Decisions with Multiple Selves: Applications of Social Choice and Game Theory to Decision Theory

Alexander Mascolo

A sub-thesis submitted in partial fulfilment of the degree of
Bachelor of Philosophy (Honours) at
The Australian National University

May 2015

Typeset in Palatino by TEX and LATEX 2ε.

Except where otherwise indicated, this thesis is my own original work.

Alexander Mascolo
27 May 2015

To my parents and brother, whom I love dearly.

Acknowledgements

A thesis, or my thesis at least, is the culmination of years of acquired knowledge; it is just the tip of an iceberg, so to speak. For this reason, I ought to thank the many people who have held my hand along this educational path. I owe my gratitude, above all, to those academics who have taken the time to supervise me on a one-to-one basis during these past years; in the chronological order of when we met, they are:

- The logicians, John Slaney and Rajeev Goré. John has been almost a fatherly figure to me: he took me under his wing when I first arrived, soon corrupted my innocent, youthful mind with all kinds on non-classical logics, then involved me in his classes which I have really enjoyed teaching; words hardly suffice to express my indebtedness. Raj is also awesome: he introduced me to modal logic, but most importantly, made me understand the distinction between proof theory and model theory, between syntactic calculi and semantics, in a way I never had before; he expressed this distinction with such clarity that it has impressed itself on my mind up to this day. His concern for my well-being and his words of support, at more or less difficult times, have been truly appreciated too.

- The AI researchers, Peter Sunehag and Marcus Hutter. Though I haven't worked with them closely, nor at length, I have benefited immensely from their nearby presence. Peter has always been encouraging towards me and very supportive of my plans, for which I am particularly grateful; it is a pity not having him around any more, now that Google DeepMind snatched him up. Marcus, instead, is one of a handful of intellectuals (alive or dead) whose ideas have profoundly shaped my view of the world; to have actually met him in person has been a great honour.

- The decision theorist, Rachael Briggs. To put it briefly, they have been an invaluable mentor to me: when I send them drafts to comment on, for example, the detail and insight of the feedback they provide are second to none; I often suspect they have put more energy into reading my essays than I have put into writing them. I am grateful for this dedication, as well as for their remarkable patience towards me, and I feel I must apologise for, perhaps, not being as diligent a student as I should have been. As if that weren't enough, Rachael's papers have left a big mark on me and, most definitely, inspired the topic of this thesis.

The initial idea for this thesis came to me a few years ago, however, during a trip to Berkeley where I was visiting the Machine Intelligence Research Institute (MIRI). I thank my host Luke Muehlhauser, for I otherwise might have ended up working on a much duller, more conventional topic.

Academic work doesn't happen in a vacuum; it needs an infrastructure of coordinated, sensible people who assist with the many administrative tasks behind the scene, and it must be supported by society at large. So I must thank John Slaney once again (this time in his capacity of honours convenor), the always available team at CECS' Student Office, and Paula Newitt who, with the assistance of Sue Wigley, successfully ran the Ph.B. (Hons) program for many years; on a related note, this undergraduate degree totally deserves credit for the unique opportunities it offers gifted students. I am extremely grateful for ANU's generous National Merit Scholarship scheme; it is far too rare, even at a university, to find awards based on academic merit rather than dubious criteria such as "leadership potential" (whatever that means). The Australian Government has also assisted me financially; as a needful recipient of social welfare, I aim to repay our society using the knowledge it has enabled me to learn.

I find it hard to express this same gratitude, though, towards the mindless bureaucrats at ANU's College of Business and Economics, with their rigid rules and such. Due to them, I have found myself in the absurd position of writing an honours thesis that is largely based on microeconomic theory, without ever having taken a single course in economics; this despite being welcomed by various lecturers to enrol in their classes. It is mainly thanks to José Rodrigues-Neto, who kindly let me audit his graduate course in game theory, that I was able acquire the necessary background to pursue my intellectual interests. Incidentally, this turned out to be my all-time favourite class alongside Edwin Bonilla's lectures on information theory; the passion and warmth these South American people radiate are wonderfully contagious.

Last but not least, my friends and family are to be thanked. I won't even try thanking my family properly as I couldn't possibly do justice to them here. Also, this isn't the place to acknowledge all of my friends, so let me just state that in Canberra, particular at the CSSA (Computer Science Students' Association), I have enjoyed the company of some of the loveliest people I ever had the fortune to cross paths with.

A special mention goes to the amazing Jan Leike for carefully reading an entire draft of this thesis and for refraining his usual enthusiasm in tearing apart other people's work (that which can be destroyed by the truth should be, but being a flawed evidentialist, I hate hearing about my faults!). And a final shout-out goes to my mate Daniel Filan for some helpful last-minute proofreading.

Abstract

> "You must bind me hard and fast, so that I canno stir from the spot where you will stand me [...] and if I beg you to release me, you must tighten and add to my bonds."
>
> The Odyssey

Social choice and game theory are formal tools usually employed to study situations where there are multiple agents present. Game theory is used to model the strategic interactions of a group of individual agents in situations of conflict or cooperation, while social choice theory is concerned with aggregating the preferences or beliefs of a group of individual agents. Both frameworks stand in contrast to decision theory, which studies how an individual agent ought to make rational decisions given its preferences and beliefs.

Recent literature, however, suggest the distinction is not so clear-cut and that social choice and game theory can be applied to some situations where only one agent is present. Building on previous results, Lattimore & Hutter (2014) treat sequential decisions as an extensive game with perfect information, and they show that a dynamically consistent agent must achieve a subgame perfect equilibrium against its future selves. Briggs (2010), instead, applies an impossibility theorem from voting theory to show that no decision procedure can meet certain desiderata. This result is achieved by treating the decision-maker's possible future selves as the voters and candidates of an election.

Details aside, the approach of these papers can be described in two steps: representing the decision-maker as a group of agents, existing at different times, each having their own preferences and beliefs; then applying tools usually reserved for multiple agents, such as social choice and game theory, to obtain results about single-agent decisions. This technique has been fruitful and, yet, no one to date has conducted a review of how tools from social choice or game theory can be applied to the decisions of an individual agent.

This thesis fills such a gap in the literature, with particular attention to implications for reinforcement learning. The main contribution is to tie together results currently scattered across separate disciplines. Open problems are also identified and discussed.

Contents

Chapter 1

Introduction

1.1 Decisions in computer science

This thesis concerns three fields: decision theory, game theory, and social choice theory. The field of decision theory studies how an agent ought to make decisions, given certain constraints such as its beliefs and values. These decisions usually involve a single agent, the decision-maker. Social choice and game theory, instead, concern groups of individuals. In game theory we develop formal models of the strategic behaviour of a group of agents in situations of conflict or cooperation, whereas social choice theory studies how to aggregate the beliefs or preferences of a group of individuals so that a consensus may be reached on collective decisions. Sometimes these fields consider descriptive questions, about how people behave in practice. We shall pay attention exclusively to normative matters, regarding how a rational agent should act.

Originally studied from a mathematical and philosophical viewpoint (Ramsey 1926), these fields soon spread to the social sciences, economics more specifically (Morgenstern & von Neumann 1944). They have now become well and truly interdisciplinary, of relevance to subjects as diverse as biology, political science, psychology, management, operations research, and, of course, computer science (Osborne & Rubinstein 1994).

Within computer science they have been applied to branches such as artificial intelligence and theoretical computer science. Terms like 'utility function' or 'expected utility maximisation' are ubiquitous in artificial intelligence and machine learning (Russell & Norvig 2009), and indeed, they originally belonged to decision theory. Reinforcement learning (Sutton & Barto 1998), which we discuss in detail later on, in particular makes use these decision-theoretic concepts. Ideas from game theory are also found in artificial intelligence, the Minimax algorithm for adversarial search perhaps being the most notable example (Russell & Norvig 2009). And in network theory, the notion of Nash equilibrium explains otherwise puzzling phenomena such as Braess' Paradox of how increasing a network's capacity can decrease its flow (Nisan et al. 2007).

In much the same way computer science adopted concepts and techniques from the decision sciences, so has the converse happened. The tools of algorithm analysis and discrete mathematics have been used to study markets and mechanism design, with algorithmic game theory becoming a major branch of theoretical computer

science (Nisan et al. 2007). Additionally, computational complexity theory has been applied to problems in social choice theory, such as analysing how difficult it is to strategically manipulate a voting procedure, and computational social choice is now an important topic in this field (List 2013).

We hope the wealth of examples above gives a taste of how closely related decision, social choice and game theory are to computer science. In this thesis, though, we will not mainly discuss content that focuses on algorithms or computation. We shall study decision problems in their own right, despite they sometimes bear little resemblance to the issues currently present in the computer science literature. But one should bear in mind that just because they do not have an immediate application to the issues of the day, it does not mean they are irrelevant to artificial intelligence or other parts computer science. Indeed, in the quest of developing machines that behave increasingly smarter, we cannot afford to ignore this important piece of knowledge that lies at the very foundations of rational decision-making.

1.2 Individuals behaving as groups

It might come across as a surprise that there are applications of social choice and game theory to decision theory. After all, the first two fields provide us with tools for studying the interactions of a group of agents, rather than the behaviour of an individual. The missing insight is that, at times, a single agent may behave as if it were a group of agents in conflict or cooperation. When this happens it can be appropriate to model its actions using social choice or game theory.

In what kinds of situation does an individual behave as many? Strotz (1955) suggests a famous example from the Odyssey.

Consider Ulysses encounter with the Sirens on his voyage back to Ithaca. He desires to hear their enchanting voices, but not at the cost drowning. Given he knows that their song leads men to madness, he makes his sailors plug their ears with wax so they will not hear the lethal call. Then, so that he cannot follow the Sirens no matter how strong the temptation, he orders his men to tie him to the ship's mast, to tighten the ropes at his every plea for freedom, and to unbind him only once the danger is over. According to the tale, he succeeds at this step of his homebound journey.

Let us analyse the situation. It is arguable that since Ulysses wants to hear the Sirens' song but does not want to drown, then he should have himself tied to the mast in anticipation of how he will behave. His pre-emptive move is to restrict the actions of his future selves, as he disapproves of what they would do of their own will. He wishes to hear the Sirens' song but not at the cost of drowning, while his future selves are willing to follow the call no matter what. It is not hard to see, now, that this situation can be modelled as a game with Ulysses and his future selves as players who take turns to compete for conflicting goals.

Hammond (1976) also provides an example similar to the one above, in which a potential addict must decide whether to take a pleasurable but highly addictive drug. We shall discuss this case in more detail later. And there are plenty of situations in

which an individual behaves as many, though it is not always clear that social choice or game theory can be applied.

Treating temporal selves as agents in their own right is a fruitful technique and there are other topics that lend themselves to this kind of analysis, for example how the beliefs of our future selves relate to our own. In recent literature on Bayesian epistemology, Briggs (2009) interprets van Fraassen's (1984) Principle of Reflection as endorsing one's future selves as epistemic experts, while Moss (2012) describes belief update as a communication between an agent's present and future selves. But as fascinating as these topics may be, we shall not consider them. We shall, instead, restrict ourselves to cases where an agent is making decisions and, therefore, there is an obvious connection to social choice or game theory.

1.3 Identity and multiple selves

So far we have suggested that an individual decision-maker be treated as group of agents no more than as a convenient modelling device. According to what view one takes on personal identity, though, this idea may even be interpreted literally.

Investigations into the notion of 'self' have long been at the centre of attention in philosophy, where they can be found in the writings of historical figures such as Locke (1689) and Hume (1739). But the view that interests us is more recent and notable due to Parfit (1971, 1984) , according to whom a person is a succession of selves in a relation of continuity and connectedness. This notion of personal identity is not just the idiosyncrasy of some philosopher and is, in fact, taken seriously by academics in disciplines such as economics and psychology (Elster 1987). Though it runs contrary to common sense, such an interpretation is actually supported by a body of empirical evidence. It receives support from behavioural evidence, regarding how we treat our future selves, and from neuroscientific evidence, regarding which parts of our brain activate when we reason about our future selves.

Pronin et al. (2008) assigned undergraduate students at Princeton University to various tasks involving decisions for one's future selves. The study was aimed at testing whether such decisions more closely resemble the choices we make for other people than those we make for our present selves. Some of these decisions were real, others required the subjects to imagine how they would choose in a hypothetical situation, and still others included conditions such as deciding for oneself at the present time or deciding for one's future selves. Across this range of experimental conditions the behaviour observed supports the hypothesis that, indeed, "decisions people make for future selves and other people are similar to each other and different from their decisions for present selves" (Pronin et al. 2008).

Jamison & Wegener (2010) advocate for us to consider our future selves as distinct people, in the literal sense, on the basis of which they go on to make a number of public policy recommendations. In support of this strong claim they cite studies in neuroscience that looked at which brain areas activate during mentalisation, that is the process of understanding someone's mental state, and prospection, that is the

process of imagining oneself in the future (which involves mentalising about one's future selves). As it happens, the system comprised of the temporal lobes, the temporoparietal junction, and the medial prefrontal cortex, which is known to be involved in mentalising, is also the brain area activated during prospection (Jamison & Wegener 2010).

Common to these works is the finding that we view or treat our future selves as if they were other people. However, we should not rush to deep conclusions about the nature of personal identity. Even accepting that we do reason about our future selves in the same way we reason about other people, and that our behaviours are similar towards these two groups, it is unreasonable to infer from this that our future selves constitute distinct people. It could be, for example, that to we employ the same mental machinery in both cases, and perhaps as a result act similarly, just to economise on resources. After all, the human brain is known to employ heuristics to solve problems from everyday life in a fast and frugal manner (Kahneman 2011).

Regardless of the position one takes, literal or figurative, it can be useful to model a decision-maker as if constituted by multiple individuals, therefore we shall do so throughout this thesis.

1.4 Overview of the thesis

In this introduction, so far, we provided the general motivation behind the thesis; we now outline the content that will be presented in successive chapters. The rest of this thesis is structured in two main chapters followed by a discussion section at the end. Each of these main chapters aims to be self-contained and considers applications of either social choice or game theory to decision theory.

Given how interdisciplinary a topic like this is, much of the relevant material is scattered across publications in different subjects. It is common for these works to adopt different formalisations for the same concept or different notations for the same formalisation, at times even within one discipline. Our intent is to spare the reader as much as possible the frustration that comes from doing interdisciplinary research. To do so we will review the basic concepts we adopt from various disciplines, and we will dutifully point out how different ways used to formalise a concept are related. Technical results will be presented, though sparingly, and emphasis will be placed on intuitive understanding over mathematical proofs. Inevitably some review sections will result rather basic according to the one's background, so we trust the reader to skim through them in a judicious manner.

The first chapter will look at an application of game theory that arises when an agent undergoes a strict reversal of preferences while carrying out a decision over time. We start by reviewing various formalisms used to model preferences, only to point out that they capture the same phenomena. We then introduce the problem of preference reversal in an informal manner, before presenting a specific instance from reinforcement learning; the necessary background to understand sequential decisions is included. In this section on reinforcement learning, we present some technical re-

sults by Lattimore & Hutter (2014), in a more accessible manner than the original paper, as well as criticise some of their limitations. It is in the next section that we follow their suggestion of using game-theoretic tools; again, any material needed is provided in the form of a review. Last, we investigate an open problem in this area and reveal that it is much harder than it may have appeared.

The second chapter will look at an application of social choice to decision theory: an impossibility theorem from voting theory is shown to have an analogue in individual decision-making. What distinguishes the decisions problems we look at here is that they consider the agent to be embedded in the environment; we will review this particular setting, with a few examples too. We then prove the impossibility result, following its original proof by Briggs (2010), but in a more general setting with stochastic environments.

The final discussion will summarise what was done up to that point, and we will suggest possible directions for future research.

Chapter 2

Sequential Games in Reinforcement Learning

2.1 Preferences, utility, and choice

This initial section is dedicated to introducing the notion of preferences, which are central to the problem of temporal inconsistency discussed in the rest of this chapter, and more generally, underpin the theory of rational decision-making.

Proofs and more detailed discussions of the results that follow can be found in standard textbooks on microeconomic theory, such as those by Mas-Colell et al. (1995) or Rubinstein (2012). Except for discount functions, though, the material on intertemporal preference is not usually part of graduate curricula in microeconomics, so references will be given below for these more specialised topics.

We state results for the simplest case where there is a finite number of elements over which preferences range, but they can be generalised to countably or uncountably infinite sets, subject to additional assumptions of course. Though this may be important elsewhere, we gloss over the particular objects of our preferences; alternatives, options, acts, outcomes, choices are all taken to be indistinguishable and we will use the terms interchangeably. The only substantial assumptions we make are that these objects are mutually exclusive, which means a decision-maker cannot pick more than one at a given time, and that they are jointly exhaustive, so they collectively make up all of the elements available in some decision situation.

Lastly, it should come as no surprise to hear that people violate many of the assumptions behind rational-choice models like the ones presented next. We will not let this concern us, though, because accurate models of human behaviour are well-beyond the scope of our investigation which, instead, concerns normative questions.

2.1.1 Instantaneous preferences

Here we introduce the three main formalisms that capture what a preference is. Different academic fields and publications in the literature may privilege an approach over another, but they have all been shown to be equivalent. The take-home message of this section is that, in fact, it does not matter which formalism we pick; we are still

talking about the same, underlying concept. For now we assume that preferences are static or that we are considering them only at a given point in time.

2.1.1.1 Preference relations

The most intuitive way of describing an agent's preferences mathematically is by using a binary relation $\succsim$. A binary relation allows us to make pair-wise comparisons between alternatives, and we interpret $x \succsim y$ to mean that the agent prefers x to y or is indifferent between them. To be a preference relation, such an ordering must satisfy a couple of conditions: completeness and transitivity. Let X be the set of possible alternatives that may come up in any decision situation, which we assume is finite. Completeness requires all alternatives to be comparable. Formally, $\forall x, y \in X$ we have that $x \succsim y$ or $y \succsim x$ must hold. While transitivity is the property that $\forall x, y, z \in X$, if $x \succsim y$ and $y \succsim z$, then $x \succsim z$. It forces some consistency on the decision-maker, since taken with completeness it disallows $x \precsim z$ which would imply that an agent has a cycle in its preferences.

We can also define a strict preference $x \succ y$ when both $x \succsim y$ and $y \precsim x$, and an indifference between alternatives $x \sim y$ when $x \succsim y$ and $y \succsim x$ both hold. Strict orderings are really just a notational alternative, as we could have constructed our non-strict preferences $x \succsim y$ by defining them as $x \succ y$ or $x \sim y$, and in turn, indifference $x \sim y$ from $x \nsucc y$ and $y \nsucc x$.

2.1.1.2 Utility functions

Another way to capture an agent's preferences is through a utility function. This is the most popular formalism in economics, for it can be quite convenient to work with numerical functions when applying optimisation techniques. It is also the approach typically employed in artificial intelligence, and will most likely be familiar to the reader, therefore I will discuss it only briefly. Utility functions are a special case of objective functions, which include loss or cost functions too. These last two are the negative of utility functions, we aim to minimise them rather than to maximise them, and they are particularly popular in machine learning of all branches of computer science. The general idea behind a utility function is that we sometimes express our preferences through judgements about the relative value of two alternatives, essentially assigning a higher or lower score to each of these alternatives. Take X to be the usual, finite set of alternatives. A function $U : X \to \mathbb{R}$ represents the preference relation $\succsim$ when $\forall x, y \in X$, $x \succsim y$ if, and only if, $U(x) \geq U(y)$. This relation is said to have a utility representation and U is called a utility function. It is a well-known result that utility representations always exists of a preference relation defined over some finite set.

Take special notice not to attribute any additional properties to the utility representation of preferences. Though each alternative is assigned a real number, it is only significant whether this number is greater or smaller than those of other alternatives; the numerical values taken by the function are irrelevant beyond that. In fact, for any

utility function U and a strictly increasing function $f : \mathbb{R} \to \mathbb{R}$, the new utility function $f(U(X))$ represents the same preference relation as U did. This shows that our utility functions only carry ordinal information, regarding the relative order of alternatives.

The richer structure of cardinal preferences does comes up in the field of reinforcement learning, discussed later, where decisions are made under uncertainty. We call lotteries the objects of our preference that involve uncertainty. A preference relation over lotteries can still be represented with a utility function, as long as it obeys additional axioms. We will not go to the trouble of reviewing this representation theorem, as nothing is to be gained by delving into such technicalities here. The details, in any case, may be found in a book like that of Morgenstern & von Neumann (1944). For our purposes, it is sufficient to remember that, in this context, a preference relation has a utility representation unique only up to positive affine transformations, which are transformations of the form $f(x) = ax + b$ with $a, b \in \mathbb{R}$ and $a > 0$.

2.1.1.3 Choice functions

One last approach is to capture a decision-maker's preferences through a mathematical model of its behaviour. We take the agent's choices to be its revealed preferences; that is to say, preferences over options that are manifested by the agent when these options are made available. Conversely, we can interpret preferences to be the agent's hypothetical choices, which are those choices that the agent would act out in a decision situation were the corresponding options to be made available.

Let X be defined as usual, so that $2^X \setminus \varnothing$ is the collection of all choice sets. A choice function C is, then, defined as $C : 2^X \setminus \varnothing \to 2^X \setminus \varnothing$ such that $\forall A \subseteq X, C(A) \subseteq A$. What it does is that it maps every choice set, representing the available alternatives in some situation, to a subset containing the elements preferred by the agent. These favourite elements may not be unique, but in accordance with our previous assumptions, the agent will have to choose one when actually carrying out a decision.

An interesting question to ask is under what conditions a decision-maker's choices coincide with some preference relation. First, though, let us consider what it means for an agent to make choices in accordance with some preference relation. To be making choices according to such a relation, whenever it is presented a set of alternatives, a decision-maker should choose those considered preferable according to that relation. One way of putting this is to say that the choice function applied to a set of alternatives should always return the maximal elements of that set according to the ordering of the preference relation. Formally, we say that a preference relation $\succsim$ rationalises a choice function C when, for all choice sets $A \in \mathcal{A}$, $C(X) = \{x \in A | x \succsim y \; \forall y \in A\}$.

It can be shown that there exists some preference relation $\succsim$ that rationalises a choice function C if, and only if, C satisfies a famous property known as the Weak Axiom of Revealed Preference (WARP). Intuitively, this axiom requires the decision-maker to be consistent with its choices; should it choose x in the presence of y, revealing x is at least as good as y according to its preferences, whenever it chooses y in the presence of x it must choose it together with x.

WARP. For $\forall A, B \subseteq X$ and $\forall x, y \in A \cap B$, if $x \in C(A)$ and $y \in C(B)$, then $x \in C(B)$.

Alternatively, the WARP is often spelt out as two properties α and β which taken together are logically equivalent to it. These properties can also be written in a number of forms, including as follows:

Property α. If $x \in A \subset B$ and $x \in C(B)$, then $x \in C(A)$.

Property β. If $x, y \in A \subset B$, $x \in C(A)$, and $y \in C(B)$, then $x \in C(B)$.

A number of everyday behaviours carried out by people have been suggested as potential counter-examples to these two properties. Once again, though, they need not need concern us. What interests us is that there are different ways of formalising an agent's preferences, but these formalisms are all equivalent under reasonable assumptions which we take to hold.

The correspondence described above is between choice functions and non-strict preference relations. There is one additional condition that a choice function must meet for being equivalent to a strict ordering: "If the choice function C corresponds to a strong ordering, then for every $A \subseteq X$, $C(A)$ is a singleton – i.e. a set with just one member. Conversely, if C corresponds to an ordering, and every choice set is a singleton, then C corresponds to a strong ordering" (Hammond 1976).

2.1.2 Preferences over time

Just as an agent may make a single decision, it can also make a series of such decisions at successive points in time. In this case, it is important that we are able to compare between what relative value the decision-maker assigns to the same alternative across time, or even, to several alternatives presented at different moments. So when we describe an agent's preference for some act, option, or outcome over another, here we speak of intertemporal preferences rather than instantaneous ones. Note that for convenience we take time to be discrete.[1]

2.1.2.1 Discounted utility functions

By far the most popular approach to intertemporal preferences, used in just about every branch of computer science including reinforcement learning, involves extending the utility function model. This is done by weighting the utility of options by a so-called discount function. A discount function $d : \mathbb{N} \to [0, 1]$ assigns to each time-step $t \in T$, which for discrete time is just $T = \mathbb{N}$, a discount rate in the interval $[0, 1]$ that measures the agent's time preference for earlier or later rewards. We will introduce later on various examples of discount functions. An intuitive way of looking at what discount functions do is as follows. In combination with a utility function, a discount function induces new preferences, or equivalently a new utility function, over each

[1] Thank you to Dr Conrad Heilmann of the Erasmus Institute for Economics and Philosophy, Rotterdam, for replying to the emails of a complete stranger like me and sending over a copy of his slides used in a very interesting talk given at the Scuola Normale Superiore di Pisa. These materials provided me with many useful pointers to the literature on axiomatic foundations of time discounting, a few results of which I mention in this thesis.

set of options considered at a different time. For example, if δ_t is the discount factor applied to time-step t and U is the underlying utility function, at this time the agent's preferences are captured by a utility function U' such that $U'(x) = \delta_t U(x)$, $\forall x \in X$ where X is the set of options. Of course, we can also compare the discounted utilities of options x and y at times t and s, respectively, by comparing the utilities $U'(x) = \delta_t U(x)$ and $U''(y) = \delta_s U(y)$.

2.1.2.2 Outcome-time relations

The algebraic model of preferences, which captures them as a relation over mathematical objects, can be extended to the intertemporal case. While we use to be able to compare two elements of a set without any context, we now need to introduce some variable that captures time. In order to do this we consider the agent's preferences over alternatives indexed by time, so quite simply, we let the relation range over outcome-time pairs (X, T) with X and T defined as before.

As a matter of historical interest, the theory of discounted utility was not derived from a model of binary relations; it was devised independently by Samuelson (1937). Only subsequently were efforts made to axiomatise it by Koopmans (1960) and Lancaster (1963). The latter derived discounted utility from the outcome-time structure described above; the former offered a more involved construction starting from preferences over sequences of outcomes in chronological order. More recently, Bleichrodt et al. (2008) have revised Koopmans analysis, using his same axioms but correcting technical issues found in the original. It is, however, the work of Fishburn & Rubinstein (1982) that remains the most instructive, though this paper differs somewhat from others since it uses a prize-time structure comprised by utility-time pairs, rather than pairs of outcomes and time. A first set of axioms captures the usual requirements we impose on preference relations, while further axioms of stationarity and time-impatience are added. These stipulate, respectively, that an agent should have a preference for receiving positive utility earlier and negative utility later, and that a time preference for outcomes should depend only on how temporally distant they are, without specific time indexes having any influence. With the additional axioms, the discount function derived is of a specific kind know as geometric; we will describe it later.

In any case, details are not particularly important to us. What all these results show is that, subject to reasonable conditions being met, the axiomatic approach coincides with the numerical representation of discounted utility.

2.1.2.3 Dynamic Choice Functions

A final, less common, way to capture intertemporal preferences is through a sequence of choice functions existing at different nodes of a decision tree. As we know, instantaneous preferences can be described by a choice function. So intertemporal preferences can simply be described with a choice function for every-time step, or even better, for every decision the agent makes. An advantage of this last approach is that, by having

a dynamic choice function defined over each decision node, we can also capture endogenously changing preferences. Whereas in outcome-time relations, preferences or tastes are assumed only to change exogenously. "That is, tastes depend on time, but not on the previous choices of the agent being considered, or, indeed, on the previous choices of other agents" (Hammond 1976). A shortcoming of this approach, however, is that it does not allow us to talk about the decision-maker's preference for some option over another at a different time; we can only compare an agent's preferences between alternatives available at the same time-step.

The idea of a dynamic choice function can be credited to Hammond (1976), who developed it to study the problem of time inconsistency which we now proceed to discuss. Much of his work in this area focused on formalising the notion of time consistency, by providing conditions on an agent's choice functions in order to make them coherent across time.

2.2 An informal description of the problem

In this brief section we introduce the problem of time inconsistency, also known as dynamic inconsistency, which is central to current chapter of the thesis. We look at how an agent may cope, or fail to cope, with its temporal inconsistency too. At no point do we employ formal notation, except for labelling options. This is done purposefully, to stress that the problem is general and does not depend in any way on particular formalisms adopted by various authors.

2.2.1 Hammond's potential addict

Let us start by looking at the following example due to Hammond (1976):

The potential addict. "Consider an individual who is contemplating a mode of behaviour which is potentially habit-forming. More specifically and clearly, suppose the individual is wondering whether or not to start taking an addictive drug. We may presume that the drug gives rise to pleasant sensations, at least initially. The individual would most prefer to take the drug infrequently, or for a short time, so that he enjoys the drug without damaging his health. It is assumed, however, that if the individual starts to take the drug, then he is certain to become an addict, and so to take the drug very much more than he had originally intended, with serious consequences. On the other hand, he can refuse to take the drug at all."

We can represent this problem schematically as shown below. Here the agent is given two opportunities for making a decision, each corresponding to one of the branching nodes. Initially, he may choose whether to try the substance or forgo this opportunity altogether. Later on, just as soon as the drug stops being pleasurable, he may revise his decision and choose whether to continue taking it or not. At any given time, the choices of consuming and not consuming the drug are, respectively, labelled by D and $\overline{D}$.

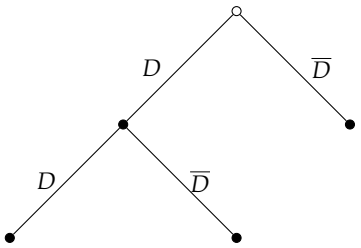

Figure 2.1: Hammond's potential addict.

This example is very simple, but it is characteristic of temporal inconsistency problems. All of these situations have the same structure, comprised of two decisions nodes and an agent whose preferences are reversed at some time in between. Unlike our example above, though, the nodes are not necessarily successive ones, and there may be more than two options involved in the decision-maker's choices.

2.2.2 Strategies

An interesting question is how should, or how does, an agent behave when it is time-inconsistent. According to Strotz (1955), such an decision-maker faces these possible courses of action:

1. It may fail to recognise it is time-inconsistent. Therefore it will plan its decisions naively, without taking into account its intertemporal conflicts.

2. It may recognise it is time-inconsistent. In which case it can solve its intertemporal conflicts by either:

   (a) Planning its decisions in a sophisticated manner, by taking into account how it will behave in future.

   (b) Binding itself to sequence of actions from which it would otherwise deviate due to its changing preferences.

Case 1. is known as inconsistent planning or, in the choice theory literature, as naive choice. While inconsistent planning may be sometimes suited to describing the behaviour of an agent, it falls short of being a norm for rational decisions. In the example above, the agent would start consuming the drug, careless of the consequences, and it would eventually end up addicted.

Case 2.(a) is known as consistent planning; choice theorists, instead, call it sophisticated choice. By planning consistently the agent chooses the best sequence of actions amongst those it will actually follow through with. In the example above, this amounts to the potential addict anticipating that he will fall for the drug and, accordingly, avoiding it in the first place.

The case 2.(b) of pre-commitment is promising. Could the agent steer his future behaviour so that he does not become addicted, perhaps by setting aside a few doses

and destroying the rest, then he would achieve his preferred outcome of consuming the drug only while it gives rise to pleasant sensations. But how can he do so? Presumably, that would mean there is some option we have omitted. By assuming our description of the problem is accurate and complete, however, we may conclude that no such option exists. As noted by Hammond (1976), more generally, pre-commitment cannot make sequential decisions time-consistent once all available options have already been taken into account.

Hammond's remarks were on choice theory specifically, whereas we will be using discount functions to describe an agent's preferences. But his observations apply, nonetheless, because dynamically inconsistent choices and inconsistent discounting are different ways of describing the same phenomenon, where agent has a reversal of preferences between two sequential decisions. If there is one lesson, above all, that the previous section was meant to teach, and that we will not tire of repeating, then it is this: it does not matter what formalism an author privileges; we are all talking about the same problem.

2.2.2.1 Resolute choice

It is worth mentioning that a further alternative, known as resolute choice, has been proposed by McClennen (1990). This theory prescribes our agent to take the drug so long as it is pleasurable, then stop consuming it at the second time-step. It advises this course of action on the basis that the decision-maker should act to resolve its intertemporal conflicts, even against whatever future preferences it may have predicted itself to have. An advocate for sophisticated choice would deem this plan unrealistic, much as they might agree with McClennen that ideally it would be preferable to follow such a route. The main point of contention is that resolute choice rejects the separability of decisions; that is to say, the assumption that choices made at a given time are independent of the preferences we hold at different times. It accepts, instead, that the choices we make at present can be in service of our future selves, or that we might be swayed by the influence of our past desires.

I am not inclined to dismiss the theory of resolute choice outright; there is plausibility to the idea that decisions cannot be separated. Still, much as I am sympathetic to this assumption, I do not believe it makes resolute choice applicable to our case. This should become clear once we spell out preferences carefully. We might take them to include whatever interpersonal desires we have, particularly intrapersonal desires that are intertemporal, or they might be of a less altruistic kind which only considers the desires of our present self. If the preferences we are speaking of are the selfish ones, then it might be that we can resolve to act against them in favour of those preferences held by our predecessors or successors who we care about. In the case where such altruistic desires towards others have already been accounted for, however, there is no way resolute choice can be applied. While it may be that resolute choice is available in a number of situations phrased similarly, our problems bear only superficial resemblance to mundane cases like that of the potential addict. They are specified formally, the agent's preferences are captured by some mathematical structure, and

that is all there is to say; any relevant desires, including those towards other temporal selves, have already been incorporated, making resolute choice infeasible.

2.3 Time-consistent discounting

Most content of this section is no way original; it is derived from work in reinforcement learning by Lattimore & Hutter (2014), some technical results of which are presented here in a more accessible manner. Our contribution is to define a more general model of time consistent discount that is history dependent, and we also explain some limitations of Legg & Hutter's (2007) and Legg's (2008) environment-based models.

2.3.1 Reinforcement Learning Preliminaries

The problem we consider is how an agent ought to make a sequence of decisions at different times. This problem is expressed formally in reinforcement learning (RL), for a notion of rationality that amounts to maximising some cumulative reward.

The RL setting consists of an agent interacting sequentially with an environment. The notation we use to describe this is that of Hutter (2005). At each time-step t the agent chooses an action $a_t \in \mathcal{A}$, whereupon it makes an observation $o_t \in \mathcal{O}$ and receives a corresponding reward $r_t \in \mathcal{R} \subseteq \mathbb{R}$. For simplicity we usually assume that $\mathcal{A}$ and $\mathcal{O}$ are finite and that $\mathcal{R} = [0, 1]$. We call *history* a sequence $a_1 r_1 o_1 a_2 r_2 o_2 \ldots$ of such actions, observations, and rewards, and we denote histories of length t and $t-1$ by $h_{1:t}$ and $h_{<t}$ respectively. In formal terms, we define the set of finite histories as $\mathcal{H} = (\mathcal{A} \times \mathcal{R} \times \mathcal{O})^*$. This notion of an *environment* with which the decision-maker interacts is also specified formally, by a probability function μ such that $\mu(r_t o_t | h_{<t} a_t)$ is the probability the agent will make the observation o_t and receive the reward r_t having chosen action a_t after history $h_{<t}$.

For convenience we represent environments as graphs. Edges correspond to the available actions, with each edge labelled by the reward received for choosing that action, and histories can be read off the paths. An environment is deterministic, like the one below, when $\mu(\cdot)$ can only take values 0 or 1.

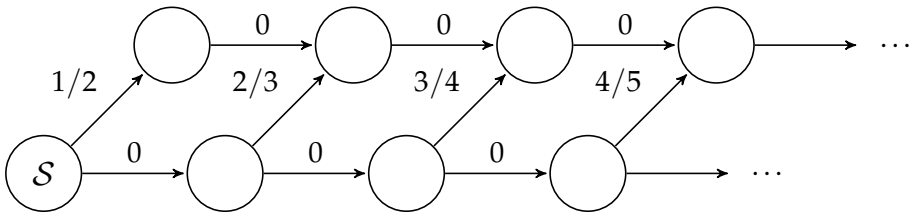

Figure 2.2: A deterministic environment (Lattimore & Hutter 2014)

The decision-maker will choose actions according to some policy. A *policy* $\pi \in \Pi$ is a function $\pi : \mathcal{H} \to \mathcal{A}$ mapping histories to actions, where Π is the set of all policies. Given a policy π and for $s \leq t$, the probability that history $h_{1:t}$ will occur once

history $h_{<s}$ already has is denoted by $P(h_{1:t}|h_{<s}, \pi)$. We define it as $P(h_{1:t}|h_{<s}, \pi) = \prod_{k=s}^{t} \mu(r_k o_k|h_{<k}\pi(h_{<k}))$. When no history has occurred, that is when $s = 1$, we omit the term $h_{<s}$ leaving $P(h_{1:t}|h_{<s}, \pi) = P(h_{1:t}|\pi)$. The expected sequence of rewards for applying a policy π after history $h_{<t}$ is $\boldsymbol{R}^{\pi}(h_{<t}) \in [0,1]^{\infty}$. While the associated reward expected at time-step k is defined as $R^{\pi}(h_{<t})_k = \sum_{h_{t:k}} P(h_{1:k}|h_{<t}, \pi) r_k$, unless $k < t$ for which we simply let $R^{\pi}(h_{<t})_k = 0$.

An agent may have a preference for receiving rewards at earlier or later times. In a traditional RL setting this preference is captured with a sequences of discount rates $d_1, d_2, \ldots$ each lying in the unit interval (Sutton & Barto 1998). For conciseness we use a discount vector $\boldsymbol{d} \in [0,1]^{\infty}$. Then the expected cumulative reward discounted by a vector $\boldsymbol{d}$, for applying a policy π after history $h_{<t}$, will be as follows:

$$V_{\boldsymbol{d}}^{\pi}(h_{<t}) = \boldsymbol{R}^{\pi}(h_{<t}) \cdot \boldsymbol{d} = \sum_{i=1}^{\infty} R^{\pi}(h_{<t})_i d_i = \sum_{i=t}^{\infty} R^{\pi}(h_{<t})_i d_i$$

As anticipated, the notion of rationality adopted amounts to maximising a cumulative reward. So the agent ought to choose an optimal policy that maximises $V_{\boldsymbol{d}}^{\pi}(h_{<t})$. This *optimal policy* is defined as $\pi_{\boldsymbol{d}}^{*} = \arg\max_{\pi \in \Pi} V_{\boldsymbol{d}}^{\pi}$ and is associated with a reward $V_{\boldsymbol{d}}^{*}(h_{<t}) = V_{\boldsymbol{d}}^{\pi_{\boldsymbol{d}}^{*}}(h_{<t})$. It belongs to the set $\Pi_{\boldsymbol{d}}^{*}$ of optimal policies that maximise $V_{\boldsymbol{d}}^{\pi}$. As the notation suggests, the optimal policy may not be unique. If that is the case then we pick an arbitrary such policy, since nothing of importance hinges on which one is chosen.

2.3.2 Discounted Utility

The standard approach for capturing the time-preferences of an agent in RL, as introduced above, is to use the discounted utility (DU) model first introduced by Samuelson (1937). In the DU model global utility is represented as the sum of local utilities, each of which is discounted by some factor. Written as an equation, the model has the following form:

$$V_k = \sum_{t=k}^{\infty} d_{t-k} r_t$$

The factor by which rewards are discounted depends only on the chronological distance of the rewards. That is to say that the discount factor d_{t-k} applied at time $t - k$ is uniquely determined by the parameter $t - k$; this observation is crucial to understanding the limitations of the model. The discount factors can be modelled as a function of time or, equivalently, using a vector indexed by time-steps.

2.3.2.1 Discount Functions

Of course, the discount function or vector may be arbitrary. We do, however, tend to restrict it to certain functions. The three main ones considered in this thesis are below. Note that all of them capture some notion impatience; discounting with these

functions, an agent will always have a preference for receiving rewards earlier rather than later.

Constant horizon. When discounting with a constant horizon, at any given time-step, an agent will only value rewards up to some limited number of steps away. This distance of time horizon does not change, and the decision-maker will care equally for all rewards that come before it. Thereafter, rewards no longer hold value for the agent. Formally, for some horizon $H \in \mathbb{N}$, we can write $d_{t-k} = [\![t - k < H]\!]$ where $[\![\cdot]\!]$ is an indicator function that takes values 1 or 0, respectively, according to whether the expression within is true or false.

Hyperbolic. A hyperbolic discount function takes the form $d_{t-k} = 1/(1 - k(t - k))$. This is a good approximation of how human beings actually discount rewards (Frederick et al. 2002).

Geometric. Geometric discounting, instead, is taken to carry some normative weight, telling agents how they should discount, and it is defined as $d_{t-k} = \gamma^{t-k}$ for some $\gamma \in (0, 1)$. In the economics literature, it is explained in terms on an agent having some chance of survival at each time-step (Mas-Colell et al. 1995). This is an underlying chance of survival, independent of whichever actions are taken. By assuming such a factor γ is uniformly distributed over time, and that rewards are irrelevant upon death with probability $1 - \gamma$, we get an agent who discounts geometrically.

2.3.2.2 Age independence

Though DU is a simple and widely adopted model of time preference, it comes with several undesirable consequences. For starters, an agent's discount function cannot change with time or, for that matter, at all. This is contradicted by empirical findings showing that discounting behaviour in humans does vary according to age.

Read & Read (2004) devised a questionnaire aimed at estimating subjects' discount rates, then presented it to a group of 123 respondents aged between 19 and 89 years of age. "The results supported the view that older people discount more than younger ones, and that middle aged people discount less than either group" (Read & Read 2004). That is to say discount rates form an inverse U-shaped curve with respect to age, skewed somewhat in the direction of decreasing age. This confirms a previous model by Sozou & Seymour (2003) based on evolutionary theory.

Harrison et al. (2002), instead, carried out a field experiment in Denmark using a nationally representative sample of 268 people from 19 to 75 years old. Though the scope of their study was much broader, as they investigated the effects of additional socio-demographic variables such as gender, primary occupation, level of educational attainment, smoking habits, and various characteristics of the subjects' households, as well as financial variables aimed at identifying market circumstances under which the discount rates might be explained. The most surprising conclusion is that "discount rates appear to decline with age, at least after middle age" (Harrison et al. 2002). However, this is consistent with Read & Read (2004) findings once we note that the elderly

sample was relatively young, taken to be people aged 50 years or greater, so the effects of old age would not be fully visible.

Harder to reconcile is a study by Green et al. (1994) in which "the rate of discounting was highest for children and lowest for older adults," with the middle-aged sitting in between. But this experiment involved only 36 participants, spread equally between three age groups. Unsurprisingly, Read & Read (2004) criticise it on the basis on the small sample, as well as for other methodological limitations such as the lack of control for confounding factors.

Though the particular shape of the relation is a matter of contention, there is a consensus that people's time preferences change as they age. Should we want a model than can account for this, then we must allow discount rates to change as a function of time.

2.3.2.3 Time Inconsistency

Another issue is that, for all except the geometric discount function, an agent will behave irrationally due to having conflicting preferences between different times (Strotz 1955). This can result in sub-optimal behaviour, as measured by the agent's cumulative reward.

To properly understand this time inconsistency we must look back at the more general phenomenon of inter-temporal preference reversal. The illustrative example, given by Hammond (1976), involved a decision-maker contemplating whether or not to take an addictive drug. If at all possible, the agent would prefer to try the drug while it is enjoyable but stop taking it thereafter. In practice this is impossible, because it was stipulated that once an agent starts consuming this particular drug it will fall into a cycle of addiction from which it cannot escape. Below is an adapted version of the example, expressed in the RL setup:

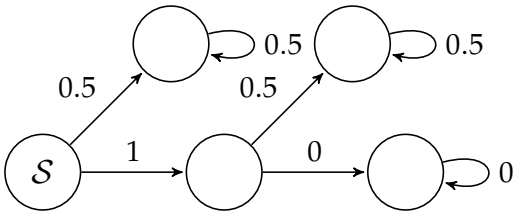

Figure 2.3: An adaptation of Hammond's potential addict.

In this environment taking the drug gives the maximum reward of 1 at first, but once the agent is addicted it receives the minimum reward of 0 regardless. There is also a baseline reward somewhere between the two, for the purpose of illustration we set it to 0.5, which correspond to the agent neither enjoying the drug nor suffering from addiction. The layout of the environment is as follows: the agent can choose to try the drug by moving to the right, then it faces the decision of continuing to take it by moving to the right again or stop taking it by moving up; the agent can also choose to never consume the drug by moving up immediately. We can easily see that

the best sequence of rewards is received by choosing to consume the drug only at the first decision node. That is to say, the sequence of rewards $[1, 0.5, 0.5, \dots]$, received by moving right then up, is preferred to $[0.5, 0.5, 0.5, \dots]$, and both are preferred to $[1, 0, 0, \dots]$. Given our assumptions about how the agent will behave after taking the drug, however, we know that following the first path is impossible. The agent's preferences between taking the drug or not have reversed, and once it becomes addicted the graph looks like this from its warped point of view:

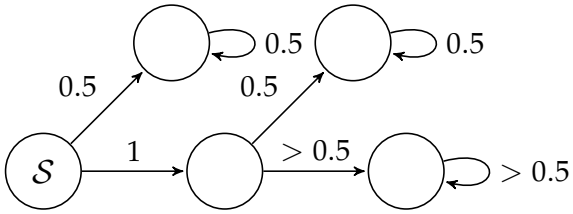

Figure 2.4: The same environment from a different perspective.

Time-consistency issues arise in the same manner, as a result of applying the discount function. Take the environment depicted in 2.3.1 and consider the behaviour of an agent with a bounded horizon of two time-steps:

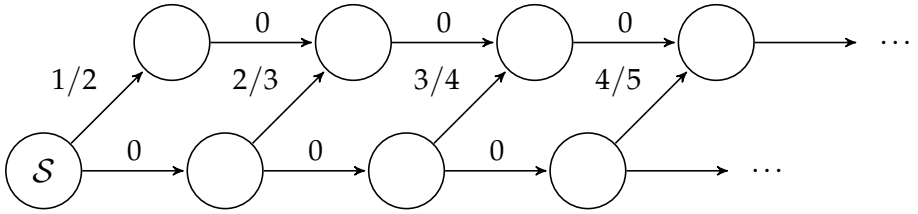

Figure 2.5: The environment from 2.3.1.

At the start, the decision-maker will delay receiving a reward in order to get the higher reward two time-steps away, then it will do the same for the next greatest reward, and so on. This agent renounces to any reward available in the false hope of receiving some greater reward at a later time-step. It follows the worst path, the only one with a cumulative reward is 0, because of the inconsistent way it discounts future utility.

What is going on? The agent's preferences at a given time-step are captured using both the utility of the rewards and the factors applied by the discount function. While the discount function does not change over time, the same factors are applied to different future rewards. So the agent's preferences change, effectively, just like in the potential addict example.

2.3.3 General Time-Consistent Discounting

To address the first limitation of the DU model, rather than a discount vector Lattimore & Hutter (2014) use a matrix to capture an agent's time preferences. This allows

the discount function to change with the age of the agent. A *discount matrix* D is a $\infty \times \infty$ matrix a discount vector $\boldsymbol{d}^k$ for its k^{th} column. We denote a generic discount vector by $\boldsymbol{d}$, while the discount vector of the agent at time k is $\boldsymbol{d}^k$ and the factor used to discount rewards at time t by that agent is d_t^k. Such a matrix looks like this:

$$D = \boldsymbol{d}^1 \boldsymbol{d}^2 \cdots \boldsymbol{d}^k \cdots = \begin{pmatrix} d_1^1 & d_1^2 & \cdots & d_1^k & \cdots \\ d_2^1 & d_2^2 & \cdots & d_2^k & \cdots \\ \vdots & \vdots & \ddots & \vdots & \ddots \\ d_t^1 & d_t^2 & \cdots & d_t^k & \cdots \\ \vdots & \vdots & \ddots & \vdots & \ddots \end{pmatrix}$$

A sequence of future rewards, as discounted by the agent at time k, will be:

$$V_k = \sum_{t=k}^{\infty} d_t^k r_t$$

2.3.3.1 More time inconsistency

Not too surprisingly, time inconsistent behaviour can still occur in this general time-consistent discounting (GTCD) model. Take an agent whose time preferences are captured by a discount matrix with vectors $\boldsymbol{d}^1 = [2, 1, 2, 0, 0, \dots]$ and $\boldsymbol{d}^2 = [*, 3, 1, 0, 0, \dots]$. Consider how it would behave in the environment below:

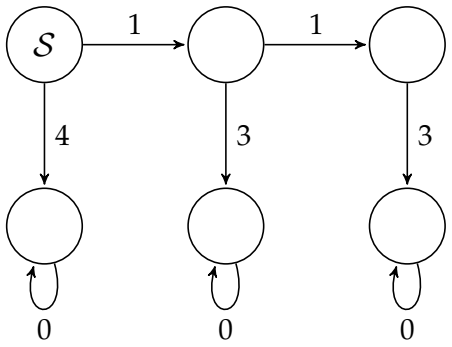

Figure 2.6: Another RL environment (Lattimore & Hutter 2014)

In this example by Lattimore & Hutter (2014), a naive agent will choose to move right, with the intent of moving right once more and receiving a cumulative reward of $\boldsymbol{d}^1 \cdot [1, 1, 3, 0, 0, \dots] = 9$. Instead, when it reaches the second node it will move down, for a lower reward of $\boldsymbol{d}^1 \cdot [1, 3, 0, 0, 0, \dots] = 5$. The reason it will choose to move down is that, at the second time-step, it is discounting rewards with respect to the vector $\boldsymbol{d}^2$. So when it reaches this stage of its plan it is actually choosing between moving down, for a reward of $\boldsymbol{d}^2 \cdot [3, 0, 0, \dots] = 9$, or moving right, for a reward of $\boldsymbol{d}^2 \cdot [1, 3, 0, 0, \dots] = 6$. If the agent sets out maximize its cumulative rewards with respect to $\boldsymbol{d}^1$, then this is hardly a successful outcome.

Of course, better strategies might be available. An agent who is aware of being dynamically inconsistent might move down immediately, to receive a reward of $\boldsymbol{d}^1 \cdot [4, 0, 0, \dots] = 8$. For it knows that if it moves right once then it will not move right again, which it would have to in order to receive a greater cumulative reward. This last plan is not one the agent will actually follow through with, so it is not taken into consideration. Alternatively, if the agent could somehow pre-commit to moving right twice, then it would be able to receive the greatest cumulative reward with respect to its discount vector $\boldsymbol{d}^1$. But it cannot do so, because there is no sequence of actions that allows it to receive this reward and from which it will have no incentive to deviate in future; there is no such path in the graph.

2.3.3.2 Time consistency and sliding matrices

What characterises an agent as having time-consistent preferences is that it has no incentive to change its plans at any point. In other words, the optimal policies for this agent will remain the same at every time-step. So a discount matrix D is *time-consistent* if, and only if, in any environment, $\Pi^*_{\boldsymbol{d}^k} = \Pi^*_{\boldsymbol{d}^j}$ for all $k, j \geq 1$. An equivalent definition of time-consistency is that a discount matrix D is time consistent if, and only if, for every k there exists $\alpha_k > 0$ such that $d^k_t = \alpha_k d^1_t$ for all $t \geq k \geq 0$ (Lattimore & Hutter 2014). This equivalence has a rather intuitive explanation, though it is not so simple to prove. All it is saying is that, for the decision-maker's time preferences to be time-consistent, the relative value of the available actions must remain the same at any given time-step. It also implies that it does not matter whether these values get scaled. In fact, the result is easy to see if we recall that cardinal utility is unique up to positive affine transformations, since the scaling applied by a discount vector is a transformation of this kind, although with no translation.

The DU model is captured by a particular kind of matrix known as sliding. A discount matrix is *sliding* if, and only if, $d^k_{k+t} = d^1_{1+t}$ for all $k \geq 1, t \geq 0$. Let $d_1, d_2, d_3, \dots$ be the discount factors of a decision-maker whose time preferences are captured by a single function as in the DU model. If the agent is making decisions sequentially then, as its discount function does not change over time, when it makes a new decision it re-adopts the old discount function by applying the same factors to different future rewards. So the discount factors $d_1, d_2, d_3, \dots$ slide down the matrix, which will look like this:

$$D = \begin{pmatrix} d_1 & * & * & \cdots \\ d_2 & d_1 & * & \cdots \\ d_3 & d_2 & d_1 & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix}$$

2.3.3.3 The consistency of geometric discounting

As anticipated, it is a know result that the only form of discounting that is time-consistent in the DU model is geometric. In our framework this is the same as showing

that a sliding discount matrix is time-consistent if, and only if, the entries are proportional to γ^t for some $\gamma \in (0, 1)$ (Lattimore & Hutter 2014). Using the conditions for time-consistency and sliding, respectively, we have $\alpha_2 d^1_{t+1} = d^2_{t+1} = d^1_t$. It follows that $\alpha_2 d^1_2 = d^1_1$, $(\alpha_2)^2 d^1_3 = \alpha_2 d^1_2 = d^1_1$, and similarly $(\alpha_2)^{t-1} d^1_t = d^1_1$. Rearranging gives $d^1_t = (\alpha_2)^{1-t} d^1_1 \propto \gamma^t$ for $\gamma = 1/\alpha_2$, which has the geometric form, and this discounting is preserved in later columns since the matrix is sliding.

2.3.4 Remarks and extensions

A couple of limitations of the DU model, as identified by Frederick et al. (2002), remain with the matrix-based approach to time discounting: utility independence and consumption independence.

Assuming global utility can be expressed as the sum of local utilities means an agent cannot have preferences for certain patterns of rewards across time. For example, provided the rewards sum to the same value, we cannot distinguish a flat well-being profile from one with greater variation between these rewards. This is because we are summing the rewards with no constraints, and addition is a commutative operation so the order of the addends does change the value of the sum. For a model of time preferences to overcome this issue it would have to apply discount weights in a way that depends on the history, more specifically on what rewards were previously received, which GTCD is unable to do.

Consumption independence, instead, is the assumption that the utility at a given time-step is independent of the utility at other time-steps. But that cannot account, as the examples usually go, for a person repeatedly eating the same dish and getting sick of it (Lattimore & Hutter 2014). Once again, there is an issue with path independence, though it is one that mostly concerns economists in this case.

More worryingly for us, we note that an agent's actions should be able to change its discount function. Counter-examples from everyday life abound, e.g. a person who decides to take a course in financial education might become more far-sighted than they would have been otherwise.[2] Furthermore, for what it is worth, there is evidence that history-dependent variables are more significantly related to people's discount rates than age is (Chao et al. 2009). Yet a discount matrix allows the discount function to change with time only; it is a small improvement over the DU model which does not allow any changes at all. Another way of looking at this is that, in a discount matrix, changes to the discount function depend on the length of histories rather than the sequences of actions, rewards, and observations that make up these histories.

2.3.4.1 Health, survival probability, and discount rates

As we discussed earlier, it is well documented that people's discount rates change as they age. The mechanisms behind changing time preference, however, are not well understood. Theories as to why our discount rates change have been advanced both in

[2] Thanks to Tor Lattimore for mentioning a clear example like this at an early stage, when I was still wrapping my head around the whole concept of time inconsistency.

economics and evolutionary psychology, sometimes invoking notions like Darwinian fitness, though there seems to be little agreement amongst the authors and even less evidence in of support of their claims (Read & Read 2004). Thankfully, some empirical work was done on the topic recently.

Chao et al. (2009) found that variables such health and survival probability are more significantly related to people's discount rates than age is. What is particularly revealing about this study is that it was conducted around South African towns, where the population has a high prevalence of HIV/AIDS. Here the risks of mortality or morbidity are not so strongly correlated with age, and indeed, this variable was not found to be a significant predictor of time preference. The authors advance the following explanation: "health and especially survival probability, not age, may be a true underlying determinant of people's SDRs [subjective discount rates]. In populations where age does correlate well with health and survival probability, the effect of these other variables on SDRs can be well-manifested by the effects of age. However, because causes of morbidity and mortality in South Africa are not necessarily related to age, age is no longer a strong predictor of health and expected survival and, hence, of SDRs" (Chao et al. 2009).

We should take findings such as these seriously. And if we do take them seriously, then we must improve our models of changing preferences so that the agent's discount rate may vary as a function of the environment, not merely of time. At the very least, we should be able to account for this kind of phenomena where discount rates are affected by the agent's chance of survival.

2.3.4.2 History dependent discounting

The limitations of GTCD are due to discounting without any consideration for the history, so an obvious way to remove them is extend the GTCD model by having discount vectors depend on histories rather than time-steps. To achieve this we can simply redefine D as a function that maps histories to discount vectors, i.e. $D : \mathcal{H} \to [0,1]^\infty$. The discount vector of the agent after history $h_{<k}$ is $\boldsymbol{d}^{h_{<k}}$, and the factor used to discount rewards at time t by that agent is $d_t^{h_{<k}}$. Though this is a slight abuse of symbols because D is no longer a matrix but, rather, a multi-dimensional array.

In any case, our notation allows the definition of time consistency to carry over naturally. Recall that what characterised an agent as having time-consistent preferences was having no incentive to change its plans at any point. In other words, the optimal policies for this agent remained the same at every time-step. So a multi-dimensional discount array D is *time-consistent* if, and only if, for any environment, $\Pi^*_{\boldsymbol{d}^{h_{<k}}} = \Pi^*_{\boldsymbol{d}^{h_{<j}}}$ for all $h_{<k}, h_{<j} \in \mathcal{H}$.

The notion of sliding can also be extended quite neatly to these multi-dimensional arrays. Recall that in sliding matrices the discount function does not change over time, which captures the DU model faithfully. As the agent is making decisions sequentially, given its discount function does not change over time, when it makes a new decision it re-adopts the old discount function by applying the same factors to different future rewards. So a multi-dimensional discount array is *sliding* if, and only

if, $d_{k+t}^{h_{<k}} = d_{1+t}^{h_{<1}}$ for all histories $h_{<k}$ and $t \geq 0$, where $h_{<1}$ is the empty history we usually omit.

Finally, it is easy to see that this model is more general than GTCD which, in turn, is more general than DU: when all histories of the same length are mapped to a unique discount vector, we have GTCD as a special case; when all histories of any length are mapped to a unique discount vector, what we have is Samuelson's DU model.

2.3.4.3 Environment-based models

One last alternative is to not apply any discounting at all. Legg & Hutter (2007) and Legg (2008) suggest the environment be used to capture time preferences, and according to Lattimore & Hutter (2014) the aforementioned limitations are overcome by these environment-based models.

A note of caution, however, is in order here. What removing a discount function achieves is avoiding the issue of time-inconsistency, as the agent cannot have changing preferences in the first place. This is fine when our goal is to build a reinforcement learning agent that learns to perform well with respect to preferences that are fixed over time, which is usually the case in computer science. Such an approach, however, does not allow us to capture more complex time preferences that change, and this may be a limiting factor in other circumstances. For example, we may want to build an AI that, given some user preferences as input, will advise what actions to take in order to maximise expected future rewards. But people are notoriously time inconsistent, so our program would need to reason about the user's changing preferences.

Why is it that environment-based models cannot capture changing preferences? To see this, let us consider how we could go about modifying an environment to capture what time preferences an agent already has over the rewards. For a time-consistent agent whose discount function is unique and fixed, the solution is simple: we can just weight rewards using factors given by the discount function. The underlying idea is to alter the environment so that, instead of the previous rewards, it now returns each reward r_t discounted by the a factor corresponding to the time-step t or, more generally, the history $h_{<t}$.

To express this new environment $\mu'(\cdot)$ rigorously, let $\mu'(h'_{<k}a_k) = \langle o'_k, r'_k \rangle$ and $\mu(h_{<k}a_k) = \langle o_k, r_k \rangle$. Then it must hold that $o'_k = o_k$ and $r'_k = \boldsymbol{d}_k \cdot r_k$ when $\forall k, t \leq k$ we let $h'_{<k} := h_{<k}$ with the substitution $\{r_t \mapsto (\boldsymbol{d}_t \cdot r_t)\}$, for some appropriately defined discount vector $\boldsymbol{d}$.

In our example, $\boldsymbol{d}_t$ is the unique factor associated with each reward produced by the environment when an action a_t is taken after history $h_{<t}$. More generally, however, the term $\boldsymbol{d}$ is not uniquely defined and there is not a clear choice to be made. In fact, considering our generalisation where temporal preferences depend on the history, we should write this term as $\boldsymbol{d}_t^{h_{<k}}$. So we can see that it depends on $h_{<k}$, which means that, for a fixed t, $\boldsymbol{d}_t^{h_{<k}}$ is still relative to which vector the discount factor for time t is sourced from. This illustrates intuitively the limitations of environment-based models: an environment alone imposes a unique time preference on the agent, whereas an agent's discounting behaviour might vary from some time-step to the next or even be

inconsistent.

2.4 Applying game theory

It is this section that we look at how game theory can be applied to reinforcement learning. We return to a problem by Lattimore & Hutter (2014) and compare it to an older puzzle from the philosophical literature. Then we introduce a novel type of policy for this and other situations of time-inconsistency, but not before covering the game-theoretic background that is needed. In the last section we talk about the problem of selecting equilibria, which limits the effectiveness subgame perfect equilibrium policies, and we reveal it to be much harder than it may have at first appeared. Though no solution is generally accepted, work in this areas has produced concepts that might well turn out to be useful in solving this open challenge; we end by discussing such criteria alongside a number of examples.

2.4.1 A case study from reinforcement learning

Recall the environment from 2.3.1 in which an agent with a bounded horizon of two time-steps would behave inconsistently:

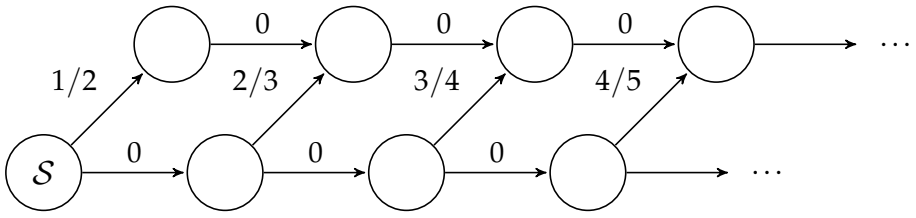

Figure 2.7: The environment from 2.3.1.

We saw that maximise expected utility greedily, at each time step, led an agent to the lowest sequence of rewards here. We also have discussed that, more generally, precommitment is not an available option in these situations of time inconsistency. The only reasonable course of action left for an agent is to select its actions consistently, by taking into account its later choices. A question that is left, then, regards how to construct such plans. The decision problem above is a rich source of insight for answering this question, so we will return to it.

2.4.1.1 Coordination failures and binding

A similar problem to one just presented has been studied in the philosophical literature. Arntzenius et al. (2004) consider a case in which the decision-maker, who is none other than Donald Trump, must make an infinite sequence of decisions:

Trumped. "Donald Trump has just arrived in Purgatory. God visits him and offers him the following deal. If he spends tomorrow in Hell, Donald will be allowed

to spend the next two days in Heaven, before returning to Purgatory forever. Otherwise he will spend forever in Purgatory. Since Heaven is as pleasant as Hell is unpleasant, Donald accepts the deal. The next evening, as he runs out of Hell, God offers Donald another deal: if Donald spends another day in Hell, he'll earn an additional two days in Heaven, for a total of four days in Heaven (the two days he already owed him, plus two new ones) before his return to Purgatory. Donald accepts for the same reason as before. In fact, every time he drags himself out of Hell, God offers him the same deal, and he accepts. Donald spends all of eternity writhing in agony in Hell. Where did he go wrong?"

This decision problem has the same underlying structure as the one we were previously considering, so it can be represented using the same graph:

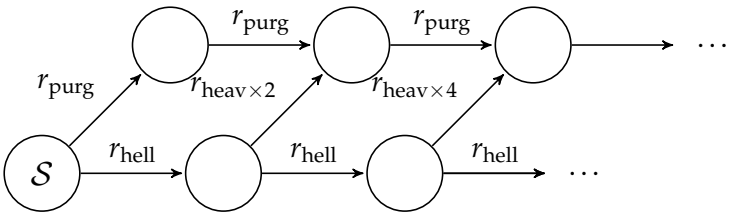

Figure 2.8: An adaptation of Arntzenius et al's problem.

Here the rewards, labelled on the edges, are given for spending a day in purgatory or hell, or several days in heaven. We need not assign precise numerical values, but we do need to have $0 \leq r_{\text{hell}} < r_{\text{purg}} < r_{\text{heav}\times 2} \leq r_{\text{heav}\times 4} \leq \ldots$. A reasonable assignment, in any case, might be $r_{\text{purg}} = 1/2$, $r_{\text{heav}\times 2} = 2/3$, $r_{\text{heav}\times 4} = 3/4$, $r_{\text{hell}} = 0$, and so on, just as in our original problem. Technically, to have several days in hell to be worse than one we would have to modify the environment and let $r_{\text{hell}} > 0$, but it is not important for this condition to hold.

The investigation carried out by Artnzenius et al. was a broader one, which looked at puzzles that arise when decision involves infinities, e.g. infinitely many options (even when they are presented sequentially) or unbounded utilities. The takeaway message of their paper is that all kind of paradoxes arise when an agent faces infinities. Perhaps, then, we should avoid using such examples whenever possible. Regardless, in our case, we have seen that the decision-maker's loss is explained entirely by how it values acts differently across time; Lattimore and Hutter's was a striking example that dramatised dynamic inconsistency, given the sheer magnitude of the rewards lost and the agent's questionable choices. In fact, infinities alone are not the issue in either of these problems, although a sequence of infinitely many options is presented to the decision-maker.

At first we may blame Trump for being myopic; his misfortune is the result of failing to consider what future offers he will accept, and it would be avoided if only he had enough foresight. This answer addresses the puzzle in its current form. Still, there is a modified version that is not so easily solved:

Satan's Apple. "Satan has cut a delicious apple into infinitely many pieces, labeled by the natural numbers. Eve may take whichever pieces she chooses. If she takes merely finitely many of the pieces, then she suffers no penalty. But if she takes infinitely many of the pieces, then she is expelled from the Garden for her greed. Either way, she gets to eat whatever pieces she has taken."

To analyse this problem, it is useful to distinguish whether Eve must make her decisions synchronically or diachronically. That is to say, whether she decides in advance how much of the apple to eat or she chooses the pieces one by one. In the former scenario, she can avoid getting banished from the Garden simply by picking some arbitrary, finite number of pieces to eat. It is true that there will always be a dominating option, of taking at least one more piece, but that is inevitable. The diachronic case is more complicated, as it depends on Eve's ability to bind herself to a course of action. Were she to possess such powers of pre-commitment, her situation would resemble the synchronic one since, in practice, she could decide in advance on which profile of offers to accept. More generally, however, unless an agent is capable of binding itself to a plan, or at least in some way causally influencing its own future choices, there is no guarantee it can coordinate to avoid the worst outcome. So, in the end, failure is to be blamed on the decision-maker's inability to self-bind; if Donald Trump cannot follow a plan, then he may remain stuck in hell indefinitely.

In Lattimore and Hutter's problem, choices are made diachronically and we also cannot rely on the decision-maker binding itself to a plan. Though we could program an agent to store a policy and force its successors to execute it, in the standard reinforcement learning setup, which we are considering, this is not assumed to be possible. An agent must, instead, select actions based only on its discount and utility functions, and its knowledge of the environment. Still, there is some leeway for the decision-maker to avoid at least the worst sequence of actions, given that discounting with a limited time-horizon helps. Artnzenius et al.'s original puzzle has many similarities, but the decision-maker has as unbounded time-horizon and no preferences for earlier rewards; it does not apply any temporal discounting. Furthermore, thanks to its perfect information about the discount matrix, our agent can anticipate in advance how its successors will behave. We shall look into all of this soon, after introducing some game-theoretic tools.

2.4.2 Basic game theory concepts

Here we review some basic concepts from game theory that are needed to understand the rest of this chapter. All material is presented along the lines of the graduate textbook by Osborne & Rubinstein (1994) and is meant as a refresher only; the reader may want to refer to such an introductory book as they go along or skip the forthcoming subsections entirely, depending on their background.

2.4.2.1 Strategic games and the Nash equilibrium

Games are modelling devices used to formalise the interactions of a group of decision-makers. A strategic game can be captured a a triple $\langle N, (A_i), (\succsim_i) \rangle$ made up of a finite set of players N, sequence of non-empty sets of actions A_i available one by one to the respective player i, and another sequence whose elements are the preference relations $\succsim_i$ each player i has over the outcomes or strategy profiles $A = \times_{j \in N} A_j$. Unless otherwise specified, we assume that the players' preference relations can be captured with utility functions and we refer to these payoffs directly. For a simple two-player game in which each player i has two moves U_i and V_i, these appear inside the grid below:

		Player 2	
		U_2	V_2
Player 1	U_1	a_{11}, b_{11}	a_{12}, b_{12}
	V_1	a_{21}, b_{12}	a_{22}, b_{22}

Figure 2.9: A generic 2×2 game.

The standard solution to strategic games is the Nash equilibrium. In a strategic game $\langle N, (A_i), (\succsim_i) \rangle$, it is defined as a profile of actions $a^* \in A$ for which it holds that $(a^*_{-i}, a^*_i) \succsim_i (a^*_{-i}, a_i)$, for every action $a_i \in A_i$ and player $i \in N$. The term a^*_{-i} is shortened notation for the rest of the profile a^* excluding a^*_i. Intuitively, this equilibrium corresponds to a stable situation in which no one has any reason to deviate so long as the other players also do not alter their moves. Let us find the Nash equilibria in the classic Prisoner's Dilemma game:

		Player 2	
		Cooperate	Defect
Player 1	Cooperate	3, 3	0, 4
	Defect	4, 0	1, 1

Figure 2.10: The Prisoner's Dilemma.

In the game above, two agents each have the option of cooperating or not with each other. They both have an incentive to defect while the other player is cooperat-

ing, so the point (Cooperate, Cooperate) is unstable , but neither has an incentive to cooperate while their opponent is defecting, which makes this second point (Defect, Defect) an equilibrium. It is, in fact, the only Nash equilibrium here.

Not all games, however, have a Nash equilibrium as defined earlier. This other classic game, colloquially referred to as Matching Pennies, does not have any:

		Player 2	
		H	T
Player 1	H	1, -1	-1, 1
	T	-1, 1	1, -1

Figure 2.11: Matching Pennies.

Although we cannot guarantee that existence of equilibria in pure strategies, of that kind described above, every game has at least one mixed strategy equilibrium. This condition holds so long as all sets of actions A_i available to the players are finite, which will always be the case in the problems we consider. The mixed strategy equilibrium is an extension of the Nash equilibrium that allows the players to randomise their moves, as opposed to having to choose them deterministically. Its construction involves defining a probability distribution over pure strategies and extending the players preferences so that they are defined over lotteries on the set of outcomes A. These details are not really crucial and can be found in Osborne and Rubinstein's (1994) textbook.

In any case, for the purpose of illustration, let us analyse Matching Pennies. We can see clearly that there is no Nash equilibrium in pure strategies, since at every cell one of the agents has an incentive to deviate unilaterally. The more interesting fact is that by playing each move with the same probability of $1/2$ the players achieve a mixed equilibrium. It gives them each an expected payoff as low as 0, but neither can increase this reward by redistributing the mass of their probability distribution across the pure strategies.

2.4.2.2 Extensive games and the subgame perfect equilibrium

In the model of a strategic game players chose their moves simultaneously, and they all made their decision independently of each other, in the sense that an agent could not observe its opponents' actions until after it had made its own choice. But what actually interests us, and we will use later, is a setting in which agents interact sequentially by taking turns. Additionally, we wish for certain criteria to be met. Players must know the structure of the game tree and the payoffs of every other player as well as their own; this amounts to complete information. Also, at any given time they

must be aware of the history of previous moves; coupled with complete information, this condition is jointly equivalent to perfect information.

An extensive game with perfect information can be defined as an ordered tuple $\langle N, H, P, (\succsim_i)\rangle$, comprised of a sequence of preferences $(\succsim_i)$ and the game form $\langle N, H, P\rangle$ made up, in turn, of a set of players N, a set of histories H, and a player function $P : H \twoheadrightarrow N$ mapping selected histories to players. Histories may or may not be finite, but the set of histories must satisfy a couple of conditions: it must contain the initial or empty history $\varnothing$, which is the starting point of the game, and H must also contain every subsequence of histories that are in it. We interpret an element a^k of the history $\left(a^k\right)_{k=1,\ldots,K+1} \in H$ as an action made by some player. This agent, who is given by the player function applied to the history h up to that point, can choose from a set of available actions $A(h) = \{a : (h, a) \in H\}$. We assume that each player i has preferences that are captured by a relation $\succsim_i$ over non-terminal histories Z, i.e. infinite histories or histories with no extension that is also in H. Note that the player function is defined everywhere except at those terminal histories, so exactly over the domain $H \setminus Z$, as it would not make sense to have a player choose moves upon completion of the game. We can draw extensive games in tree format as done below:

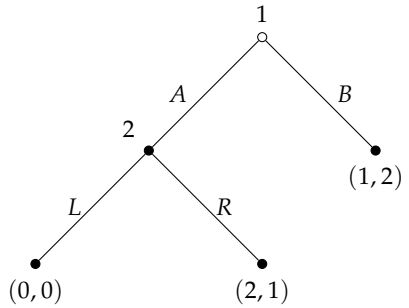

Figure 2.12: A two-player extensive game (Osborne & Rubinstein 1994).

A strategy s_i for a player $i \in N$ is function mapping non-terminal histories $h \in H \setminus Z$, such that $P(h) = i$, to unique actions in $A(h)$. While a profile of such strategies $s = (s_i)_{i \in N}$ determines the outcome $O(s)$ of a game, which is the terminal history that results when each player acts as prescribed by its strategy in this profile. Like before, we can construct mixed strategies by providing a probability distribution on sets of pure strategies as just defined. A strategy can be thought of as a complete contingent plan, as it accounts for all eventualities and, thus, is defined even over histories that would not be reached should the strategy actually be played.

The notion of Nash equilibrium also carries over directly and is now defined as a strategy profile s^* with the property that $O\left(s^*_{-i}, s^*_i\right) \succsim_i O\left(s^*_{-i}, s_i\right)$, for every strategy s_i of every player $i \in N$. However, it is no longer a reasonable choice for a rational agent because it fails to account for the sequential structure of the extensive games.

Consider the figure above which could depict, say, a game with two firms: one choosing to enter a market or not; the other, who is is already present in said market, choosing whether to remain and compete. There are Nash equilibria at (A, R) and

(B, L) with payoffs of $(2, 1)$ and $(1, 2)$ respectively. The former is an equilibrium because the second player would not want an outcome of $(0, 0)$ once node 2 has been already reached, nor could the first firm prefer to choose B when its opponent would play R after A is chosen. The latter is an equilibrium too because the player with the initial move prefers the outcome $(1, 2)$ over $(0, 0)$, so it would not opt for A when it is followed by L, and the same firm would not choose B should the other one reply with R. These claims are trivial to verify by calculating the outcomes of four possible profiles. Though both points are Nash equilibria, (B, L) is unrealistic since, unless we stipulate otherwise, there is no credible way the second firm can enforce a threat of playing L in retaliation to A. This is because it would no longer have an incentive to act in that way once the second node is reached.

Issues like these motivate a refinement of the Nash equilibrium known as the subgame perfect equilibrium. To introduce it we must first define formally the notion of a subgame, which for $\Gamma = \langle N, H, P, (\succsim_i) \rangle$ is the extensive game $\Gamma(h) = \langle N, H|_h, P|_h, (\succsim_i |_h) \rangle$ that follows after history h. The elements of this new quadruple define the tree that is contained after history h, and they include preference relations that are consistent with the new game. So all players remain the same, the set of histories $H|_h = \{h' : (h, h') \in H\}$ gets restricted to continuations of h, the new player function accordingly is $P|_h = P(h, h')$ for $\forall h' \in H|_h$, and each preference relation is now defined as $h' \succsim_i |_h h''$ if, and only if, $(h, h') \succsim_i (h, h'')$. Let O_h be the outcome function for the subgame $\Gamma(h)$. A subgame perfect equilibrium is a strategy profile s^* with the property that $O_h(s^*_{-i}|_h, s^*_i|_h) \succsim_i |_h O_h(s^*_{-i}|_h, s_i)$, for every strategy s_i of every player $i \in N$, and additionally, in every subgame $\Gamma(h)$. More simply, this subgame perfect equilibrium is a strategy profile that induces a Nash equilibrium on every subgame of the full tree.

In the first example, only (A, R) was a subgame perfect as well as a Nash equilibrium, so (B, L) is no longer considered a solution. The subgame perfect equilibrium concept addresses the issue of non-enforceable threats, but it has been criticised for other shortcomings which can be seen in the famous Centipede Game, originally by Rosenthal (1981), represented here in its six-period version:

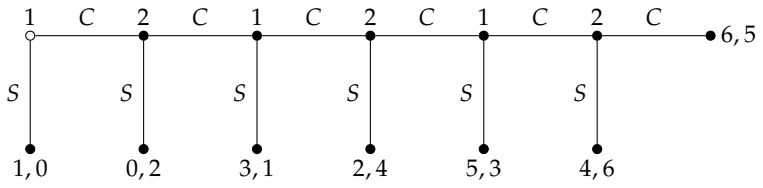

Figure 2.13: The Centipede Game (Osborne & Rubinstein 1994).

In this game players alternate taking moves, and at each time-step one chooses whether to quit and receive whatever the current payoffs are, or alternatively, play on for a chance to gain a higher reward. The catch is that receiving a higher reward later depends on the opponent being willing to continue as well, rather than preferring to quit when it is most advantageous for them. There is only one subgame perfect

equilibrium, which consists of each player choosing option S at every period (Osborne & Rubinstein 1994); this can be found by backwards induction. So the solution is supposedly to take the rewards immediately. But imagine the first player decides to play on, as people frequently do when asked to participate in this game (McKelvey & Palfrey 1992). Would it still be sensible to stick with the subgame perfect equilibrium? This strategy was based on assumptions about the opponent's behaviour that have just been violated, and in any case, following it now would amount to disregarding observations of past actions which suggest that the other player is cooperative and willing to play on for a mutually beneficial outcome. Perhaps, then, the subgame equilibrium is not the optimal profile of strategies after all.

It is unclear whether these last considerations are at all relevant us, but it is only fair we mention them, lest we give a too favourable impression of the subgame perfect equilibrium. They are not issues that relate closely to our case study from reinforcement learning, but they make this solution concept less appealing overall.

2.4.3 Game-theoretic equilibria as policies

Having acquired the right mathematical tools, we can finally apply game theory to our problems from reinforcement learning.

2.4.3.1 Subgame perfect equilibrium policies

Even though we established that consistent planning was the only viable option for an agent who is time-inconsistent and wishes to avoid intertemporal conflicts, we are still left with the task of formalising this notion and constructing adequate policies. Thankfully, extensive games and the subgame perfect equilibrium provide us with what we need.

The way Lattimore & Hutter (2014) go about this is by treating the sequential decision problems of RL as a game. More precisely, as an extensive game with perfect information. Players in this game are the decision-maker and its future selves, and each player will choose actions according to its own preferences. That is to say, each agent existing at time-step t will act to maximise expected future rewards as discounted by the vector $\boldsymbol{d}^t$. The sequential nature of the RL setting makes it appropriate for the extensive form, rather than the strategic one, and both requirements of perfect information are met. The agent has full information about all strategic possibilities at every point, which are the actions available at each time-step, and it knows the preferences of other players, as it shares its underlying preferences with them and also has access to their discount vectors through the discount matrix. So complete information is satisfied. The decision-maker is always fully aware of the history. By construction it returns actions based on previous histories, so each of its temporal selves must be informed about the previous moves made during the game. Therefore the requirement of perfect information is satisfied too.

As we have seen, the textbook solution for extensive games with perfect information is to play a subgame perfect equilibrium. A policy π^*_D is a *subgame perfect*

equilibrium if, and only if, for every t and $h_{<t}$, $V^{\pi_D^*}_{\boldsymbol{d}^t}(h_{<t}) \geq V^{\tilde{\pi}}_{\boldsymbol{d}^t}(h_{<t})$, where $\tilde{\pi}$ is any policy such that $\tilde{\pi}(h_{<i}) = \pi_D^*(h_{<i})$ for all $h_{<i}$ with $i \neq t$. Our new definition does not differ substantially from the more common phrasings found in the game theory literature and is simply translated to the RL setting.

What assurances do we have that an agent will always be able to play a subgame perfect equilibrium against its future selves? Well, we do know that for all environments μ and discount matrices D there exists a policy π_D^* that is subgame perfect equilibrium (Lattimore & Hutter 2014); the existence is subject to some technical assumptions which can be found in the same paper. This result appeases our concerns, though it comes with a caveat: the policy may not be unique. The non-uniqueness of such equilibria is not a feature of our formalisation but of the concept of subgame perfect equilibrium itself, and it is rather problematic. As for the other issues that have drawn criticism to this solution concept, we can brush them aside. Similarities in the tree shapes notwithstanding, the Centipede Game resembles Lattimore and Hutter's environment only superficially, with the main difference being that our problem involves more than two players.

2.4.3.2 Non-uniqueness of the policies

As noted by Lattimore & Hutter (2014), their game below has two subgame perfect policies in pure strategies. At least technically, there are also infinitely many mixed strategies that qualify as subgame perfect equilibria. These profiles, however, agree on all but what move the initial agent should play, which means there is a unique mixed subgame perfect equilibrium modulo the first player's move. As the game does not have a finite horizon, we cannot use backwards induction to prove our claims. So we shall give arguments.

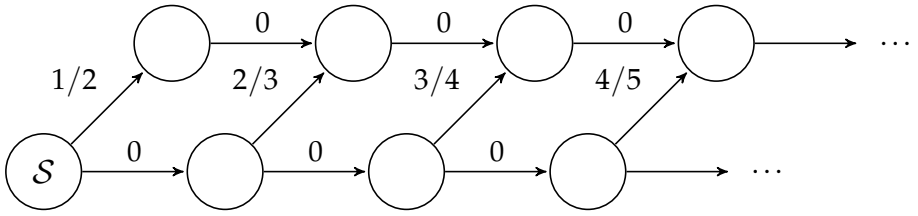

Figure 2.14: The environment from 2.3.1.

Let us start by ignoring the top nodes, as they each have one available action, therefore all strategy profiles coincide there. Looking at the bottom nodes only, then, we can see that a subgame perfect equilibrium cannot have two consecutive agents both moving up or both moving right. The predecessor of an agent who moves right must move up, for if it also moved right it would miss out on rewards altogether; recall that our players only cares about rewards that are at most two nodes away. Similarly, the predecessor of an agent who moves up must move right, since it would otherwise miss out on the larger reward received by having its successor move up at the next node.

So far, we have narrowed possible equilibria down to profiles with agents at the bottom nodes alternating between moving up and right, of which there are two according to the initial action taken. It is easy to see that both these strategy profiles are subgame perfect equilibria.

Without loss of generality, let us pick the profile of strategies in which the first player moves up. Consider an arbitrary player whose strategy is to move up and, keeping fixed all other players' strategies, suppose it decided to move right instead. Since we know that of all future rewards it only values the immediately successive one, and its successor will move right, by moving right itself it would end up with a discounted cumulative reward of zero. That, however, is inferior to the non-zero reward it would have received otherwise, so it has no incentive do deviate. By the same line of reasoning, an arbitrary agent whose prescribed strategy is to move right has no incentive to unilaterally deviate. This is because its successor will be moving up, which would give it a higher cumulative reward. Therefore both the alternating strategy profiles are, indeed, subgame perfect equilibria.

So much for pure strategies. Now, we shall proceed to find the mixed subgame perfect equilibria, and let us, once again, ignore the top nodes. Label with r_n the reward that the n^{th} player receives for moving up, i.e. $r_1 = 1/2$, $r_2 = 2/3$, and so forth. Also, let p_n be the probability that the n^{th} player moves up in some given mixed strategy. It is rational for a decision-maker to play a mixed strategy only when it expects this mixture of moves to yield rewards that are no lower than the payoffs given by either move; otherwise it should play a pure strategy directly. For the n^{th} player, whose successor moves up with probability p_{n+1}, this gives us the condition $r_n = p_{n+1} \cdot r_{n+1}$, which implies its successor should move upwards with probability $p_{n+1} = r_n / r_{n+1}$. So we have fixed a precise value for the probability with which each player should move up according to a mixed strategy in subgame prefect equilibrium, except for the agent at the initial decision node who is not subject to any constraints. It is easy to verify that an equilibrium is preserved regardless of the first player's move, and the degree of freedom that gives us uncountably many equilibria comes from here.

Perhaps, it may be a cause for worry that our mixed strategies are in some sense unstable.[3] While each agent has no incentive to change its move, it also has no strict preference to stick with the equilibrium. That is to say, any player can deviate unilaterally from its equilibrium strategy without incurring penalties. The instability is a feature of all equilibrium points in mixed strategies (Harsanyi 1973), not just the ones of this particular game, and it is only apparent. Mixed strategies cease to be unstable once we make a very plausible background assumption, that games we interpret as having fixed payoffs are actually perturbed. In other words, the payoffs are randomly fluctuating, with fluctuations, perhaps extremely small, that are given by some irreducible level of uncertainty each player has about the other players' payoffs. This uncertainty regarding a player's utility function may be explained in terms of

[3] Thanks to my supervisor Rachael Briggs for raising this concern and, in doing so, making me reflect at length about the usefulness of mixed strategies. Credit also goes to them for pointing out the mixed equilibrium in this game.

whatever external factors influence an agent's preferences, and such sources of randomness will exist in any realistic scenario. So we can rest assured knowing that mixed strategies are stable once this assumption is taken into account, as justified and proved rigorously by Harsanyi (1973, 1988).

What is truly concerning is the presence of multiple equilibria. Not simply because there are several of them, but because none make for a clear choice on which the agents can coordinate.

Without even considering all the mixed equilibria, let us just focus on the profiles of pure strategies. The cumulative rewards for playing these profiles differ from agent to agent. For example, the agent at the initial node will receive rewards of $\boldsymbol{d} \cdot [1/2, 0, 0, \dots] = 1/2$ and $\boldsymbol{d} \cdot [0, 2/3, 0, \dots] = 2/3$, with $\boldsymbol{d} = (1, 1, 0, \dots)$, in the equilibrium where it moves up or right respectively. This is where matters get interesting. The decision-maker with the first move wishes to maximise its own utility, so between those equilibria just mentioned it prefers the one with the greatest cumulative reward for itself, i.e. the one associated with a reward of $2/3$. As it has the first move we might be tempted to conclude that it should play according to that equilibrium. The situation is more complicated, though, because the player cannot force subsequent agents to act accordingly. In fact, the second player prefers the other of the two equilibria, for which it receives a reward of $\boldsymbol{d} \cdot [0, 0, 3/4, \dots] = 3/4$ instead of $\boldsymbol{d} \cdot [0, 2/3, 0 \dots] = 2/3$, with $\boldsymbol{d}$ now equal to $(0, 1, 1, \dots)$. The next player prefers the original equilibrium which gives it the higher reward in a particular subgame, and so on. Eventually, should all the intertemporal selves play according to their preferred equilibria, then the result is that no equilibrium gets played and nobody receives any rewards.

Agents are sometimes able to coordinate thanks to a public signal, achieving a so-called correlated equilibrium (Aumann 1974). Other games have focal points that are obvious coordination targets due their distinctive characteristics (Schelling 1960). Unfortunately, none of these tricks are available here; we need a solution to the underlying problem of how to select between multiple equilibria.

2.4.4 The problem of equilibrium selection

We have just seen that, for some discount matrices and in some environments, there are multiple subgame perfect equilibrium policies. Furthermore, these policies may appear identical with respect to any desirable characteristic we want them to satisfy, but for our solution concept to be satisfactory it needs to provide an unequivocal, unique plan for the agent to follow. Lattimore & Hutter (2014) are aware of this, and at the end of their paper, they identified it as an open question that needs to be solved. Though they note that "the problem of how players might choose a sub-game perfect equilibrium policy appears surprisingly understudied," and they were correct in doing so, what they may not have realised is that this problem runs much deeper and affects more than just the usefulness of the subgame perfect equilibrium concept.

This problem of equilibrium selection is actually a major issue that lies at the very heart of game theory. As we shall see, multiple Nash equilibria can occur even in

very simple 2×2 matrix games. First, though, consider this extreme case devised by Harsanyi & Selten (1988):

Bargaining. "Two players have to agree on how to divide \$100; the money is lost to them if they cannot agree. (We will assume that both players have linear utility functions for money.) This game can be represented by the following bargaining model: Each player has to name a real number, representing his payoff demand. The numbers named by players 1 and 2 will be called x_1 and x_2, respectively. If $x_1 + x_2 \leq 100$ (if the two players' payoff demands are mutually compatible), then both will obtain their payoff demands, with $u_1 = x_1$ and $u_2 = x_2$. In contrast, if $x_1 + x_2 > 100$ (if their payoff demands are incompatible), they will receive zero payoffs $u_1 = u_2 = 0$ (as this will be taken to mean that they could not reach an agreement)."

Every pair (x_1, x_2) such that $x_1, x_2 \geq 0$ and $x_1 + x_2 = 100$ is an Nash equilibrium, so if the players are allowed to pick demands in the real numbers, then there are infinitely many equilibria for this game. Taking the problem description more literally, say by restricting the players' payoffs to integer amounts of dollars, we are still left with at least 101 equilibrium points. A theory that only tells so much is hardly informative; clearly, we need a stronger solution concept.

One would be justified in thinking that the problem of equilibrium selection has received notable attention, given its gravity, but sadly this is not the case. Some economists have gone as far as remarking that: "Despite its central role in game theory and economics, the literature offers very few formal approaches to the problem of equilibrium selection. For instance, most solution concepts proposed in the literature on refinements of Nash equilibria [...] have nothing to say about selection among strict Nash equilibria" (Matsui & Matsuyama 1995). Indeed, as a cursory read through a graduate textbook in game theory like that of Osborne & Rubinstein (1994) is sufficient to reveal, there are plenty of solution concepts such as the subgame perfect equilibrium we have already discussed at length. These refinements provide us with sets of points that are usually more restricted than the Nash equilibria of a same game, which is due to the new solutions having to satisfy additional properties. However useful they may be for solving specific classes of games, though, these concepts neglect to tell us how we should select amongst several, possibly symmetric, equilibria of the new kind.

Challenging the long-standing solution concept for non-cooperative games is no easy feat. But one serious attempt at this gargantuan task, of providing a new concept to replace the Nash equilibrium, was undertaken by Harsanyi & Selten (1988).

Although the authors believe that an acceptable solution is too complex for an axiomatic characterisation, they do suggest a few axioms as starting points: isomorphism invariance, payoff monotonicity, subgame and truncation consistency, and best-reply invariance. First is the indispensable requirement that a solution concept should not depend on the agents' names or how we label their moves, which is to say it should be invariant under isomorphisms of these variables, and it should also remain the same when scaling payoffs by positive linear transformations. Their second

desideratum is the intuitive requirement that increasing a payoff at what is already an equilibrium should preserve this point as a solution; according to Harsanyi and Selten, the property is no longer convincing once we consider games with three or more players, and therefore, should be discarded. Subgame and truncation consistency force the solution concept to return the same equilibria at certain portions of the game that are assumed to be strategically self-contained; Harsanyi (1995) will later reject this condition and propose a solution that violates it. Last, best-reply invariance is the property that the removal of an inferior choice should not change the solution of a game; to be satisfied in general, however, this requirement needs to be weakened substantially.

We will not discuss the specifics of Harsanyi and Selten's work any further as it is extremely technical, well beyond the level of mathematical sophistication of this thesis. It also has never been adopted by the mainstream literature, with one of the authors himself, Harsanyi (1995), retracting the original theory. Still, it has been an investigation fruitful enough to produce some concepts that have been appreciated by the field at large; we shall introduce these now.

2.4.4.1 Equilibrium selection criteria

There are a couple of general ways we can go about selecting equilibria. Our approach is a deductive one which assumes rational agents have beliefs that are consistent with playing according to a certain equilibrium. Another possibility would be to use principles that rely on learning or evolutionary dynamics to select a one-point solution. In a number of realistic scenarios, a decision-maker plays the same game or very similar games repeatedly, and these interactions may provide guidance on which equilibria to coordinate on. Though Lattimore and Hutter's problem does involve a number of almost identical subgames, it is as if different agents were playing them, therefore we cannot rely on these inductive principles to select a unique solution. We restrict ourselves to the deductive ones, instead, and for convenience we will explore them in the context of simple 2×2 games.

In some cases, there is an equilibrium that all players would prefer to be the outcome of the game. The simplest selection criteria involves picking this point. Consider the example below:

		Player 2	
		U_2	V_2
Player 1	U_1	9, 7	0, 1
	V_1	7, 2	8, 6

Figure 2.15: A game with payoff dominance (Harsanyi & Selten 1988).

The game above has two equilibria at $U = (U_1, U_2)$ and $V = (V_1, V_2)$. Their payoffs are $(9, 7)$ and $(8, 6)$, respectively, so we can see quite clearly that each player receives a higher payoff at the former point. The equilibrium at V is said to be Pareto dominated, as there is another outcome, namely U, that would make at least one player better off and no players worse off. This dominating equilibrium seems like the rational solution to the game since the players have an incentive to play according to it, given that they all prefer the payoffs at U over those at V.

It may be tempting to make payoff dominance our selection criterion, at least for those games that have some Pareto dominated equilibrium. After all, it is a point that no one has an incentive to deviate from, just like Nash's solution, and it is also stable in a counterfactual sense: should this outcome occur, no agent would express regrets for how things might have turned out instead. There are other considerations that need to be balanced, however, as illustrated by cases like this one:

		Player 2	
		U_2	V_2
Player 1	U_1	9, 9	0, 8
	V_1	8, 0	8, 8

Figure 2.16: Conflicting payoff and risk dominance (Harsanyi & Selten 1988).

Just as before, the game has two equilibria at U and V. The outcome U is still Pareto dominant since $a_{11} = 9 > 8 = a_{22}$ and $b_{11} = 9 > 8 = b_{22}$, but it may be risky to play according to this equilibrium when each agent is unsure of how the other will behave. Player 1 might expect Player 2 to coordinate with it on U, and vice versa, though they cannot be certain of how the other will behave. If either agent is mistaken about this then it will pay a high price, ending up with no reward. On the other hand, by aiming for V everyone is guaranteed a more than reasonable payoff of 8. These considerations are accentuated by the fact that neither agent has a particularly strong incentive to play according to U given how close the rewards are to V.

An alternative selection criterion might, perhaps, involve selecting the equilibrium that guarantees the highest minimum payoff should an adversary fail to coordinate. Such a solution would work for the example above, but it is not appealing more generally. Consider this last game:

Player 1 \ Player 2	U_2	V_2
U_1	99, 49	0, 0
V_1	0, 0	1, 51

Figure 2.17: A game with risk dominance only (Harsanyi & Selten 1988).

Once again we have two Nash equilibria, though neither of them are payoff dominant, nor does playing according to any of them guarantee a higher minimum reward since the rows of Player 1 and the columns of Player 2 all contain an entry of 0. Nonetheless, there is quite clearly a more conservative profile of strategies that can, and arguably should, be played here. That is to say the profile corresponding to the equilibrium point U. Although the second agent knows it will miss out on its highest possible reward $b_{22} = 51$, which in this case is not so much greater than $b_{11} = 49$ anyway, by playing according to U it is taking less of a risk. This is because the first player has a much stronger incentive to aim for $a_{11} = 99$ over $a_{22} = 1$, of course.

All these considerations are captured formally by the notion of risk dominance. Let u_i and v_i be the losses that Player i would incur by deviating unilaterally from the equilibria U and V respectively, i.e. $u_1 = a_{11} - a_{21}$, $u_2 = b_{11} - b_{12}$, $v_1 = a_{22} - a_{12}$, and $v_2 = b_{22} - b_{21}$. This solution concept prescribes agents to play according to what equilibrium maximises the product of the deviation losses (Harsanyi & Selten 1988), also known as the Nash-product, which is given by u_1u_2 at U and v_1v_2 at V. In our example, we can see that the point U has, in fact, a higher Nash-product than V since $u_1u_2 = (a_{11} - a_{21})(b_{11} - b_{12}) = 99 \cdot 49 > 51 \cdot 1 = (a_{22} - a_{12})(b_{22} - b_{21}) = v_1v_2$.

Risk dominance might seem like the solution we were looking for. Not only does it offer anecdotal results that are intuitively appealing, it can also be motivated more rigorously in more than one way. When introducing it, Harsanyi and Selten appeal to three of the axioms we mentioned earlier: isomorphism invariance, payoff monotonicity, and best-reply invariance. Additionally, they derive it based on the so-called tracing procedure, which is a Bayesian technique to selecting specific equilibrium points.

Sadly, risk dominance has its shortcomings. In some cases where it runs in an opposite direction than payoff dominance, it is not so clear which criterion should be privileged. In fact, Harsanyi and Selten's original solution usually picks a payoff dominant equilibrium over a risk dominant one when both of these are present (Matsui & Matsuyama 1995). The example we gave above was rather extreme and devised to justify risk dominance in the face of payoff dominance. More generally, however, there are plenty of borderline situations for which a choice is tougher; we could easily modify the same game so that the payoffs are no longer tilted in favour of the risk-dominant equilibrium. Risk dominance is also pair-wise notion in that it applies properly to two equilibria, but in its current form it is intransitive, i.e. U

risk-dominating V and V risk-dominating W does not guarantee that U risk dominates W too. In his later work, Harsanyi (1995) addresses this issue with the idea of multi-lateral risk dominance, which he uses instead of payoff dominance as the central concept behind his new theory of equilibrium selection. Ultimately, though, none of these solutions have gained anywhere near widespread adoption or provided a credible substitute for the Nash equilibrium.

Chapter 3

Impossibility Theorems in Expected Utility Theory

3.1 General background

Social choice theory studies how to aggregate collections of individual inputs, usually preferences, into unified outputs. These results that are generated from the judgements of many constitute some kind of aggregate opinion, a societal consensus so to speak, which may be used to guide collective decisions. Social choice theory is a vast field in which many techniques have been applied to study different, though related, questions about the possibilities, but especially the limits, of collective decision-making. In fact, as popularised by the ground-breaking work of Arrow (1951), a major tradition in this field has focused on so called impossibility theorems regarding aggregation methods. The main component of these theorems are a set of desiderata that we intuitively regard as tenets of rationality, at least when taken in isolation, but that are then shown to be contradictory when taken together. Such results are perplexing and have major implications for socially meaningful issues such as the possibility of fair, democratic voting systems.

In any case, we promised this thesis would look at applications of social choice to decision theory, so we will focus on that. In particular, we will discuss how impossibilities from social choice theory arise in situations with a single decision-maker present. Unlike in the previous chapter where we reviewed a large body of literature, here there is little existing work that closely relates to relevant questions; we will jump almost straight into the crux of the matter. Nor will we need any background knowledge to prove an impossibility result later on, as the mathematical proof is self-contained. The reader, however, is invited to consult a reference textbook like that of Gaertner (2009) if need be.

As just mentioned, the literature that lies at the intersection of social choice and decision theory is small and rather sparse. Some recent research has been carried out by Macaskill (2014): he looks at how voting rules can be used in situations of moral uncertainty to combine the precepts of different moral theories, some of which may be otherwise incomparable. The only other work I am aware of in this area, which is the one that interests us, is due to Briggs (2010) who discovered an impossibility theorem

in expected utility theory.

Expected utility theory is a normative theory of rational choice (Briggs 2014) that stipulates how a rational agent should make decisions under uncertainty. A major point of disagreement between many scholars regards how we should calculate expected utility in situations where the decision-maker is part of the environment, with two sides debating for either causal or evidential decision theory. Artificial intelligence, and computer science more generally, have not yet caught up with this philosophical debate; the only work that I know of is a draft currently being written by Everitt et al. (2015), which looks at extending reinforcement learning to this new setting. Nonetheless, I am confident it will increasingly draw the attention of computer scientists: traditional RL problems assume some form of dualism, in the sense that they model the agent as separate from the environment; this simplifying assumption becomes untenable, for example, in decision problems where multiple agents can observe each other's behaviour and make inferences about their adversary's source code. For these reasons, I felt comfortable including such an unorthodox topic in this computer science thesis of mine.

Our goal will be to prove that impossibility theorem for expected utility theories in a less restricting setting with stochastic environments, rather than assuming the world is deterministic as done originally. In accordance with Briggs' (2010) paper, our approach to modelling environments is based on counterfactuals and is the one started by Gibbard & Harper (1981); we follow Lewis' (1981) suggestions in extending it to the generalised case that has indeterminism.

3.2 Technical preliminaries

The problems we look at here involve one-off decisions, rather than sequences of decisions as previously considered. As usual, we assume the agent's goal is to maximise what value it attains through its actions. This time, however, rather than rewards or utility functions, we will use proposition of the form $V = v$ to indicate that the agent has obtained an outcome of value v.

A decision situation can be described as a triple $\langle \mathcal{A}, \mathcal{K}, C \rangle$. $\mathcal{A}$ is the set of actions available to the agent, which we assume to be finite. $\mathcal{K}$ is a finite set of dependency hypotheses, i.e. environments, which describes the counterfactual relations that the agent believes might hold between acts and value propositions. Any of its elements $K \in \mathcal{K}$, a dependency hypothesis, is defined as $K = \{A \mathbin{\square\!\!\rightarrow} (P = p_{A,K}) : A \in \mathcal{A}\}$. That is to say, it contains exactly one proposition of the form $A \mathbin{\square\!\!\rightarrow} (P = p)$ for each each act $A \in \mathcal{A}$. These propositions can be interpreted as saying "if I were to take action A then '$P = p$' would hold," where P is a random variable that can equal some distribution p over V. This probability function p captures the chance present in the world, which is an irreducible level of uncertainty given not by our imperfect knowledge, but by the natural laws that govern the world. Finally, we have the credence function C. Our credence in a proposition is our subjective belief that the proposition holds; this is distinct from chance, but should agree with it when no further uncer-

tainty is present. The function C maps propositions to the unit interval according to the strength of our belief in them, with the usual constraints of probability functions.

For given A and K, let $p_{A,K}$ be the probability distribution p such that $A \mathrel{\square\!\!\rightarrow} (P = p) \in K$. The expectation of the random variable V with respect to a generic distribution p is $E_p[V] = \sum_v p(V = v)\ v$, so the expectation for $p_{A,K}$ will be $E_{p_{A,K}}[V] = \sum_v p_{A,K}(V = v)\ v$. We set the value of an expected outcome $V(A\&K)$ equal to this last expectation, i.e. $V(A\&K) = \sum_v p_{A,K}(V = v)\ v$. An outcome is a state of affairs of the world; we associate each of these outcomes with a value v. In a deterministic setting an outcome is identified by the conjunction $A\&K$. We assume, however, that the world is stochastic with outcomes affected by objective chance, usually distinct from our credences, so performing an act $A \in \mathcal{A}$ will not lead to a definite outcome. It could lead to several possible outcomes, with a likelihood given by the probability distribution $p_{A,K}$ unique for each dependency hypothesis $K \in \mathcal{K}$. Out of convenience, we use propositions of the form $V = v$ to state an outcome with value v comes about, which avoids us ever having to refer directly to these various outcomes.

Our credence that some value v has been realised, conditional on the fact that we have performed action A, is denoted by $C(V = v|A)$, which in turn is equivalent to $\sum_K C(K|A)\ p_{A,K}(V = v)$. To see why this is the case, consider that when we know the distribution of outcomes given by chance, finding out which action was taken will reduce our subjective uncertainty about the value realised exactly down to the likelihood given that objective chance. So, within a dependency hypothesis K, it will hold that $C(V = v|A) = p_{A,K}(V = v)$. Adding another layer of uncertainty to model our ignorance about which dependency hypothesis is true, we have that $C(V = v|A) = \sum_K C(K|A)\ p_{A,K}(V = v)$.

Our credence that performing an act A will bring about an outcome of value v is, instead, $C_A(V = v)$. The chance that this action will lead to an outcome of such value is $p_{A,K}(V = v)$ for each dependency hypothesis K, so weighting these chances by our beliefs in the hypothesis gives us $C_A(V = v) = \sum_K C(K)\ p_{A,K}(V = v)$. Again, when we are fully confident in a dependency hypothesis K the equality reduces to $C_A(V = v) = p_{A,K}(V = v)$. This, incidentally, shows that when we are certain of which dependency hypothesis is true, the two decision theories that will be under consideration agree, and they will return the same verdict regarding an action's expected value since $V_E(A) = \sum_v C(V = v|A)\ v = \sum_v p_{A,K}(V = v)\ v = \sum_v C_A(V = v)\ v = V_C(A)$.

This concludes our technical preliminaries. All assumptions but one are the same as Briggs', and the notation we adopted is in line with that of the original paper. The change we made is that dependency hypotheses are now stochastic, whereas they used to be of the form $K = \{A \mathrel{\square\!\!\rightarrow} (V = v) : A \in \mathcal{A}\}$. Although we no longer claim that each action and dependency hypothesis give rise to a unique outcome, we do assume that there is a unique probability distribution that corresponds to the pair; this assumption is our most substantial one and has been discussed by Lewis (1981).

3.2.1 Remarks about a credence term

We have used $C_A(V = v)$ where Briggs uses $C(A \boxrightarrow (V = v))$. Both credence terms in these equations aim to capture the notion of your belief that an outcome of value v will come about on the supposition that you take action A, in the causal sense that your action brings about this outcome. But there are good reasons for us to adopt a different notation, though you may be tempted to reason as in the following paragraphs.

Your credence in $A \boxrightarrow (V = v)$ is your belief that "if I were to perform A then an outcome of value v would occur". Denote this proposition by φ, i.e. $\varphi := A \boxrightarrow (V = v)$.

If you know the dependency hypothesis, and with background assumptions about how credences are related to objective chances, then your belief that φ holds should be the chance $p_{A,K}(V = v)$. That is to say, when the true K is known, $C(A \boxrightarrow (V = v)) = p_{A,K}(V = v)$.

When you ignore which dependency hypothesis is the true one you must spread your beliefs over the various hypotheses K, with appropriate weights, so your belief in φ will be, for each hypothesis K, $C(K)\ p_{A,K}(V = v)$. That is to say, in general, $C(A \boxrightarrow (V = v)) = \sum_K C(K)\ p_{A,K}(V = v)$.

The issue, however, is that the notation $A \boxrightarrow (V = v)$ can be interpreted as stating "if I were to perform A then an outcome of value v would *certainly* occur." In fact, this is how counterfactuals are usually evaluated in the literature. Under such an interpretation, your credence in φ should no longer be as described above; it will be zero when you know the world is chancy. So we have opted for $C_A(V = v)$, instead, to avoid ambiguity with the established usage of this term.

3.3 Decision theories and problems

This review section will be devoted to introducing two theories knows as evidential decision theory and causal decision theory, henceforth referred to as EDT and CDT, and to looking at some classical problems from the related literature; our analysis will follow closely that of Briggs' paper. We start by discussing the different ways that EDT and CDT assign expected value to actions, then look at how these decision rules perform on actual problems.

3.3.1 Two kinds of expected value

The expected value of an act is given, over every value that an outcome may take, by your belief that your act will lead to an outcome of that value multiplied by the value itself. In this context your belief captures both your own ignorance and the inherent uncertainty of the world, as previously discussed. We can distinguish between two kinds of expected value according to how exactly the credence term is defined.

The first kind is evidential expected value, which captures the principles of evidential reasoning and is formalised as follows:

$$V_E(A) = \sum_v C(V = v|A)\ v$$

The other kind, instead, is causal expected value, which captures the principles of causal reasoning and is formalised as follows:

$$V_C(A) = \sum_v C_A(V = v)\ v$$

The distinction between EDT and CDT is, perhaps, best understood by reformulating the definitions so that we consider the value of an act in terms of the dependency hypotheses. So, for each dependency hypothesis, an act's value will be given by its expected value in that dependency hypothesis, weighted by your belief in the hypothesis on the supposition that you are taking the act. The formulae will then look as follows:

$$V_E(A) = \sum_K C(K|A)\ V(A\&K)$$

$$V_C(A) = \sum_K C(K)\ V(A\&K)$$

When calculating the expected value of an act, EDT conditionalises your credence function on this action whereas CDT does not. What this means, in practice, is that EDT treats your choice of an act as evidence of which dependency hypothesis is true. This may seem a natural thing to do; after all, for an embodied agent, taking an action is the result of some mechanism that is part of the environment. As we will see, however, for either theory there are situations where this difference leads to unfortunate choices.

3.3.1.1 Dependency hypothesis forms

The equivalences we claimed held between the first equations and those other definitions, expressed in terms of the dependency hypotheses, can be proven in a few steps. Below we show that the two forms of EDT coincide:

$$\begin{aligned} V_E(A) &= \sum_v C(V = v|A)\ v \\ &= \sum_v \sum_K C(K|A)\ p_{A,K}(V = v)\ v \\ &= \sum_K C(K|A) \sum_v p_{A,K}(V = v)\ v \\ &= \sum_K C(K|A) V(A\&K) \end{aligned}$$

Also, the equivalence for CDT can be proven in much the same way:

$$\begin{aligned} V_C(A) &= \sum_v C_A(V = v)\ v \\ &= \sum_v \sum_K C(K)\ p_{A,K}(V = v)\ v \\ &= \sum_K C(K) \sum_v\ p_{A,K}(V = v)\ v \\ &= \sum_K C(K)\ V(A\&K) \end{aligned}$$

3.3.2 Decision problems

The two decision theories will sometimes disagree about which actions are preferable. Sometimes it will be EDT that advises we take the most sensible act, other times it will be CDT that returns sound advice. Let us analyse a couple of these situations in which there is discordant advice given by EDT and CDT.

3.3.2.1 The Smoking Lesion

Here is a situation in which it is CDT that recommends what we intuitively consider right.

The Smoking Lesion. "Susan is debating whether or not to smoke. She believes that smoking is strongly correlated with lung cancer, but only because there is a common cause—a condition that tends to cause both smoking and cancer. Once we fix the presence or absence of this condition, there is no additional correlation between smoking and cancer. Susan prefers smoking without cancer to not smoking without cancer, and she prefers smoking with cancer to not smoking with cancer."

In this example, Susan should smoke because it makes her better off no matter what. The choice of smoking may well indicate the she is likely to have a cancerous lesion, though it will not make her any worse off. Of course, the assumption that

there is no further correlation between smoking and cancer is at odds with our everyday knowledge the world. This is somewhat unfortunate, but we will stick with the example since it is widely adopted in the literature, including Briggs' paper.

To formalise the problem above, we will denote the dependency hypotheses by R and $\neg R$ for the hypotheses, in this order, that the cancer-inducing condition is present and that it is absent. Also, to label the actions we will use S and $\neg S$ in place of, respectively, smoking and not smoking. Picking suitable numbers for the agent's value function, we can then represent it with the following table:

$V(\cdot)$	R	$\neg R$
S	-90	10
$\neg S$	-100	0

And, picking other suitable numbers, we can represent the agent's credence function with another table:

$C(\cdot)$	R	$\neg R$
S	0.37	0.13
$\neg S$	0.13	0.37

We can now proceed to calculate the expected values of the actions according to each theory, starting from EDT.

$$\begin{aligned} V_E(S) &= C(V = -90|S)\ (-90) + C(V = 10|S)\ 10 \\ &= 0.74\ (-90) + 0.26 \cdot 10 \\ &= -64 \end{aligned}$$

$$\begin{aligned} V_E(\neg S) &= C(V = -100|\neg S)\ (-100) + C(V = 0|\neg S)\ 0 \\ &= 0.26\ (-100) + 0.74 \cdot 0 \\ &= -26 \end{aligned}$$

So $V_E(\neg S) = -26 > -64 = V_E(S)$. Smoking is seen as negative news because, all thing being equal, it informs you that are more likely to have a lesion that is also associated with cancer. EDT is led astray by this confounding variable, and it advises you to avoid the pleasurable experience of smoking despite the fact that, in this example, it does not have a causal impact on cancer. CDT, on the other hand, manages to steer clear of spurious correlations.

$$\begin{aligned} V_C(S) &= C(R)\ V(R\&S) + C(\neg R)\ (\neg R\&S) \\ &= 0.5\ (-90) + 0.5 \cdot 10 \\ &= -40 \end{aligned}$$

$$\begin{aligned} V_C(\neg S) &= C(R)\ V(R\&\neg S) + C(\neg R)\ (\neg R\&\neg S) \\ &= 0.5\ (-100) + 0.5 \cdot 0 \\ &= -50 \end{aligned}$$

So $V_C(S) = -40 > -50 = V_C(\neg S)$, showing that CDT, as its name suggests, reasons correctly about the causal relation between smoking and cancer in this example.

3.3.2.2 The Psychopath Button

In this other example, however, it is EDT that advises what is arguably the most sensible action.

The Psychopath Button. "Paul is debating whether to press the "kill all psychopaths" button. It would, he thinks, be much better to live in a world with no psychopaths. Unfortunately, Paul is quite confident that only a psychopath would press such a button. Paul very strongly prefers living in a world with psychopaths to dying."

First of all, it is clear that Paul should not press the button. Pressing the button tells him that he is most likely a psychopath and, therefore, that he will most likely die as a result of this action. As dying is his least preferred outcome, this is not a happy ending.

Next, we can formalise the problem and see what actions are prescribed by our decision theories. There are two dependency hypotheses that the agent should consider, which agree on everything except whether he is a psychopath or not; let us call them P and $\neg P$ respectively. There are also two actions available to this decision maker: pressing the button, labelled as B, and not pressing it, labelled as $\neg B$. We can then represent the value function with the following table:

$V(\cdot)$	P	$\neg P$
B	-100	20
$\neg B$	0	0

And we can represent the agent's credence function with another table:

$C(\cdot)$	P	$\neg P$
B	0.09	0.01
$\neg B$	0.01	0.89

Now, we just need to calculate the expected values according to each decision theory. Starting from EDT we have what follows.

$$\begin{aligned} V_E(B) &= C(V = -100|B)\ (-100) + C(V = 20|S)\ 20 \\ &= 0.9\ (-100) + 0.1 \cdot 20 \\ &= -88 \end{aligned}$$

$$\begin{aligned} V_E(\neg B) &= C(V = 0|\neg B) \cdot 0 \\ &= 1 \cdot 0 \\ &= 0 \end{aligned}$$

So $V_E(\neg B) = 0 > -88 = V_E(B)$, and EDT gives the answer we would hope for. But what about CDT?

$$\begin{aligned} V_C(B) &= C(P)\ V(P\&B) + C(\neg P)\ (\neg P\&B) \\ &= 0.1 \cdot (-100) + 0.9 \cdot 20 \\ &= 8 \end{aligned}$$

$$\begin{aligned} V_C(\neg B) &= C(P)\ V(P\&\neg B) + C(\neg P)\ (\neg P\&\neg B) \\ &= 0.1 \cdot 0 + 0.9 \cdot 0 \\ &= 0 \end{aligned}$$

So $V_C(B) = 8 > 0 = V_C(\neg B)$. CDT, unfortunately, fails to provide sensible advice in this case. The reason is that, unlike EDT, it does not consider the evidence that is gotten through the act of pushing the button; this can be seen clearly by comparing the definitions of the two theories in dependency hypothesis form.

3.4 Applying social choice theory

So far we have encountered puzzling situations in which a decision theory or another fails to provide sensible advice, without, however, having an understanding of why this is the case. Now, we will go through an unusual analysis of decision-theoretic paradoxes: we will see that these problems have the same structure as voting paradoxes, which means that, conveniently, results from social choice theory can be applied here.

3.4.1 Decision problems as elections

Briggs suggest that decision problems be recast as elections, in which our possible future selves are both the voters and candidates. A rationale for this can be found by

going back to those notions of identity that consider us comprised of multiple selves. Under this view, decision rules are ways of choosing an action for the well-being of the people we might become.

In the original construction, there is a possible future self S^A for every available action $A \in \mathcal{A}$. All of these future selves can be assumed to share our values, while their individual beliefs must be updated on the corresponding action that would have been taken. So S^A's value function is unchanged, i.e. $V^A = V$, and its credence function is derived by conditionalising on A which gives $C^A = C(\cdot|A)$. This leaves only one way for each S^A to cast its vote on an action B that we are considering, which is as follows:

$$V^A(B) = \sum_K C(K|A)\ V(B\&K)$$

Accordingly, both definitions of expected value can be reformulated in voting form:

$$V_E(A) = V^A(A)$$

$$V_C(A) = \sum_B C(B)\ V^B(A)$$

Decision theories rank available actions by their degree of choice-worthiness, given by expected value which may be of one kind or another, and the agent is advised to choose an action that is maximally choice-worthy. A decision rule is mathematically equivalent to a social-welfare functional, so it can be defined as a function from the set $\{V^A : A \in \mathcal{A}\}$ to an expected value function E over the members of $\mathcal{A}$.

3.4.1.1 Voting representation

We can easily prove the aforementioned voting form equivalences. For EDT the voting form can be read off directly from the original definition.

$$\begin{aligned} V_E(A) &= \sum_K C(K|A)\ V(A\&K) \\ &= V^A(A) \end{aligned}$$

Proving the equivalence for CDT involves rearranging some terms of a summation, but it should still be quite straightforward to see that this works.

$$\begin{aligned} V_C(A) &= \sum_K C(K)\ V(A\&K) \\ &= \sum_K \sum_B C(B\&K)\ V(A\&K) \\ &= \sum_K \sum_B C(B)\ C(K|B)\ V(A\&K) \\ &= \sum_B C(B) \sum_K C(K|B)\ V(A\&K) \\ &= \sum_B C(B)\ V^B(A) \end{aligned}$$

3.4.1.2 Non-uniqueness of each voter

There is a minor detail we glossed over when introducing the voting interpretation of decision rules. Recall that in Briggs' paper dependency hypotheses were assumed to be deterministic, whereas we removed that restriction. So an action may lead to a number of possible outcomes and, consequently, to as many future selves. This makes interpreting the term S^A tricky, because there is no longer a unique future self who corresponds to that voter.

One response might be to think of S^A as a sub-electorate. More specifically, a committee whose members are the possible selves who might come into existence were action A to be performed. Based on how the value function was defined, we are weighting each of their votes by the likelihood of the voter's existence upon A being performed. Note that all these voters will have slightly different credence functions given by their differing beliefs about objective probability, which are going to agree only on the world no longer being chancy; each member of the sub-electorate believes in a deterministic dependency hypothesis, because from their perspective they already know what the outcome would be if the present self were to perform act A.

Another take on this matter is that there is unique S^A but we do not know who it will be. Although we cannot be certain of this person's identity, we still know with some likelihood who it is going to be. So we might value actions as to minimise the expected disagreement with the judgements of the possible individuals who S^A could be. I strongly suspect, though have not verified, that this approach gives the same result as the previous suggestion.

Either way, concerns like these should not bother us too much. As stated at the outset of this thesis, we are employing the tools of social choice theory as a convenient modelling device; it is up to the reader how literally to take our analogies about multiple selves.

3.4.2 Desiderata

Briggs (2010) identifies two properties that are desirable for a decision rule to satisfy. One of them is a Pareto condition **P**, while the other is a self-sovereignty condition **S**. These conditions are usually written in different notation, as they come from the

social choice literature, but here we will state them as follows, in accordance with our voting representation:

P. For any actions A_1 and A_2 in $\mathcal{A}$, if for all $B \in \mathcal{A} : V^B(A_1) \geq V^B(A_2)$ and there is some $B \in \mathcal{A}$ such that $V^B(A_1) > V^B(A_2)$, then $E(A_1) > E(A_2)$.

S. For any decision situations $D = \langle \mathcal{A}, \mathcal{K}, C \rangle$ and $D' = \langle \mathcal{A}, \mathcal{K}', C' \rangle$ involving the same set of actions $\mathcal{A}$, if $\mathcal{A}$ contains a pair of actions A_1 and A_2 such that $V^{A_1} = V'^{A_1}$ and $V^{A_2} = V'^{A_2}$, then where $E(A)$ is the expected value assigned to A in D and $E'(A)$ is the expected value assigned to A in D', $E(A_1) > E(A_2)$ if and only if $E'(A_1) > E'(A_2)$.

P is a strong Pareto condition. It states that if you believe that you will value an act no less than any other, and you might value it more for all that you know, then you should perform this act; a very reasonable requirement. It is "strong" in the sense that it has a weaker antecedent than the Pareto condition yielding the same consequent but requiring strict dominance at every point. Our Pareto principle is, perhaps, best understood in relation to the dominance property it implies of a chosen action A_1 compared to any other:

Weak dominance. A_1 weakly dominates A_2 if, and only if, for all $K \in \mathcal{K} : V(A_1\&K) \geq V(A_2\&K)$ and there is some $K \in \mathcal{K}$ such that $C(K) > 0$ and $V(A_1\&K) > V(A_2\&K)$.

What the definition above is saying is that a chosen action A_1 dominates another A_2 when you believe that it will make you no worse off and, for all that you know, it might even make you strictly better off. In the smoking lesion problem, smoking dominates every other option so we intuitively prefer it for such a reason. This is where CDT gets its appeal from: CDT satisfies **P**, therefore it privileges dominant options, unlike EDT which prescribes dominated ones in cases like this.

S is the analogous of a self-sovereignty principle in voting theory. We can interpret it as saying that, when comparing the values of two actions A_1 and A_2, we should base our decision solely on what our desires and beliefs would be after taking those actions A_1 and A_2. This condition might sound a bit obscure at first, but consider how CDT fails to satisfy **S**.

In the psychopath button problem, CDT led us astray because it based its verdict on the prior likelihood of the dependency hypotheses, disregarding what evidence is given by the act under consideration; it ignored the information you would get by pushing the button, i.e. that you are a psychopath. Each dependency hypothesis stipulates the possible future selves who you might become. So, put in other words, CDT takes into account the beliefs and desires of all your possible future selves, even those who you are certain will never come into existence after your actions. It worries about the welfare of people who will never exist!

Of course, other hypotheses are irrelevant once you are certain of the true one. Accordingly, you should give less weight to the votes of successors who are unlikely

to ever exist, and in the limit case, you should disregard altogether the advice of those future selves who will never come to be. This is what **S** demands and precisely what EDT does. In fact, EDT gives the right answer in the psychopath button problem by conditionalising on what evidence is available and, therefore, listening only to the possible future selves whose opinions are relevant. Though, of course, it still fails to satisfy the Pareto condition **P** which may be required elsewhere.

3.4.3 Proofs

Below are proofs of the claims made during the last section. Our proof that CDT satisfies **P** was not present in Briggs' paper, while the rest are structured differently but based on the same ideas. Nothing important will be lost to the reader if they skim through this part, or even skip it entirely.

3.4.3.1 The Pareto condition privileges dominant actions

We want to prove that weak dominance of an action A_1 over another A_2, taken with the condition **P**, implies $E(A_1) > E(A_2)$. First, we show that for all $K \in \mathcal{K} : V(A_1\&K) \geq V(A_2\&K)$ implies that for all $B \in \mathcal{A} : V^B(A_1) \geq V^B(A_2)$. Second, we prove that if there is some $K \in \mathcal{K}$ such that $C(K) > 0$ and $V(A_1\&K) > V(A_2\&K)$, then there is some $B \in \mathcal{A}$ such that $V^B(A_1) > V^B(A_2)$. Together with **P**, these two conditions give us $E(A_1) > E(A_2)$.

$$\begin{aligned} V^B(A_1) - V^B(A_2) &= \sum_K C^B(K)V(K\&A_1) - \sum_K C^B(K)V(K\&A_2) \\ &= \sum_K C^B(K)\left[V(K\&A_1) - V(K\&A_2)\right] \\ &\geq \sum_K C^B(K)\left[V(K\&A_1) - V(K\&A_1)\right] \\ &= 0 \end{aligned}$$

So $V^B(A_1) - V^B(A_2) \geq 0$ and, therefore, $V^B(A_1) \geq V^B(A_2)$ for all $B \in \mathcal{A}$.

By weak dominance, there is also some $K \in \mathcal{K}$ with $C(K) > 0$ such that $V(A_1\&K) > V(A_2\&K)$. For the same K there must be some $B \in \mathcal{A}$ with the property that $C^B(K) > 0$, which can be seen as follows. As $\sum_{A\in\mathcal{A}} C(A\&K) = C(K) > 0$, there is some B in $\mathcal{A}$ such that $C(K\&B) > 0$. But since $C(B) \geq C(K\&B)$, then also $C(B) > 0$. So the ratio $C(K\&B)/C(B)$ must be positive, and from that we have $C(K|B) > 0$. Given we assumed that the agent updates its beliefs by conditionalising on the evidence it observes, this last inequality coincides with $C^B(K) > 0$. So for this particular B it holds that $C^B(K)\ V(K\&A_1) > C^B(K)\ V(K\&A_2)$ and, thus, $V^B(A_1) > V^B(A_2)$.

We have proved that for all $B \in \mathcal{A} : V^B(A_1) \geq V^B(A_2)$ and that there is some $B \in \mathcal{A}$ such that $V^B(A_1) > V^B(A_2)$. This means the antecedent of **P** is met, completing our proof that $E(A_1) > E(A_2)$.

3.4.3.2 CDT satisfies P

Take a pair of actions A_1 and A_2 in $\mathcal{A}$, and assume that for all $B \in \mathcal{A} : V^B(A_1) \geq V^B(A_2)$. Furthermore, assume there is some $B \in \mathcal{A}$ such that $V^B(A_1) > V^B(A_2)$; denote this particular action as B'. Then we can show what follows:

$$\begin{aligned}
V_C(A_1) - V_C(A_2) &= \sum_B C(B)\ V^B(A_1) - \sum_B C(B)\ V^B(A_2) \\
&= \sum_B C(B) \left[V^B(A_1) - V^B(A_2)\right] \\
&= C(B') \left[V^{B'}(A_1) - V^{B'}(A_2)\right] + \sum_{B \in \mathcal{A} \setminus B'} C(B) \left[V^B(A_1) - V^B(A_2)\right] \\
&\geq C(B') \left[V^{B'}(A_1) - V^{B'}(A_2)\right] + \sum_{B \in \mathcal{A} \setminus B'} C(B) \left[V^B(A_1) - V^B(A_1)\right] \\
&= C(B') \left[V^{B'}(A_1) - V^{B'}(A_2)\right] \\
&> 0
\end{aligned}$$

So $V_C(A_1) - V_C(A_2) > 0$ and, therefore, $V_C(A_1) > V_C(A_2)$.

3.4.3.3 EDT satisfies S

Consider two decision situations $D = \langle \mathcal{A}, \mathcal{K}, C \rangle$ and $D' = \langle \mathcal{A}, \mathcal{K}', C' \rangle$, with $\mathcal{A}$ containing a pair of actions A_1 and A_2 such that $V^{A_1} = V'^{A_1}$ and $V^{A_2} = V'^{A_2}$. It holds that $V_E(A_1) = V_{A_1}(A_1) = V'_{A_1}(A_1) = V'_E(A_1)$ and, similarly, $V_E(A_2) = V_{A_2}(A_2) = V'_{A_2}(A_2) = V'_E(A_2)$. So, substituting $V'_E(A_1)$ and $V'_E(A_2)$ for $V_E(A_1)$ and $V_E(A_2)$ respectively, we have that $V_E(A_1) > V_E(A_2)$ if and only if $V'_E(A_1) > V'_E(A_2)$.

3.4.4 A decision-theoretic impossibility theorem

CDT fulfils the requirements of the Pareto property **P**, though fails to satisfy **S**. The converse holds for EDT: it satisfies **S** but not **P**. Could there a third alternative that meets both these conditions? Surprisingly, the answer is no; we are about to see that no decision theory can do such a thing.

3.4.4.1 The counter-example

To show that decision theories cannot satisfy **P** and **S** at the same time, we construct a pair of situations in which no theory can satisfy **P** for each while satisfying **S** for both.

Consider the decision situations $D = \langle \mathcal{A}, \mathcal{K}, C \rangle$ and $D' = \langle \mathcal{A}, \mathcal{K}', C' \rangle$, with $\mathcal{A}$ containing a pair of actions A_1 and A_2 such that $V^{A_1} = V'^{A_1} = V^{A_2} = V'^{A_2}$. Let $\mathcal{F} = \mathcal{A} \setminus \{A_1, A_2\}$ be non-empty, i.e. $\mathcal{F} \neq \varnothing$. Also, assume that $\forall F \in \mathcal{F} : V^F(A_1) > V^F(A_2)$ and $V'^F(A_1) < V'^F(A_2)$.

First off, we have that $\forall B \in \mathcal{A} : V^B(A_1) \geq V^B(A_2)$; the condition holds as an equality for the actions in A_1 and A_2, and as an inequality for the remaining actions in $\mathcal{F}$. Those actions in $\mathcal{F}$, of which there is at least one, satisfy additionally

$V^B(A_1) > V^B(A_2)$. So the preconditions of **P** are met, which means to satisfy this Pareto condition it must hold that $E(A_1) > E(A_2)$.

Conversely, we have that $\forall B \in \mathcal{A} : V'^B(A_1) \leq V'^B(A_1)$; the condition holds as an equality for the actions in A_1 and A_2, and as an inequality for the remaining actions in $\mathcal{F}$. Those actions in $\mathcal{F}$, of which there is still at least one, satisfy additionally $V'^B(A_1) < V'^B(A_2)$. So the preconditions of **P** are once again met, and this time, to satisfy the Pareto condition, it must hold that $E'(A_1) < E'(A_2)$.

Since $V^{A_1} = V'^{A_1}$ and $V^{A_2} = V'^{A_2}$, however, **S** requires that $E(A_1) > E(A_2)$ if and only if $E(A_1) > E(A_2)$. But this violates the above conditions, so there cannot exist a decision theory that satisfies both **P** and **S** here.

3.4.4.2 Remarks and objections

We have finished proving an impossibility theorem for decision theories, this time in a more general setting than considered by the original paper: our dependency hypotheses were assumed to be stochastic, rather than merely deterministic.

The implications of this impossibility result, in either form, are quite significant: if we are right about no decision rule being able to do everything we want, then a number of long-standing paradoxes have been dispelled and the ongoing philosophical debates they have given rise to can be ceased. Still, these debates are ongoing and there are reasonable objections that can be raised to our result. Briggs (2010), themself, suggests that EDT and CDT will agree "in all but a handful of unlikely situations," making the impossibility theorem generally inapplicable. In any case, I wish to conclude by reflecting on a couple of other issues that come to mind.

When interpreting decision rules as voting procedures, we asserted that our various future selves could be polled about what value they assign to the set of available actions $\mathcal{A}$. These values were calculated causally; that is to say, they were derived from the equations of CDT. There are technical reasons for this, as actions in $\mathcal{A}$ form a partition, so they are pair-wise incompatible and, without additional assumptions, we cannot conditionalise a credence function on two of them at once; this last step would be required to calculate the evidential value our future selves assign to actions (Briggs 2010). Setting aside the difficulties of using EDT, the choice of CDT seems arbitrary and we do not have conclusive argument to offer as to why it is the right one. This, admittedly, is a glaring omission that needs to be addressed.

Also, the conditions **P** and **S** are purely ordinal. These principles restrict in what ways we are allowed aggregate the opinions of our future selves to derive an expected value function E. But the kind of constraints they impose do not force anything beyond a relative scale for E, and most importantly, nor do they care about what values an agent might assign to these actions on some absolute scale. If the voters were to have cardinal preferences, however, our impossibility theorem might be avoided, as is the case for so many other such results (List 2013). Regardless, this is a more general criticism, directed not at our specific result but at most works in the literature on social choice theory.

Chapter 4

Conclusion

This thesis looked at two ways in which social choice or game theory can be applied to guide the decisions of an individual agent, and in doing so, it tackled a couple of open questions: "How should an agent select between multiple subgame perfect equilibrium policies?" and "Does the decision-theoretic impossibility theorem hold for stochastic environments?".

We started by looking at reinforcement learning. In this setting, there is an agent who makes sequential decisions with the aim of maximising some cumulative reward. The agent may be subject to changes of preference; that is to say, at the time of a later decision it might value some act over another it preferred during an earlier decision. Policies that neglect to account for such changes can lead to inconsistent behaviour which, ultimately, results in the agent receiving lower rewards than it otherwise would have; inconsistent behaviour being intended, here, as the pursuit of conflicting goals at different times.

The kind of scenario just described is not hypothetical. The usual formulation of reinforcement learning contains several assumptions about agents' preferences: it assumes that preferences can be captured by a utility function together with a discount function; furthermore, the discount function might not be geometric. This last assumption, in particular, is problematic, so we introduced the model of general time-consistent discounting. Still, a question remains: how can an agent who knows it is time-inconsistent make decisions that maximise its overall, discounted reward?

This question is more general than reinforcement learning and has been considered elsewhere, especially in the economics literature; we argued as much by showing that preference relations, and utility and choice functions are two sides of the same coin. One answer, which appears in existing papers, is for an agent to treat its possible future selves as adversaries in an extensive game with perfect information. The solution concept for these games is the subgame perfect equilibrium. However, it has the severe shortcoming that the policies it gives rise to may not be unique, with no obvious choice of an equilibrium point for an agent's various temporal selves to coordinate on.

In all honesty, I must admit that I set out to write my thesis with the intention of solving the problem of equilibrium selection. Of course it sounds foolish now, with the benefit of hindsight, as this open problem lies at the very foundations of game theory and is much harder than it first appeared. So hard, in fact, that John Harsanyi himself,

a Nobel Prize winner and one of the greatest game theorists to have ever lived, only came up with a of couple suggestion for selection criteria. If such a distinguished man failed to provide an accepted solution to this problem, then I feel excused for not doing so in my thesis. Also, scientific research does not always give positive results; it often returns negative findings that are valuable nonetheless (however unappreciated they may be in the world of academic publishing, leading to issues of publication bias).

In the second chapter, we considered social choice theory. Our investigation into extending an impossibility theorem for decision rules was, arguably, more successful: we were able to show that the result holds in probabilistic environments, as well as deterministic ones. This proved to be a fruitful area of enquiry, which suggests it is worth looking, more generally, into decision-theoretic interpretations of impossibility theorems. There are a number of such results scattered around the economics literature; it is plausible that at least some of them have meaningful translations to a single agent setting.

On conclusion of these reflections, we have come a long way. I hope you, the reader, have been entertained.

Bibliography

Arntzenius, F., Elga, A. & Hawthorne, J. (2004), 'Bayesianism, Infinite Decisions, and Binding', *Mind* **113**(450), 251–283.

Arrow, K. J. (1951), *Social Choice and Individual Values*, Wiley.

Aumann, R. J. (1974), 'Subjectivity and Correlation in Randomized Strategies', *Journal of Mathematical Economics* **1**(1), 67–96.

Bleichrodt, H., Rohde, K. I. & Wakker, P. P. (2008), 'Koopmans constant discounting for intertemporal choice: A simplification and a generalization', *Journal of Mathematical Psychology* **52**(6), 341–347.

Briggs, R. (2009), 'Distorted Reflection', *Philosophical Review* **118**(1), 59–85.

Briggs, R. (2010), 'Decision-Theoretic Paradoxes as Voting Paradoxes', *Philosophical Review* **119**(1), 1–30.

Briggs, R. (2014), Normative Theories of Rational Choice: Expected Utility, *in* E. N. Zalta, ed., 'The Stanford Encyclopedia of Philosophy', fall 2014 edn.

Chao, L.-W., Szrek, H., Pereira, N. S. & Pauly, M. V. (2009), 'Time Preference and Its Relationship with Age, Health, and Survival Probability', *Judgment and Decision Making* **4**(1), 1.

Elster, J. (1987), *The Multiple Self*, Cambridge University Press.

Everitt, T., Leike, J. & Hutter, M. (2015), Universal Sequential Decision Theory: Causal and Evidential. Manuscript.

Fishburn, P. C. & Rubinstein, A. (1982), 'Time Preference', *International Economic Review* **23**(3), 677–694.

Frederick, S., Loewenstein, G. & O'Donoghue, T. (2002), 'Time Discounting and Time Preference: A Critical Review', *Journal of Economic Literature* **40**(2), 351–401.

Gaertner, W. C. (2009), *A Primer in Social Choice Theory*, revised edn, Oxford University Press.

Gibbard, A. & Harper, W. L. (1981), Counterfactuals and Two Kinds of Expected Utility, *in* W. L. Harper, R. Stalnaker & G. A. Pearce, eds, 'Ifs: Conditionals, Belief, Decision, Chance, and Time', Vol. 15 of *The University of Western Ontario Series in Philosophy of Science*, Springer Netherlands, pp. 153–190.

Green, L., Fry, A. F. & Myerson, J. (1994), 'Discounting of Delayed Rewards: A Life-Span Comparison', *Psychological Science* **5**(1), 33–36.

Hammond, P. J. (1976), 'Changing Tastes and Coherent Dynamic Choice', *The Review of Economic Studies* **43**(1), 159–173.

Harrison, G. W., Lau, M. I. & Williams, M. B. (2002), 'Estimating Individual Discount Rates in Denmark: A Field Experiment', *The American Economic Review* **92**(5), 1606–1617.

Harsanyi, J. C. (1973), 'Games with Randomly Disturbed Payoffs: A New Rationale for Mixed-Strategy Equilibrium Points', *International Journal of Game Theory* **2**(1), 1–23.

Harsanyi, J. C. (1995), 'A New Theory of Equilibrium Selection for Games with Complete Information', *Games and Economic Behavior* **8**(1), 91–122.

Harsanyi, J. C. & Selten, R. (1988), *A General Theory of Equilibrium Selection in Games*, MIT Press.

Hume, D. (1739), *A Treatise of Human Nature*, 2000 reprint edn, Oxford University Press.

Hutter, M. (2005), *Universal Artificial Intellegence: Sequential Decisions Based on Algorithmic Probability*, Springer.

Jamison, J. & Wegener, J. (2010), 'Multiple Selves in Intertemporal Choice', *Journal of Economic Psychology* **31**(5), 832–839.

Kahneman, D. (2011), *Thinking, Fast and Slow*, Farrar, Straus and Giroux.

Koopmans, T. C. (1960), 'Stationary Ordinal Utility and Impatience', *Econometrica* **28**(2), 287–309.

Lancaster, K. (1963), 'An Axiomatic Theory of Consumer Time Preference', *International Economic Review* **4**(2), 221–231.

Lattimore, T. & Hutter, M. (2014), 'General Time Consistent Discounting', *Theoretical Computer Science* **519**, 140–154.

Legg, S. (2008), Machine Super Intelligence, PhD thesis, University of Lugano.

Legg, S. & Hutter, M. (2007), 'Universal Intelligence: A Definition of Machine Intelligence', *Minds and Machines* **17**(4), 391–444.

Lewis, D. (1981), 'Causal Decision Theory', *Australasian Journal of Philosophy* **59**(1), 5–30.

List, C. (2013), Social Choice Theory, *in* E. N. Zalta, ed., 'The Stanford Encyclopedia of Philosophy', winter 2013 edn.

Locke, J. (1689), *An Essay Concerning Human Understanding*, 1995 reprint edn, Prometheus Books.

Macaskill, W. (2014), Normative Uncertainty, PhD thesis, University of Oxford.

Mas-Colell, A., Whinston, M. D. & Green, J. R. (1995), *Microeconomic Theory*, Oxford University Press.

Matsui, A. & Matsuyama, K. (1995), 'An Approach to Equilibrium Selection', *Journal of Economic Theory* **65**(2), 415–434.

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

McKelvey, R. D. & Palfrey, T. R. (1992), 'An Experimental Study of the Centipede Game', *Econometrica: Journal of the Econometric Society* pp. 803–836.

Morgenstern, O. & von Neumann, J. (1944), *Theory of Games and Economic Behavior*, 2007 reprint edn, Princeton University Press.

Moss, S. (2012), 'Updating as Communication', *Philosophy and Phenomenological Research* **85**(2), 225–248.

Nisan, N., Roughgarden, T., Tardos, E. & Vazirani, V. V. (2007), *Algorithmic Game Theory*, Cambridge University Press.

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

Parfit, D. (1971), 'Personal Identity', *The Philosophical Review* **80**(1), 3–27.

Parfit, D. (1984), *Reasons and Persons*, Oxford University Press.

Pronin, E., Olivola, C. Y. & Kennedy, K. A. (2008), 'Doing Unto Future Selves As You Would Do Unto Others: Psychological Distance and Decision Making', *Personality and Social Psychology Bulletin* **34**(2), 224–236.

Ramsey, F. P. (1926), Truth and Probability, *in* A. Eagle, ed., 'Philosophy of Probability: Contemporary Readings', 2011 reprint edn, Routledge.

Read, D. & Read, N. (2004), 'Time discounting over the lifespan', *Organizational Behavior and Human Decision Processes* **94**(1), 22–32.

Rosenthal, R. W. (1981), 'Games of Perfect Information, Predatory Pricing and the Chain-Store Paradox', *Journal of Economic Theory* **25**(1), 92–100.

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

Russell, S. J. & Norvig, P. (2009), *Artificial Intelligence: A Modern Approach*, third edn, Prentice Hall.

Samuelson, P. A. (1937), 'A Note on Measurement of Utility', *The Review of Economic Studies* **4**(2), 155–161.

Schelling, T. C. (1960), *The Strategy of Conflict*, Harvard University Press.

Sozou, P. D. & Seymour, R. M. (2003), 'Augmented discounting: interaction between ageing and time–preference behaviour', *Proceedings of the Royal Society of London B: Biological Sciences* **270**(1519), 1047–1053.

Strotz, R. H. (1955), 'Myopia and Inconsistency in Dynamic Utility Maximization', *The Review of Economic Studies* **23**(3), 165–180.

Sutton, R. S. & Barto, A. G. (1998), *Introduction to Reinforcement Learning*, MIT Press.

van Fraassen, B. C. (1984), 'Belief and the Will', *The Journal of Philosophy* **81**(5), 235–256.