most underlying statistical data are collected either by personal observation or from information received. How such data are obtained, classified, collated, and used by agents is a matter of ongoing psychological research (Chater and Oaksford 2008). Much must be assumed to depend on sampling of accidental experience and hence to involve significant chance variation. A system of intuitive classification presumably reflects, in part, long-term historic and prehistoric experience somewhat as assumed in evolutionary psychology, as described in §3.1. It is likely, therefore, to be only partly appropriate to current conditions. Even if classification remains appropriate, statistical records may become obsolete. A process of forgetting is needed to correct this. Since storage is limited, statistical recording must involve some perhaps stochastic granularity.

The resulting epistemic system is part of a larger one that generates value driven decision making. Classification and data gathering are selectively deployed within this larger system, mainly via the mechanism of selective attention as described in §4.7. Since values vary both between individuals and over time, classification and data gathering may also vary, quite aside from any variation in inferred probabilities, due to value differences that emerge during planning. The effect is complicated by the fact that value-relevant effects may arise indirectly, from the interaction of conditions that are not themselves value-relevant. Hence to the extent that data-gathering is driven directly by current instrumental valuation, some potentially relevant data may escape notice. This justifies a wider range of exploratory activity driven by non-instrumental valuation, including the deliberate sampling of unusual environments, as reflected in the usual contrast in behavioural psychology between exploration and exploitation (Wilson et al. 2014).

Regardless of its adaptively driven origins it is clear that established human methodology is far from providing a settled and generally successful predictive account of the context within which agency is effective. It is clear, in particular, that successful prediction is not uniformly distributed over environments but is more frequent in what we may call good cases – cases in which many recognized patterns are frequently detected. It follows that it is generally of adaptive benefit to focus the development of descriptive resources on those that discriminate with a high level of certainty in good cases, in which prediction is likely to be relatively successful, and to allow uncertainty to persist elsewhere, where greater precision would yield little in terms of successful prediction.

One particular issue that affects the variable reliability of probability estimates is identified in the notion of goodness-of-fit in classification. This effect is often referred to as if it was a probability, as in ‘that is probably a pink-spotted flycatcher’. But, if so, it is of a different type.

Variable goodness-of-fit applies everywhere in observation-based classification but less often in abstract or predictive classification. For example, it makes little sense to attribute to a sandwich that I envisage making a probability of its being a sandwich. More importantly, although an evolved quantification may be calibrated against typical data, judgements of goodness-of-fit do not depend on assumptions of typicality but on pattern-matching data generated, in context, by the perceptual or measurement procedure used (Feldman 2003, Goldman 2012: 52-66). And, most importantly, assigned quantities do not necessarily sum to one over discriminable alternatives. This is seen most clearly in the classification of ambiguous images such as the duck-rabbit shown in Figure 11, in which perception assigns a relatively high goodness-of-fit for each of two incompatible alternatives. A converse effect may occur in, for example, an attempt to identify a letter in the bottom row of a optician’s chart, in which all alternatives are assigned a low or very low goodness-of-fit. An argument that such quantities ought to sum to one has no purchase if there is no uniquely defined partition of admissible alternatives.

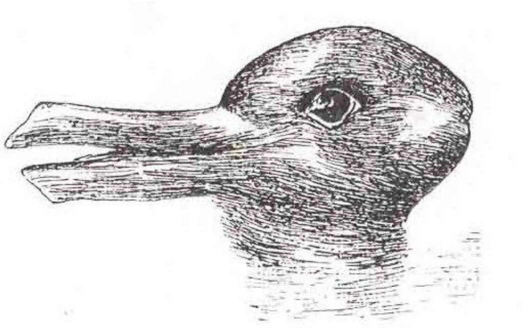

Figure 11: Duck-rabbit image

A similar form of unreliability is captured in the notion of precision in measurement, as in ‘that is probably, or approximately, 1.2 metres long’.

Again, the degree of assumed imprecision depends on details of the measurement procedure as used in the current context or, where what is at issue is a design specification, on constraints assumed to apply in its realization – for example, manufacturing tolerances.

Variable goodness-of-fit may generate vagueness in classification. As remarked, successful prediction is not uniformly distributed over cases. If variable goodness-of-fit reflects variably successful prediction it is of little adaptive benefit to focus epistemic resources so as to enhance classificatory discrimination in relatively bad cases rather than in those that admit more successful prediction. Categories that involve an intractable subset of bad cases are standardly classified as vague.

A commonly cited example of a vague category is a heap. In good cases, recognizing something as a heap allows a number of reliable predictions, such as that it is relatively stable but unstructured, that it tends to disperse by random interaction, that it is associated with functionally coherent animate activity, and that it can be cleared by shovelling. But the category does not have precise boundaries. Bad cases also occur. The commonly discussed issue is numerosity – does removing one grain from a heap leave a heap? – but there are other more predictively significant issues. For example, a heap of cement will, before long, turn onto a solid mass. A heap of snow may melt. Bricks may form a pile rather than a heap if their arrangement is sufficiently regular. A mass of rocks may have accumulated geologically rather than by animate activity, and so on. In such cases some otherwise reliable predictions are likely to prove unreliable. For example, an attempt to shovel solidified cement is likely to fail. Imposing a precise definitional boundary will not correct this. On this basis most intuitive categories are in some degree vague.

Vagueness is widely admitted in philosophy (e.g. Williamson 1994, Keefe 2000) but it is often treated as a defect of classification that can be corrected on the basis that every statement of category membership ought to be analysable in terms of a set of truth conditions that every putative instance must either satisfy of fail to satisfy. This, however, assumes that it is possible to specify truth conditions in a way that does not itself reintroduce vagueness. The present account makes no such assumption. On the contrary, as Williamson observes, “Ignorance is a natural human state. … Our knowledge stands more in need of explanation than does our ignorance” (1994: 216). Knowledge rests ultimately on evolved but imperfectly adapted descriptive resources that, as outlined above, record features of the current environment, each with an implicitly associated degree of uncertainty. In what Williamson calls borderline cases the level of certainty assigned by prevailing discriminative procedures is low, but it is of little or no adaptive benefit to modify them so as to increase it unless such a modification improves the scope of some predictively successful methodology. This is as true in scientific as in intuitive categories. For example, the category of a spring for the purposes of Hooke’s law or of a gas for the purposes of Boyle’s law must be vague since these laws fail progressively. Fixing boundaries does not avoid this.

This ubiquitous variation in reliability is reflected in judgements of statistical probability in two ways. Firstly it leads to uncertainty in the collection of statistical data, since cases may be misclassified. In particular, borderline cases are likely to be undersampled. And secondly, assumptions of typicality are likely to be often imperfectly realized. What occurs is not as assumed. A third, rather unexpected, source of predictive uncertainty arises from the fact that the sample data associated with a given data type may be sparse. The effect occurs as follows. Generally, the more specific the type the smaller the sample and hence the greater the implied sampling error. So, counterintuitively, the reliability of predictive inference often decreases as the quantity of relevant information about a case increases. It is a version of the reference class problem described in §6.1.

Finally, any predictive system is at the mercy of unforeseen events. Even assuming causal determinism we have no reason to believe that any system that is part of the universe is capable of modelling, or accurately predicting, the remainder. Sooner or later, unforeseen or catastrophic change is likely to render some existing classification and associated data obsolete and hence to defeat the justified assignment of related statistical probabilities. Similarly, changed boundary conditions are likely defeat nomological prediction. I will discuss these effects more fully in §6.7.

Prediction does not always involve estimated probabilities. A simpler method involves direct extrapolation of currently observed conditions or trends based on current classification. It depends in part on the notion of a specious present, in which the content of 'now' extends to an interval admitting second order features such as continuity, discontinuity, and patterns of qualitative or quantitative fluctuation (Lockwood 2005: 366). Extrapolation is most commonly used in short range prediction. Statistical prediction, which rests on an assumption of typically rather than continuity, is more commonly used in longer range prediction. A third method, increasingly used in formal contexts, involves theory-based dynamical modelling. It requires an explicitly developed nomology, assumptions about relevant boundary conditions, and appropriate methods of logical inference. I will discuss the methodology of nomological development in the following section. Other methods such as revelation, divination, and clairvoyance are also used but are generally unreliable. A final method involves seeking expert advice. It is parasitic on at least one of the other methods mentioned.

6.5 Nomology

The account of statistical probability proposed above tacitly assumes that an agent registers the flux of experience in terms of features variously present in a series of distinguishable environments. An environment is best conceptualized as an internal record, using available descriptive resources, of everything that the agent notices or envisages on an occasion or in a relevantly bounded interval. A type of environment is a class of environments having a characteristic set of fixed and/or variable features. An agents collects and stores statistical data on the frequency, by type, of environments, of transitions between environments, and of features and combinations of features of environments.

For human agents the usual intuitive typology of environments is evidently of great complexity and subtlety. A tiny fragment of it was investigated by Wittgenstein in his analysis of games. A typical intuitive classification might recognize environments characteristic of not only types of games – chess, card games, football matches, golf, etc. – but also states of play within a game, of types of shop and interior states and locations within a shop, varying sea states in sailing, road conditions when driving, states arrived at during conversation, calculation, and puzzle solving, and so on.

A set of transition relations between environments classified by type constitutes a landscape. Both actual transitions and specific features of successor environments are often more or less predictable from accumulated landscape data – as recorded, for example, during the playing of a game, or on a journey, or when moving from room to room in a house. If the data relating to a type of environment are sparse it can often be supplemented by data from nearby environments adjusted to accommodate apparent trends.

A landscape, as a set of transition relations among environments, is quasi-Markovian, but the logic is inverted. In Markov theory, descriptive resources, classification, and probabilities are fixed so as to define a unitary stochastic process. Here it is assumed that descriptive resources and classification are open to dynamic modification so as to allow the system, via its inferred probabilities, to achieve improved predictive success.

A landscape provides data for the construction of a causal model of the world as assumed in §4.2 but is not itself such a model. Most notably it lacks creativity. It is, in effect, a generalized history of sequential patterns of actual experience. It serves as a direct guide to future experience only insofar as the latter reproduces a prior pattern. The relevance of this distinction is seen most clearly in the contrast between scenarios S1 and S3 as described on page 134. In S1 the agent envisages a possible action. This may involve reproducing a prior pattern, although its achievability in the current circumstances and its likely consequences may need to be established by separate nomological analysis. But in S3 the agent envisages a future condition independently, not on the basis that it is connected to the current environment by an established sequential pattern. The task in planning is to create such a pattern. This may be done by searching the landscape for possible paths, but it often involves devising an entirely new path based, in part, on current nomological assumptions. If the envisaged condition is unprecedented, this is the only viable method.

The inductive step from landscape to nomology is not logically simple. A landscape maps out all sorts of moderately reliable sequences that answer to the intuitive idea of processes. Finding general principles and/or mechanisms that account for their occurring and persisting as they do in the then-prevailing circumstances and that can be used to infer new patterns in other circumstances requires an additional step that historical experience suggests can be justified only by cumulative predictive success. No successful quasi-deductive methodology is known. Some relevant relationships may be suggested by patterns of observed statistical coincidence as proposed by Pearl (2009) but, as noted in §3.6, any inference to directionality requires active intervention, as represented by the *do*(·) operator. More radical nomological development almost always requires significant conceptual innovation by abstraction, extension, or analogy, underdetermined conjecture, and some explicit or implicit predictive testing of assumed consequences, approximately as described by Popper (1963). Since testing is of tokens not types and may be of only partial or peripheral consequences, a resulting nomological system may be very imperfectly justified, especially in its assumed ontic commitments.

It seems likely that for humans the bulk of this developmental work hitherto has been done by biological and cultural evolution, establishing the very varied intuitive nomology of interacting objects, substances, forces, motion, health and illness, growth and decay, action, information, social order, and other minds that is typical of humans (e.g. Norton 2007, Elga

2007, Helman 2007, Ratcliffe 2007). Scientific development depends, in addition, on deliberate experimentation and hence on the intuitive nomology of records and agency described in Chapter 3.

As noted, nomological assumptions often supplement statistical data in probability estimation and prediction. Statistical data, although dependent on classification and sampling, usually reflect observable contingencies at least approximately. But nomology, although sometimes enormously more powerful, is significantly more prone to gross error. The history of bizarre causal beliefs is difficult to comprehend, and it cannot reasonably be doubted that many current causal beliefs are almost equally erroneous. Resulting errors of prediction may be mitigated by preferential reliance on statistical data where available but, conversely, theoretical bias may distort classification, inhibit data gathering, and impair the judgement of typicality. The net effect of misguided causal beliefs may not appear as uncertainty – on the contrary, it may appear as misplaced certainty – but it ought, in terms of either adaptation or rational effectiveness to be deprecated.

One nomological scheme that is of very wide significance is that of other minds, often referred to as folk psychology, as mentioned above and in §5.6. At its simplest the issue is that when attempting to predict the behaviour of others humans seldom rely merely on statistical probability. They rely instead on a concept of motivated choice standardly involving assumed beliefs and desires. It proves possible, by treating the implied beliefs and desires of others as predictable either from communicated information or from then-prevailing circumstances, and assuming that these beliefs and desires determine action in a standard way, to predict action nomologically rather than statistically. This is often, in practice, relatively successful and, perhaps more importantly, it serves as a basis for achieving effective coordination in collective action via the communication of relevant belief-desire information. The scheme is often also applied, presumably by analogy, to other biological beings and even, sometimes, to inanimate objects and natural phenomena, evidently with more limited success.

Importantly, this methodology adds a novel type of uncertainty in future outcomes. Since beliefs and desires in others, although bounded by partly accessible personal histories, are inferred from observed behaviour – including communicative behaviour – the vector of inference is open to intentional manipulation. Agents may deliberately mislead. This introduces an array of relational concepts including trust, dishonesty, and deception, and what amounts to an evolving technology aimed at promoting compliance or anticipating non-compliance, the effectiveness of which is not guaranteed. I will discuss this issue at greater length in §6.6.

Cases in which prediction is of effects generated by human social interaction involve exceptionally high variability in predictive reliability. In conventionally structured social situations, patterns of interaction are often highly predictable by extrapolation over quite long intervals. In other words people, both individually and collectively, often tend to continue doing what they are doing. This is particularly so where there is an established system of roles and duties as discussed in §5.6. In cases involving authoritative decision making the actions of subordinates are often significantly predictable by reference to a prior authoritative proposal or instruction. In consequence, an authoritative decision maker generally has access to a more reliable predictive methodology than is available to an independent agent. But in independent action, or where an established system of roles and duties or authoritative decision making breaks down, extrapolatory prediction of others' actions is, for reasons to be discussed further in §6.6, generally much less reliable.

This discussion of predictive reliability has omitted one key issue, namely an agent's prediction of their own actions. This is a central point of controversy in the contrast between the Bayesian decision theories of Savage (1954) and Jeffery (1965), and it is a central topic in the theory of predictive processing as described in §4.7. In Savage's theory, and in most standard psychological theory, actions and consequences are treated as fundamentally different. For Savage, actions are mappings from states of the world to outcomes, and in most standard psychological theory actions, or behaviours, are the fundamental explananda whereas outcomes are otherwise-explained contingent effects. But for Jeffrey and advocates of predictive processing, an action is a particular kind of consequence to which the agent may assign a probability that depends on a current internal or intentional state. This considerably expands the range of cases for which an appropriate probability model can be constructed (Schwarz 2021, Hohwy 2013). The present proposal makes a similar assumption, but it focuses more explicitly on the way such predictions are generated and responded to. It assumes that an agent's predictions of their own actions do not generally involve either simple extrapolation of observed effects, or inference based on statistical probability. Instead they involve inference from internally available data relating to the assumed achievability of an envisaged option via an actually or hypothetically available method, given the currently expected cost and/or benefit associated with its envisaged realization by that method as outlined in §5.1, including a current level of commitment as outlined in §5.5. In short, it assumes an implicit inferential process approximating to the analysis of decision making proposed above. A similar process is not generally available for predicting the actions of others owing to a lack of data. It is typically replaced by the simplified scheme described on page 186 or by an assumption of role performance as in §5.6.

Evidently the process of generating predictions is complicated and imperfect and predictions are of varying reliability. But the level of reliability is itself to some extent predictable since it depends on the resources available and how they are used. Hence insofar as adaptation depends on maximizing predictive accuracy, a rational agent's established decision making processes can be expected to accommodate or compensate for this variable reliability where possible.

6.6 Secular Trends

Uncertainty about future conditions and events arises not only from local variability but also from long-term fluctuations. The latter are also to some extent predictable, and hence a well adapted action system may be expected to be adjusted to accommodate not only the implied uncertainty but also any discernible trends. I will discuss such trends in this and the next section.

As described in §5.6, cooperation, or partial cooperation, based on the realization within a group of constraints implicit in a shared system of roles and duties is ubiquitous. The aggregate advantage gained by cooperation is obviously immense. Virtually all the goods we enjoy depend on the approximate predictability of cooperation, both now and in the past. Indeed, the human species owes its biological dominance very largely to the evolution of such cooperation. However, cooperation depends on communication which, in the present state of the world, is generally open to falsification. Furthermore, the advantages gained by cooperation tend to be unequally shared. Almost always, some participants or onlookers can gain greater advantage by partial or superficial participation or by intermittent or continual subversion, and any process instituted to deter this requires cooperation and is itself open to the same effect. The overall result is a sort of arms race in which defection and/or subversion are frequently suppressed by cooperative action but continually re-emerge. Even if stability can be transiently established, exogenous events are likely to disrupt it.

This enduring instability typically drives the development of a soft technology of detection and deterrence involving the definition and allocation of roles and duties based on convention, moral principles, and law which is, however, only ever partly effective. It plausibly accounts for the way in which the dyadic value system described in §4.4 dominates social relations. For whilst it is plausible that many instances of defection and subversion are transient and opportunistic, such is the need to detect and deter these effects that the semi-permanent labelling of others as, in various degrees, cooperators or non-cooperators is selectively preferred as the most efficient and robust way currently available to maintain approximate stability within a generally cooperative group. The reduction in uncertainty generally outweighs the cost.

Despite the perceived irrationality of persistent instability and evidence that general levels of public disorder have decreased in well governed societies in the modern era (Pinker 2011), there is little evidence that it can easily be eliminated. The historical record shows relatively orderly intervals separated by spasms of extreme conflict. The proliferation in the modern era of increasingly destructive technology and the increase in competitive global interaction plausibly makes extreme conflict increasingly dangerous even if it occurs less frequently. It is tempting to assume in the intervals between conflict that a more consensual era has finally been reached (e.g. Macaulay 1848, Gooch 1911, Bell 1960, Fukuyama 1992). Experience unfortunately suggests otherwise.

Nevertheless, despite the frequency of conflict, secular change is to a large extent cumulative and irreversible. Hence the aggregate long-term effect is often less divergent than might be expected from extrapolation of short-term trends. The stabilizing effect is due to in part a process of negative feedback and adaptive selection that, as in biological evolution, tends to oppose cumulative maladaptation. A case that is of particular relevance in the current analysis arises in entrepreneurial market economics, as investigated in the classic work of Knight (1921).

Knight’s key insight is that profit is a consequence not of predictable probabilities but of uncertainty. Different entrepreneurial responses to uncertainty typically generate different levels of profit. Hence to the extent that profitability operates as a selective criterion it selects among competitors in favour of the successful management of uncertainty, however this is achieved. It does not, a priori, favour any particular methodology. It operates merely via relative success over a diverse set of attempts. A similar effect occurs in many other developmental processes where no a priori methodology guarantees success, for example in innovative science and technology. This provides a evolutionary vindication of predictive success as a selection principle. Historical experience of its effect plausibly accounts for the prevalence in some communities of an assumption that active amelioration of currently perceived problems, especially shared economic and epistemic problems, is generally achievable in the medium term – an attitude that might be described as broadly optimistic.

The principal problem in predicting long-term patterns of complex social interaction is not the simple extrapolation of current trends but the anticipation of rare but highly disruptive events. The underlying problem is that successful extrapolation based on assumed beliefs and desires or roles and duties cannot be long extended, that statistical prediction depends on an assumption of typicality and on the relevant type having sufficient associated statistical data, and that for large-scale grossly disruptive events the latter condition is almost never satisfied. The issue has been investigated most thoroughly, especially in economic affairs, for which there exist reasonably precise long-term data, by Taleb (2009, 2010, 2012, Taleb and Goldstein 2012). Recent events strongly confirm his contention that standard modelling based on extrapolation of contemporary trends grossly underestimates intertemporal instability. Whilst this bias cannot be precisely estimated or corrected, rational decision making ought, presumably, to attempt to mitigate its more significant effects. I will consider this issue again in Chapter 7.

6.7 Chaos, Entropy, Open Systems

Unpredictability is also an ever present feature of the physical world. As described in §3.5, causal underdetermination appears ubiquitous. Even where a system of deterministic laws can justifiably be assumed, the complexity of interactions and the sensitivity of dynamic systems to initial conditions means that the level of observational precision required for even approximate prediction over any moderate time interval frequently exceeds feasible limits. Many physical systems, termed chaotic, exhibit divergent patterns of behaviour from approximately identical initial conditions and hence allow effective prediction only of aggregate effects (Gleick 1988). And quantum mechanics introduces another source of uncertainty. For example, probabilistic interference patterns are observed in the inertial motion of even quite large molecules (e.g. Eibenberger et al. 2013, Cotter et al. 2017). Indeed, the Everettian or many-worlds interpretation of quantum mechanics – which, although controversial, is not easy to refute – characterizes the world as involving a continual bifurcation into apparently contradictory actual futures (Hughes 1989: 289-64, Wallace 2008: 39-52).

On the other hand, as described in §3.5, thermodynamic systems tend toward macroscopic uniformity via an increase in microscopic disorder. This produces a stabilizing physical effect oddly analogous to adaptation. Its effect is to enhance rather than reduce macroscopic predictability. For example, cosmologists, with some reservations, can confidently predict the ultimate heat death of the universe (Frautsci 1982). The tendency to decay is partly concealed but not reversed by evolutionary adaptation, for the latter is not a counterexample to the former. Net entropy does not decrease. It increases, but the excess is dumped in a vast highly disordered quasi-Malthusian residue. As Gould remarks, "The price of perfect design is … relentless slaughter" (1990: 8). Ultimately, disorder rules.

Just as the historical experience of general resilience grounded in evolutionary adaptation plausibly justifies a broadly optimistic attitude to the medium-term future, the historical experience of decay, and particularly of death, plausibly justifies an assumption that all active amelioration must eventually fail – an attitude that might be described as ultimately pessimistic. The former is expressed in the aphorism, 'While there's life there's hope'; the latter in, 'All things must pass'. The principle that the former may or must eventually give way to the latter is implicit in the experientially justified human evaluation of future probabilities and hence is a significant factor in long-term rational decision making.

Whilst chaos, thermodynamic decay, and other forms of local unpredictability occur somewhat predictably in closed systems there is another source of uncertainty that is more problematical. It arises from the fact that, ultimately, there are no closed systems. Wherever the boundary of a system is drawn, events outside it may disrupt its development. Most dramatically, because of the light cone structure implicit in spatiotemporal relativity, relatively simultaneous events are causally inaccessible to each other, yet either may causally effect the subsequent history of the other. As Rovelli is widely quoted as saying, “A strong burst of gravitational waves could come from the sky and knock down the rock of Gibraltar” (1997: 193). Given that gravitational waves propagate at the speed of light, this effect is likely to be entirely unpredictable to observers on the rock. Less radically, whatever boundary a decision maker implicitly places around the presumably relevant causal factors in any given situation, an event may occur outside that boundary that turns out to be significant. The approximate frequency of such events may perhaps be estimated but their precise impact is inherently unpredictable.

Hence in relation to long-term prospects four conclusions appear justified. One is that a direction of time can be defined relative to cosmological expansion. A second is that complex structure tends to decay, via an effect that leads, in any disruptive change, to increasing microscopic statistical uniformity. A third is that inherently unpredictable disruptive change is unavoidable. A fourth is that, nevertheless, in currently prevailing conditions on Earth, adaptation tends to produce continuing effective evolution of relatively well adapted structural forms. The result is that both a broadly optimistic and an ultimately pessimistic attitude to the future is plausibly justified.

6.8 Conclusion

In this chapter I have attempted to answer the third question on page 10, ‘How are probabilities fixed?’ In order to do so it has proved necessary to resolve a rather fundamental problem, namely to reconcile the usual intuitive and manifestly useful attribution of probabilities to envisaged conditions with the discovery that all relevant judgements of probability assume a reference class that is open to alternative modes of specification. The proposed solution is to assert that the system of probability judgements is shaped by the adaptive value of successful prediction, that prediction must often be made in the absence of some relevant information, and that in such cases the best policy is for the agent to assume that the case in question is typical of a class of similar cases for which there exist associated statistical data and to assign a probability distribution over envisaged alternatives that reflects the available data, subject to any relevant nomological assumptions available. A large number of conclusions follow from this proposal. One is that, except in rare cases such as radioactive decay where there is a uniquely preferred type, nothing answering to the usual notion of objective probability exists. This adds significantly to doubts expressed by, for example, Preyer and Siebert (2001: 12-16) and Hall (2004) about the viability of the Principal Principle – the principle that subjective probabilities ought to equal objective probabilities.

The main conclusion is that unreliability of prediction is ubiquitous and arises, partly predictably, from various sources. These include vague concepts, limited data, individual and group bias, evaluative bias, erroneous causal modelling, interpersonal deception, and physical indeterminacy or causal unpredictability. Within this envelope, two key issues are identifiable. One is that rare but more or less catastrophic forms of uncertainty may overwhelm prediction, especially prediction based on simple extrapolation of current trends. The other is that two kinds of apparently opposing long-term stabilization can be identified, namely adaptation and thermodynamic decay. These justify opposing attitudes that we may refer to as optimistic and pessimistic, both of which moderate sheer unpredictability but in opposite directions.

Having assembled responses to the three questions posed on page 10 the stage is now set to examine in detail how rational decision making depends, in various circumstances, on the variable futurity of envisaged consequences and, in particular, the extent to which this dependence entails an effect that answers to the notion of hyperbolic discounting.

Chapter 7 Analysis and Implications

7.1 Adaptation Revisited

This analysis of human decision making began by examining the standard approach of formal modelling employed in neoclassical and behavioural economics. The latter, explicitly or implicitly, assumes an ‘as if’ standard of representational adequacy, or what was previously called ‘saving the phenomena’. There appears to be no a priori reason why it should not succeed. Many precedents can be cited, including Kepler’s laws, the gas laws, and even quantum mechanics. A similar modelling methodology is used throughout applied science and engineering and in predicting aggregate effects in human behaviour, such a traffic management or purchasing behaviour. Nevertheless, as amply demonstrated in cited research it has not been conspicuously successful as an account of individual decision making.

The deficiency emerges most strongly in relation to the role of futurity. This is relatively easy to account for. Other factors in the choice situation are, at least in some sense, either observably present or generically defined. But futurity, ipso facto, applies only to envisaged possibilities. A theory or model that lacks a distinct category of envisaged possibilities has, strictly speaking, nothing to attach degrees of futurity to. It has to fake the connection by arbitrary modification of an effect that is, by assumption, already generically defined – for example, the utility function.

That this is a particular problem arising in the analysis of rational decision making is shown by comparing the three processes of adaptive evolution that bear, directly or indirectly, on human behaviour – biological, cultural, and cognitive. A concept of futurity does not, even intuitively, enter into either of the first two, except insofar as effects are mediated by the third – as in planned animal breeding or advertising strategies. Natural biological evolution, for example, is entirely backward looking. Nothing answers to the notion of envisaged possibilities.

The difference arises because these three processes differ not only in their typical velocity but, more fundamentally, in the selective effect that gives direction to each. The discriminator in biological evolution is hereditary survival. The discriminator in cultural evolution is social acceptance. The discriminator in cognitive evolution is accurate prediction. In each, an underlying consequentialist principle is realized by a process of selection among persistent but varying methodological components on a trial and error basis, if at differing rates. But a specific notion of envisaged possibilities, and hence futurity, is essential in the third because it is a necessary ingredient of the notion of accurate prediction. Even cases of delusion or paranoia depend, ultimately, on a notion of erroneous prediction.

To reiterate, according to the theory proposed here it is the adaptive selection of decision making methodology on the basis of successful prediction that drives the emergence of what is standardly identified as rationality. The latter is grounded in and hence operates within an envelope defined by biological and cultural evolution over the more or less recent past but it is nomologically separate from them. It contributes to overall adaptation on the basis that it frequently gives agents a significant adaptive advantage in virtue of facilitating a very rapid and, for the most part, increasingly effective response to environmental contingencies as judged against current rational, cultural, and ultimately biological criteria.

7.2 Problems and Issues

In this scheme, decision making is interpreted as part of a system of predictively successful adaptation. The product of this process is a partly hereditable system of embodied, mostly intuitive, heuristic methods, existing within a broader cognitive system that admits tolerably reliable processes of, for example, sensory discrimination, concept formation, memory storage and retrieval, inductive and deductive inference, and fluent motor control. These are, we know, realized in an evolved neurophysiological mechanism. But they are analysable for present purposes in terms of functional features and patterns of processing in a way that approximates, very roughly, to the explanatory analysis of the operation of a conventional digital computer in terms of program features and associated patterns of contingent processing rather than in terms of either machine code or electronics – or, similarly, of biological evolution in terms of phenotypes rather than genomics or biochemistry.

At this level, decision making methodology can be seen as an evolved and evolving response to three characteristic problems – value attribution, option selection, and probability assignment. However, the quality of the response in terms of, for example, its speed and efficiency typically depends not only on current methodology but also on certain broad features of the agent's situation that differentially facilitate or inhibit its effectiveness in situ. For example, catastrophic change inhibits effective processing. For reasons to be explained I will identify these broad features as 'issues'. This reflects an underlying principle that, in general, the tendency of evolutionary selection is to modify differential effectiveness not uniformly but within the constraints imposed by currently prevailing issues.

There is a significant analytical difference between problems and issues. Problems are approximately equivalent to what are usually called needs, except that needs are generally attributed contingently to particular bearers whereas problems are a priori. Issues, on the other hand, are facts about the situation. Catastrophic change may occur. A methodology of, for example, value attribution must handle it in some way, or else agency fails. Crucially, some issues are characteristically time dependent. For reasons that will become apparent it is these that I shall be primarily concerned with.

I am not in a position to provide a complete catalogue of issues that significantly affect the evolution of human decision making methodology. To do so would require considerable extra research. Rather, I will identify six that appear on initial inspection to be of wide significance and discuss the impact of each on the effectiveness of the existing intuitive methodology in relation to each of the three core problems identified and the consequent bias, or apparent bias, in decision making induced. To set the scene I will start with one that appears not to be characteristically time varying.

1 Descriptive complexity

Value attribution, option selection, and probability assignment all require the descriptive characterization of relevant features of the world. Indeed, descriptive characterization is, perhaps, not exclusive to organisms having a decision making capacity of the type considered here. Hence it ought, perhaps, to be added to the list of more general biological or quasi-biological problems set out in §3.3 – namely, nutrition, excretion, escaping predation, infection control, sexual contact, dispersal of progeny, and communicating with allies. However, I will not at present extend the analysis in this direction.

Relevant features of the world are of enormous variety. It seems that the evolved methodology used in human decision making is dyadic, in that it involves identifying and classifying apparently distinct objects or other individuated loci and attributing one or more variable, often quantifiable, qualities to each. This creates a combinatorial system that is able to handle arbitrarily complex patterns systematically, as reflected in subject-predicate structure of language and in the vector representation commonly used in science and data management. In decision making it is exploited variously in the methodology associated with each of the three core problem – particularly in the assignment of affinity and salience in valuation, in combinatorial option description, and in the accumulation of statistical data and derived probabilities by type.

It is perhaps worth remarking that despite its evident effectiveness in human decision making we cannot be certain that this type of dyadic structure is universally appropriate to the characterization of actual effects. A Kantian argument from intuitive usage is not compelling, and more recent scientific developments, including the adoption of tensor rather than vector calculus in general relativity and sum-over-histories in quantum electrodynamics suggests that there are grounds for doubt. Nevertheless, its effectiveness in human decision making is unquestionable.

Evidently, the combinatorial complexity of encountered decision making situations as so described varies markedly. Most games, for example, are deliberately contrived to admit a sufficient description using only a quite limited set of agreed constructs, and most laboratory research is conducted on what are, so far as possible, closed systems in order that extraneous descriptive features can be ignored. Given this simplification, decision making is likely to be especially effective in these cases. It is rational, therefore, for agents to seek out or construct similarly simplified situations, where wider consideration of value allow. Such action is typically motivated by the negative salience generally attached to descriptive uncertainty. The trend towards rule-based social organization can be explained partly on this basis. The differential effect of varying combinatorial complexity is not generally futurity-dependent – except, perhaps, insofar as agents envisage an impending descriptive transformation via, for example, the emergence of a political or religious utopia. Hence, with these unusual exceptions, it is not a generally significant factor in futurity related bias.

2 Internal resources

A second significant issue is variation in internal resources. Such variation occurs most generally in the accumulation of descriptive resources. In value attribution it occurs in adjustments to assigned levels of affinity and salience and in the addition of learned proxies. In option selection it arises in gain or loss of skill or fluency and in the extension or consolidation of learned methods. And in probability assignment it arises in the accumulation of statistical data and the development of significant nomology.

All of these are, to various degrees, typically futurity dependent but the effect varies markedly with age. Levels of skill and the repertoire of learned methods and accumulated data are generally increasing, but rates of change are greatest in young people. In later life levels are more often static or decreasing. Hence it is rational for young people, in particular, to delay long-term decisions and, provided that relevant community values do not diverge significantly, to transfer short-term decision making and agency to others vicariously. By contrast, as remarked in §4.3, there seems to be little evidence that agents of any age anticipate and adjust to probable future value change – nor, perhaps, is there any rational justification for doing so. Hence value change, in itself, entails little futurity bias.

3 External resources

A third issue is variation in external resources. In value attribution this mainly involves intertemporal variation in current reference levels. Since valuation adjusts to accommodate revised reference levels, the effect of this is difficult to anticipate and may sometimes appear paradoxical. For example, increasing community affluence may lead to divergent levels of satisfaction. Hard times may be remembered fondly. The colostomy example described on page 109 is instructive. Hence there is no clear futurity bias in value attribution arising from variation in external resources.

Variation in external resources is much more significant in option selection, most notably in already satisfied prerequisite conditions. The discovery of already satisfied prerequisite conditions tends to vindicate achievability, abbreviate planning, and reduce the need for additional action to satisfy such conditions, with the attendant risk of unanticipated costs. This produces, in most circumstances, a very strong bias in favour of pursuing an acceptable option promptly, especially insofar as the continuing satisfaction of relevant prerequisite conditions cannot be relied on. It is, for example, obviously rational to get on the bus when it arrives rather than waiting for even a minute or two.

In probability assignment, the most important external resource subject to significant variation is the supply of public data. Since the supply of primary statistical data is limited by personal experience and small data sets create a problem of sampling error, a supply of public data is valuable. Predictably, a very considerable effort goes into collecting it. Hence an agent may reasonably assume that more data will be available in the future and delay decision making in the meantime.

4 Uncertainty

A fourth issue arises in variable sources of uncertainty. In value attribution this is chiefly a matter of obscure or initially hidden conditions. Since such conditions may have significant short-term value consequences, rapid detection or disambiguation is often valuable. This tends to lead to a proliferation of proxies and consequent short-term double counting, and urgent epistemic activity. The result is often a very marked immediacy bias.

In option selection there is a similar issue concerning prerequisite conditions and unanticipated consequences. Insofar as the selection process is itself urgent, this urgency is exacerbated. An opposite delaying effect arises as a result of uncertainty in the time-to-completion of a chosen but still pending or discontinuous course of action directed towards a non-urgent outcome. The effect is strongest if adventitious satisfaction of prerequisite conditions is a possibility or a deadline is not clearly fixed or other motivated actions may intervene. The result may be to delay the start or continuation of relevant activity so as to leave a minimum apparently feasible completion trajectory. In the extreme it may entail unresolved procrastination, in which the uncertainty-weighted marginal benefit of relevant action never exceeds that of concurrent alternatives.

In probability assignment there is a very specific uncertainty effect described on page 182, namely that uncertainty increases as the quantity of relevant information about a case increases, due to a lack of specifically relevant statistical data. So, for example, as a presumably advantageous occurrence approaches and is anticipated in increasing detail, nervousness increases as the range of more positive and negative possible realizations becomes less probabilistically constrained by available data. This creates a temporary negatively valued condition that entails a paradoxical decline in the net value of the relevant option in the interim.

5 Cooperation

A fifth issue arises in variable cooperation. Most human action involves a cooperative element grounded in affinity and culturally given resources, but the level of relevant cooperation is not fixed. In value attribution the key issue is reciprocity. This tends to amplify the degree of cooperation and hence shared valuation. It is not reliably futurity-dependent, except insofar as it is anticipated by extrapolation.

In option selection and probability assignment, the availability of cooperative methods, especially based on shared assumptions about roles and duties, increases the scope and local predictability of actions and outcomes whilst, ipso facto, limiting individual choice. Hence, insofar as modern society has tended to become increasingly cooperative, the differential effect of variable futurity on decision making is, except where conventionally specified timing prevails, much reduced.

6 Secular trends

Finally, there is a more general issue of rationally expected secular trends. In value attribution, an expected trend in reference levels may lead to re-evaluation of expected long-term outcomes. In option selection, expected trends in technological innovation and in economic, institutional, and organizational change may radically modify the options likely to be available. And in probability assignment, two significant effects can be inferred – that simple extrapolation of current trends is likely to become increasingly inaccurate and that, nevertheless, in the long term, as described in §6.7, two contrasting trends, arising from assumptions of either continuing adaptation or thermodynamic equilibration, justify contrasting attitudes that can be described, generically, as optimistic and pessimistic.

This brief survey of the interaction of problems and issues in decision making is not intended to be either comprehensive or formally precise. To make it so would require a body of research considerably beyond the current project. Nevertheless it gives, I hope, some indication of the complexity of the topic and an indication of the more significant explanatory relationships involved. It shows a number of ways in which envisaged futurity affects the typical pattern of human decision making. I will examine these patterns more formally in the following two sections.

7.3 Futurity Effects

The preceding analysis has identified several distinguishable types of bias that may arise in decision making owing to issues typically involving significant variation with respect to futurity. These biasing effects are general, being inherent in the constraints that modulate any methodology approximately like that which humans intuitively use, as described in Chapters 4, 5, and 6. They are, in this sense, not products of the particular heuristic methods that happen to have emerged but features or inherent consequences of the design brief that these heuristic methods have evolved to satisfy. They can be circumvented without loss of adaptive benefit only by adjustments in prevailing circumstances or by the adoption of a radically different methodology. I will discuss this possibility in §7.6.

1 Promptness

One key type of bias involves a choice being made and any immediately implied action being performed as soon as relevant prerequisite conditions are satisfied. I will call this promptness. It realizes, inter alia, what Ericson and Laibson (2018) identify as present-focused preferences and what is described in (A14) as the immediacy effect, except in that it may extend backwards into planning. For example, a sprinter may not merely push off from the blocks as soon as the starting pistol fires but plan to do so.

Its ultimate basis lies in the often transient avoidability of evolutionary risks in the encountered environment and, equivalently, in the transient availability of opportunities, exacerbated in both cases by imperfect information. In value attribution the latter leads to a characteristic problem in detecting hidden valued conditions, the solution to which is the establishment of a wide variety of independently valued proxies, as described in §4.6. Since in many cases there are several relevant proxies variously correlated, this leads to a widespread problem of double counting. The typical effect is to produce an initial over-valuation of hidden or partly hidden conditions – characterized by Loewenstein as visceral – and hence an exaggerated promptness bias. Evolved heuristics may alleviate or compensate for this effect but they cannot eliminate it without some loss of adaptive functionality.

Promptness is also exhibited in option selection in that the satisfaction of conditions vindicating achievability may be similarly transient and hence need to be exploited promptly. In other words, its is rational to select an option, plus an associated method, as soon as information is available indicating that relevant prerequisite conditions are satisfied and, indeed, to act promptly during planning to secure such information. It includes, for example, securing access to a suitable system of interpersonal cooperation.

2 Deliberation

A second effect is one in which a choice, and any immediately implied action, may be delayed within an extended interval without the total expected value of remaining alternatives being significantly reduced. This is, in a sense, not a bias, but it can easily appear so, especially by contrast to promptness. Its basis is that, during planning, several methods and/or scheduling arrangements may appear to be available as alternative means to realize a currently preferred option. For example, there may be several methods of getting access to a book, such as buying a copy in a bookshop, buying a copy on line, borrowing it from a library, buying an e-book, reading it on line, and so on. And, similarly, it may be done today, or tomorrow, or at some other time. Each method and/or schedule depends on the satisfaction of various prerequisite conditions. Given that the satisfaction of any of these conditions, unless they are already satisfied, may involve some partly unpredictable costs, the plan with the highest expected value at any time may be one that involves waiting to find out how events unfold. In extreme cases the original option may be achieved by other means or it may be rendered irrelevant by intervening events. For example, I may find that I already have the book or that the task for which I needed it has been cancelled. Deliberation is, therefore, a strategy to avoid wasted effort. Its extreme equivalent in industry is just-in-time methodology.

Although, in principle, deliberation is merely an optimization strategy, in some cases it may involve serious risks and hence is more obviously a kind of bias. For example, in just-in-time manufacturing the disruption of a relatively trivial part of the supply chain can bring the entire manufacturing process to a halt. In ordinary human decision making, deliberation can take the form of procrastination. The implicit risk is that the tacit move to avoid wasted effort may result, via prediction error, in executive failure.

Procrastination is widely interpreted as a result of akrasia. This, however, rests on an implicit folk theory that interprets decision making as an exercise of ‘will power’, the assumed motivational mechanism of which is, at least, opaque. In the theory assumed here, the intuitive impression cited as evidence of the exercise of will power is explained as arising from awareness of the motivating effect arising from the value attributed to commitment, as discussed in §5.5. Procrastination, on this basis, is an exceeding of commitment by other contrary value considerations. In answer to the question posed by Ericson and Laibson (2018) of why people typically underestimate their own tendency to procrastination, two possibly overlapping answers are available. One is that in generalizing from a large number of relatively trivial past choices they fail to recognize the rather moderate value attributed to commitment. The other is that they fail to anticipate the transient effects of double counting. The latter, in situ, is often interpreted intuitively as temptation. Insofar as it is predicted to exceed commitment it can generally be defeated only by some external constraint, as in the story of Ulysses and the Sirens (Elster 1977).

3 Second thoughts

A third effect involves the same initial over-valuation of hidden or partly hidden conditions as often underpins promptness but it also involves a period of forced delay before a positive decision can become effective – before or instead of an explicit or implicit deadline. During the period of delay the initial over-valuation may be diluted by additional considerations including, for example, a reduction of some double counting – including of visceral effects – or a recognition of additional prerequisites or possible consequences. In cases in which the initial valuation is sufficient to entail a prompt positive decision the extent of the subsequent dilution may be such as to outweigh the resulting commitment, leading to a reversal of the decision before it becomes effective. This is commonly recognized as 'having second thoughts'. Such a delay, or cooling-off period, may be deliberately incorporated into the decision making context to facilitate this effect and hence to reduce promptness bias in cases where it is recognized as potentially dysfunctional. The constraint reputedly imposed on Ulysses is of this type, in that it prevents his expected decision to visit the Sirens' island from being promptly effective.

4 Transitional uncertainty

A fourth type of bias is described in §7.2 and originally in §6.4. It is a side effect of the reference class problem. As described in §6.3, the assignment of probabilities to envisaged future conditions and events generally depends on the assumption that some relevant object or event is a typical member of a particular class. As more information becomes available, the appropriate class becomes increasingly specific. But the quantity of statistical data associated with a class generally declines with increasing specificity and hence predicable sampling error increases. The surprising effect is that as more information is gathered about a prospective case, probability estimates tend to become increasingly uncertain. Since uncertainty is, for good epistemic reasons, negatively valued, during an interval in which information approaches a maximum, an option typically becomes apparently less positively valued. Receiving test results is a typical example. It may produce an effect colloquially described as ‘chickening out’.

It is worth noting that this effect occurs only when prediction of future conditions or events depends on typicality-based estimated probability rather than on either simple extrapolation of current trends or nomology. The former tends to reinforce promptness, given that its implications become increasingly uncertain with increasing futurity. The latter, after allowing for uncertainty in measuring boundary conditions, is largely independent of futurity although strongly dependent on the theory’s validity. Hence the decline in preferability due to transitional uncertainty is felt chiefly in nomologically non-standard departures from the default future.

5 Forced choice, delayed benefit

A fifth effect occurs in many of the experiments described in the Appendix and in many ordinary economic transactions. It involves the promise of a future outcome contingent on an earlier expression of choice. The connection is often, but need not be, mediated by social convention. Planting seed, for example, offers the promise of a future crop. The key feature is that the interval is non-negotiable. Many preparatory activities can be interpreted as realizing this pattern.

Typically, the motivating value of the envisaged outcome at the point of choice is not equal to the value at the point of promised realization. At least, a period of uncertainty, perhaps transitional uncertainty, usually intervenes. If the promised realization occurs, the earlier motivating value will usually be lower, given that it is attenuated by uncertainty. If the promised realization fails to occur – if the promise was, in effect, fraudulent – the earlier motivating value will be higher. An agent choosing accordingly will be disappointed.

6 Long-term projection

A sixth effect that may perhaps be classified as a type of bias concerns the prediction of broadly successful amelioration of perceived problems in the medium-term and long-term future. As described, there are reasons to adopt each of two generic attitudes that I have termed optimistic and pessimistic. The effect of the former is to set, by default, a generally positive probable upper bound on the ameliorative effect of envisaged action. The effect of the latter is to reduce this upper bound to zero. A plausible resolution is to package the optimistic within the pessimistic – that in the end amelioration will be impossible but in the meantime problems generally can be solved. This places a premium on preserving the adaptive system and circumvents the otherwise intractable problem of computing infinitely remote outcomes.

7 Cooperation

Finally, there is an effect that, although not even loosely a type of intertemporal bias, has a sufficiently widespread effect to warrant special mention – namely social cooperation or, in its more extreme form, vicarious achievement. Its significance lies in the fact that it changes radically the basis on which other factors, especially options, are evaluated and hence it changes envisaged time horizons.

One issue discussed extensively in §5.6 is that an established system of roles and duties both expands and systematizes the range of options typically available to an agent and tends to make temporally remote outcomes more predictable. On the other hand, divergent or extraneous values among participants typically lead to some degree of non-cooperation. Since cooperation generally depends on reciprocity there is an ongoing possibility of an increase or decrease in cooperation that may be envisaged, in the extreme, as ending in a state of anarchy, or of political or religious utopia. Agents may plan to promote or exploit such outcomes. Conversely, the experience of a rigid or permanently supportive social context, in which outcomes are not dependent on personal choice, may typically lead to institutional dependency in which all but the most short-term outcome-led planning atrophies.

7.4 Aggregate Analysis

It is now possible to return to the original question posed on page 9 – To what extent if any does, or should, the relative value of the possible consequences of a course of action depend on their relative futurity? But by now it is apparent that the answer to either the descriptive or the prescriptive version of this question is complicated. There are, at least, several different futurity-dependent effects that need to be admitted. Since these effects may occur in various combinations we need a method of mapping out the motivational implications of each through time in a way that admits aggregation.

A convenient place to start is to reconsider Figure 3, reproduced here:

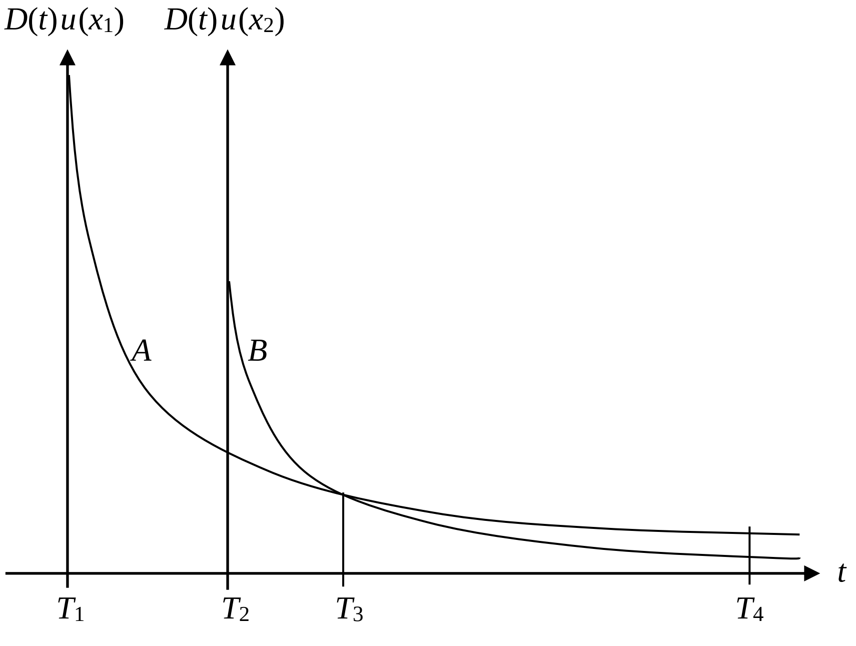

Figure 3: Preference reversal (copied from §2.2)

This figure is normally interpreted, as in Loewenstein and Elster (1992), as showing the relative motivational force of two alternative outcomes consequent on a choice at some time t of either of two alternative options. However, in this respect it has some awkward features. Firstly, contrary to the usual convention, time runs from right to left. Secondly, it is unclear what interpretation attaches to the discontinuities at T_1 and T_2. In the preceding Figure 2 on page 51 the equivalent discontinuity represents value attributed *now*, and the null section to the left represents the irrelevant past. A similar interpretation is available in Figure 3 only if each outcome is assumed to be a completed event of zero duration at a precisely predicted moment. Thirdly, it is unclear whether any intertemporal variation in probability is allowed for. If it is, it must be incorporated into the ordinate magnitudes since they are interpreted as representing motivational force, but this is not made explicit in their derivation from an assumption of hyperbolic discounting. This is the worry behind much of the research cited on page 22. And finally, perhaps most seriously, choice is modelled as if it were between outcomes whereas it is actually between options. It is on this account that no consideration is given to motivational effects arising from the possibility of predictable collateral costs or benefits or from satisfied or unsatisfied prerequisite conditions.

It is possible to resolve these difficulties in a modified representation, as in Figure 12. I will first describe a generic case and then show how it can be adapted to suit various different futurity-dependent effects as described in §7.3 and, ultimately, how aggregate implications are derivable.

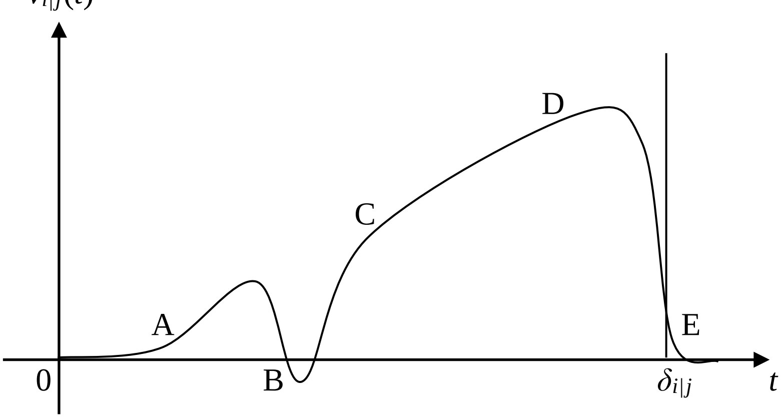

Figure 12: Generic graph of outcome value against futurity

Figure 12 shows a notional relationship between the effective value $v_{i|j}(t)$ of a beneficial outcome *i* under option *j* chosen at time *t* and *t*. I will call $v_{i|j}(t)$ the 'outcome value' of *i* under *j* at *t*. Measured time runs from left to right, as usual. $t = 0$ is fixed as convenient, generally at or before the start of the interval under consideration. An option *j* is a possible action of some sort that may be chosen in preference to the current default, as described in Chapter 5. An outcome *i* is a distinguishable differential consequence of choosing *j* rather than the current default. The outcome value $v_{i|j}(t)$ of outcome *i* under option *j* at time *t* is the value attributable at *t* to all currently assumed differential features attributable to *i*, as described in Chapters 4 and 5, including specifically associated collateral costs and benefits, attenuated by currently assigned probability, as described in Chapter 6. Outcome *i* is beneficial under option *j* insofar as $v_{i|j}(t)$ is generally positive. It may involve either the realization of a condition not expected by default or the prevention or mitigation of a condition expected by default, including by simple extrapolation. Both are accounted for similarly. As a measure of motivational force, $v_{i|j}(t)$ corresponds almost exactly to the notion of expected utility except that its relativity to a default future is explicit and its variability with time is not only admitted but is the focus of analysis.

In Figure 12 a number of typical features of this variable relationship are shown. It is assumed that at $t = 0$ the possibility of either *i* or *j* has not yet arisen. At A, *i* is conjectured to be both beneficial and perhaps achievable under evolving option *j*. At B a serious objection is discovered, sufficient that choosing *j* becomes, in terms of *i*, apparently disadvantageous. At C the objection is resolved. Between C and D various prerequisite conditions are resolved such that achievability without significant additional cost is increasingly probable. $t = \delta_{i|j}$ represents the deadline for realizing *i* via *j*. Subsequently, *i* is not realizable via *j*. Hence from E onwards $v_{i|j}(t) = 0$. The non-infinite gradient near E represents uncertainty concerning the precise location of the deadline. In other cases this may be either more or less steep.

An interesting feature of the above analysis is its vindication of the notion of a deadline. This, in part, answers the question about the discontinuities at T_1 and T_2 in Figure 3. In many decision making cases there is a point beyond which an envisaged option is ineffective and hence beyond which the outcome condition, whatever it is, contributes no motivational value to that option. It is important to notice that this effect is option dependent. Missing the bus, for example, is usually of negative value. But once it has occurred, what follows cannot modify the motivation to catch the bus. All the relevant motivational force is packaged up, in anticipation, up to the point of either catching the bus or missing it. The motivational force inherent in the missed-it condition contributes to a different set of options, such as getting a taxi or having a drink with a friend who appears by chance. This packaging-up is a very significant feature of human valuation in general, indicated in the colostomy example described on page 109, and it plausibly accounts, in part, for the objection to quantifying utility by integration, as illustrated in §2.2. It is a key aspect of the temporal granularity of units of action discussed in §4.3.

Figure 12 is a generic illustration. Of greater theoretical interest are the patterns associated with each of the ubiquitous effects described in §7.3. Figure 13 illustrates promptness:

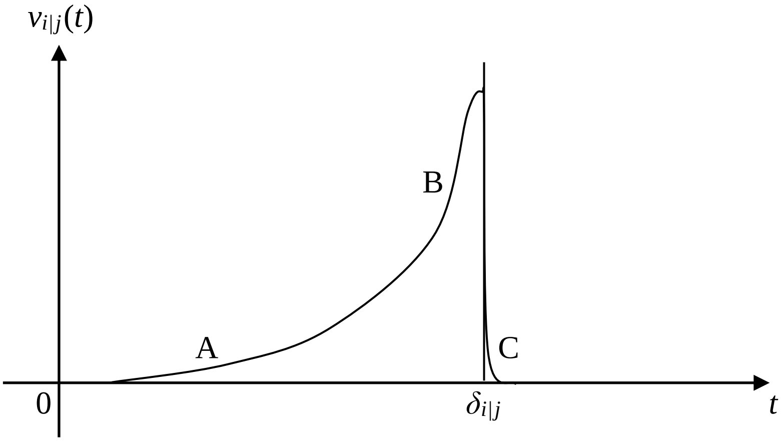

Figure 13: Promptness

Here, as in Figure 12, at $t=0$ the possibility of either i or j has nor yet arisen. At A it is envisaged, but indications of either risk or opportunity are slight or ambiguous. At B evidence is increasing, perhaps supplemented by double counting. At C it is too late, either because the beneficial opportunity is lost or because it was taken and hence is no longer an option. The latter involves virtual deadline, one that is fixed by the act of choosing.

Figure 14 illustrates deliberation:

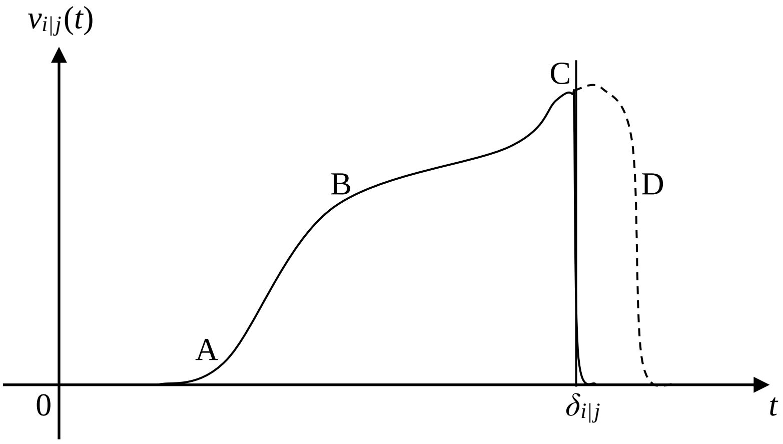

Figure 14: Deliberation

Here again, $i|j$ is envisaged at A. Its feasibility is well established at B but there is no urgency to choose. However, if not already chosen it becomes urgent at C as the deadline approaches. At D it is still highly desirable, but too late. The curve is accordingly shown dashed.

Figure 15 illustrates second thoughts:

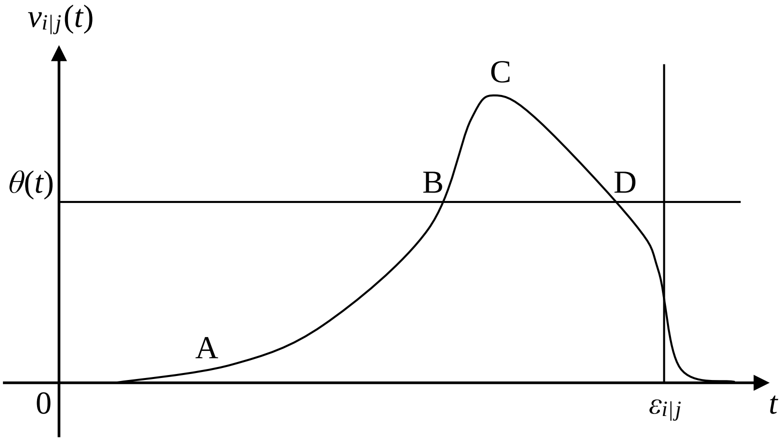

Figure 15: Second thoughts

Again, $i|j$ is envisaged at A. Its outcome value increases rapidly, as in Figure 13, so as to exceed the current threshold, represented notionally by $\theta(t)$, at B. This results in a prompt but not immediately effective choice to pursue $i|j$ at B. $t = \varepsilon_{i|j}$ represents the moment of delayed effectiveness. Between C and D, second thoughts arise such that the outcome value at D, despite being augmented by the value attributed to commitment, falls below the current threshold – or, in more extreme cases, below zero. Hence the choice is reversed at D.

Figure 16 illustrates transitional uncertainty:

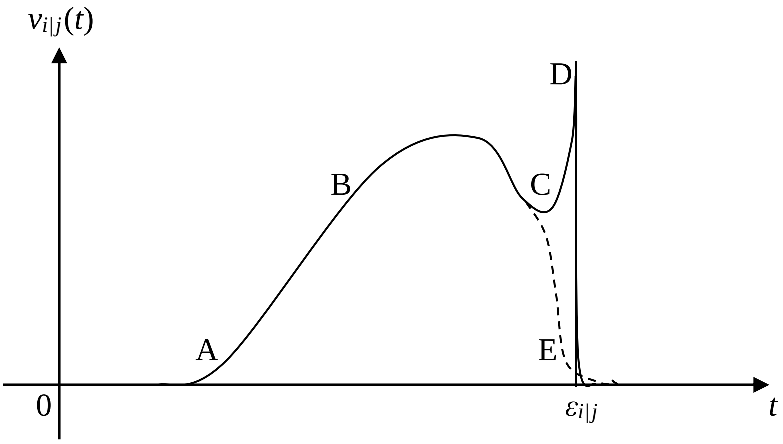

Figure 16: Transitional uncertainty

In this figure, transitional uncertainty occurs at C. Thereafter, achievability may be vindicated, as indicated at D, or the uncertainty at C may prove well founded, as indicated at E.

Figure 17, below, illustrates forced choice, delayed benefit. In this figure, $t = \chi_{i|j}$ represents the moment of forced choice whereas $t = \varepsilon_{i|j}$ represents the moment of delayed effect. $v_{i|j}$ between B and C represents the outcome value of a conceivably delayed or revised choice. $v_{i|j}$ at B depends on $v_{i|j}$ at C but is depleted by intervening uncertainty. The promise may be vindicated as at C or broken as at D.

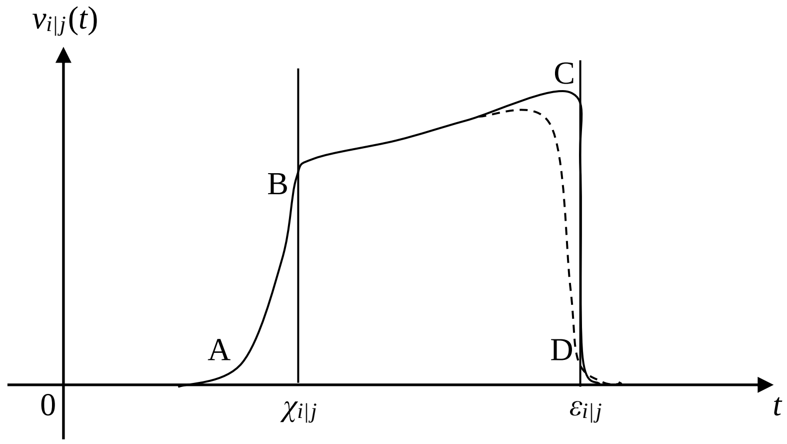

Figure 17: Forced choice, delayed benefit

Finally, Figure 18 illustrates long-term projection:

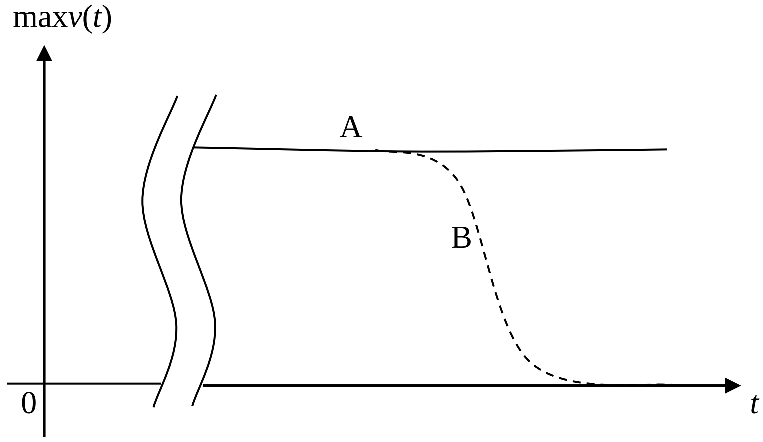

Figure 18: Long-term projection

In this figure the vertical axis represents a generic upper bound on beneficial outcome value over possible options. It illustrates, very approximately, the idea that a generally optimistic assumption of successful long-term adaptation, at A, is terminated at B by a collapse into disorder.

There appears to be no characteristic intertemporal effect associated with cooperation but, in general, so long as cooperation is predictable, time horizons are extended and uncertainty decreased. This is partly reflected in the relatively low rate of increase in outcome value between B and C in Figure 17 and perhaps also in Figure 14, and the survival of optimism to A in Figure 18. Cooperation also typically creates extra options.

These graphs are intended to show, approximately, the relationship between outcome value and futurity in several characteristic cases. The theory requires two additional assumptions mirroring those of expected utility theory: that the total value $\phi_j(t)$ of option j at time t is equal to the sum over all relevant outcomes, $\overset{i}{\Sigma} v_{i|j}(t)$, and that option j is chosen at time t only if $\phi_j(t)$ is greater than $\phi_{j'}(t)$ for every other envisaged option j' at t, including the current default – for which $\phi(t) = 0$. The theory is predictively complete if it is assumed that option j is chosen at t if, in addition, $\phi_j(t)$ exceeds a current threshold $\theta(t)$ that varies with felt urgency and that transient fluctuation dynamically eliminates tied values.

7.5 Hyperbolic Form

It is clear that even if the foregoing analysis is only very approximately correct there are grounds to doubt that there is a straightforward general relationship between motivational value and futurity, either in detail or as a broad idealization, and either descriptively or prescriptively, except at most in a very limited class of cases such as financial investment decisions. This conclusion is, in reality, not a surprise. It is implicit in the very wide variety of experimental and observational results referred to above and described in the Appendix and elsewhere in the literature. The persistence of models that assume a simple relationship seems to have more to do with the attractions of conceptual simplicity and mathematical tractability than with any known empirical justification. This is understandable but ought to be deprecated.

However, it does not follow, as appears to be assumed by many theorists, especially in neuroeconomics, that only an analysis formulated in terms of particular heuristic procedures or neurophysiological mechanisms – a constructive rather than a principle theory – is viable. A principle theory cannot be viable if it significantly contradicts what is procedurally possible but, as the above analysis shows, a principle theory based on assumed adaptation to generic evolutionary problems is not out of the question. Such a theory has the enormous philosophical merit, if well founded, of illuminating key questions not only of observed human choice and action but also of rationality, valuation, and epistemology. It is the focus on futurity, and hence on diachronic structure including hyperbolicity, that clarifies the viability of such an account.

The hyperbolic and quasi-hyperbolic forms discussed in Chapters 1 and 2 and widely elsewhere in the literature are not generally assumed, a priori, to have to satisfy a strict mathematical form. Their key characteristic, as illustrated in Figure 2, is that whilst both hyperbolic and exponential forms decrease monotonically with futurity – and therefore increase with proximity – the hyperbolic form is significantly more convex. It is this excess convexity that gives rise to the ordinal contrast illustrated in Figure 3 and hence the inference to preference reversal.

However, there is a complication in the standard derivation. It is that the inference to preference reversal depends on an inference to expected utility, not utility, and hence it requires an additional assumption of constant probability. If this assumption is relaxed the overall conclusion remains but without the implication that it arises as a result of value discounting. Other effects, including not only varying probability but also varying uncertainty, discovered costs, discovered benefits, and satisfied prerequisites may intervene. This defeats any immediate inference of irrationality.

This is the route I have followed. The conclusions are set out in §7.4. In these, hyperbolic segments are ubiquitous, most notably in Figures 13 and 15. Hence apparent preference reversal is to be expected very widely with no implication of inconsistency. For example, if during some ongoing or planned activity a prospective condition requiring a prompt remedy arises, relevant remedial action will typically interrupt it. The effect is illustrated in Figure 19. Here, at A and B, deliberation regarding option *j* as a response to prospective outcome *i* dominates. But at C, consideration of a possible response *j´* to the more urgent prospective outcome *i´* intervenes. If the current threshold $\theta(t)$ is exceeded, interruption will occur at C. This may but need not ultimately defeat *j*.

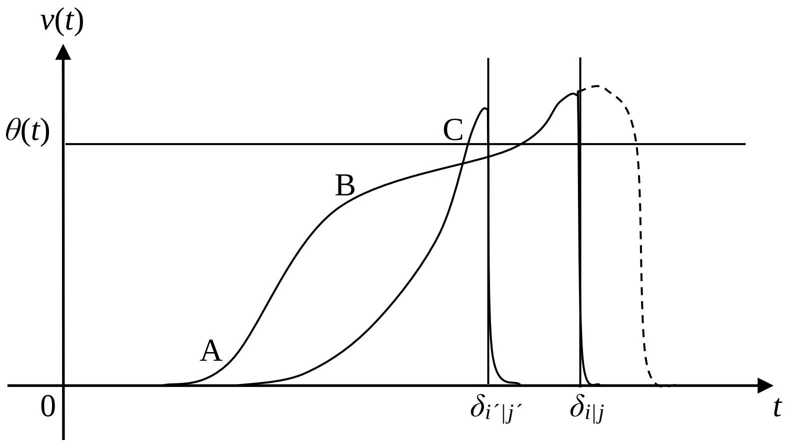

Figure 19: Deliberation versus promptness

Many similar cases can be expected. The ubiquity of hyperbolic segments can be accounted for by the observation that the typical upswing in outcome value prior to a choice depends more on the discovery or contingent satisfaction of prerequisite conditions, which accelerates with directed attention, directed action, and other local circumstances, rather than on varying estimates of probability. It is a consequence, largely, of the agent being active rather than passive.

There is, however, the possibility of a more nearly exponential profile, as illustrated, after any necessary disaggregation, in the segment between B and C in figure 16 – that is, between the occasions of forced choice and delayed benefit – in which little relevant occurs other than speculation concerning the likelihood of failure. It is plausible that, ceteris paribus, the intervening uncertainty is approximately proportional to the delay, which would entail an exponential trajectory, as in bank interest rates. Interestingly, this pattern – of forced choice, delayed benefit – is the typical format used in experimental economics as illustrated in the Appendix and is otherwise chiefly characteristic of economic transactions in a predictably cooperative social context. It is, then, not surprising that an exponential model would be attractive to economists and, after allowing for other effects such as risk aversion, is approximately confirmed in many standard experiments.

Conversely, the above analysis suggests that preference reversal is not a result only of hyperbolicity. It follows more immediately, as illustrated in Figure 16, from the transitional uncertainty that may arise in an evolving option but not in the corresponding default, where the latter is derived by extrapolation rather than statistical inference. It is a result, loosely speaking, of proximal anxiety rather than temptation, in which the security of the relatively abstract is replaced by sharply distinguished immediate alternatives. It may, for example, be difficult to enter any one of a series of unfamiliar expensive-looking restaurants despite the settled intention to do so because each, when approached, generates an uncomfortable degree of uncertainty as to the outcome.

7.6 Dynamic Normativity

This observation raises once again the question of the relationship between evolved methodology and rationally justified normative standards. Several conclusions can be drawn.

Evolved methods are criticizable on rational grounds. The generic rational standard is predictive accuracy. Action is inherently predictive in that its business is to realize envisaged conditions or events. A methodology may, in a given class of contexts, be more or less regularly successful in achieving predicted outcomes regardless of the value attributed by the agent or others to the conditions or events realized. Hence, regardless of whether a methodology is biologically, culturally, or cognitively evolved, it is criticizable on grounds of its being insufficiently successful, either in general or in some relevant class of contexts. For example, the methodology that results in a failure to enter restaurants is criticizable on the grounds that it results in outcomes contrary to the prior prediction, implicit in the relevant settled intention, that a restaurant will be entered. And, conversely, the prior planning methodology is criticizable on the grounds that it produces a prediction that cannot be reliably realized by the available means.

Preference reversal as usually defined may also be criticizable, but it is not thereby inherently irrational. Every choice is made in some particular context, and no two contexts are exactly alike. Hence if any two choices appear mutually inconsistent it is always possible that some difference in their contexts accounts for the difference. A cat that almost always eats meat may sometimes eat grass. No irrationality can be inferred merely from its being unusual. A detailed analysis of decision making methodology in relation to current values, or behavioural evidence of perceived error, would be necessary to substantiate the claim of irrationality. Only the latter is likely to be currently feasible in the case of cats.

Value change is criticizable insofar as it arises from a misperceived or unsafe association. It is, for example, at the mercy of fraudulent correlation and propaganda. This is a particular problem in the evolution of proxies. In addiction, for example, the addictive condition becomes, by association, a fraudulent proxy for an underlying valued condition – usually a valued but transiently hidden physiological condition. Fake news may create a similarly fraudulent association.

Even where all proxies are appropriate, double counting is criticizable, as in the Milford Haven example cited in §4.6. Similarly, particular versions of the dyadic value system described in §4.4 are criticizable on the grounds that they are relatively poor predictors of contingently associated value-relevant conditions or events. Not all snakes, for example, are dangerous. Statistical data can be criticized on similar grounds. Crucially, all such criticism rests ultimately on loss of some relevant predictive accuracy, not merely on disputed value attribution.

Given that methodology can be criticized there is a motive to effect some amendment, either as part of or supplementing the usual process of evolutionary mutation, recombination, and selection. The fact that the evolutionary process has so far produced a very sophisticated and in many ways effective system does not exclude rationally planned improvement. Much modern innovation in language, science, technology, political, legal, and administrative methodology and cultural convention can be seen as working in this direction. Some modifications are broad-based and largely intuitive, some are theory based, and many are local fixes to particular perceived problems. Not all are equally effective. An innovation may be based on unsound theory. Many political and economic programmes, such as forced collectivization, are of this type. A local fix may have unintended consequences. For example, high frequency trading tends to produce excessive market volatility. The tendency of democratic methodology to produce relatively resilient political systems suggests that shared intuitive judgement, in situ, should not be quickly overruled by inadequately tested theory, however seemingly well argued.

This analysis vindicates a principle of dynamic normativity. In this, rationally justified amendment consists not only in correcting errors or inconsistencies in current beliefs and valuations but also in refining the methodology via which beliefs and valuations, and consequent decisions, are realized. It may involve, for example, adjusting what is assumed to be ordinarily feasible, as described on page 97, or deliberately contriving educational experiences via which values are modified, as described on page 129, or imposing a cooling-off period, as described on page 206.

On this basis, the extent to which the relative value of the possible consequences of a prospective course of action ought, by then-current normative standards, to depend on their relative futurity – the second part of the original query as posed on page 9 – is generally not fixed a priori but is open to evolutionary adjustment based on the possibility of achieving improved predictive success in realizing currently desirable outcomes. Deadlines, cooling-off periods, and patterns of forced choice, as well as collective, consensual, and majoritarian decision making arrangements and other legal and ethical innovations, the use of which may be difficult to account for in a standard theory in which normative justification is assumed to rest only on the consistency of preferences, are, on this basis, best viewed as novel methodological devices justified in terms of their contributing to dynamic normativity, rather than as intrinsically desirable.

Normative optimization is a process, not an end. The key problem is not to formulate, nor to satisfy, a set of currently consistent preferences, but to develop an evolving agential methodology so as to achieve generally improved outcomes over time, as concurrently evaluated, in an environment full of uncertain contingencies and complex positive and negative feedback. An analysis of ideal decision making patterns or principles may be helpful in this investigation but it cannot provide a means to eliminate or even to definitively minimize actual uncertainty. Patterns of response that vary with futurity, including those described in §7.4 and §7.5, appear unavoidable.

7.7 Testing

The theory described here diverges from much philosophical analysis in aiming to be explanatory and, therefore, to be open to empirical testing. It is unlikely that the research described in the Appendix is ideally suited to this task but it gives a place to start.

A preliminary point should be made. An experimental result provides a reasonably conclusive test of a theory only if it is such that the theory offers no possible explanation of it. This is rarely the case. More often the theory offers a possible explanation, but the explanation rests on one or more assumptions that themselves need to be tested. If they are found not to be satisfied then there is a problem. This is the repeated pattern described in Chapter 2. Where results already exist the immediate issue is, therefore, whether they can be explained. If they can be, this is a good start. But there remains the problem of testing the implied assumptions. This will usually require additional research. I will discuss this further in §7.8. In the meantime, conclusions must remain tentative.

Considering the relationship between the results described in the Appendix and the theory proposed above, a number of general issues emerge. The reported results often conceal a great deal of individual variation. This creates a problem of interpretation, since most of the models being created or tested have no obvious way of justifying this variation. An illustration is given in (A30), in which Fehr-Duda et al. (2010) find 27% ‘EUT types’ and 73% ‘non-EUT types’. It raises a general question of the explanatory significance of statistical variation in the absence of any relevant account of its possible origin – in terms of, for example, explicable parametric variation within a population. The present theory assumes, inter alia, that individual human agents’ values and judgements of probability vary systematically depending, generally, on their past experience. This imposes a very significant constraint on aggregate data.

In most cases described in the Appendix, the choice available to an agent is presented as a direct choice between remote hypothetical outcomes, as verbally described. In many early cases there is not even a convincing promise that the chosen outcome will be realized. There is no systematic recognition that what is actually chosen is not a described outcome but a current verbal response embedded in a complex system of epistemic and social assumptions. Attempts are often made to bridge this obvious gap, such as by emphasizing the trustworthiness of the experimenter, but no corresponding decision making effect is generally admitted in the analysis. The present theory makes this a key issue, as in Figure 17.

In most cases, hypothetical outcomes are described as exact numerical quantities and analysed as if they are significant exactly in proportion to these quantities. This applies to both outcome size and probability, and, where relevant, delay. Little attention is paid to plausible granularity, either of absolute quantities or of differences, nor, usually, to plausible relativity in scaling. This approach is obviously problematical outside a fairly narrow range of quasi-economic contexts in which relevant quantities are standardly defined and, although it has the merit of often yielding precise predictions, it has the disadvantage that these predictions are almost always false. The present theory admits a predictable degree of subjective variation.

From a theoretical point of view, perhaps the greatest difference between the modelling assumed in most of the studies described in the

Appendix and the present theory is that the present theory assumes that variations in attitude characterized as, for example, risk aversion, ambiguity aversion, or preference for gains versus losses, certainty, ownership, increased wellbeing, variety, or fairness are additionally valued conditions rather than distortions of the value assigned to a single primary condition. The effect is to add additional dimensions to the outcome space. Although the results reported do not provide a direct test of this assumption, the general difficulty in accounting for the variety of results in (A6) to (A24) tends to support something of the kind. The Ellsberg paradox (A10) and the increasing sequences effect (A20) provide obvious examples.

Finally, there is clear evidence of a relationship between uncertainty and delay, particularly in (A14) and (A15). This is the subject of many other recent studies, as cited on page 22. In the present theory it is assumed to arise straightforwardly from the ever-present possibility of unanticipated change, transient ignorance or misjudgement, evolving cooperation, and external disruption. Most of the models referenced in the Appendix cannot easily accommodate this possibility because of a methodological assumption that probabilities are fixed, usually by stipulation.

Several of the examples listed in the Appendix warrant more careful consideration. For example, Loewenstein and Elster (1992) (A1) observe that, “We construct shoddy highways … [but] eschew nuclear waste disposal sites … .” These examples are relevant and thought provoking but they raise an important question about assumed rationality, for they carry the implicit suggestion that they demonstrate irrationality by virtue of inconsistency. But in terms of human decision making this is by no means clear. An obvious first question is who are “we”. The construction of highways or the selection of nuclear waste disposal sites is an extremely complex socially mediated process in which no individual’s decision making generally dominates. Individuals object to nuclear waste disposal sites near their homes for obvious reasons – reasons that are plausibly explicable in the theory proposed here. If the result is that no site is chosen despite a general acceptance within the population that a site should be found somewhere, this is an argument for an improved collective decision making methodology, not a demonstration of individual irrationality. The literature on collective choice (e.g. Sen 1970) shows how difficult the problem is.

Amongst the most impressive sets of results are those demonstrating the endowment effect. Kahneman et al. (1990) (A6) reports that merely having possession of a randomly provided mug increased the median price that a subject demands for someone else to take possession of it from $3.12 to $7.12, whilst the median price a subject would pay to take possession of the same mug was $2.87. This is difficult to account for unless it is assumed, as in the current theory, that possession of a valued object is additionally valued. The effect is intuitively obvious in what is commonly called sentimental value, as is its typical amplification by association with positively valued events and its attenuation or reversal by association with negatively valued events.

The case of the New York cabdrivers (A7) warrants further mention. Camerer et al. (1997) interpret the results as indicating that drivers engage in daily income targeting rather than maximizing total earnings. The result can be explained as follows. For independent cabdrivers – who explicitly choose not to work fixed hours – deciding to stop working for the day is, presumably, a deliberative decision, as in Figure 14. Hence it is decided on the basis that some relevant outcome value exceeds a threshold, as in Figure 19. The latter must be set relative to some currently relevant condition. This must usually be the day’s earnings, there being, in general, no other relevantly envisageable daily outcome. Longer term aggregate earnings, for example, do not supply such a threshold since they are not an outcome for which there is any chooseable daily option. The only option for which long-term aggregate earnings would be a relevant outcome is a long-term policy-setting option such as is typically developed by organized businesses – which is, presumably, why organized businesses typically compute their quarterly or annual earnings and plan accordingly whilst independent cabdrivers do not. The resulting effect is, incidentally, no more irrational than an animal eating what it requires for the day and then resting.

The Ellsberg paradox (A10) poses the question whether, except in sophisticated gambling and statistical modelling, cases involving verbally described probabilities are treated as standardly assumed rather than as representative cases of known types, as in §6.3. A human subject might, for example, treat the 50:50 case as typical of cases in which there is no best strategy but in which you may win if you are lucky, and the other as typical of obscure cases in which there is often a hidden snag. That the experimenter assures the subject that there is no snag does not prevent this, since that may be part of the snag. The subject's assigned probabilities are then those associated with the assumed types on the basis of prior data, not the probabilities stated. Gamblers who happen to make early wins or over-record wins versus losses will, on this basis, overestimate the probability of winning even if odds are described accurately, whilst those offered insurance may underestimate the probability of disaster (A12).

Loewenstein's (1987) study (A17) is particularly interesting. He suggests that the preference to delay receiving a kiss from a movie star may perhaps arise from the added value of pleasurable anticipation. If so, it might be interpreted as a type of displaced double counting. But such an effect lacks the usual adaptive justification and, if admitted generally, appears to warrant a preference to delay all delayable rewards, contrary to what is generally observed. The account offered by the present theory is that it is a case of transitional uncertainty. An imminent kiss involves considerable uncertainty, indicated by a sensation of nervousness, whereas a remote kiss remains, more abstractly, a member of a broad type. The effect can be inferred almost exactly from Figure 16.

Finally, several studies, including (A25) and (A26) – the much discussed cases of proportions of people being saved or dying and of Linda the bank teller – tend to demonstrate the importance of probability assignment or valuation based on classification by prominent features rather than as implied by the working out of an implicit inferential calculus. This may occasionally produce results standardly classified as irrational, but it may, in context, be usually both effective and efficient and therefore adaptively preferred. Problematic cases can be resolved by adopting a more sophisticated methodology.

Two more general remarks on reported results may also be made. One is the enormous diversity of estimates of an assumed intertemporal discount rate, reported in Frederick et al. (2002 Table 1) to range for -6% to ∞ and that later results show no greater consistency than earlier ones. On the other hand, studies show relatively consistent differences between individuals as measured on single tasks. For example, gamblers generally display higher apparent discount rates. Rates generally appear to decrease with age except that older adults show the greatest rate of all when the expected delay is between 3 and 10 years (Chabris et al. 2010). These results are compatible with a theory that attributes apparent discounting mainly to available information and assumed risk or uncertainty rather than to futurity as such.

Many other features of the theory are open to testing. Perhaps the strongest is the structure and aetiology of values. For example, the dyadic system of affinity and salience and the system of proxies and double counting together impose very strong constraints on the valuation of novel conditions, whether envisaged or experienced. Similarly the theory of statistical probability imposes strong constraints on the relation between estimated probability and prior data. An assumption of rank-order quantification is also strongly constraining. Threshold dynamics, unanalysed in Figures 15 and 19, requires both further theoretical investigation and testing. For example, a declining threshold can be expected to result in otherwise insufficiently prominent options being chosen, as in trivial leisure activity, and an increasing threshold can be expected to suppress otherwise urgent activity. I leave these issues for future research.

7.8 Final Observations

This study started out as an investigation into a specific technical issue in behavioural economics. But in the end I wish to defend it, at least equally, as a contribution to philosophy more generally.

The underlying thought is as follows. Prior to the mid-19th century a remarkable general theory was developed, culminating in deism, that attributed the intelligibility of the universe and the existence of value within it to its having been designed on rational principles and valued accordingly and that humans can, no doubt imperfectly, participate in appreciating these principles and values by virtue of their own rationality. It is difficult to overstate the extraordinary breadth, coherence, and apparent explanatory success of this theory. It was the basis, inter alia, of Newtonian mechanics. Hence even when its theistic basis was disbelieved its central thesis was not generally doubted. The only problem is that, at least if the rational principles and values assumed are the ones that we humans intuitively subscribe to, it is clearly false. Whatever principles and values, if any, underlie the design of the universe, they are not ones we readily appreciate.

This line of thought makes it difficult to justify what appears to be a common assumption in 20th century analytic philosophy, namely that the fundamental categories in terms of which human thought and action are properly to be analysed, such as knowledge, belief, desire, and preference, and the corresponding descriptive categories identifying objects, qualities, and states of the world more generally, are fixed a priori and, if rigorously investigated, are intuitively discernible as such. It is implicit, for example, in the Fregean assumption that the actual extension of an intuitive predicate is generally well defined and in the Lewisian assumption that the set of possible worlds – not merely a set of toy worlds – is definable in some unspecified but conventional descriptive algebra.

The converse possibility – that explanatorily relevant concepts may be, at least initially, radically counterintuitive – is a problem that the natural sciences are familiar with. The apparently successful response, which in §1.6 I have termed Galilean, is well documented (Toulmin 1953, Hanson 1958, Lakatos and Musgrave 1970, Nersessian 1992, 2008). It involves, inter alia, rigorous dynamic modelling of related, often prototypical, members of an expanding class of observable phenomena, and progressive reconstruction of models and of supporting nomological principles and concepts in response to persistent modelling failure, based – usually – on a quite detailed but ultimately agnostic attitude to related ontic assumptions. It is an admitted feature of this response that all analytical categories, however apparently well justified, are open to revision.

It is often assumed that this provisionality is specifically characteristic of scientific knowledge. But since it is also generally assumed that scientific knowledge supersedes intuitive or folk knowledge wherever the two are clearly incompatible there is an obvious argument that it extends equally to this less rigorously justified intuitive base. It extends, in particular, to intuitive empirical categories and epistemic assumptions, including those that are, presumably, characteristic of human agency. This is the assumption underlying the methodology described in §1.3.

If intuitive empirical categories and epistemic assumptions are not a priori, the question arises as to their origin and justification. The post-Darwinian account accepted here is that they are features of an evolved adaptive response to encountered environmental constraints. The standard objection to this claim in its usual sociobiological formulation is that it is inconsistent with the form, and particularly the velocity, of evolutionary change typical of rational adaptation. But there is a reply, as set out in §3.2, namely that rational adaptation involves the evolution of a cognitive and agential methodology that is capable of generating and testing empirical predictions and that its evolution is shaped not solely by hereditary survival but by predictive success within the envelope of hereditary survival. This accounts both for the partial but imperfect biological appropriateness of the evolved methodology and for its openness to rational revision in the light of evidence. The fundamental implication is that knowledge is consequent on action, action on value, and value on adaptation, not vice versa.

It is not necessary to defend every constituent claim made in the course of the above analysis in order to accept its broad conclusions. As described in §1.5, the possibility of coherent explanatory analysis depends on adopting at least a provisional response to each key analytical problem. The least that can be said of the proposals described above is that they appear consistent with current evidence; but they are obviously open to possible amendment and are, therefore, not beyond probable criticism.

Admitting this, the broad epistemic implications are both positive and negative. On the positive side, truth, whilst provisional, is not merely relative. Evidence can be adduced, claims tested, values criticized, action may succeed, methods can be improved, and disruptive emotions accommodated. Indeed, the theory offers the prospect that human action may be significantly more explicable, and predictable, than hitherto assumed, and hence that the generality of human values may in due course be more consensually accommodated. On the negative side, our conceptual grip on the structure of the world is marginal, being only approximately adapted to our ecological niche and extended speculatively, and the prospect of an ideal system of knowledge or logic or practical reason or political organization or social welfare is accordingly diminished.

Since the proposed theory assumes a computational mechanism, a major analytical problem is to develop a representational system supporting relevant logical inference, and natural language, that does not rest on Fregean or idealist ontology. It offers the possibility of a constructive theory to complement the present principle theory. I hope to return to this in due course.

A key supporting idea is the methodology described in §1.3. In a sense, this merely regularizes common practice. But it adds an explicit provision that is both liberating and constraining. It is that intuitions can be rejected, but only if they can be explained. This provides a principled method of adjudicating between disputed intuitions – or, indeed, of moving to a completely novel theory as in, say, quantum mechanics. The legitimacy of this kind of step is otherwise opaque. It would be a liberation in philosophy if people who wished to admit the viability of a novel theory were obliged to show how, under it, they can account non-trivially for the vast bulk of intuitive impressions and, conversely, their opponents were obliged to show how the intuition of its possibility can be explained by normal epistemic means despite the claim that it is non-veridical. In this way some long-standing proposals might be set aside as no longer worthy of debate. It might remove some of the unjustified intellectual protection afforded to obviously unsubstantiated conspiracy theories and to the imagined validity of 'your truth'.

Behind this is the question of whether it is assumed that philosophy can be explanatory. It once was. It is true that, typically, the resulting explanations, where convincing, have been spun off to form new scientific fields, but this is no argument against their philosophical origin. Experience shows that fundamental explanatory problems are not solved by launching into unstructured empirical research. Some deep thinking about concepts and processes is needed. Philosophers over the past hundred years seem to have become increasingly pessimistic about the prospects of this kind of activity and retreated into a perpetual re-analysis of existing concepts. Perhaps their pessimism is justified. But it may just be that the remaining problems, revolving largely around the status of cognition and agency, are difficult, and the favoured methodology has been unsuitable. Perhaps the present work may prove an exception or, at least, a model for a subsequent exception.

Assuming that it has merit, the proposed analysis has both explanatory and practical applications. As indicated at various points during the presentation it appears to be capable of explaining a number of existing observations about human behaviour. Moreover, if sufficient information becomes available about held values and experiential data, quite detailed but wholly intelligible explanations of individual human conduct ought to be feasible. At a practical level, an artificial system built on the implied model ought to provide a much more situationally and evaluatively aware and flexible pattern of response than current AI systems. It admits the possibility that ethical principles might be incorporated as core values rather than as design constraints on computed outcomes.

On another practical level, planned human activity might better respond to the whole spectrum of implicit human values, especially with respect to future risk. A feature of the current decision making system, especially in politics, is that it responds to those values that happen to be prominent by virtue of being currently unsatisfied or recognized as currently at risk. This is rational, but it has the consequence that the resulting action may lead to the dissatisfaction of other currently non-prominent values. So, for example, inequality in the context of an adequate supply of housing leads to rent controls, which lead to a reduction in the housing supply and hence to homelessness. Greater attention to consequences evaluated in terms of what we may call latent values might avoid errors of this type. This effect can be seen in the eventual response to, for example, just-in-time technology and alcohol prohibition. No policy is without some negative value consequences, but probable outcomes ought to be evaluated against latent as well as overt values.

Ultimately there is the problem described by Taleb, namely that some rare events have such large value consequences that, if weighted proportionately, they dominate all others and hence that no standard plan can accommodate them equivalently. A rational policy must be to build enough resilience into the system to reduce the all-things-considered size of such consequences to an acceptable level, in situ. A better understanding of probability, and of the location and interaction of values and of their modification in response to changing circumstances, might make this more feasible.

Appendix Anomalies

Empirical research into economic decision making has for several decades focused largely on the investigation of phenomena that are, or appear to be, inconsistent with some current theory. Dhami attributes this to the influence of Popper and Lakatos (2016: 5-6). They are ‘anomalies’. For example, Camerer (2000: 149 Table 5.1) describes ten phenomena from field research that appear to be inconsistent with standard expected utility theory, such as that stock returns are too high relative to bond returns and that purchases are more sensitive to price rises than to cuts.

Such anomalies are of interest in the present context for three reasons. They provide evidence that is relevant in the testing and revision of proposed models, as Dhami suggests. They offer a more general insight into the relative significance of and interaction among the various factors that appear to influence decision making, especially as they involve the three key variables, utility, probability, and futurity, irrespective of the precise technicalities of current modelling. And they test the relationship between normative and descriptive theory, particularly insofar as a commonly observed effect may continue to appear rationally defective and hence remain normatively anomalous even where a revised descriptive theory renders it formally predictable.

In this appendix I will describe a range of anomalies discussed in the literature, as conventionally classified. In order to maintain as much distance as possible between the phenomena and any particular theoretical interpretation, and to emphasize that what we have here are only data not confident explanatory analysis I will, in most cases, briefly describe a specific experiment or observation as actually reported. Most of these experiments or observations are classics, much discussed in the literature and often repeated. In total, and in their variety, they give an indication of the difficulties that have been encountered in devising a theory of decision making based on standard assumptions that is not clearly inadequate.

A1 Political choice

> We construct shoddy highways, refuse to switch to the metric system, even type on a keyboard that was designed to slow the typist down … But we also invest in basic research whose payoff is remote and eschew nuclear waste disposal sites because they may cause problems centuries hence (Loewenstein and Elster 1992: x).

A2 The common consequence effect

The common consequence effect is usually illustrated in the comparison of two pairs of probabilistic options, or 'lotteries'. A lottery is a mutually exclusive and exhaustive set of outcomes each with a specified probability which an agent is offered and may accept or reject as a single option. It is standardly expressed in the form (x_1, p_1; … ; x_n, p_n), where x_i is an outcome promised with probability p_i, and $\sum_{i=1}^{n} p_i = 1$. A compound lottery is a lottery of lotteries. Assuming linearity, a compound lottery is always reducible to a simple lottery and so is an admissible option.

Suppose that *A*, *B*, *C*¸ and *D* are lotteries, and an agent is offered (1) a choice between (*A*, *p*; *C*, 1-*p*) and (*B*, *p*; *C*, 1-*p*) and (2) a choice between (*A*, *p*; *D*, 1-*p*) and (*B*, *p*; *D*, 1-*p*). Since *C* is common to both options in (1) and *D* is common to both options in (2), according to a plausible rational principle embodied in a standard axiom of independence – or, slightly differently, in Savage's 'sure thing principle' – the order of preference in both (1) and (2) should depend only on the relative preferability of *A* versus *B*. Hence the order should be the same in both cases. However, Kahneman and Tversky (1979) report an experiment in which 72 subjects choose between options involving the following promised payments, in Israeli pounds, with the indicated probabilities:

(1) (£0, 0; £2400, 1) or (£0, 0.01; £2400, 0.66; £2500, 0.33)

(2) (£0, 0.66; £2400, 0.34) or (£0, 0.67; £2500, 0.33).

These can be decomposed into the form defined above with

p=0.34; A=(£2400,1); B=(£2500,$^{33}/_{34}$; £0,$^{1}/_{34}$); C=(£2400,1); D=(£0,1).

Hence the order of preference should be the same in both. However, Kahneman and Tversky report that 82% of subjects chose the first option in (1) and 83% chose the second option in (2). Similar results have been obtained in many later experiments.

A3 The common difference effect

The common difference effect is usually illustrated in the comparison of two pairs of time-indexed outcomes. A time-indexed outcome, standardly expressed in the form (x, t), is the promise of an outcome x to be realized with probability one after a specified delay t which an agent is offered and may immediately accept or reject.

Suppose that x, x' are specified outcomes and t, $t' \geq 0$ are specified delays and $\tau > 0$ is an additional delay, and an agent is offered (1) a choice between (x, t) and (x', t') and (2) a choice between (x, $t+\tau$) and (x', $t'+\tau$). Since the additional delay τ is added equally to both options in (2), a plausible principle of preference consistency embodied in the axiom of stationarity requires that if (x, t) is preferred to (x', t') then (x, $t+\tau$) should be preferred to (x', $t'+\tau$). However, Thaler (1981) reports an experiment in which about 80 subjects are asked to say what would be currently acceptable compensation in place of a foregone future reward, over 36 combinations of outcome and delay, from -\$250 to +\$3000 and from one month to ten years. The results strongly contradict the principle of temporal uniformity. Again, the effect has been reproduced many times.

A4 The common ratio effect

The common ratio effect is usually illustrated by the comparison of two pairs of outcomes with proportionately differing probabilities. An agent is offered (1) a choice between (x, p) and (x', p') and (2) a choice between (x, kp) and (x', kp'), where $k>1$ and kp, $kp'\leq 1$. Since (1) is equivalent to a k^{th} partition of (2), a plausible principle of preference consistency again embodied in the axiom of independence requires that if (x, p) is preferred to (x', p') then (x, kp) should be preferred to (x', kp'). However, Kahneman and Tversky (1979) again report to the contrary. 95 subjects were asked to choose between options involving the following promised payments, in Israeli pounds, with the indicated probabilities:

(1) (£3000, 0.25) or (£4000, 0.2).

(2) (£3000, 1) or (£4000, 0.8)

Since the probability ratios are identical, with $k = 4$, the order of preference should be the same in both. However, the authors report that 80% of subjects chose the first option in (2) and 65% chose the second option in (1). Again, many similar results have been reported.

A5 Non-transitive choices

Non-transitive choices are often observed in the comparison of two pairs of outcomes, H a high-probability-low-payout option and L a low-probability-high-payout option, and C_H a sure payout judged equally preferable to H, and C_L a sure payout judged equally preferable to L. Since C_H is equivalent to H, and C_L is equivalent to L, a plausible principle of preference consistency embodied in the axiom of ordering requires that if H is preferred to L then C_H should be preferred to C_L. Tversky et al. (1990) report to the contrary. 198 subjects each chose between pairs of options drawn from 18 triples each consisting of an H and an L – with previously elicited values C_H and C_L – and a third ‘for sure’ option X. Overall, in approximately 50% of comparisons in which C_L was greater than C_H, H was preferred to L, or vice versa. Most were of the former type. Approximately 10% of comparisons of H, L, and X exhibited non-transitive triples. Again, many similar results have been reported. Almost identical non-transitive choices involving delay rather than probability are also observed.

A6 The endowment effect

The endowment effect involves a characteristic asymmetry in the response to perceived gains versus losses relative to a current context-specific base value or, more specifically, a characteristic difference between the amount a person is willing to pay for an item (WTP) and the, usually larger, amount they are willing to accept in payment (WTA). Associated effects are referred to as reference dependence, the sign effect, the absolute magnitude effect, loss aversion, gain-loss asymmetry, and status quo bias. Kahneman et al. (1990), for example, reports a complex series of experiments aimed at trying to discover the scope of the endowment effect. They involve creating markets for the exchange of various goods, including intrinsically worthless tokens with an exchangeable cash value and similarly exchangeable mugs and pens. In two key experiments, 194 subjects were divided into sellers, choosers, and buyers. Both sellers and choosers were asked how much they would accept for an item, whereas buyers were asked how much they would pay for an item. The decisive difference was that sellers but not choosers had physical possession of the relevant item at the time. The endowment effect was observed in sellers but not in choosers, who acted more like buyers. Median valuations were, respectively, $7.12, $3.12, and $2.87. Similar results have been reported in experiments involving many other types of goods (Dhami 2016: 219).

A7 Choice bracketing

Options may be nested within each other in various ways. This allows various levels of granularity in an agent’s valuation of expected outcomes. For example, a player in a tournament may try to win on every play or optimize the outcome over each game or over the entire tournament. The effect is standardly referred to as choice bracketing. The decisions made may vary accordingly.

A classic example is described in Camerer et al. (1997). It concerns the daily working hours of New York city cabdrivers. Drivers typically lease their cabs at a fixed daily rate, work an optional number of hours each day (up to 12), and keep the profit. They record their hours and earnings on daily trip sheets. The authors analysed over 1800 trip sheets from over 1200 drivers. They found greater variability in hours worked per day than in money earned and a negative correlation between the two. They interpret this as indicating that drivers engage in daily income targeting. They calculate that total income would be 7.8% higher if drivers worked a uniform number of hours per day.

A8 Convex utility function for losses

Bernoulli's response to the St. Petersburg paradox was to postulate a logarithmic relation between objective value and utility. This relativizes marginal changes to a current base value, but it implies infinitely large negative utilities where the envisaged objective outcome-value is zero and a concave utility function for both gains and losses. Evidence suggests that, on the contrary, gains and losses are treated unequally and that the rule of diminishing increments applies to both, as in Figure 5. Abdellaoui et al. (2007) elicited risk profiles from 48 subjects over eleven substantial but hypothetical monetary losses relative to their known income, evaluating points of indifference from binary choice data. The results show a convex – that is, a marginally diminishing – utility profile in the domain of losses for 33 of the 48 subjects and a mixed profile for 11 others.

A9 The certainty effect

The certainty effect, also called the extreme probability anomaly, involves a characteristic over or under-response to outcomes having a probability close to either zero or one. For example, Bruhin et al. (2010) reports the results of three experiments conducted between 2003 and 2006 involving 448 subjects, designed to test the relation between specified or objective probability and subjective probability. About 20% of subjects displayed a linear relationship, as assumed in standard expected utility theory, about 30% displayed substantial departures from linearity near zero, overweighting probabilities near zero by a factor of up to 300% and underweighting probabilities near one similarly. The remainder showed significant departures from linearity but less than the second group. This classification was robust over repeated trials. Fehr-Duda and Epper (2012) lists other studies with similar results.

A10 Ambiguity aversion

A distinction is standardly made between risk, uncertainty, and ambiguity (Camerer and Lowenstein 2004: 18-21, Dhami 2016: 79). The distinction can be traced historically. Originally, theorists were interested in games of chance, in which outcomes were assumed to have objective probabilities. These are cases of risk. Agents, however act as if they assign probabilities to outcomes even where there are no agreed objective probabilities. These are cases of uncertainty. Later, cases were discovered that cannot be interpreted on the basis that agents assign any consistent probabilities to outcomes. These are cases of ambiguity.

In 1961 Ellsberg described the following scenario. A subject is faced with two urns. Urn *A* contains 50 red balls and 50 green balls. Urn *B* contains 100 balls, each either red or green in an unspecified ratio. Both urns are well mixed. The subject is asked to choose an urn and a colour and to draw a ball from that urn. If and only if it matches their chosen colour they win a prize. Subjects typically choose urn *A*. No consistent assignment of probabilities to possible outcomes can account for this.

A11 Rabin's paradox

As noted above, in standard expected utility theory risk aversion is accounted for by assuming a concave utility function. Rabin (2000) showed that if the assumed non-linearity is sufficient to account for observed risk aversion over small stakes it predicts very high risk aversion over high stakes. Dhami (2016: 105) gives an illustrative calculation. This type of extreme risk aversion is not observed.

A12 Gambling and underinsurance

Under standard expected utility theory a rational gambler must overestimate the probability of winning. The same effect should apply to potential losses, leading to a high take-up of insurance. Kunreuther et al. (1978) show that, on the contrary, people typically fail to purchase adequate insurance against natural hazards despite being encouraged to do so.

A13 The low probability discontinuity

As described in §2.3, many studies show that low probabilities are commonly overestimated. But in other cases, as Kahneman and Tversky (1979) observe, evidently possible outcomes are treated as if they had zero probability of occurrence. Some low probability risks are defended against and some low probability prizes are sought, but others are discounted. This contrast is intelligible but not easy to account for (Dhami 2016: 32, 193-6).

A14 The immediacy effect

The immediacy effect involves a characteristic tendency to prefer immediately available outcomes disproportionately. For example, Keren and Roelofsma (1995), in the first part of an experiment designed to investigate the relative importance of immediacy versus certainty, asked 60 subjects whether they would prefer to receive (*A*) Dutch Fl.100 now or Fl.110 in four weeks, and (*B*) Fl.100 in 26 weeks or Fl.110 in 30 weeks. 82% preferred the first option in *A*. 37% preferred the first option in *B*. Experiments of this type are controversial, partly because of the difficulty in isolating or controlling the various potentially relevant factors. Halevy (2015: 350-1) compares a number of alternatives methods. There is evidence of significant variation among subjects (Dhami 2016: 673-7). Nevertheless, an immediacy effect is widely reported.

A15 Impatience

Impatience usually refers to a more general tendency to prefer earlier gains to equal but later gains. Experiments comparing preferences often omit an immediate outcome to separate impatience from immediacy. For example, Attema et al. (2016) describes a method of measuring impatience without assuming any measurement of utility. In a study involving 96 subjects they report an average annual discount rate of 35% between weeks 1 and 52. Almost all the 42 studies listed in Frederick et al. (2002: Table 1) show a significant, although varying, degree of impatience.

A16 Decreasing impatience

Standard discounted utility theory implies that an agent's level of impatience, as measured by the inferred discount rate, ought to be constant. Many studies have tested this claim. Their results show that, on the contrary, impatience generally decreases with time. For example, Benzion et al. (1989) asked 204 subjects how much they would pay to postpone or expedite an expected but delayed payment of a specified amount. In all cases the discount rate decreased with delay, approximately halving over the interval from 6 months to 4 years. Average results are shown in Loewenstein and Prelec (1992: 135).

A17 Delayed gratification

Although impatience appears ubiquitous, Loewenstein (1987) describes a contrary result. 30 subjects were asked how much they would pay to obtain a specified outcome immediately or after various periods of delay from 3 hours to 10 years. One option was 'a kiss from a movie star of your choice'. Respondents assigned this an increasing value up to 3 days and a value greater than its immediate value up to one year. One possible interpretation is that the state of anticipation has a separate positive value. Loewenstein (1992: 28) describes pleasurable anticipation as a factor in the purchase of lottery tickets.

A18 Subadditive discounting

Read and Roelofsma (2003) note that the standard method of evaluating patience relates differences in patience to differences in absolute delay rather than to the interval between prospective outcomes. Two experiments, involving 141 subjects in total, were designed to distinguish these alternatives. They found that most of the apparent increase in patience can be attributed to subadditive discounting – to an effect in which differences in discounted utility between outcomes separated by a short interval are disproportionately large. This is similar to the subproportionality that appears to characterize subjective probability. The findings are not consistent with standard hyperbolic discounting.

A19 Intertemporal risk aversion

Miao and Zhong (2012) reports an experiment to investigate the response to risk relative to delay. Consider two possible outcomes, one earlier and one later, such that either (*A*) both occur for sure, or (*B*) either both or neither occurs, with 50% probability, or (*C*) either one or the other occurs but not both, each with 50% probability, or (*D*) each occurs with 50% probability independently. 46 subjects were given tokens, to be redeemed later, to allocate to each option under various conditions. The results show a similarly greater aversion to delay in conditions *C* and *D*, where risk is distributed in time, than in conditions *A* and *B*.

A20 The increasing sequences effect

Under standard expected utility theory the value of a series of payments should not depend on the order in which they are received. Under standard discounted utility theory a decreasing sequence should be preferred to an equivalent increasing sequence since later outcomes are increasingly discounted. However, Hsee et al. (1991) present evidence of the reverse. 96 subjects were asked to rate various prospective salary profiles. Increasing sequence were generally preferred. The experiment was designed primarily to distinguish contextual conditions that heightened or reduced the effect, but the effect is marked in all cases.

A21 Other sequence effects

Loewenstein and Prelec (1993) reports a study in which 51 museum visitors were asked to say which of each of two pairs of options they would prefer for a series of meals over the next five weekends:

$$A=(F, H, H, H, H) \text{ or } B=(H, H, F, H, H),$$
$$C=(F, H, H, H, L) \text{ or } D=(H, H, F, H, L),$$

where *F*=‘dinner at a fancy French restaurant’, *H*=‘eat at home’, and *L*=‘an exquisite lobster dinner at a 4-star restaurant’. 88% preferred *B* to *A*. 51% preferred *C* to *D*. This suggests that the sequence of anticipated outcomes may affect their preferability.

A22 Delay-speedup asymmetry

Loewenstein (1988) reports the following experiment. 66 subjects were divided into two groups. All were asked questions from the set (*A*) how much you would pay for a VCR recorder to be delivered today, (*B*) how much you would accept to delay receipt for one year, (*C*) how much you would pay for a VCR recorder to be delivered in one year, and (*D*) how much you would pay to speed up delivery from one year to today. Members of one group were asked *A*, *B*, and *C*. Members of the other group were asked *C*, *D*, and *A*. The mean response to *B* was $126. The mean response to *D* was $54. This shows an asymmetry between delay and speedup.

A23 Intertemporal gain-loss interaction

Rao and Li (2011) presents a series of results that appear to be at variance with all current models. For example, 93 subjects were asked to choose (1) whether they preferred (*A*) to gain 5 apples now and lose 6 apples tomorrow or (*B*) to lose 6 apples tomorrow and gain 8 apples in 1 week, and (2) whether they preferred (*C*) to gain 5 apples now or (*D*) to gain 8 apples in 1 week. 84% preferred *B* to *A*. 34% preferred *D* to *C*. Similarly, 118 subjects were asked to choose (1) whether they preferred (*A*) to gain ¥1,000,000 now or (*B*) to gain ¥5,000,000 in 10 years, and (2) whether they preferred (*C*) to gain ¥1,000,000 now and ¥6,000,000 in 1 year or (*D*) to gain ¥6,000,000 in 1 year and ¥5,000,000 in 10 years, and (3) whether they preferred (*E*) to gain ¥1,000,000 now and lose ¥2,000,000 in 11 years or (*F*) to gain ¥5,000,000 in 10 years and lose ¥2,000,000 in 11 years. 72% preferred *A* to *B*. 48% preferred *C* to *D*. 50% preferred *E* to *F*. Rao and Li doubt whether any conventional discount function can account for these results.

A24 Time-sensitivity

Ebert and Prelec (2007) reports a series of experiments aimed at measuring the extent to which the variation in a person's discount rate depends upon contextual conditions. 309 subjects, divided randomly into six groups, were asked to name the present cash value of one or more of a set of $80 prizes to be paid in 1 day, 1 week, 1 month, 3 months, or 1 year. Members of group (1) named all five in a random order. Members of groups (2) to (6) named only one each. Group (1) showed significantly greater variation in mean discount rate over the set of prizes than the others combined. In a second similar experiment using restaurant tokens, subjects who had more time to evaluate a set of options showed greater variation than those who had less time. The authors interpret their results as indicating that time judgements are 'fragile'.

A25 Framing

There is considerable evidence that the way options are described affects resulting choice. Tversky and Kahneman (1981) reports an experiment in which 307 subjects were asked, concerning a prospect in which a disease 'is expected to kill 600 people' in the U.S., which of two programmes they would favour. 152 were told that if *A* is adopted ‘200 people will be saved’ and if *B* is adopted ‘there is a $^1/_3$ probability that 600 people will be saved and a $^2/_3$ probability that no people will be saved’. The others were told that if *C* is adopted ‘400 people will die’ and if *D* is adopted ‘there is a $^1/_3$ probability that nobody will die and a $^2/_3$ probability that 600 will die’. 72% of the first group chose *A*. 78% of the second group chose *D*.

A26 The conjunction fallacy

Tversky and Kahneman (1983) reports a related experiment. 173 subjects were given a questionnaire describing ‘Linda … 31 years old … single … very bright … deeply concerned with issues of discrimination and social justice …’, and listing eight descriptive statements including (*A*) ‘Linda is a bank teller’ and (*B*) ‘Linda is a bank teller and is active in the feminist movement’. They were asked to rank the eight statements in order of probability. 88% ranked *B* as more probable than *A*. Subjects classified as logically sophisticated scored almost as highly as those classified as logically naïve.

A27 Fairness, Altruism, Enmity

Standard expected utility theory is often interpreted as assuming that decision makers act only in their own immediate self-interest at each point during any period of interaction (Dhami 2016: 339). However,

> … people may sometimes choose to “spend” their wealth to punish others who have harmed them, reward those who have helped them, or to make outcomes fairer (Camerer and Lowenstein 2004: 26).

Numerous experimental demonstrations exist. For example, in the Centipede Game two players each start with $2. They play in turn. On any turn the player whose turn it is may steal $2 from the other and play ends. Otherwise that player receives $1 and play continues, up to a maximum of 100 plays. The conventionally rational policy is to steal on the last available turn. But this applies at every turn including the first. Hence by backward induction the rational policy is to steal at the first opportunity, forgoing all later receipts. McKelvey and Parlfrey (1992) find that at most 15% of players follow this strategy. In the ultimatum game for two players *A* and *B* (Güth et al. 1982), *A* is given an integer sum of money and may offer any proportion to *B*. If *B* accepts, both keep the proceeds. Otherwise both get nothing. The optimal self-interested policy is for *A* to offer the smallest possible sum and *B* to accept it. The experiment has been run many times. The mean proportion offered is found to be between 0.3 and 0.4 (Dhami 2016: 349). The dictator game is the same except that *B* must accept whatever is offered. The mean proportion offered is smaller but not zero – typically between 0.15 and 0.2. It increases if play is or appears to be observed. Experimental investigation of enmity seems to be rarer, but given what we know about the history of the world its significance can hardly be doubted.

A28 Self-control

Since undesired outcomes may occur, disappointment and regret are ubiquitous. Disappointment occurs when the outcome of a particular choice is worse than expected. Regret occurs when an outcome is worse than the expected outcome of an alternative not chosen (Zeelenberg et al. 1998). Either can be anticipated as a possible outcome and included as a factor in decision making, with possibly inconsistent consequences (D’Arms and Jacobson 2009, Kahneman 2012, Loomes and Sugden 1987, Gul 1991). This poses an analytical problem of self-control, or commitment (Elster 1977, Ainslie and Haslam 1992b, Bermúdez 2018). Several problematic effects, notably procrastination, addiction, and compulsive behaviour are typically marked by an alternation of action and regret, leading to cycles of repeatedly abandoned plans for reform (Ainslie and Haslam 1992a: 79, O’Donoghue and Rabin 2001: 121, Steel 2007: 7).

A29 Collective action

It is usual in decision theory to model collective action as an interactive conjunction of the actions of individuals using the methodology of game theory (e.g. Gintis 2009). But elsewhere in economics and in social science more generally it is widely assumed that institutional or organizational action-patterns exist independently of individual choice (Arrow 1994, Hodgson 2007). Organizations are treated as distinct agents, with values, aims, and policies of their own and with internal control processes operating via assumptions about duty and authoritative instruction rather than personal choice. It is notable that Dhami (2016), despite its length and scope, contains no index entry to institutional or collective action or organization, nor is it apparent how the models it describes might capture the notion of institutional structure.

A30 Individual differences

Reported results often show wide variation between individuals or between otherwise comparable cases. For example (Frederick et al. 2002: Table 1) reports discount rates from -6% to ∞. It is often found that decision makers fall into several distinct types. For example, Fehr-Duda et al. (2010) identify a clear distinction between ‘EUT types’ (27%) and ‘non-EUT types’ (73%). Many studies, e.g. Ebert and Prelec (2007), discard some subjects for apparently inconsistent or inappropriate responses. As Frederick et al. comment (2002: 377), there is little evidence as yet of the increasing regularity characteristic of physical science.

Bibliography

Abdellaoui, M., Bleichrodt, H., & Paraschiv, C. (2007) Loss aversion under prospect theory: a parameter-free measurement. *Management Science* 53(10), 1659-74.

Abdellaoui, M., Kemela, E., Paninb, A., & Vieiderc, F. (2019). Measuring time and risk preferences in an integrated framework. *Games and Economic Behavior* 115, 459-69.

Achinstein, P. (2002). Is there a valid experimental argument for scientific realism? *Journal of Philosophy* 99(9), 470-95.

Aghanim, N., et al. (2020). Planck 2018 results-VI. Cosmological parameters. *Astronomy & Astrophysics* 641, A6.

Aharonov, Y., & Vaidman, L. (2008). The two-state vector formalism: an updated review. *Lect. Notes Phys*. 734, 399-447.

Ahmed, A. (2007). Agency and causation. In Price, H., & Corry, R. (eds.). *Causation, physics, and the constitution of reality*. Oxford University Press, 120-55.

Ahmed, A. (2018a). Self-control and hyperbolic discounting. In Bermúdez, J. (ed.). *Self-control, decision theory, and rationality: New Essays*, 96-120.

Ahmed, A. (2018b). Rationality and Future Discounting. *Topoi* https://doi.org/10.1007/s11245-018-9539-3.

Ainslie, G. (1974). Impulse control in pigeons. *Journal of the Experimental Analysis of Behavior* 21, 485-9.

Ainslie, G., & Haslam, N. (1992a). Hyperbolic discounting. In Loewenstein, G., & Elster, J. (eds.). *Choice over time*. Russell Sage Foundation, 57-92.

Ainslie, G., & Haslam, N. (1992b). Self-control. In Loewenstein, G., & Elster, J. (eds.). *Choice over time*. Russell Sage Foundation, 177-209.

Akerlof, G., & Kranton, R. (2005) Identity and the economics of organizations. *Journal of Economic Perspectives* 19(1), 9-32.

Albert, D. (2000). *Time and chance*. Harvard University Press.

Albert, D. (2014). the sharpness of the distinction between the past and the future. In Wilson, A. (ed.). *Chance and temporal asymmetry*. Oxford University Press, 159-74.

Allan, L. (1979). The perception of time. *Perception & Psychophysics* 26(5), 340-54.

Ambrus, A, & Rozen, K. (2015). Rationalizing choice with multi-self models. *The Economic Journal* 125(585), 1136-56.

Anderson, J. (1983). *The architecture of cognition*. Harvard University Press.

Anderson, J. (2009). *How can the human mind occur in the physical universe?* Oxford University Press.

Anderson, M., & Chemero, T. (2013). The problem with brain GUTs: conflation of different senses of “prediction” threatens metaphysical disaster. *Behavioral and Brain Sciences* 36(3), 204-05.

Andreoni, J., & Sprenger, C. (2012). Risk preferences are not time preferences. *American Economic Review* 102 (7), 3357–76.

Andreoni, J., & Sprenger, C. (2015). Risk preferences are not time preferences: reply. *American Economic Review* 105(7), 2287-93.

Anscombe, G. (1957). *Intention*. Blackwell.

Antony, L. (2007). Everybody has got it: a defense of non-reductive materialism. In McLaughlin, B., & Cohen, J. (eds.). *Contemporary debates in philosophy of mind*. Blackwell, 143-59.

Arrow, J. (1994). Methodological individualism and social knowledge. *American Economic Review* 84(2), 1-9.

Attema, A., Bleichrodt, H., Gao, Y., Huang, Z., & Wakker, P. (2016). Measuring discounting without measuring utility. *American Economic Review* 106(6), 1476-94.

Axten, N., & Fararo, T. (1977). The information processing representation of institutionalised social action. In Krishnan, P. (ed.). *Mathematical models of sociology. Sociological Review Monograph 24*. University of Keele, 35-77.

Axten, N., & Skvoretz, J. (1980). Roles and role-programs. *Quality and Quantity*, 14(4), 547-583.

Baas, A., & Bihan, B. (2020). What does the world look like according to superdeterminism. *arXiv preprint arXiv:2009. 13908*.

Beebee, H. (2007). Hume on causation: the projectivist interpretation. In Price, H., & Corry, R. (eds.). *Causation, physics, and the constitution of reality*. Oxford University Press, 224-49.

Bell, D. (1960). *The end of ideology*. Harvard University Press.

Bell, J. (1987). Free variables and local causality. In *Speakable and unspeakable in quantum mechanics: collected papers in quantum philosophy*. Cambridge University Press, 100-4.

Benjamin, D., Heffetz, O., Kimball, M., & Rees-Jones, A. (2012). What do you think would make you happier? What do you think you would choose? *American Economic Review* 102(5), 2083-110.

Benzion, U., Rapoport, A., & Yagil, J. (1989). Discount rates inferred from decisions: an experimental study. *Management Science* 35, 270-84.

Bermúdez, J. (ed.). (2018). *Self-control, decision theory, and rationality: new essays*. Cambridge University Press.

Bicchieri, C. (2010). Norms, preferences, and conditional behavior. *Politics, philosophy & economics* 9(3), 297-313.

Binmore, K., & Shaked, A. (2010). Experimental economics: Where next? *Journal of Economic Behavior & Organization* 73, 87-100.

Binmore, K., Swierzbinski, J., & Proulx, C. (2001). Does minimax work? An experimental study. *Economic Journal* 111(473), 445-64.

Blackmore, S. (1999). *The meme machine*. Oxford University Press.

Bradford, W., Dolan, P., & Galizzi, M. (2019). Looking ahead: subjective time perception and individual discounting. *Journal of Risk and Uncertainty* 58, 43-69.

Bradley, R. (2017). *Decision theory with a human face*. Cambridge University Press.

Bratman, M. (2009a). Intention, belief, and instrumental rationality. In Sobel, D., & Wall, S. (eds.). *Reasons for action*. Cambridge University Press, 13-36.

Bratman, M. (2009b). Intention, belief, practical, theoretical. In Robertson, S. (ed.). *Spheres of reason*. Oxford University Press, 29-61.

Brighton, H., & Gigerenzer, G. (2008). Bayesian brains and cognitive mechanisms: harmony or dissonance? In Chater, N., & Oaksford, M. (eds.). *The probabilistic mind: prospects for Bayesian cognitive science*. Oxford University Press, 189-208.

Brink, D. (1997). Moral motivation. *Ethics* 108(1), 4-32.

Broome, J. (1991a). *Weighing goods*. Blackwell.

Broome, J. (1991b). Utility. *Economics and Philosophy* 7, 1-12.

Broome, J. (2004). *Weighing lives*. Oxford University Press.

Brown, H. (2005). *Physical relativity: space-time structure from a dynamical perspective*. Oxford University Press.

Brown, J. (2013). Immediate justification, perception, and intuition. In Tucker, C. (ed.). *Seemings and justification: new essays on dogmatism and phenomenal conservatism*. Oxford University Press, 71-88.

Bruhin, A., Fehr-Duda, H., & Epper, T. (2010). Risk and rationality: uncovering heterogeneity in probability distortion. *Econometrica* 78(4), 1375-412.

Bruni, L., & Sugden, R. (2007). The road not taken: how psychology was removed from economics, and how it might be brought back. *The Economic Journal* 117, 146-73.

Callender, C. (2010). The past hypothesis meets gravity. In Ernst, G., & Hüttemann, A. (eds.). *Time, chance, and reduction: philosophical aspects of statistical mechanics*. Cambridge University Press, 34-58.

Callender, C. (2018). The normative standard for future discounting.

Camerer, C. (2000). Prospect theory in the wild: evidence from the field. Reprinted in Camerer, C., Lowenstein, G., & Rabin, M. (eds.). (2004). *Advances in behavioral economics*. Russell Sage Foundation, 148-61.

Camerer, C., & Lowenstein, G. (2004). Behavioral economics: past, present, future. In Camerer, C., Lowenstein, G., & Rabin, M. (eds.). *Advances in behavioral economics*. Russell Sage Foundation, 3-51.

Camerer, C., Babcock, J., Loewenstein, G., & Thaler, R. (1997). Labor supply of New York city cabdrivers: one day at a time. *Quarterly Journal of Economics* 112(2), 407-41.

Camerer, C., Ho, T., & Chong, J. (2004). A cognitive hierarchy model of games. *Quarterly Journal of Economics* 119(8), 861-98.

Camerer, C., Lowenstein, G., & Rabin, M. (eds.). (2004). *Advances in behavioral economics*. Russell Sage Foundation.

Carnap, R. (1950). *Logical foundations of probability*. University of Chicago Press.

Carroll, S., & Chen, J. (2004). Spontaneous inflation and the origin of the arrow of time. *arXiv:hep-th/0410270*.

Carruthers, P. (2006). *The architecture of mind*. Oxford University Press.

Cartwright, N. (1999). *The dappled world: a study of the boundaries of science*. Cambridge University Press.

Caster, O., & Ekenberg, L. (2012). Combining second-order belief distributions with qualitative statements in decision analysis. In *Managing safety in heterogeneous systems*. Springer, 67-87.

Cattaneo, M. (2011). Belief functions combination without the assumption of independence of the information sources. *International Journal of Approximate Reasoning* 52(3), 299-315.

Chabris, C., Laibson, D., & Schuldt, J. (2010). Intertemporal choice. In Durlauf, S., & Blume, L. (eds.). *Behavioural and experimental economics*. Palgrave Macmillan, 168-77.

Chakraborty, A., Halevy, Y., & Saito, K. (2019). The relation between behavior under risk and over time. Mimeographed, California Institute of Technology.

Chambon, V., & Haggard, P. (2013). Premotor or ideomotor: how does the experience of action come about? In Prinz, W., Beisert, M., & Herwig, A. (eds.). *Action science: foundations of an emerging discipline*. MIT Press, 359-80.

Chang, R. (2009). Voluntarist reasons and the sources of normativity. In Sobel, D., & Wall, S. (eds.). *Reasons for action*. Cambridge University Press, 243-71.

Chater, N., & Oaksford, M. (eds.). (2008). *The probabilistic mind: prospects for Bayesian cognitive science*. Oxford University Press.

Chomsky, N. (1957). *Syntactic structures*. Mouton.

Chomsky, N. (1964). *Current issues in linguistic theory*. Mouton.

Chomsky, N. (1965). *Aspects of the theory of syntax*. MIT Press.

Chung, S., & Herrnstein, R. (1967). Choice and delay of reinforcement. *Journal of the Experimental Analysis of Behavior* 10(1), 67-74.

Churchland, P. (1981). Eliminative materialism and the propositional attitudes. *Journal of Philosophy* 78(2), 67-90.

Churchland, P. (2000). Cognitive activity in artificial neural networks. In Cummins, R., & Cummins, D. (eds.). *Minds, brains, and computers: The foundations of cognitive science*. Blackwell, 198-216.

Churchland, P. (2007). The evolving fortunes of eliminative materialism. In McLaughlin, B., & Cohen, J. (eds.). *Contemporary debates in philosophy of mind*. Blackwell, 160-81.

Clapin, H. (ed.). (2002). *Philosophy of mental representation*. Oxford University Press.

Clark, A. (1998). Embodied, situated, and distributed cognition. In Bechtel, W., & Graham, G. (eds.). *A companion to cognitive science*. Blackwell, 506-17.

Clark, A. (2013). Whatever Next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences* 36, 181-253.

Clark, A. (2016). *Surfing uncertainty: prediction, action, and the embodied mind*. Oxford University Press.

Clark, A. (2017). Predictions, precision, and agentive attention. *Consciousness and Cognition* 56, 115-119.

Clark, A., & Chalmers, D. (1998). The extended mind. *Analysis* 58(1), 7-19.

Clark, P. (2001), Statistical mechanics and the propensity interpretation of probability. In Bricmont, J., Dürr, D., Galavotti, M., Ghirardi, G., Petruccione, F., & Zanghì, N. (eds.). *Chance in physics: foundations and perspectives*. Springer, 271–81.

Clark, P. (2009). Mackie's motivational argument. In Sobel, D., & Wall, S. (eds.). *Reasons for action*. Cambridge University Press, 200-18.

Colombo, M., & Wright, C. (2017). Explanatory pluralism: an unrewarding prediction error for free energy theorists. *Brain and Cognition* 112, 3-12.

Cosmides, L., & Tooby, J. (1997). Evolutionary psychology: a primer.

Cotter, J., Brand, C., Knobloch, C., Lilach, Y., Cheshnovsky, O., & Arndt, M. (2017). In search of multipath interference using large molecules. *Science Advances* 3(8), e1602478.

Cramer, J. (1986). The transactional interpretation of quantum mechanics. *Reviews of Modern Physics* 58, 647-87.

Creswell, M. (2004). Adequacy conditions for counterpart theory. In Jackson, F., & Priest, G. (eds.). *Lewisian themes: the philosophy of David K. Lewis*. Oxford University Press, 29-42.

Cullis, J., & Jones, P. (2008). How big should government be? In Lewis, A. (ed.). *The Cambridge handbook of psychology and economic behaviour*. Cambridge University Press, 281-303.

Cummins, R. (1991). The role of mental meaning in psychological explanation. In McLaughlin, B. P. (ed.). *Dretske and his critics*. Blackwell, 102-117.

Cummins, R., & Cummins, D. (eds.). (2000). *Minds, brains, and computers: The foundations of cognitive science*. Blackwell.

D’Arms, J., & Jacobson, D. (2009). Regret and irrational action. In Sobel, D., & Wall, S. (eds.). *Reasons for action*. Cambridge University Press, 179-99.

Darwall, S. (2009). Authority and second-personal reasons for acting. In Sobel, D., & Wall, S. (eds.). *Reasons for action*. Cambridge University Press, 134-54.

Darwin, C. (1859). *On the origin of species by means of natural selection*. John Murray.

Davidson, D. (2004). *Problems of rationality*. Oxford University Press.

Davies, P. (1994). Stirring up trouble. In Halliwell, J., Pérez-Mercader, J. and Zurek, W. (eds.). *Physical origins of time asymmetry*. Cambridge University Press, 119-30.

Daw, N., Courville, A., & Dayan, P. (2008). Semi-rational models of conditioning: the case of trial order. In Chater, N., & Oaksford, M. (eds.). *The probabilistic mind: prospects for Bayesian cognitive science*. Oxford University Press, 431-52.

De Finetti, B. (1937). La prévision: ses lois logiques, ses sources subjectives. translated as, Foresight: its logical laws, its subjective sources. In Kolz, S., & Johnson, N. (eds.). (1992). *Breakthroughs in statistics: foundations and basic theory*. Springer, 134-74.

De Herdt, T. (2003). Cooperation and fairness: the Flood-Dresher experiment revisited. *Review of Social Economics* 61(2), 183-210.

Dehling, H. (1997). Daniel Bernoulli and the St. Petersburg paradox. *Nieuw archief voor wiskunde* 15, 223-8.

Dennett, D. (1987). *The intentional stance*. MIT Press.

De Regt, H. (2005). Scientific realism in action: molecular models and Boltzmann’s Bildtheorie. *Erkenntnis* 63(2), 205-30.

Dhami, S. (2016). *The foundations of behavioral economic analysis*. Oxford University Press.

Dicken, P., & Lipton, P. (2006). What can Bas believe? Musgrave and van Fraassen on observability. *Analysis* 66(3), 226-33.

Dietrich, F., & List, C. (2016). Mentalism versus behaviourism in economics: a philosophy-of-science approach. *Economics and Philosophy* 32(2), 249-81.

Dreber, A., Fudenberg, D., Levine, D., & Rand, D. (2016). Self-control, social preferences and the effect of delayed payments.

Dretske, F. (1988). *Explaining behavior: reasons in a world of causes*. MIT Press.

Dretske, F. (1997). *Naturalizing the mind*. MIT Press.

Driver, J. (2014). The history of utilitarianism. In Zalta, E. (ed.). *The Stanford encyclopedia of philosophy* (Winter 2014 edition).

Dupré, J. (2012). *Processes of life*. Oxford University Press.

Eagle, A. (2004). Twenty-one arguments against propensity analyses of probability. *Erkenntnis* 60, 371-416.

Eagle, A. (2016). Probability and randomness. In Hájek, A., & Hitchcock, C. (eds.). *Oxford handbook of probability and philosophy*. Oxford University Press, 440-59.

Earman, J. (1986). *A primer on determinism*. Reidel.

Earman, J. (2006). The past hypothesis: not even false. *Studies in History and Philosophy of Modern Physics* 37(3), 399-430.

Ebert, J. (2010). The surprisingly low motivational power of future rewards: Comparing conventional money-based measures of discounting with motivation-based measures. *Organizational Behavior and Human Decision Processes* 111, 71-92.

Ebert, J., & Prelec, D. (2007). The fragility of time: time-insensitivity and valuation of the near and far future. *Management Science* 53(9), 1423-38.

Egan, A. (2004). Second-order predication and the metaphysics of properties. In Jackson, F., & Priest, G. (eds.). *Lewisian themes: the philosophy of David K. Lewis*. Oxford University Press, 49-67.

Eibenberger, S., Gerlich, S., Arndt, M., Mayor, M., & Tüxen, J. (2013). Matter-wave interference of particles selected from a molecular library with masses exceeding 10000 amu. *Physical Chemistry Chemical Physics* 15(35), 14696-700.

Eiler, B., Kallen, R., & Richardson, M. (2017). Interaction-dominant dynamics, timescale enslavement, and the emergence of social behavior. In Vallacher, R., Read, S., & Nowak, A. (eds.). *Computational social psychology*. Routledge, 105-26.

Elga, A. (2007). Isolation and folk physics. In Price, H., & Corry, R. (eds.). *Causation, physics, and the constitution of reality*. Oxford University Press, 106-19.

Elster, J. (1977). Ulysses and the sirens: a theory of imperfect rationality. *Social Science Information* 16(5), 469-526.

Elster, J. (2015). *Explaining social behavior: more nuts and bolts for the social sciences*. Cambridge University Press.

Epper, T., Fehr-Duda, H., & Bruhin, A. (2011). Viewing the future through a warped lens: why uncertainty generates hyperbolic discounting. *Journal of Risk and Uncertainty* 43,169-203.

Ericson, K., & Laibson, D. (2018). Intertemporal choice. NBER Working Paper No. 25358. National Bureau Of Economic Research.

Ericson, K., White, J., Laibson, D., & Cohen, J. (2015). Money earlier or later? Simple heuristics explain intertemporal choices better than delay discounting. NBER Working Paper No. 20948.

Ernst, G., & Hüttemann, A. (eds.). (2010). *Time, chance, and reduction: philosophical aspects of statistical mechanics*. Cambridge University Press.

Evans, P. (2011). A study of time in modern physics. University of Sydney.

Evans, P. (2015). Retrocausality at no extra cost. *Synthese*. 192(4), 1139-55.

Fararo, J. (2009). Generativity. In Cherkaoui, M., & Hamilton, P. (eds.). *Boudon: a life in sociology, vol. 2*, Bardwell Press, 393-410.

Fehr, E., Kremhelmer, S., & Schmidt, K. (2008). Fairness and the optimal allocation of ownership rights, *Economic Journal* 118(531), 1262-84.

Fehr-Duda, H., & Epper, T. (2012). Probability and risk: foundations and economic implications of probability-dependent risk preferences. *Annual Review of Economics* 4, 567-93.

Fehr-Duda, H., Bruhin, A., Epper, T., & Schubert, R. (2010). Rationality on the rise: why relative risk aversion increases with stake size. *Journal of Risk and Uncertainty* 40(2), 147-80.

Feldman, F. (2004). *Pleasure and the good life: concerning the nature, varieties, and plausibility of hedonism*. Oxford University Press.

Feldman, R. (2003). *Epistemology*. Prentice-Hall.

Ferreira, F. (2005). Psycholinguistics, formal grammars, and cognitive science. *The Linguistic Review* 22, 365-80.

Fetzer, J. (1982). Probabilistic explanations. *PSA: Proceedings of the biennial meeting of the Philosophy of Science Association* 1982(2), 194-207.

Feynman, R. (1965). *The character of physical law*. MIT Press.

Feynman, R. (1985). *QED: the strange theory of light and matter*. Princeton University Press.

Fine, A. (1984). The natural ontological attitude. In Lepin, J. (ed.). *Scientific realism*. University of California Press, 83-107.

Fisher, R. (1937). *The design of experiments, 2nd edition*. Oliver and Boyd.

Fisher, R. (1956). *Statistical methods and scientific inference*. Oliver and Boyd.

Flood, M. (1958). Some experimental games. *Management Science* 5, 5-26.

Fodor, J. (1975). *The language of thought*. Thomas Crowell.

Fodor, J. (1983). *The modularity of mind*. MIT Press.

Fodor, J. (1998). *Concepts: where cognitive science went wrong*. Oxford University Press.

Fodor, J. (2000). *The mind doesn't work that way: the scope and limits of computational psychology*. MIT Press.

Fodor, J., & McLaughlin, B. (2000). Connectionism and the problem of systematicity: why Smolensky's solution doesn't work. In Cummins, R., & Cummins, D. (eds.). *Minds, brains, and computers: the foundations of cognitive science*. Blackwell, 273-85.

Fracchia, J., & Lewontin, R. (2005). The price of metaphor. *History and Theory* 44(1), 14-20.

Francis, G. (2012). Publication bias and the failure of replication in experimental psychology, *Psychon Bull Rev* 19(6), 975-91.

Frautsci, S. (1982). Entropy in an expanding universe. *Science* 217(4560), 593-9.

Frederick, S., Loewenstein, G., & O'Donoghue, T. (2002). Time discounting: a critical review. Reprinted in Camerer, C., Lowenstein, G., & Rabin, M. (eds.). (2004). *Advances in behavioral economics*. Russell Sage Foundation, 162-222.

Frenkel, D. (1993). Order through disorder: entropy strikes back. *Physics World* 6(2), 24-5.

Friedman, M. (1953). *The methodology of positive economics*. University of Chicago Press.

Frigg, R. (2008). A field guide to recent work on the foundations of statistical mechanics. In Rickles, D. (ed.). *The Ashgate companion to contemporary philosophy of physics*. Ashgate Publishing, 99-196.

Frigg, R. (2010). Probability in Boltzmannian statistical mechanics. In Ernst, G., & Hüttemann, A. (eds.). *Time, chance, and reduction: philosophical aspects of statistical mechanics*. Cambridge University Press, 92-118.

Frisch, M. (2007). Causation, counterfactuals, and entropy. In Price, H., & Corry, R. (eds.). *Causation, physics, and the constitution of reality*. Oxford University Press, 351-96.

Frisch, M. (2010). Does a low-entropy constraint prevent us from influencing the past? in Ernst, G., & Hüttemann, A. (eds.). *Time, chance, and reduction: philosophical aspects of statistical mechanics*. Cambridge University Press, 13-33.

Frisch, M. (2014). Why physics can't explain everything, in Wilson, A. (ed.) *Chance and temporal asymmetry*. Oxford University Press, 221-40.

Friston, K. (2010). The free-energy principle: a unified brain theory? *Nature Reviews Neuroscience* 11(2), 127.

Friston, K., & Kiebel, S. (2009). Predictive coding under the free-energy principle. *Philosophical Transactions of the Royal Society B: Biological Sciences* 364(1521), 1211-21.

Friston, K., FitzGerald, T., Rigoli, F., Schwartenbeck, P., & Pezzulo, G. (2016). Active inference and learning. *Neuroscience and Biobehavioral Reviews* 68, 862-79.

Friston, K., Mattout, J., & Kilner, J. (2011). Action understanding and active inference. *Biological Cybernetics* 104(1-2), 137-60.

Friston, K., Thornton, C., & Clark, A. (2012). Free-energy minimization and the dark-room problem. *Frontiers in Psychology* 3, 130.

Frith, C. (2007). *Making up the mind: how the brain creates our mental world*. Blackwell.

Froese, T., & Ikegami, T. (2013). The brain is not an isolated “black box,” nor is its goal to become one. *Behavioral and Brain Sciences* 36(3), 213-214.

Fudenberg, D., & Levine, D. (2016). Whither game theory? Towards a theory of learning in games. *Journal of Economic Perspectives* 30(4), 151-70.

Fukuyama, F. (1992). *The end of history and the last man*. Free Press.

Fumagalli, R. (2013). The futile search for true utility. *Economics and Philosophy* 29, 325-47.

Fumagalli, R. (2016). Economics, psychology, and the unity of the decision sciences. *Philosophy of the social sciences* 46(2), 103-28.

Fumagalli, R. (2019). (F)utility exposed. *Philosophy of Science* 86, 955-6.

Fumagalli, R. (2020a). How thin rational choice theory explains choices. *Studies in the History and Philosophy of Science Part A* 83, 63-74.

Fumagalli, R. (2020b). On the individuation of choice options. *Philosophy of the social sciences* 50(4), 338-65.

Gaifman, H. (1986). A theory of higher order probabilities. In Halpern, J. (ed.). *Theoretical aspects of reasoning about knowledge, Proceedings of the 1986 conference*. Morgan Kaufmann, 275-92.

Gibbard, A., & Harper, W. (1978). Counterfactuals and two kinds of expected utility. In Hooker, A., Leach, J., & McClennen, E. (eds.). *Foundations and applications of decision theory*. Kluwer, 125-62.

Gigerenzer, G. (1991). How to make cognitive illusions disappear: beyond "heuristics and biases". *European Review of Social Psychology* 2(1), 83-115.

Gilboa, I. (2009). *Theory of decision under uncertainty*. Cambridge University Press.

Gillett, C., & Loewer, B. (eds.). (2001). *Physicalism and its discontents*. Cambridge University Press.

Gilovich, T., Griffin, D., & Kahneman, D. (2002). *Heuristics and biases: the psychology of intuitive judgment*. Cambridge University Press.

Gintis, H. (2009). *The bounds of reason: game theory and the unification of the social sciences*. Princeton University Press.

Gintis, H. (2011). Gene-culture coevolution and the nature of human sociality. *Philosophical Transactions of the Royal Society* 366, 878-88.

Gleick, J. (1988). *Chaos: making a new science*. Heinemann.

Godfrey-Smith, P. (1994). A modern history theory of functions. *Noûs* 28(3), 344-62.

Godfrey-Smith, P. (2009). *Darwinian populations and natural selection*. Oxford University Press.

Goeree, J., & Holt, C. (2004). A model of noisy introspection. *Games and Economic Behavior* 46(2), 365-82.

Goldman, A. (2012). *Reliabilism and Contemporary Epistemology*. Oxford University Press.

Gooch, G. (1911). *History of our time, 1885-1911*. H. Holt.

Gould, S. (1990). Darwin and Paley meet the invisible hand. *Natural History* 99(11), 8-12.

Gould, S., & Lewontin, R. (1979). The spandrels of San Marco and the Panglossian paradigm: a critique of the adaptationist programme. *Proceedings of the Royal Society of London B* 205(1161), 581-98.

Grayling, A. (2003). *What is good?* Weidenfeld & Nicolson.

Greene, P., & Sullivan, M. (2015). Against time bias. *Ethics* 125(4), 947-70.

Green, S. (2014). A philosophical evaluation of adaptationism as a heuristic strategy. *Acta Biotheoretica* 62(4), 479-98.

Greenwood, J. (2015). *A conceptual history of psychology: exploring the tangled web, 2nd edition*. Cambridge University Press.

Griffiths, T., & Yuille, A. (2008). A primer on probabilistic inference. In Chater, N., & Oaksford, M. (eds.). *The probabilistic mind: prospects for Bayesian cognitive science*. Oxford University Press. 33-57.

Guala, F. (2019). Preferences: neither behavioural nor mental. *Economics & Philosophy* 35(3), 383-401.

Gul, F. (1991). A theory of disappointment aversion. *Econometrica* 59(3), 667-86.

Güth, W., Schmittberger, R., & Schwarze, B. (1982). An experimental analysis of ultimatum bargaining. *Journal of Economic Behavior and Organization* 3(4), 367-88.

Gyntelberg, J., & Hansen, F. (2004). Expected utility theory with "small worlds". Institute of Economics, University of Copenhagen.

Hafner, R., Hertweck, T., Klöppner, P., Bloesch, M., Neunert, M., Wulfmeier, M., Tunyasuvunakool, S., Heess, N., & Riedmiller, M. (2020). Towards general and autonomous learning of core skills: a case study in locomotion. *arXiv preprint arXiv:2008.12228*.

Hájek, A. (1997). 'Mises Redux' – Redux: fifteen arguments against finite frequentism. *Erkenntnis* 45, 209-27.

Hájek, A. (2003). What conditional probability could not be. *Synthese* 137(3), 273-323.

Hájek, A. (2007). The reference class problem is your problem too. *Synthese* 156, 563-85.

Hájek, A. (2009). Fifteen arguments against hypothetical frequentism. *Erkenntnis* 70, 211-35.

Hájek, A. (2011). Conditional Probability. In Bandyopadhyay, P., & Forster, M. (eds.). *Philosophy of Statistics*. Elsevier, 99-135.

Hájek, A. (2019). Interpretations of probability. In Zalta, E. (ed.). *The Stanford encyclopedia of philosophy* (Fall 2019 edition).

Hájek, A., & Hitchcock, C. (eds.). (2016). *The Oxford handbook of probability and philosophy*. Oxford University Press

Halevy, Y. (2014). Some comments on the use of monetary and primary rewards in the measurement of time preferences, Technical Report, University of British Columbia.

Halevy, Y. (2015). Time consistency: stationarity and time invariance. *Econometrica* 83(1), 335-52.

Hall, N. (2004). Two mistakes about credence and chance. In Jackson, F., & Priest, G. (eds.). *Lewisian themes: the philosophy of David K. Lewis*. Oxford University Press, 94-112.

Han, R., & Takahashi, T. (2012). Psychophysics of time perception and valuation in temporal discounting of gain and loss. *Physica A* 391, 6568-76.

Hanson, N. (1958). *Patterns of discovery*. Cambridge University Press.

Hansson, P., Juslin, P., & Winman, A. (2008). The naïve intuitive statistician: organism-environment relations from yet another angle. In Chater, N., & Oaksford, M. (eds.). *The probabilistic mind: prospects for Bayesian cognitive science*. Oxford University Press, 237-259.

Hardisty, D., & Pfeffer, J. (2017). Intertemporal uncertainty avoidance: when the future is uncertain, people prefer the present, and when the present is uncertain, people prefer the future.

Harsanyi, J. (1977). *Rational behavior and bargaining equilibrium in games and social situations*. Cambridge University Press.

Hart, O., & Moore, J. (2008). Contracts as reference points. *Quarterly Journal of Economics* 123(1), 1-48.

Hauptmann, A., & Adler, J. (2020). On the unreasonable effectiveness of CNNs. *arXiv preprint arXiv:2007.14745*.

Hawkins, G., Forstmann, B., Wagenmakers, E., Ratcliff, R., & Brown, S. (2015). Revisiting the evidence for collapsing boundaries and urgency signals in perceptual decision-making. *Journal of Neuroscience* 35(6), 2476-84.

Helman, C. (2007). *Culture, health and illness, 5th edition*. Hodder Arnold.

Helmholtz, H. (1878). The facts of perception. From *Selected Writings of Hermann Helmholtz*. Wesleyan University Press.

Hertwig, R., & Pleskac, T. (2008). The game of life: how small samples render choice simpler. In Chater, N., & Oaksford, M. (eds.). *The probabilistic mind: prospects for Bayesian cognitive science*. Oxford University Press, 209-35.

Heyes, C. (2019). Enquire within: cultural evolution and cognitive science. *Philosophical Transactions of the Royal Society B* 373(1743), 20170051.

Hilton, D. (2008). Theory and method in economics and psychology. In Lewis, A. (ed.). *The Cambridge handbook of psychology and economic behaviour*. Cambridge University Press.

Hinton, G. (2017). https://www.axios.com/artificial-intelligence-pioneer-says-we-need-to-start-over-1513305524-f619efbd-9db0-4947-a9b2-7a4c310a28fe.html.

Hitchcock, C. (2007). What Russell got right. In Price, H., & Corry, R. (eds.). *Causation, physics, and the constitution of reality*. Oxford University Press, 45-65.

Hodgson, G. (2007). Meanings of methodological individualism. *Journal of Economic Methodology* 14(2), 211-26.

Hofer-Szabó, G., Rédei, M., & Szabó, L. (2013). *The principle of the common cause*. Cambridge University Press.

Hohwy, J. (2013). *The predictive mind*. Oxford University Press.

Hohwy, J. (2016). The self-evidencing brain. *Noûs* 50(2), 259-285.

Holyoak, K., & Morrison, R. (eds.). (2005). *The Cambridge handbook of thinking and reasoning*. Cambridge University Press.

Hong, C., & Sagi, J. (2003). Small worlds: modelling attitudes towards sources of uncertainty.

Horst, S. (2011). *Laws, mind, and free will*. MIT Press.

Hossenfelder, S. (2020). Superdeterminism: a guide for the perplexed. *arXiv preprint arXiv:2009. 13908*.

HSE. (1997). *The explosion and fires at the Texaco refinery, Milford Haven, 24 July 1994*. HSE Books.

Hsee, C., Abelson, R., & Salovey, P. (1991). The relative weighting of position and velocity in satisfaction. *Psychological Science* 2(4), 263-6.

Huber, D. (2008). Causality in time: explaining away the future and the past. In Chater, N., & Oaksford, M. (eds.). *The probabilistic mind: prospects for Bayesian cognitive science*. Oxford University Press, 351-76.

Hughes, R. (1989). *The structure and interpretation of quantum mechanics*. Harvard University Press.

Hull, C. (1943). *The principles of behavior*. Appleton-Century-Crofts.

Hume, D. (1739). *A treatise on human nature*.

Hurka, T. (2019). More seriously wrong, more importantly right. *Journal of the American Philosophical Association* 5(1), 41-58.

Hutchins, E. (1995). *Cognition in the wild*. MIT Press.

Hutto, D., & Ratcliffe, M. (eds.). (2007). *Folk psychology re-assessed*. Springer.

Ioannidis, J. (2005). Why most published research findings are false. *PLoS Med*, 2(8): e124.

Jackson, T. (2008). Sustainable consumption and lifestyle change. In Lewis, A. (ed.). *The Cambridge handbook of psychology and economic behaviour*. Cambridge University Press, 335-62.

Jeffrey, R. (1965). *The logic of decision*. McGraw-Hill.

Jenkins, E., DeChant, M., & Perry, E. (2018). When the nose doesn’t know: canine olfactory function associated with health, management, and potential links to microbiota. *Frontiers in Veterinary Science* 5(56).

Joyce, J. (1999). *The foundations of causal decision theory*. Cambridge University Press.

Joyce, J. (2010). A defence of imprecise credences in inference and decision making. *Philosophical Perspectives* 24, 281-323.

Kahneman, D. (2012). *Thinking fast and slow*. Penguin.

Kahneman, D., & Tversky, A. (1979). Prospect theory: an analysis of decision under risk. *Econometrica* 47(2), 263-91.

Kahneman, D., Knetsch, J., & Thaler, R. (1990). Experimental tests of the endowment effect and the Coase theorem. *Journal of Political Economy* 98(6), 1325-48.

Kaidesoja, T. (2012). The DBO theory of action and distributed cognition. *Social Science Information* 51(3), 311-37.

Katz, J., Redelmeier, D., & Kahneman, D. (1997). Memories of painful medical procedures. Paper presented at the American Pain Society 15th Annual Meeting.

Keefe, R. (2000). *Theories of vagueness*. Cambridge University Press.

Keren, G., & Roelofsma, P. (1995). Immediacy and certainty in intertemporal choice. *Organizational Behavior and Human Decision Processes* 63(3), 287-97.

Keynes, J. (1921). *A treatise on probability*. Macmillan.

Kiefer, A., & Hohwy, J. (2018). Content and misrepresentation in hierarchical generative models. *Synthese* 195(6), 2387-415.

Kiefer, C. (2010). Quantum gravity and the arrow of time. In Ernst, G., & Hüttemann, A. (eds.). *Time, chance, and reduction: philosophical aspects of statistical mechanics*. Cambridge University Press, 59-67.

Kim, B., & Zauberman, G. (2009). Perception of anticipatory time in temporal discounting. *Journal of Neuroscience, Psychology, and Economics* 2(2), 91-101.

Kim, J. (2001). Mental causation and consciousness: the two mind-body problems for the physicalist. In Gillett, C., & Loewer, B. (eds.). *Physicalism and its discontents*. Cambridge University Press, 271-83.

Kim, J. (2005). *Physicalism, or something near enough*. Princeton University Press.

Kitcher, P. (1985). *Vaulting ambition: socio-biology and the quest for human nature*. MIT Press.

Klumpp, V., & Hanebeck, U. (2009). Bayesian estimation with uncertain parameters of probability density functions. In *2009 12th international conference on information fusion*. IEEE, 1759-66.

Knight, F. (1921). *Risk, uncertainty and profit*. Houghton Mifflin.

Kogo, N., & Trengove, C. (2015). Is predictive coding theory articulated enough to be testable? *Frontiers in Computational Neuroscience* 9, 111.

Kolmogorov, A. (1933). *Grundbegriffe der Wahrscheinlichkeitrechnung*. Translated as *Foundations of the theory of probability* (1950). Chelsea Publishing Company.

Konstantinidis, E., van Ravenzwaaij, D., Güney, S., & Newell, B. (2018). Now for sure or later with a risk? Modeling risky inter-temporal choice as accumulated preference. *Decision*.

Koopmans, T. (1960). Stationary ordinal utility and impatience. *Econometrica* 28, 287-309.

Krabbe, P., Salomon, J., & Murray, J. (2007). Quantification of health states with rank-based nonmetric multidimensional scaling. *Medical Decision Making* 27, 395.

Kroes, P. (1985). *Time: its structure and role in physical theories*. D. Reidel.

Kunreuther, H., Ginsberg, R., Miller, L., Sagi, P., Slovic, P., Borkan, B., & Katz, N. (1978). *Disaster insurance protection: public policy lessons*. Wiley.

Kwisthout, J., & van Rooij, I. (2015). Free energy minimization and information gain: the devil is in the details. *Cognitive Neuroscience* 6(4), 216-18.

Ladyman, J. (2002). *Understanding philosophy of science*. Routledge.

Ladyman, J., & Ross, D. (2007). *Every thing must go*. Oxford University Press.

Laibson, D. (1994). Essays in hyperbolic discounting. *PhD diss. MIT*.

Lakatos, I. (1970). Falsification and the methodology of scientific research programmes. In Lakatos, I., & Musgrave, A. (eds.). *Criticism and the growth of knowledge*. Cambridge University Press, 91-196.

Lakatos, I., & Musgrave, A. (eds.). (1970). *Criticism and the growth of knowledge*. Cambridge University Press.

Lam, V., & Esfeld, M. (2012). The structural metaphysics of quantum theory and general relativity. *Journal for General Philosophy of Science* 43(2), 243-58.

Larson, E. (2004). *Evolution: the remarkable history of a scientific theory*. Random House.

Larsson, J. (2014). Loopholes in Bell inequality tests of local realism. *Journal of Physics A: Mathematical and Theoretical* 47(42), 424003.

Laudan, L. (1981). A confutation of convergent realism. *Philosophy of Science* 48(1), 19-49.

Laughlin, R. (2005). *A different universe*. Basic Books.

Law, J. (2016). *Oxford dictionary of business and management, 6th edition*, Oxford University Press.

Lawrence, M., O'Connor, M., & Edmundson, B. (2000). A field study of sales forecasting accuracy and processes. *European Journal of Operational Research* 122(1), 151-60.

Lea, S. (2008). Evolutionary psychology and economic psychology. In Lewis, A. (ed.). *The Cambridge handbook of psychology and economic behaviour*. Cambridge University Press, 512-26.

Lebowitz, J. (1993). Boltzmann's entropy and time's arrow. *Physics Today* 46(9), 32-8.

Lebowitz, J. (1994). Time's arrow and Boltzmann's entropy. In Halliwell, J., Pérez-Mercader, J., & Zurek, W. (eds.). *Physical origins of time asymmetry*. Cambridge University Press, 131-46.

Leeds, S. (2001). Possibility: physical and metaphysical. In Gillett, C., & Loewer, B. (eds.). *Physicalism and its discontents*. Cambridge University Press, 172-93.

Levitt, S., & List, J. (2007). What do laboratory experiments measuring social preferences reveal about the real world? *Journal of Economic Perspectives* 21(2), 153-74.

Lewandowsky, S., & Farrell, S. (2011). *Computational modeling in cognition: principles and practice*. Sage Publications.

Lewens, T. (2015). *Cultural evolution: conceptual challenges*. Oxford University Press.

Lewis, A. (ed.). (2008). *The Cambridge handbook of psychology and economic behaviour*. Cambridge University Press.

Lewis, D. (1986). A subjectivist's guide to objective chance, with postscripts. In *Philosophical papers 2*. Oxford University Press, 83-132.

Li, W. (2002). Zipf's law everywhere. *Glottometrics* 5, 14-21.

Lin, J. (1995). The Needham Puzzle: why the industrial revolution did not originate in China. *Economic Development and Cultural Change* 43(2), 269-92.

Little, D. (1984). Reflective equilibrium and justification. *The Southern Journal of Philosophy* 22(3), 373-387.

Lockwood, M. (2005). *The labyrinth of time*. Oxford University Press.

Loewenstein, G. (1987). Anticipation and the value of delayed consumption. *The Economics Journal* 97, 666-84.

Loewenstein, G. (1988). Frames of mind in intertemporal choice. *Management Science* 34(2), 200-14.

Loewenstein, G. (1992). The fall and rise of psychological explanations in the economics of intertemporal choice. In Loewenstein, G., & Elster, J. (eds.). *Choice over time*. Russell Sage Foundation, 3-34.

Loewenstein, G. (1996). Out of control: visceral influences on behavior. *Organizational Behavior and Human Decision Processes* 65(3), 272-92.

Loewenstein, G. (2010). Insufficient emotion: soul-searching by a former indicter of strong emotions. *Emotion Review* 2(3), 234-9.

Loewenstein, G., & Elster, J. (eds.). (1992). *Choice over time*. Russell Sage Foundation.

Loewenstein, G., & Prelec, D. (1992). Anomalies in intertemporal choice: evidence and an interpretation. In Loewenstein, G., & Elster, J. (eds.). *Choice over time*. Russell Sage Foundation, 119-45.

Loewenstein, G., & Prelec, D. (1993). Preferences for sequences of outcomes. *Psychological Review* 100(1), 91-108.

Loewenstein, G., Weber, R., Flory, J., Manuck, S., & Muldoon, M. (2001). Dimensions of time discounting. Presented at conference on survey research on household expectations and preferences, Ann Arbor, Nov. 2–3.

Loewer, B. (2007). Counterfactuals and the second law. In Price, H., & Corry, R. (eds.). *Causation, physics, and the constitution of reality*. Oxford University Press, 293-326.

Loewer, B. (2012a). The emergence of time's arrows and special science laws from physics. *Interface Focus* 2, 13-9.

Loewer, B. (2012b). Two accounts of laws and time. *Philosophical Studies* 160(1), 115-37.

Lohrenz, T., & Montague, P. (2008). Neuroeconomics: what neuroscience can learn from economics. In Lewis, A. (ed.). *The Cambridge handbook of psychology and economic behaviour*. Cambridge University Press, 457-92.

Loomes, G., & Sugden, R. (1987). Some implications of a more general form of regret theory. *Journal of Economic Theory* 41(2), 270-87.

Luce, R., & Suppes, P. (1965). Preference, utility, and subjective probability. In Luce, R., Bush, R., & Galanter, E. (eds.). *Handbook of mathematical psychology, vol. III*. Wiley, 249-410.

Luckman, A., Donkin, C., & Newell, B. (2017). People wait longer when the alternative is risky: the relation between preferences in risky and inter-temporal choice. *Journal of Behavioral Decision Making* 30(5), 1078-92.

Macaulay, T. (1848). *The history of England from the accession of James II*. Porter & Coates.

McKelvey, R., & Parlfrey, T. (1992). An experimental study of the centipede game. *Econometrica* 60(4), 803-36.

McLaughlin, B. (ed.). (1991). *Dretske and his critics*. Blackwell.

McLaughlin, B., & Cohen, J. (eds.). (2007). *Contemporary debates in philosophy of mind*. Blackwell.

McTaggart, J. (1908). The unreality of time. *Mind* 17, 457-73.

Maher, P. (2006). The concept of inductive probability. *Erkenntnis* 65(2), 185-206.

Maudlin, T. (2002). *Quantum non-locality and relativity, 2nd edition*. Blackwell.

Mazur, J. (1987). An adjusting procedure for studying delayed reinforcement. In Commons, J., Mazur, J., Nevin, J., & Rachlin, H. (eds.). *Quantitative analyses of behavior V: the effect of delay and of intervening events on reinforcement value*. Erlbaum.

Mellor, D. (1971). *The matter of chance*. Cambridge University Press.

Menary, R. (2007). *Cognitive integration: mind and cognition unbounded*. Palgrave Macmillan.

Menzies, P., & Price, H. (1993). Causation as a secondary quality. *British Journal for the Philosophy of Science* 44(2), 187-203.

Miao, B., & Zhong, S. (2012). Separating risk preference and time preference. Department of Economics, National University of Singapore.

Miao, B., & Zhong, S. (2015). Risk preferences are not time preferences: separating risk and time preference: comment. *American Economic Review* 105(7), 2272-86.

Miller, G. (1956). The magical number seven, plus or minus two: some limits on our capacity for processing information. *Psychological Review* 63(2), 81-97.

Miller, G., Galanter, E., & Pribram, K. (1960). *Plans and the structure of behavior*. Holt, Rinehart & Winston.

Mischel, W., Shoda, Y., & Rodriguez, M. (1992). Delay of gratification in children. In Loewenstein, G., & Elster, J. (eds.). *Choice over time*. Russell Sage Foundation, 147-64.

Moulines, C. (2010). The crystallization of Clausius's phenomenological thermodynamics. . In Ernst, G., & Hüttemann, A. (eds.). *Time, chance, and reduction: philosophical aspects of statistical mechanics*. Cambridge University Press, 139-58.

Nagel, T. (1970). *The possibility of altruism*. Oxford University Press.

Naur, P. (2007). Computing versus human thinking. *Communications of the ACM* 50(1), 85-94.

Nelson, J. (2008). Towards a rational theory of human information acquisition. In Chater, N., & Oaksford, M. *The probabilistic mind: prospects for Bayesian cognitive science*. Oxford University Press, 143-63.

Nersessian, N. (1992). How do scientists think? Capturing the dynamics of conceptual change in science. *Cognitive models of science* 15, 3-44.

Nersessian, N. (2008). *Creating Scientific Concepts*. MIT Press.

Newell, A. (1994). *Unified theories of cognition*. Harvard University Press.

Newell, A., & Simon, H. (1972). *Human problem solving*. Prentice-Hall.

Newell, A., Shaw, J., & Simon, H. (1958). Elements of a theory of human problem solving. *Psychological review* 65(3), 151-166.

Ng, Y. (2005). Intergenerational impartiality: replacing discounting by probability weighting. *Journal of Agricultural and Environmental Ethics* 18, 237–257.

Nickerson, R. (2002). The production and perception of randomness. *Psychological Review* 109(2), 330-57.

Niven, R. (2005). Exact Maxwell-Boltzmann, Bose-Einstein and Fermi-Dirac statistics. *Physics Letters* 342(4), 286-93.

Norton, J. (2007). Causation as folk science. In Price, H., & Corry, R. (eds.). *Causation, physics, and the constitution of reality*. Oxford University Press, 11-44.

O’Donoghue, T., & Rabin, M. (2001). Choice and procrastination. *Quarterly Journal of Economics* 116(1), 121-60.

Okasha, S. (2016). On the interpretation of decision theory. *Economics and Philosophy* 32(3), 409-433.

Okasha, S. (2018). *Agents and goals in evolution*. Oxford University Press.

Ord, T. (2009). Beyond action: applying consequentialism to decision making and motivation. PhD diss. University of Oxford.

Paivio, A. (2007). *Mind and its evolution: a dual coding theoretical approach*. Lawrence Erlbaum Associates.

Parfit, D. (1984). *Reasons and persons*. Oxford University Press.

Park, B., & Rothbart, M. (1982). Perception of out-group homogeneity and levels of social categorization: memory for the subordinate attributes of in-group and out-group members. *Journal of Personality and Social Psychology* 42(6), 1051-68.

Patel, J., & Read, C. (1996). *Handbook of the normal distribution, 2nd ed*. Marcel Dekker.

Patil, Y., Chakram, S., & Vengalattore, M. (2015). Measurement-induced localization of an ultracold lattice gas. *Physical Review Letters* 115, 140402.

Paul, L. (2014). Experience and the arrow. In Wilson, A. (ed.). *Chance and temporal asymmetry*. Oxford University Press, 175-93.

Pearl, J. (2009). *Causality: models, reasoning, and inference, 2nd edition*. Cambridge University Press.

Pearl, J., & Mackenzie, D. (2018). *The book of why: the new science of cause and effect*. Penguin.

Penrose, R. (1989). *The emperor's new mind: concerning computers, minds and the laws of physics*. Oxford University Press.

Penrose, R. (2004). *The Road to Reality*. Jonathan Cape.

Peters, O. (2019). The ergodicity problem in economics. *Nature Physics* 15, 1216-21.

Phelps, E., & Pollak, R. (1968). On second best national savings and game equilibrium growth. *Review of Economic Studies* 35(2), 185-99.

Piccinini, G., & Bahar, S. (2013). Neural computation and the computational theory of cognition. *Cognitive Science* 37(3), 453-88.

Pinker, S. (1997). *How the mind works*. W. W. Norton.

Pinker, S. (2011). *The better angels of our nature: why violence has declined*. Viking Books.

Plotkin, H. (1994). *Darwin machines and the nature of knowledge*. Penguin.

Plott, C. (1996). Rational individual behavior in markets and social choice processes: the discovered preference hypothesis. In Arrow, K., Colombatto, E., Perlman, M., & Schmitt, C. (eds.). *The rational foundations of economic behavior*. St. Martin's Press, 225-50.

Popper, K. (1957). *The poverty of historicism*. Routledge.

Popper, K. (1959). The propensity interpretation of probability. *British Journal of Philosophy of Science* 10(37), 25-42.

Popper, K. (1963). *Conjectures and refutations: the growth of scientific knowledge*. Routledge.

Prelec, D. (1998). The probability weighting function. *Econometrica* 66(3), 497-527.

Prelec, D. (2004). Decreasing impatience: a criterion for non-stationary time preference and "hyperbolic" discounting. *Scandinavian. Journal of Economics* 106(3), 511-32.

Prelec, D., & Loewenstein, G. (1991). Decision making over time and under uncertainty: a common approach. *Management Science* 37(4), 770-86.

Preyer, G., & Siebert, F. (2001). Reality and Humean supervenience: some reflections ob David Lewis's philosophy. In Preyer, G., & Siebert, F. (eds.). *Reality and Humean supervenience*. Rowman & Littlefield, 12-16.

Price, H. (1996). *Time's arrow and Archimedes' point*. Oxford University Press.

Price, H. (2007). Causal perspectivalism. In Price, H., & Corry, R. (eds.). *Causation, physics, and the constitution of reality*. Oxford University Press, 250-92.

Price, H., & Corry, R. (eds.). (2007). *Causation, physics, and the constitution of reality*. Oxford University Press.

Prinz, W., Beisert, M., & Herwig, A. (eds.). (2013). *Action science: foundations of an emerging discipline*. MIT Press.

Pylyshyn, Z. (1999). Is vision continuous with cognition? The case for cognitive penetrability of vision. *Behavioral and Brain Sciences* 22(3), 341-423.

Pylyshyn, Z. (ed.). (1987). *The robot's dilemma: the frame problem in artificial intelligence*. Ablex.

Quattrone, G., & Jones, E. (1980). The perception of variability within in-groups and out-groups: implications for the law of small numbers. *Journal of Personality and Social Psychology* 38(1), 141-52.

Quiggin, J. (1982). A theory of anticipated utility. *Journal of Economic Behavior and Organization* 3(4), 323-43.

Rabin, M. (2000). Risk aversion and expected utility theory: a calibration theorem. *Econometrica* 68(5), 1281-92.

Rachlin, H. (2018). In what sense are addicts irrational? In Bermúdez, J. (ed.). *Self-control, decision theory, and rationality: new essays*. Cambridge University Press, 147-66.

Rachlin, H., & Raineri, A. (1992). Irrationality, impulsiveness, and selfishness as discount reversal effects. In Loewenstein, G., & Elster, J. (eds.). *Choice over time*. Russell Sage Foundation, 93-118.

Rae, J. (1834/1905). *The sociological theory of capital*. Macmillan.

Railton, P. (2009). Practical competence and fluent agency. In Sobel, D., & Wall, S. (eds.). *Reasons for action*. Cambridge University Press, 81-115.

Ramsey, F. (1931). Truth and probability. In Braithwaite, R. (ed.) *Foundations of mathematics and other logical essays*. Routledge and Kegan Paul, 156-98.

Ransom, M., & Fazelpour, S. (2015). Three problems for the predictive coding theory of attention. *Midas Online*.

Ransom, M., Fazelpour, S., & Mole, C. (2017). Attention in the predictive mind. *Consciousness and Cognition* 47, 99–112.

Rao, L., & Li, S. (2011). New paradoxes in intertemporal choice. *Judgment and Decision Making* 6(2), 122-9.

Ratcliffe, M. (2007). *Rethinking commonsense psychology*. Palgrave Macmillan.

Rauscher, T., & Patkós, A. (2010). Origin of the chemical elements. *arXiv preprint arXiv:1011.,5627*.

Rawls, J. (1971). *A theory of justice*. Harvard University Press.

Raz, J. (2009). Reasons: practical and adaptive. In Sobel, D., & Wall, S. (eds.). *Reasons for action*. Cambridge University Press, 37-57.

Read, D., & Roelofsma, P. (2003). Subadditive versus hyperbolic discounting: a comparison of choice and matching. *Organizational Behaviour and Human Decision Processes* 65(2), 140-53.

Reichenbach, H. (1949). *The theory of probability*. University of California Press.

Reichenbach, H. (1956). *The direction of time*. University of California Press.

Rideout, D., & Sorkin, R. (1999). Classical sequential growth dynamics for causal sets. *Physical Review D* 61(2), 024002.

Ridge, M. (2009). The truth in ecumenical expressivism. In Sobel, D., & Wall, S. (eds.). *Reasons for action*. Cambridge University Press, 219-42.

Robb, A. (1914). *A theory of time and space*. Cambridge University Press.

Robertson, K. (2016). Can the two-time interpretation of quantum mechanics solve the measurement problem? *Studies In History and Philosophy of Modern Physics* 58, 54-62.

Roeper, P., & Leblanc, H. (1999). *Probability theory and probability logic*. University of Toronto Press.

Ross, D. (2013). Action-oriented predictive processing and the neuroeconomics of sub-cognitive reward. *Behavioral and Brain Sciences* 36(3), 225-226.

Ross, W. (1930). *The right and the good*. Oxford University Press.

Rouse, J. (1988). Arguing for the natural ontological attitude. *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association 1988, Volume 1: Contributed Papers*, 294-301.

Rovelli, C. (1997). Halfway through the woods: contemporary research in space and time. In Earman, J., & Norton, J. (eds.). *The cosmos of science*. University of Pittsburgh Press, 180-223.

Rugh, S., & Zinkernagel, H. (2009). On the physical basis of cosmic time. *Studies in History and Philosophy of Modern Physics* 40(1), 1-19.

Rupert, R. (2009). *Cognitive systems and the extended mind*. Oxford University Press.

Russell, B. (1913). On the notion of cause. *Proceedings of the Aristotelian Society* 13, 1-26.

Ryle, G. (1949). *The concept of mind*. Hutchinson.

Salmon, W. (1984). *Scientific explanation and the causal structure of the world*. Princeton University Press.

Samuelson, P. (1937). A note on the measurement of utility. *Review of Economic Studies* 4, 155-61.

Savage, L. (1954). *The foundations of statistics*. Wiley.

Schmidt, U. (2014). Risk preferences may be time preferences: A comment on Andreoni and Sprenger (2012) (No. 1942). Kiel Working Paper.

Schwartenbeck, P., FitzGerald, T., Dolan, R., & Friston, K. (2013). Exploration, novelty, surprise, and free energy minimization. *Frontiers in Psychology* 4, 710.

Schwarz, W. (2018). No interpretation of probability. *Erkenn* 83(6), 1195-212.

Schwarz, W. (2021). Objects of choice. *Mind* 130(517), 165-97.

Searle, J. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences* 3(3), 417-24.

Secchi, D. (2010). *Extended rationality: understanding decision making in organizations*. Springer.

Sen, A. (1970). *Collective choice and social welfare*. Holden-Day.

Shankland, R. (1964). Michelson-Morley experiment. *American Journal of Physics* 32(1), 16-35.

Shannon, C. (1948). A mathematical theory of communication. *Bell System Technical Journal* 27(3), 379-423.

Shea, N. (2013). Perception versus action: the computations may be the same but the direction of fit differs. *Behavioral and Brain Sciences* 36(3), 228-29.

Shefrin, H., & Thaler, R. (1992). Mental accounting, saving, and self control. In Loewenstein, G., & Elster, J. (eds.). *Choice over time*. Russell Sage Foundation, 287-330.

Sher, S., & McKenzie, C. (2008). Framing effects and rationality. In Chater, N., & Oaksford, M. (eds.). *The probabilistic mind: prospects for Bayesian cognitive science*. Oxford University Press, 79-96.

Shultz, D. (2019). Mystery solved? Why cats eat grass. *Science,* doi:10.1126/science.aaz0485.

Sidgwick, H. (1907). *The methods of ethics, 7th edition*. Macmillan.

Silverstein, T. (1998). The real reason why oil and water don't mix. *Journal of Chemical Education* 75(1), 116-8.

Simon, H. (1956). Rational choice and the structure of the environment. *Psychological Review* 63(2), 129-38.

Simon, H. (1982). Theories of bounded rationality. In *Models of bounded rationality, Volume 2: behavioral economics and business organization*. MIT Press, 408-23.

Simon, H. (1990). A mechanism for social selection and successful altruism. *Science* 250(4988), 1665-8.

Sklar, L. (1993). *Physics and chance: philosophical issues in the foundations of statistical mechanics*. Cambridge University Press.

Sklar, L. (2002). Physics, metaphysics, and method in Newton's dynamics. In Gale, R. (ed.). *The Blackwell guide to metaphysics.* Blackwell, 1-18.

Skow, B. (2015). *Objective becoming*. Oxford University Press.

Skvoretz, J., Fararo, T., & Axten, N. (1980). Role-programme models and the analysis of institutional structure. *Sociology* 14(1), 49-67.

Smith, A. (1776). *An enquiry into the nature and causes of the wealth of nations*. W. Strahan and T. Cadell.

Smith, M. (1987). The Humean theory of motivation. *Mind*, 36-61.

Smith, M. (2009). The explanatory role of being rational. In Sobel, D., & Wall, S. (eds.). *Reasons for action*. Cambridge University Press, 58-80.

Smolensky, P. (2000). Connectionism, constituency, and the language of thought. In Cummins, R., & Cummins, D. (eds.). *Minds, brains, and computers: the foundations of cognitive science*. Blackwell, 286-306.

Smolin, L. (2006). *The trouble with physics*. Allen Lane.

Sobel, D., & Wall, S. (eds.). (2009). *Reasons for action*. Cambridge University Press.

Spash, C. (2008). Contingent valuation as a research method: environmental values and human behaviour. In Lewis, A. (ed.). *The Cambridge handbook of psychology and economic behaviour*. Cambridge University Press, 429-53.

Stanford, P. (2006). *Exceeding our grasp: science, history, and the problem of unconceived alternatives*. Oxford University Press.

Stanovich, K. (2012). On the distinction between rationality and intelligence: implications for understanding individual differences in reasoning. In Holyoak, K., & Morrison, R. (eds.). *The Oxford handbook of thinking and reasoning*. Oxford University Press, 343-65.

Stanovich, K., & West, R. (2000). Individual differences in reasoning: implications for the rationality debate. *Behavioral and Brain Sciences* 23(5), 645-726.

Starmer, C. (2000). Developments in non-expected utility theory: the hunt for a descriptive theory of choice under risk. *Journal of Economic Literature* 38, 332-82.

Steel, P. (2007). The nature of procrastination. *Psychological Bulletin* 133(1), 65-94.

Stehle, P. (1993). Least-action principle. In Parker, S. (ed.). *McGraw-Hill encyclopedia of physics, 2nd edition*. McGraw-Hill, 670.

Sterelny, K., & Griffiths, P. (1999). *Sex and death: an introduction to the philosophy of biology*. University of Chicago Press.

Stewart, N., & Simpson, K. (2008). A decision-by-sampling account of decision under risk. In Chater, N., & Oaksford, M. (eds.). *The probabilistic mind: prospects for Bayesian cognitive science*. Oxford University Press, 261-276.

Steyvers, M., & Griffiths, T. (2008). Rational analysis as a link between human memory and information retrieval. In Chater, N., & Oaksford, M. (eds.). *The probabilistic mind: prospects for Bayesian cognitive science*. Oxford University Press, 329-49.

Stich, S. (1996). *Deconstructing the mind*. Oxford University Press.

Stich, S. (1983). *From folk psychology to cognitive science*. MIT press.

Stroltz, R. (1955). Myopia and inconsistency in dynamic utility maximization. *Review of Economic Studies* 23(3), 165-80.

Ströltzner, M. (1994). Action principles and teleology. In Altmanspacher, H., & Dalenoort, G. (eds.). *Inside versus outside.* Springer, 33-62.

Suhler, C., & Callender, C. (2012). Thank goodness that argument is over: explaining the temporal value asymmetry. *Philosopher's Imprint* 12(15), 1-16.

Sullivan, M. (2018). *Time biases: a theory of rational planning and personal persistence*. Oxford University Press.

Takahashi, T. (2011). Psychophysics of the probability weighting function. *Physica A* 390(5), 902-90.

Taleb, N. (2009). Errors, robustness, and the fourth quadrant. *International Journal of Forecasting* 25(4), 744-59.

Taleb, N. (2010). *The black swan: the impact of the highly improbable, 2nd edition*. Random House.

Taleb, N. (2012). The future has thicker tails than the past: model error as branching counterfactuals. *arXiv preprint arXiv:1209.2298.*

Taleb, N., & Goldstein, D. (2012). The problem is beyond psychology: the real world is more random than regression analyses. *International Journal of Forecasting* 28(3), 715-6.

Temple, R. (1986). *The genius of China: 3,000 years of science, discovery, and invention*. Simon and Schuster.

Thaler, R. (1981). Some empirical evidence of dynamic inconsistency. *Economics Letters* 8(3), 201-7.

Thaler, R. (1985). Mental accounting and consumer choice. *Marketing Science* 4, 199-214.

Thaler, R. (1999). Mental accounting matters. Reprinted in Camerer, C., Lowenstein, G., & Rabin, M. (eds.). (2004). *Advances in behavioral economics*. Russell Sage Foundation, 75-103.

Thaler, R., & Shefrin, H. (1981). An economic theory of self-control. *Journal of Political Economy* 89(2), 392-406.

Thorndike, E. (1911). *Animal intelligence*. Macmillan.

Thura, D., Beauregard-Racine, J., Fradet, C., & Cisek, P. (2012). Decision making by urgency gating: theory and experimental support. *Journal of Neurophysiology* 108(11), 2912-30.

Timmerman, T. (2020). Meghan Sullivan, Time biases: a theory of rational planning and personal persistence. *Journal of Moral Philosophy* 17(6), 690-4.

Torretti, R. (1983). *Relativity and geometry*. Pergamon Press.

Toulmin, S. (1953). *The philosophy of science*. Hutchinson.

Tucker, C. (2013). Seemings and justification: an introduction. In Tucker, C. (ed.). *Seemings and justification: new essays on dogmatism and phenomenal conservatism*. Oxford University Press, 1-29.

Turan, A. (2019). Intentional time inconsistency. *Theory and Decision* 86, 41-64.

Turing, A. (1950). Computing machinery and intelligence. *Mind* 59(236), 433-60.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: heuristics and biases. *Science* 185, 1124-30.

Tversky, A., & Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science* 211, 453-8.

Tversky, A., & Kahneman, D. (1983). Extensional versus intuitive reasoning: the conjunction fallacy in probability judgment. *Psychological Review* 90(4), 293-315.

Tversky, A., & Kahneman, D. (1992). Advances in prospect theory: cumulative representation of uncertainty. *Journal of Risk and Uncertainty* 5(4), 297-323.

Tversky, A., Slovic, P., & Kahneman, D. (1990). The causes of preference reversal. *American Economic Review* 80(1), 204-17.

Uffink, J. (2010). Irreversibility in stochastic dynamics. In Ernst, G., & Hüttemann, A. (eds.). *Time, chance, and reduction: philosophical aspects of statistical mechanics*. Cambridge University Press, 180-207.

Urbina, S. (2011). Tests of intelligence. In Sternberg, R., & Kaufman, S. (eds.). *The Cambridge handbook of intelligence*. Cambridge University Press, 20-38.

Usher, M., Elhalal, A., & McClelland, J. (2008). The neurodynamics of choice, value-based decisions, and preference reversal. In Chater, N., & Oaksford, M. (eds.). *The probabilistic mind: prospects for Bayesian cognitive science*. Oxford University Press, 277-300.

Vallacher, R., Read, S., & Nowak, A. (eds.). (2017). *Computational social psychology*. Routledge.

Vallino, J., Algar, C., González, N., & Huber, J. (2013). Use of receding horizon optimal control to solve MaxEP-based biogeochemistry problems. In Dewar, R., Lineweaver, R., Niven, R., & Regenauer-Lieb. (eds.). *Beyond the second law: entropy production and non-equilibrium systems*. Springer, 337-360.

Van Fraassen, B. (1995). Belief and the problem of Ulysses and the Sirens. *Philosophical Studies* 77(1), 7-37.

Venn, J. (1888). *The logic of chance, 3rd edition*. Macmillan.

Visser, M. (2003). Essential and inessential features of Hawking radiation. *International Journal of Modern Physics D* 12(4), 649-61.

Volchan, S. (2002). What is a random sequence? *The American Mathematical Monthly* 109(1), 46-63.

Von Neumann, J., & Morgenstern, O. (1944). *Theory of games and economic behavior*. Princeton University Press.

Vovk, V. (2019). Testing randomness. *arXiv preprint arXiv: 1906.09256.*

Wakker, P., & Deneffe, D. (1996). Eliciting von Neumann-Morgenstern utilities when probabilities are distorted or unknown. *Management Science* 42(8), 1131-50.

Wald, R. (1984). *General relativity*. University of Chicago Press.

Wallace, D. (2008). Philosophy of quantum mechanics. In Rickles, D. (ed.). *The Ashgate companion to contemporary philosophy of physics*. Ashgate Publishing, 16-98.

Wallace, D. (2011). The logic of the past hypothesis.

Wallace, D. (2013). What statistical mechanics actually does.

Watson, G. (2009). Promises, reasons, and normative powers. In Sobel, D., & Wall, S. (eds.). *Reasons for action*. Cambridge University Press, 155-78.

Watson, J. (1913). Psychology as the behaviorist views it. *Psychological Review* 20(2), 158-77.

Watson, J. (1968). *The double helix*. Weidenfeld and Nicolson.

Watson, M. (2019). Games. In Stein, F., Lazar, S., Candea, M., Diemberger, H., Robbins, J., Sanchez, A, & Stasch, R. (eds.). *The Cambridge encyclopedia of anthropology*.

Weber, B., & Chapman, G. (2005). The combined effects of risk and time on choice: Does uncertainty eliminate the immediacy effect? Does delay eliminate the certainty effect? *Organizational Behavior and Human Decision Processes* 96(2), 104-18.

Webley, P., & Nyhus, E. (2008). Inter-temporal choice and self-control: saving and borrowing. In Lewis, A. (ed.). *The Cambridge handbook of psychology and economic behaviour*. Cambridge University Press, 105-54.

Wiese, W. (2017). What are the contents of representations in predictive processing? *Phenomenology and the Cognitive Sciences* 16(4), 715-736.

Wilkinson, N. (2008). *An introduction to behavioral economics*. Palgrave Macmillan.

Williams, B. (1981). Persons, character, and morality. In Williams, B (ed.). *Moral luck*. Cambridge University Press, 1-19.

Williamson, T. (1994). *Vagueness*. Routledge.

Williamson, T. (2000). *Knowledge and its limits*. Oxford University Press.

Wilson, A. (ed.). (2014). *Chance and temporal asymmetry*. Oxford University Press.

Wilson, R., Geana, A., White, J., Ludvig, E., & Cohen, J. (2014). Humans use directed and random exploration to solve the explore-exploit dilemma. *Journal of Experimental Psychology* 143(6), 2074-81

Witt, U. (2008). Evolutionary economics and psychology. In Lewis, A. (ed.). *The Cambridge handbook of psychology and economic behaviour*. Cambridge University Press, 493-511.

Wittgenstein, L. (1953). *Philosophical investigations*. Macmillan.

Wrangham, R. (2009). *Catching fire: how cooking made us human*. Profile Books.

Wright, R. (1994). *The moral animal: evolutionary psychology and everyday life*. Pantheon.

Yule, P., Fox, J., Glasspool, D., & Cooper, R. (2013). *Modelling high-level cognitive processes*. Psychology Press.

Zee, A. (2013). *Einstein gravity in a nutshell*. Princeton University Press.

Zeelenberg, M., van Dijk, W., Manstead, S., & der Pligt, J. (1998). The experience of regret and disappointment. *Cognition & Emotion* 12(2), 221-30.