TÉCNICO LISBOA

The Fall of *Homo Economicus*

The Role of Cognitive Biases and Theory of Mind in Human Coordination

Pedro Alexandre Caetano Lopes Ferreira

Thesis to obtain the Master of Science Degree in

Mathematics and Applications

Supervisors: Prof. Francisco João Duarte Cordeiro Correia dos Santos
Prof. Conceição Amado

Examination Committee

Chairperson: Prof. António Manuel Pacheco Pires
Supervisor: Prof. Francisco João Duarte Cordeiro Correia dos Santos
Member of the Committee: Prof. José Alberto Rodrigues Pereira Sardinha

October 2019

Acknowledgments

I would like to thank my advisor Prof. Francisco Santos and unofficial advisor Prof. Sérgio Pequito for teaching me the essence of *research* and guiding me when I felt most lost. I would also like to give a warm thanks to Matilde Farinha, for her valuable discussions and emotional support throughout the development of this work. Without you, my work would not be possible.

My family and friends – who have provided me with the means and motivation to finish this chapter of my life – you have made not only my work, but also my academic life entirely worth it. I have had the best days of my life with you.

I also would like to thank the Dean of the School of Engineering Shekhar Garde, Head of Department William (Al) Wallace and (again) Prof. Sérgio Pequito for welcoming me into the United States and providing valuable time, advice, patience and living expenses to conduct most of this research at the Industrial and Systems Engineering (ISE) department of the Rensselaer Polytechnic Institute.

Last, but not least, I would also like to thank the researchers at RPI that dedicated their valuable time to attend our seminars, discuss insightful ideas and provide their unique perspective on this work and how it relates to their fields.

Abstract

The cognitive mechanics of human decision-making is affected by a large set of high-level processes. Some of these processes, called cognitive biases, are often regarded as failures, since they prescribe behavior which is not deemed as rational. Furthermore, in social settings, humans employ a process known as theory of mind which enables them to create and manage a dynamic model of the mental states of others, allowing for the prediction of future actions to better inform current behavior. Can cognitive biases promote coordination? Can increasingly sophisticated levels of theory of mind promote coordination? In this thesis, we answer these questions by showing how coordination among agents measuring value using the prescriptive Expected Utility Theory (EUT) differs from the coordination among agents measuring value using the descriptive Cumulative Prospect Theory (CPT), in two experimental settings: a normal-form stag hunt game allows us to study how coordination differs when agents use EUT and CPT as theories of value, while a Markov game of stag hunt focuses on studying the effects of increasingly sophisticated policies among both EUT- and CPT-agents, using the recursive theory of mind level-k model that captures bounded rationality. We show that CPT-agents are better able to coordinate in both experiments, compared to EUT-agents. Furthermore, in the Markov stag hunt, while coordination with both EUT and CPT stand to gain from increasingly sophisticated policies, CPT-agents do not require as much sophistication as EUT-agents do to coordinate to the same extent. We can thus conclude that, while some of these cognitive biases are viewed as failures in individual decision-making, they actually make social interaction easier.

Keywords

Coordination; Cognitive Bias; Theory of Mind; Expected Utility Theory; Cumulative Prospect Theory; Level-k; Bounded Rationality.

Resumo

Os mecanismos cognitivos do processo de decisão humano é afetada por uma grande quantidade de processos de alto-nível. Alguns destes processos, chamados tendências cognitivas, são muitas vezes vistos como falhas, visto que levam a comportamentos que não é considerado racional. Para além disso, em cenários sociais, os humanos usam um processo conhecido como teoria da mente que os permite criar e gerir um modelo dinâmico de estados mentais de outros, permitindo a previsão de ações futuras para melhor informar comportamentos presentes. Podem as tendencias cognitivas ajudar na coordenação? Podem níveis mais sofisticados de teoria da mente ajudar na coordinação? Nesta tese, respondemos a estas perguntas mostrando como a coordinação entre agentes que medem os valores das suas ações usando a prescritiva teoria da utilidade esperada (TUE) difere da coordenação entre agentes que medem os valores das suas ações usando a descritiva teoria da perspetiva cumulativa (TPC), em dois cenários experimentais: um jogo de caça ao veado em forma normal permite-nos estudar as diferenças na coordenação quando os agentes usam a TUE e a TPC, enquanto que um jogo de Markov da caça ao veado se foca em estudar os efeitos de políticas cada vez mais sofisticadas entre agentes que usam ambas teorias de valor, usando o modelo de nível-k como uma teoria da mente recursiva que captura efeitos de racionalidade limitada. Nesta tese, demonstramos que os agentes-TPC são melhores a coordenar em ambas as experiências, comparados com os agentes-TUE. Para além disso, no jogo da caça ao veado de Markov, embora ambas as teorias de valor mostrem aumentos na coordenação com o aumento da sofisticação da política, os agentes-TPC não necessitam de tanta sofisticação quanto os agentes-TUE para coordenar na mesma medida. Podemos então concluir que, embora algumas destas tendências cognitivas são vistas como falhas em processos de decisão

individuais, elas facilitam a interação social entre humanos.

Palavras Chave

Coordenação; Tendência Cognitiva; Teoria da Mente; Teoria da Utilidade Esperada; Teoria da Perspetiva Cumulativa; Nível-k; Racionalidade Limitada.

Contents

1 Introduction **1**

1.1 Cognitive Mechanics . . . 2

1.1.1 Framing and Risk Attitude . . . 3

1.1.2 Theory of Mind . . . 3

1.2 Problem Statement and Research Questions . . . 4

1.3 Main Contributions . . . 5

1.4 Outline of the Thesis . . . 5

2 Background Theory **7**

2.1 Theories of Value . . . 8

2.1.1 Expected Utility Theory . . . 9

2.1.2 Prospect Theory . . . 11

2.1.3 Cumulative Prospect Theory . . . 13

2.2 Game Theory . . . 14

2.2.1 Normal-Form Game . . . 15

2.2.2 The Stag Hunt Game . . . 17

2.2.3 Markov Game . . . 17

2.2.3.A Markov Chain . . . 18

2.2.3.B Markov Decision Process . . . 19

2.2.3.C Markov Game . . . 21

2.3 Level-k Bounded Rationality - A Theory of Mind Model . . . 22

3 Related Work **25**

3.1 Risky Decisions . . . 26

3.2 Risky Games . . . 27

3.3 Intuitive Psychology . . . 28

3.4 Summary of Related Work . . . 29

4 Model **31**
4.1 Normal-form Stag Hunt . . . 32
4.2 Markov Game with CPT-Value . . . 33
4.2.1 Markov Stag Hunt . . . 34
5 Results **37**
5.1 Normal-form Stag Hunt . . . 38
5.2 Markov Stag Hunt . . . 38
5.3 Limitations . . . 42
6 Conclusion **45**
6.1 Summary . . . 46
6.2 Suggestions for Future Work . . . 47
6.3 Implications . . . 48
A Analysis of Game Theory of Mind **57**
A.1 Description . . . 57
A.2 Sequential versus Simultaneous Model . . . 59
B Dynamic Programming **63**
B.1 Backwards Induction in the Centipede Game . . . 63
B.2 Dynamic Programming in Markov Decision Processes . . . 64

List of Figures

2.1 Common CPT utility function, $u(x) = x^{0.85}$, for $x \geq 0$, and $u(x) = 2|x|^{0.85}$, for $x < 0$, with reference point at 0. Both negative and positive parts are sublinear, demonstrating risk-sensitivity, and the negative part is steeper, demonstrating loss aversion by weighting losses more heavily than gains. 12

2.2 Common CPT probability weighting function, $w(p) = \exp\{-\delta(-\log(p))^{\gamma}\}$, with $\delta = 0.5$ and $\gamma = 0.9$. Low probabilities are overweighted significantly. CPT transforms cumulative probabilities, effectively overweighting rare and extreme events. 12

2.3 A (simplified) map of games and stochastic models. There is a myriad of games studied under game theory, with different assumptions on the information structure that specify a game. The Markov game is at the intersect between game theory and stochastic processes, a generalization of the dynamic game, by allowing stochastic transitions between states, and of Markov decision process, by extending it to the multi-agent case. 14

2.4 The transition diagram of the Markov chain modelling the discrete line with 16 states. . . . 19

2.5 The transition matrix of the Markov chain modelling the discrete line with 16 states. 20

2.6 Stationary distribution of the drunkard's random walk. It is unsurprisingly skewed to the left, since there is a higher probability of going to the left than staying or going to the right. 20

2.7 The transition diagram of the Markov chain resulting from choosing action L, in the Markov decision process example. 21

2.8 The transition diagram of the Markov chain resulting from choosing action S, in the Markov decision process example. 21

2.9 The transition diagram of the Markov chain resulting from choosing action R, in the Markov decision process example. 21

2.10 The optimal value for the MDP example, with discount factor $\beta = 0.9$. 21

2.11 The optimal policy for the MDP example, with discount factor $\beta = 0.9$. 21

4.1 The transition matrices of the Markov game, for the three available actions L, S, and R, where L stands for going left, S stands for staying in the same state and R stands for going right. An agent can move to adjacent states or remain by choosing one of these three available actions, but the outcome is not deterministic, i.e. there is a chance of failure. For instance, choosing L does not guarantee that the agent moves to the left and may end up staying in the same place or going the opposite direction. 35

4.2 The reward functions of both agents. They are symmetrical like the payoff matrix of the stag hunt. State 3 gives reward of $+1$ to an agent in it regardless of the state of the other agent, to represent the hares that can be captured alone but offer little reward. State 11 gives reward of $+5$ to the two agents if they are both in state 11. All other states give a reward of 0. 35

5.1 Resulting EUT and CPT policies of agent 1 as functions of the agent states, for sophistication levels $k = 1, 2, 3, 4$. Due to the symmetry of the game, the policies for agent 2 are the transpose of these policies. 39

5.2 EUT- and CPT-values as functions of the agent states, s_1 and s_2, for sophistication levels $k = 1, 2, 3, 4$. We assumed reference points $b_1 = b_2 = 0$, discount factors $\beta_1 = \beta_2 = 0.9$, utility function $u(x) = x$ and weighting function $w(x) = x$ for EUT and $w(x) = e^{-0.5(-\log(x))^{0.9}}$ for CPT 40

5.3 Stationary distributions of the resulting Markov chains obtained by conditioning the Markov game to increasingly sophisticated policies, $k = 1, 2, 3, 4$, for EUT- and CPT-agents. We assumed reference points $b_1 = b_2 = 0$, discount factors $\beta_1 = \beta_2 = 0.9$, utility function $u(x) = x$ and weighting function $w(x) = x$ for EUT and $w(x) = e^{-0.5(-\log(x))^{0.9}}$ for CPT. 40

5.4 Stationary distributions of the resulting Markov chains obtained by conditioning the Markov game to increasingly sophisticated policies, $k = 1, 2, 3$ and 4, for EUT-agents and CPT-agents. We assumed reference points $b_1 = b_2 = 0$, utility function $u(x) = x$ and weighting function $w(x) = x$ for EUT and $w(x) = e^{-0.5(-\log(x))^{0.9}}$ for CPT. (Left) Stationary distribution for EUT- and CPT-agents using discount factor $\beta = 0.85$. (Right) Stationary distribution for EUT- and CPT-agents using discount factor $\beta = 0.85$. 41

5.5 Stationary distributions of the resulting Markov chains obtained by conditioning the Markov game to increasingly sophisticated policies, $k = 1, 2, 3$ and 4, for CPT-agents with several reference points $b = -1, 0, 1, 2$. We assumed discount factors $\beta_1 = \beta_2 = 0.9$, utility function $u(x) = x$ and weighting function $w(x) = e^{-0.5(-\log(x))^{0.9}}$. 42

A.1 An illustration of the QRE (black line) in the Stag Hunt game in Table 2.2, as a function of λ. Calculating QREs often involves solving a transcendental equation, which, in this case, we bypass by plotting a surrogate function and observing its zeros. 58
A.2 Results of the simultaneous level-k in [1]. (Left) Value function of agent 1. (Middle) Value function of agent 2. (Right) Stationary distribution of agents. Each row corresponds to the results of a particular level $k = 1, 2, 3, 4$. 61
A.3 Results of the proposed sequential level-k. (Left) Value function of agent 1. (Middle) Value function of agent 2. (Right) Stationary distribution of agents. Each row corresponds to the results of a particular level $k = 1, 2, 3, 4$. 62

B.1 The tree diagram of the centipede game, an example of an extensive-form game. The name of the game stems from the first appearance where $n = 100$ steps. 64

List of Tables

2.1 The payoff matrix of the prisoner's dilemma. The rows are actions decided by agent 1 and columns are actions decided by agent 2, with C stands for confess and S stands for remaining silent. The reward of both agents is written in the cell of the corresponding joint action. The number on the left is the reward of agent 1 and the one on the right is the reward of agent 2. 17

2.2 The payoff matrix of the stag hunt. The rows are actions decided by agent 1 and columns are actions decided by agent 2, where S stands for hunting stag and H stands for hunting hare. The reward of both agents is written in the cell of the corresponding joint action. The number on the left is the reward of agent 1 and the one on the right is the reward of agent 2. 17

Acronyms

EUT	Expected Utility Theory
PT	Prospect Theory
CPT	Cumulative Prospect Theory
FSD	First-order Stochastic Dominance
SSD	Statewise Stochastic Dominance
DNE	Deterministic Nash Equilibrium
SNE	Stochastic Nash Equilibrium
MDP	Markov Decision Process
LMDP	Linearly-solvable Markov Decision Process
QRE	Quantal Response Equilibrium
MAS	Multi-Agent System

1

Introduction

Contents

1.1 Cognitive Mechanics 2
1.2 Problem Statement and Research Questions 4
1.3 Main Contributions 5
1.4 Outline of the Thesis 5

"Selection shapes brains that maximize the number of offspring who survive to reproduce themselves. This is very different from maximizing health or longevity. It is also different from maximizing matings. That is why organisms do things other than having sex. Especially humans. Having the most offspring requires allocating plenty of thought and action to getting resources other than mates and matings, especially social resources, such as friends and status. Everyone else is doing the same thing, creating constant conflict, cooperation, and vast social complexity whose comprehension requires a huge brain."

Randolph M. Nesse

Even with our inherently human quirks, we are the only species on this planet to have dominated it, for better or worse. This would, of course, not be possible without our remarkable ability to cooperate and coordinate our efforts towards a common goal. When studying coordination, however, researchers often assume that agents, the mathematical representations of humans, are rational. This is likely due to the parsimony of the resulting models and the ease with which conclusions can then be drawn from them. The commonality of this assumption in scientific fields such as game theory and economics eventually coined the term *homo economicus*, a play on the taxonomic name of our species to capture the essence of the purely rational agent.

Homo economicus, much like Santa Claus or free lunches, does not exist. This thesis is thus motivated by creating a better model for human decision-making, based on cognitive mechanisms of humans, in order to gain a better understanding of human coordination.

1.1 Cognitive Mechanics

The systematic deviations from rational behavior are called **cognitive biases**. It has been argued that cognitive biases are useful for understanding human decision-making in inherently human domains such as finance [2]. This is also our argument, that understanding the apparent failures of humans due to cognitive biases may allow us better understand why we behave the way that we do, specifically in social settings.

In addition, humans employ a theory of mind to predict the behavior of others by reasoning about "what I think that you think that I think that..." and so on. The notion of bounded rationality comes into play, because this reasoning is not performed *ad infinitum* by humans. *Homo economicus* would indeed be able to do this, being an all rational being but, when in scenarios where others are not rational, it may be sub-optimal and therefore irrational to have a "rational" (i.e. reasoning about "what I think that you think that I think that..." *ad infinitum*) theory of mind. It is therefore worthwhile to study the effects of increasingly sophisticated theory of mind on the coordination of agents.

1.1.1 Framing and Risk Attitude

Take, for example, the **framing effect**: a cognitive bias that describes how people's decisions is based on the perceptual appearance of the semantics of outcomes [3]. Humans tend to frame outcomes into subjective gains and losses and this framing depends on several factors. Outcomes are then evaluated and a decision is made. The evaluation of outcomes is an important part of decision-making, which we will discuss further in the next chapter. The perception of risk is an inherently animal characteristic, not only of humans. **Risk** is defined as the possibility of losing something of value. Together with the framing effect, losing is a subjective concept. It depends on your own wealth, the relative value of outcomes, and many other factors. Risk attitude is the way humans attempt to lower uncertainty when exposed to it. To illustrate this, consider the two options: a certain outcome of receiving a gift of 1M€ and a gamble in which 2M€ and 0€ are received with equal probability. The expected outcome of the gamble is equal to the value of the certain outcome. A person is said to be:

- **Risk-averse** if they would prefer the gift, even if the gift was slightly lower than the 1M€,
- **Risk-neutral** if the gift and the gamble are equally preferred,
- **Risk-seeking** if they would prefer the gamble, even if the gift was slightly higher than the 1M€.

Humans tend to be risk-averse for outcomes perceived as gains and risk-seeking for outcomes perceived as losses [3]. This realization, together with the non-linear perception of probabilities, proved useful in the creation of **cumulative prospect theory** [4], a theory about making decisions under risk, which we will formally introduce and use to model human decision-making in proceeding chapters.

We've seen how humans are complex in how we decide and now wish to contextualize risk-sensitivity when multiple agents are present. To accomplish this task, we must once again understand how humans do it.

1.1.2 Theory of Mind

Psychologist Daniel Goleman wrote in his famous book: *"We are wired to connect"* [5]. Humans are naturally sociable to the extent that we are biologically dependent on social interaction to live a healthy life. Sociability is not exclusive to humans but the complexity of social interactions is. A close relative to humans, the chimpanzee, also exhibits complex behavior we attribute to developed emotional and social intelligence. A highly cited article from 1978 sought to find if chimpanzees were capable of imputing mental states to others and at the same time offered insights into how we humans intuitively interact with each other [6]. **Theory of mind is the ability to attribute mental states to oneself and to others, and to acknowledge that these may be different from ones own**.

Humans develop a theory of mind around age 4, and is crucial to child development; it enables internal behaviors such as perspective taking and action prediction that informs many decisions we do in our every-day lives. The lack of a theory of mind in humans can be indicative of brain disorders such as autism and Alzheimer's disease. Therefore, agents without a working theory of mind are lacking an essential cognitive mechanism to inform decisions and produce human-like behavior. For this reason, **we will equip our agents with a working theory of mind model**.

1.2 Problem Statement and Research Questions

Cognitive biases and theory of mind play an important role in human decision-making. Coordination between humans is a well-studied topic in game theory, but it is commonly done by invoking axioms of rationality and drawing conclusions which would only apply to a rational agent, the *homo economicus*.

In this thesis we wish to topple the long-standing reign of the *homo economicus* as the fundamental building block in the study of social dilemmas and create a better, more human-like agent, by equipping it with some cognitive biases and a working theory of mind. Specifically, **we create agents with risk-sensitivity and theory of mind and study how they coordinate in normal-form and Markov stag hunt games**. The rest of the thesis focuses on answering the following questions:

- **Q1** - Can cognitive biases concerning risk promote coordination?

- **Q2** - Can increasingly sophisticated levels of theory of mind promote coordination?

We show that both of these questions are answered with a resounding **yes**. This indicates that, while these mechanisms often create sub-optimal individual behavior, they greatly facilitate human coordination.

The emergence of coordination and self-organization in nature, and indeed human societies, remains an open question in many scientific areas, from evolutionary biology to psychology. As such, this thesis offers a small contribution towards the understanding of these phenomena, showing that cognitive biases and theory of mind may provide collective advantage which may be selected by evolution.

On the other hand, the understanding of the mechanisms of self-organization of collective action in multi-agent systems remains one of the most important questions in artificial intelligence. In this context, this work indicates that cognitive biases and theory of mind are two main ingredients that may promote the cooperation between not only machines, but also in hybrid populations (with humans and machines) that may exist in the near future [7].

1.3 Main Contributions

In this thesis, we combine a theory from economics about how people, with their cognitive biases, value uncertain outcomes (i.e., using cumulative prospect theory) with a theory from psychology about how people predict the actions of other people by managing an internal model of others (i.e., using recursive theory of mind mimicking bounded rationality). The resulting outcome of this thesis is a novel framework based on Markov games that allows for the study of agent interaction in stochastic environments where time plays an important role, where the agents more realistic than the current paradigm, which generally uses expected utility theory and perfect rationality.

The present thesis resulted in a conference paper under review, thus contributing to various scientific disciplines of artificial intelligence, namely multi-agent systems and opponent modelling.

1.4 Outline of the Thesis

Chapter 1 described the setting of the rest of the work, and defined the problem statement as a set of research questions to be answered at the end of the thesis. Specifically, we discussed the complexity of human cognitive mechanics and how we propose to contribute to the creation of human-like agents and to the understanding of coordination among these agents. We do so by making agents sensitive to risk and equipping them with a working theory of mind. The remainder of this thesis is organized as follows: The necessary background in individual decision theory, game theory, and theory of mind is introduced in Chapter 2. Chapter 3 provides an overview of relevant work on risk perception, game theory with risk-sensitivity and coordination among machines equipped with theory of mind, which form the theoretical basis of this thesis. In Chapter 4, we detail two experimental setups to study the coordination of risk-sensitive agents equipped with theory of mind. The results of the two experimental models are discussed in Chapter 5. A summary of the results and implications is provided in Chapter 6, as well as suggestions for future research.

2

Background Theory

Contents

2.1 Theories of Value 8

2.2 Game Theory 14

2.3 Level-k Bounded Rationality - A Theory of Mind Model 22

The way people attribute value to objects or events, such as economic goods and services, is a very complicated subject which has been tackled well before the field of economics and game theory were created. The attribution of value is formally called a **theory of value**. Today, marketing theories of value are subjective in the sense that customers choose to buy on the basis of perceived value and, therefore, perception (e.g. risk-sensitivity) has a lot to do with how humans decide.

When in groups, agents must value the outcome of their actions in accordance to what others will do. **Classical game theory** studies these strategic interactions by assuming agents are rational. **Behavioral game theory** is a descriptive theory of behavior which relies on studying games between agents which are assumed to have bounded rationality.

In this chapter, we will describe how theories of value have changed over the years and introduce current value theories, so that we may use these models to create agents that perceive value as humans do. Furthermore, we will introduce concepts in game theory that will enable us to study coordination in two different scenarios: single decision and decisions over time.

2.1 Theories of Value

Due to their heavy focus on human choices, theories of value are at the intersect of economic theory and philosophy. In fact, the theory of value can be thought of as "ethics", in a philosophical sense; what people deem to be good or bad, irrespective of whether the entity under scrutiny is tangible, such as a person or object, or intangible, such as an idea or event. While the concept of value has been in the minds of humans for as long as there has been trade of commodities, the first recorded origins of theories of value can be traced back to Greece [8], during the Classical period (*circa* 480 BC to 323 BC). Plato and his pupil, Aristotle, began the nearly two-millennia-long discussion on what it means for things to have value. Plato thought of value as a quality intrinsic to an object whereas Aristotle offered two points of view on value: **use value** and **exchange value**. Use value expressed the usefulness of objects or ideas in a practical sense, e.g. "My horse is valuable because it does heavy labor for me". Exchange value was meant to capture the essence of how people regard objects of trade; it arises when some people have too much and some people too little, e.g. a liter of water is much valuable in a desert than it is in a tropical zone. These two perspectives were not meant to be mutually exclusive since a shoe can both be worn and traded, but the use value and the exchange value need not be equal. In a transaction, Aristotle hypothesised that a fair trade would be one in which all parties involved would be no better or worse after the transaction than before the transaction. In other words, a shoe is as valuable as the value of the labour of the shoe-maker when it is to be traded, and as valuable as one wants his feet warm if it is to be used. According to this **labour theory of value**, a trade should only occur between that shoe and another commodity whose labour value is equal to that of the shoe-maker

to make the shoe.

This idea of fairness in price did not evolve for many centuries. During the Middle Ages, feudalism was the predominant socio-political system and, as such, the theory of value became less about the labour and more about the social status of the labourer. Europe was going through an economic transformation with the rise of the merchant class and this tore a gap between the consumer and the producer. The producer, author of the labour, was now more distant from the sale of the goods and thus it became harder for people to estimate the value of things. During this time, the Roman Catholic Church dominated Europe and because of that, most knowledge was owned and produced from theologians. The difficulty of ascribing value to goods was alarming to the Church, as they deemed the growing materialism as spiritually dangerous. The Church therefore adopted a form of Aristotelian theory of value combined with medieval labour theory and Christian theology. An important character of this time was theologian Thomas Aquinas, who deemed the selling of goods at a higher price than they are valued as immoral. His views were a mix of Aristotle's labour theory of value and his religious views originating the notion of a **just price** which, in his view, was an intrinsic property of a commodity not necessarily correlated to the price it was sold for. Thus, he showed difference between price and value.

As mercantilism kept growing in Europe, the theory of value began diverging from a labour perspective to one focused on the **utility** and quantity of goods. A physician-turned-builder called Nicholas "If-Christ-had-not-died-for-thee-thou-hadst-been-damned" Barbon, named after his Puritan father Praise-God "Unless-Jesus-Christ-Had-Died-For-Thee-Thou-Hadst-Been-Damned" Barebone (or Barbon), wrote pamphlets about the idea that a market value is determined by the supply and demand of goods [9, p. 63]. Many people such as William Petty, John Locke and Adam Smith (in his book *Wealth of Nations*) were responsible for the advancement of mercantilism up until the 18th century. By that time, a new school of thought emerged, the *physiocrats*, which turned back to the idea of value through usefulness, or utility, and that there was no intrinsic value to things.

While the term *homo economicus* was first used in the late nineteenth century to describe a "dollar-hunting animal", in this thesis we refer to the same term to describe agents which use Expected Utility Theory (EUT) – the prescriptive model of behavior.

2.1.1 Expected Utility Theory

The **St. Petersburg paradox** discovered by Nicolas Bernoulli [10] was one of the first instances of a problem now tackled in economics. The problem is as follows: a person (the entrant) is to determine the fair value he should pay to enter a game offered by someone else (the host) in which a (presumably unbiased) coin is repeatedly tossed and once a heads comes up, the host pays the entrant 2^N coins, where N is the number of tosses until a heads comes up. A fair value would be the number of coins such that both the entrant and the host have similar expected earnings. The paradox comes up when

trying to compute the expected earnings of the entrant:

$$\mathbb{E}\big[2^N\big] = \sum_{n=0}^{\infty} 2^n \, \mathbb{P}(N = n) = \sum_{n=0}^{\infty} 2^n \left(\frac{1}{2}\right)^n = \sum_{n=0}^{\infty} 1 = \infty. \tag{2.1}$$

This means the entrant would rationally be willing to pay any sum to enter the game since he is expected to earn infinite coins at the end, while the host would not rationally accept any finite payment of coins from the entrant. The way one computes the expected earnings as a direct expectation of a random variable is called **expected value theory**.

In correspondence with his cousin, Daniel Bernoulli paved way for what was later called **expected utility theory** whilst solving this problem:

> "The determination of the value of an item must not be based on the price, but rather on the utility it yields. . . . There is no doubt that a gain of one thousand ducats is more significant to the pauper than to a rich man though both gain the same amount."

As such, he proposed a *logarithmic utility* solution, which included the entrant's wealth w and cost to enter c:

$$\mathbb{E}\big[u(2^N, w, c)\big] = \sum_{n=0}^{\infty} u(2^n, w, c) \, \mathbb{P}(N = n) = \sum_{n=0}^{\infty} (\ln(w + 2^n - c) - \ln(w)) \left(\frac{1}{2}\right)^n < \infty. \tag{2.2}$$

The idea of using a utility function enabled the representation of a measure of **risk** through **diminishing marginal returns**, a fundamental concept in economics. It is well known that humans are not risk-neutral. In general, we have some predisposition to risk, determined by several factors including our personalities, mental state and financial means. This is what expected utility theory, and every theory of value, tries to describe.

The scientific revolution and end of feudalism increased the amount of scientific research in Europe but, by that time, the philosophical debate had turned into an economic one for many years. Ethics were not longer the issue, money was. It was only many years later that expected utility theory was formalized by John von Neumann and Oskar Morgenstern, with the advent of **classical game theory**, which will be discussed in the next section. In their book [11], they use **preference ranking of outcomes** as a way to describe rational decision-making, for which there exists a utility function that is able to replace the abstract ordering by a ranking of real numbers. These outcomes may be uncertain themselves, in which case they are called **lotteries** or **prospects**. If outcome X is preferred over outcome Y then $X \succ Y$. If outcome Y is preferred over outcome X then $X \prec Y$. If outcome X is indifferent to outcome Y then $X \sim Y$. A preference ranking of outcomes is rational if it satisfies the following four **axioms of choice**:

- Completeness: For any two prospects X,Y, either $X \succ Y$ or $X \prec Y$ or $X \sim Y$.

- Transitivity: If $X \succ Y$ and $Y \succ Z$ then $X \succ Z$ and similarly for $\sim$.
- Continuity: If $X \succcurlyeq Y \succcurlyeq Z$ then $\exists p \in [0,1]$ such that $pX + (1-p)Z \sim Y$.
- Independence: If $X \succcurlyeq Y$ then for all Z $\exists p \in [0,1]$ such that $pX + (1-p)Z \succcurlyeq pY + (1-p)Z$.

The conclusions drawn from classical game theory rests on these four axioms. Thus, classical game theory describes the strategic interaction between rational decision-makers. Depending on the domain in which classical game theory is applied, and thus on the nature of the agents it tries to model, this notion of rationality may or may not apply. In applications wherein rationality of preferences is determined by immutable rules (e.g. the laws of physics), such as modelling the growth of several types of bacteria in a petri dish, classical game theory sees all four axioms satisfied and conclusions drawn can be trusted with relatively high confidence. However, in scenarios where agents are abstract representations of humans, classical game theory is challenged as an accurate description of decision-making [4, 12]. The famous Allais paradox [13] and Ellsberg paradox [14] are two such demonstrations of the divergence between the axiomatic of classical game theory and human decision-making.

Behavioral game theory attempts to create models to describe human decision-making [15]. The major difference between classical and behavioral game theory is that the former describes the actions of rational agents while the latter gets rid of the notions of rationality and focuses instead on how humans actually decide. Models of human behavior are harder to obtain however, due to the previously discussed complex cognitive mechanics which are an impediment to mathematical tractability.

2.1.2 Prospect Theory

Daniel Kahneman and Amos Tversky proposed, in 1979, a behavioral theory of value called Prospect Theory (PT). In [12], an argument is made about the irrationality of people's decisions and, with behavioral data, they propose a theory of value that models a few cognitive biases observed in the data.

Definition 2.1 (Prospect)**.** Let R be a discrete random variable representing an outcome, with supp$(R) = \{r_1, ..., r_n\}$. A prospect $(r_1, p_1; ...; r_n, p_n)$ is a gamble that yields outcome r_i with probability $\mathbb{P}(R = r_i) = p_i$, where $\sum_{i=1}^{n} p_i = 1$.

Prospect theory allows an agent to consider prospects and make a decision. It does so in two phases: the **editing phase** and the **evaluation phase**. In the editing phase, agents pre-process prospects according to their perception and cognitive mechanics. One of these is the **framing effect**, which divides the prospect into losses and gains by considering a reference point. The perceived prospect can then be written as $(r_1, p_1; ...; r_k, p_k; ...; r_n, p_n)$, where k is the index of the reference point. Then, in the evaluation

phase, the agent evaluates the perceived prospect as follows:

$$V(R) = \sum_{i=k+1}^{n} u^+(r_i)w(p_i) + \sum_{i=1}^{k} u^-(r_i)w(p_i). \tag{2.3}$$

Intuitively, the value of a prospect is a sum of "distorted" expected values of the "split" random value R via the reference point. Outcomes r_i are divided into gains and losses and are perceived, as in expected utility theory, by their respective utility functions u^+ and u^-. To reflect the risk attitude of people about gains and losses, u^+ is concave and u^- is convex, making these agents **risk-averse in gains** and **risk-seeking in losses**. Furthermore, u^- is also taken to be steeper than u^+, effectively considering losses more heavily than gains, a cognitive bias called **loss aversion**. Probabilities are distorted by a weighting function w which non-linearly transforms probabilities to model our perception of these and the so-called **certainty effect**.

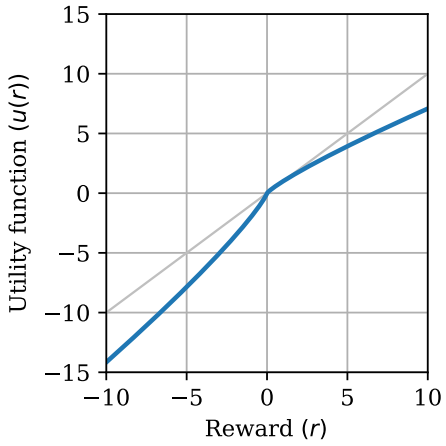

Figure 2.1: Common CPT utility function, $u(x) = x^{0.85}$, for $x \geq 0$, and $u(x) = 2|x|^{0.85}$, for $x < 0$, with reference point at 0. Both negative and positive parts are sublinear, demonstrating risk-sensitivity, and the negative part is steeper, demonstrating loss aversion by weighting losses more heavily than gains.

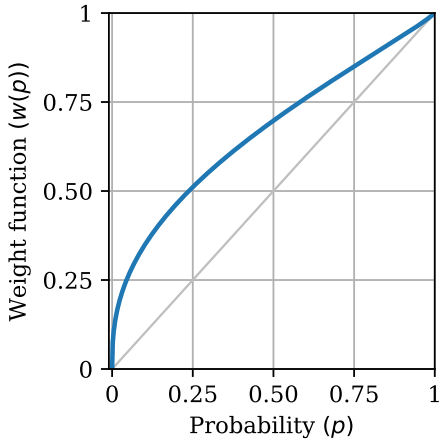

Figure 2.2: Common CPT probability weighting function, $w(p) = \exp\{-\delta(-\log(p))^\gamma\}$, with $\delta = 0.5$ and $\gamma = 0.9$. Low probabilities are overweighted significantly. CPT transforms cumulative probabilities, effectively overweighting rare and extreme events.

Some theorists take issue with this value theory because it does not satisfy first-order stochastic dominance [4]. Stochastic dominance establishes an ordering of random variables based on their distribution functions over the set of outcomes [16].

Definition 2.2 (First-order Stochastic Dominance (FSD))**.** Let X and Y be random variables with distribution functions F_X and F_Y, respectively. X dominates Y, in the first-order stochastic sense, if $\mathbb{P}(X \geq r) \geq \mathbb{P}(Y \geq r)$ for all r, and for some r, $\mathbb{P}(X \geq r) > \mathbb{P}(Y \geq r)$.

In other words, X dominates Y in the first-order stochastic sense, if the following holds: for any outcome r, X gives at least as high a probability of receiving at least r as does Y, and for some r, X gives a higher probability of receiving at least r.

Definition 2.3 (Statewise Stochastic Dominance (SSD))**.** Let X and Y be random variables with distribution functions F_X and F_Y, respectively. Then, X dominates Y in the statewise stochastic sense, if $x \geq y$ for all x, y and $x > y$ for some x, y.

In other words, X dominates Y, in the statewise stochastic sense, if the following holds: all outcomes of X are at least as good as all outcomes of Y and at least one outcome of X is strictly better than all outcomes of Y.

Therefore, FSD is a special case of SSD. A simple example of how prospect theory violates SSD, and consequently, FSD can be done with a cast of a die.

Example 2.1.1 (Prospect theory violation of SSD)**.** Consider the prospect of casting an unbiased 6-sided die. The outcome of a cast is represented by a discrete uniform random variable R, with support $\{1, 2, 3, 4, 5, 6\}$. Taking all these outcomes to be gains by setting the reference point to be 0, we can calculate the value using prospect theory as follows:

$$V(R) = \sum_{i=1}^{6} u(r_i)w(p_i) = w\left(\frac{1}{6}\right)(u(1) + u(2) + u(3) + u(4) + u(5) + u(6)). \tag{2.4}$$

Assuming $u(r) = r^{0.85}$ and $w(p) = \exp\left\{-0.5(\log(p))^{0.9}\right\}$, the value of this prospect is 7.35, which is larger than the maximum possible outcome which is 6. In fact, an agent which behaves according to prospect theory would choose this gamble over the certainty of a gift of 7; thus, violating the SSD.

2.1.3 Cumulative Prospect Theory

To overcome the limitation presented in Example 2.1.1, Kahneman and Tversky improved on their theory in 1992 with Cumulative Prospect Theory (CPT) [4]. By weighing cumulative probabilities instead, first-order stochastic dominance is satisfied. The value of a prospect R is obtained as follows:

$$V(R) = \sum_{i=k+1}^{n} u^{+}(r_i)[w^{+}(\mathbb{P}(R \geq r_i)) - w^{+}(\mathbb{P}(R > r_i))] + \sum_{i=1}^{k} u^{-}(r_i)[w^{-}(\mathbb{P}(R \geq r_i)) - w^{-}(\mathbb{P}(R > r_i)). \tag{2.5}$$

The relative dependency of the outcomes on the probabilities translates the fact that people overweight extreme but unlikely events, whereas with PT agents would overweight unlikely events independently of the outcome.

Henceforth, we distinguish two types of agents: EUT-agents and CPT-agents, depending on which theory of value they use, i.e., EUT and CPT, respectively.

2.2 Game Theory

Game theory is the study of strategic thinking. It makes use of **games** to model group interactions between agents[1].

A **game** is specified by four elements:

- **Agents** - these are the decision makers that will perform the actions,
- **Actions** - a specification of what agents can decide,
- **Information** - the basis on which agents can inform their decisions,
- **Rewards** - the objective of the agents, the motivation for their decisions.

Solving a game (i.e. finding what players will decide) requires further assumptions on the rationality of players and on external influences (e.g. are agreements binding or self-enforced?). Common knowledge of rationality creates games that are studied under **classical game theory**, whereas **behavioral game theory** studies games wherein no such assumption is made. Classical game theory is divided into two major areas: **cooperative game theory** and **noncooperative game theory**. Cooperative game theory studies games in which externally enforced coalitions can be formed and agents employ some degree of trust in one another. Noncooperative game theory, on the other hand, makes no such assumptions; it studies games where agents have only their individual self-interest at heart[2]. For the purpose of

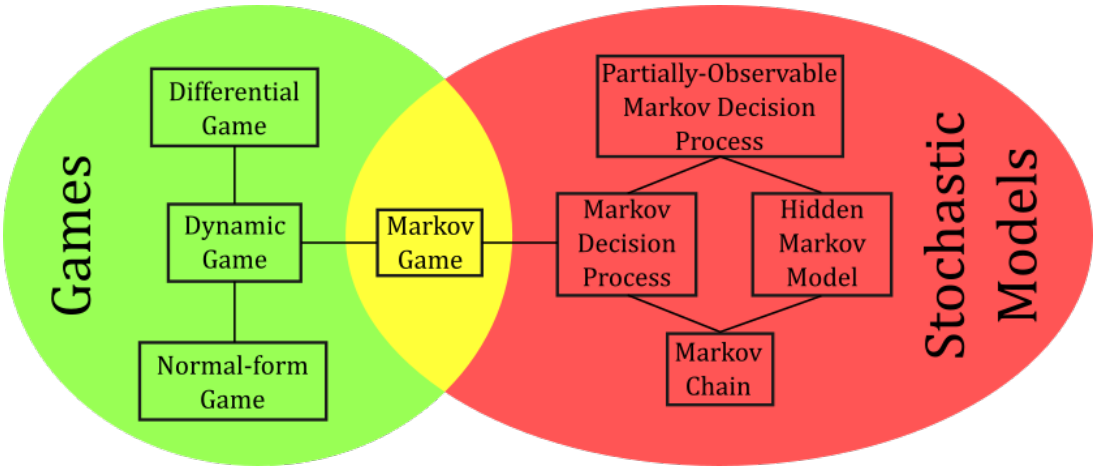

Figure 2.3: A (simplified) map of games and stochastic models. There is a myriad of games studied under game theory, with different assumptions on the information structure that specify a game. The Markov game is at the intersect between game theory and stochastic processes, a generalization of the dynamic game, by allowing stochastic transitions between states, and of Markov decision process, by extending it to the multi-agent case.

[1] A distinction is sometimes made between the terms *agent* and *player*, to distinguish between the decision makers in cooperative and noncooperative games. We will stick with the (arguably) more general term *agent*, since *player* assumes there is always an opponent.

[2] This does not exclude cases where individual self-interest is a function of the well-being of others.

this thesis, we will review normal-form games and the Nash equilibrium concept in order to understand how to study coordination. To broaden our discussion of coordination under risk, we will also strive to understand the trade-off between long-term and short-term rewards in a game-theoretical setting. To accomplish this we will review some basic concepts on Markov chains, which we will see as a model of discrete time evolution of a passive agent. We will then add control and motivation to the agent, resulting in a Markov decision process. Lastly, we will create a Markov game, a model that allows several of these agents to interact.

2.2.1 Normal-Form Game

A normal-form game is a tuple $(N, \mathcal{A}, R)$ with the following elements:

- A set of n agents $N = \{1, ..., n\}$,
- A collection of sets of action spaces $\mathcal{A} = \{\mathcal{A}_1, ..., \mathcal{A}_n\}$, with joint action space $\boldsymbol{\mathcal{A}} = \underset{i \in N}{\times} \mathcal{A}_i$, and
- A collection of reward functions $R = \{r_1, ..., r_n\}$.

In a normal-form game, the process of decision-making is simultaneous (i.e., agents decide at the same time) or, equivalently, decisions are made with no additional information of the player's behavior before it is executed. A play of a normal-form game has the following steps:

1. Each agent i simultaneously decides which action a_i to perform, out of the $\mathcal{A}_i$ possible actions, creating a joint action $\boldsymbol{a} = (a_1, ..., a_n) \in \boldsymbol{\mathcal{A}}$, and
2. Each agent i receives a reward $r_i(\boldsymbol{a})$, as a function of the actions of all agents.

A rational agent seeks to perform the maximization of the received reward

$$\max_{a_i} r_i(a_i, \boldsymbol{a}_{-i}), \tag{2.6}$$

where we write $\boldsymbol{a} = (a_i, \boldsymbol{a}_{-i})$ to emphasize the perspective of agent i, with $\boldsymbol{a}_{-i}$ being the joint action of all agents except that of agent i. If every agent is performing this maximization, then $\boldsymbol{a}_{-i}$ is being chosen accordingly by every other player. The common knowledge of rationality assumption, made by classical game theory, means that everyone knows all agents are rational, that everyone knows that everyone knows that all agents are rational, and so on *ad infinitum*. This allows agents to break free of this recursion and calculate what is called a Nash equilibrium.

Definition 2.4 (Deterministic Nash Equilibrium (DNE))**.** A joint action $\boldsymbol{a}^* = (a_1^*, ..., a_n^*)$ is a deterministic (or pure) Nash equilibrium if, for every agent i,

$$r_i(a_1^*, ..., a_i^*, ..., a_n^*) \geq r_i(a_1^*, ..., a_i, ..., a_n^*), \tag{2.7}$$

for all actions $a_i \in \mathcal{A}_i$.

In some cases, agents may be allowed to choose a distribution over actions, instead of a single action. In these cases, the behavior of an agent i is determined by a distribution function called **policy** $\pi_i : A_i \times [0, 1]$, with $\sum_{a \in A_i} \pi_i(a) = 1$, where $\pi_i(a)$ is the probability that agent i chooses action $a \in A_i$. The space of all possible policies of agent i is denoted by Π_i such that $\pi_i \in \Pi_i$. It is worthy of note that deterministic decision-making can be represented by a degenerate probability distribution and we will refer to these as **deterministic policies**, so that the term policy becomes the umbrella term for behavior. In this case, the notion of a Nash Equilibrium must be different, since the reward function depends on the actions played, which are now random variables. Therefore, the agents must find a way to evaluate their actions. In game theory, rational agents calculate the expected value of the rewards, given the joint policy of the other agents. We will denote the value of a joint policy $\boldsymbol{\pi} = (\pi_i, \boldsymbol{\pi}_i)$ from the perspective of agent i by $V_i(\pi_i, \boldsymbol{\pi}_i)$.

Definition 2.5 (Stochastic Nash Equilibrium (SNE))**.** A joint policy $\boldsymbol{\pi}^* = (\pi_1^*, ..., \pi_n^*)$ is a stochastic (or mixed) Nash equilibrium if, for every agent i,

$$V_i(\pi_1^*, ..., \pi_i^*, ..., \pi_n^*) \geq V_i(\pi_1^*, ..., \pi_i, ..., \pi_n^*), \tag{2.8}$$

for all policies $\pi_i \in \Pi_i$.

In other words, a Nash equilibrium[3] is a joint action from which no agent will be better off by unilaterally switching their individual action/policy. Thus, Nash equilibria prescribe the possible behavior of rational agents under the assumption that there is common knowledge of rationality.

Noncooperative game theory uses normal-form games (albeit others games are used as well) to study social dilemmas. A **social dilemma** is a situation between agents in which selfish behavior is profitable only if it is not adopted by everyone. These are modelled with normal-form games such as the **prisoner's dilemma** [18], one of the most famous of all normal-form games. In the prisoner's dilemma, two prisoners accused of robbing a bank are jailed in two separate cells. The police interrogates them and offers them a deal. If they both stay silent and say nothing, they both serve a sentence of 1 year. If they both confess to the crime, and consequently snitch on their partner, they both serve a sentence of 5 years. However, if one of them stay silent and the other confesses, the silent partner goes to jail for 10 years and the snitch goes free. Only one Nash equilibrium exists in the prisoner's dilemma, both agents choose to confess. **This is an example of why completely rational agents would not choose to stay silent and minimize the total years spent in prison**.

[3] For a study of the epistemic conditions of the Nash equilibrium, see [17].

		Agent 2	
		C	S
Agent 1	C	$-5, -5$	$0, -10$
	S	$-10, 0$	$-1, -1$

Table 2.1: The payoff matrix of the prisoner's dilemma. The rows are actions decided by agent 1 and columns are actions decided by agent 2, with C stands for confess and S stands for remaining silent. The reward of both agents is written in the cell of the corresponding joint action. The number on the left is the reward of agent 1 and the one on the right is the reward of agent 2.

		Agent 2	
		S	H
Agent 1	S	$5, 5$	$0, 1$
	H	$1, 0$	$1, 1$

Table 2.2: The payoff matrix of the stag hunt. The rows are actions decided by agent 1 and columns are actions decided by agent 2, where S stands for hunting stag and H stands for hunting hare. The reward of both agents is written in the cell of the corresponding joint action. The number on the left is the reward of agent 1 and the one on the right is the reward of agent 2.

2.2.2 The Stag Hunt Game

While one can find analogies to the prisoner's dilemma in several domains, there is another normal-form game which encompasses a different type of dilemma. The game of the **stag hunt** tells a story of two hunters going out to hunt a stag [19]. On the way, they both see a hare running off right in front of them. Both hunters now have to make a decision to either keep hunting the stag or switch to hunting the hare. They both know that hunting the stag alone is an unfruitful endeavor, while the hare can be hunted solo. This presents the hunters with a dilemma regarding individual safety and social cooperation; the hare is a sure-thing but is a small prey and the stag is more rewarding but requires **coordination**.

Definition 2.6 (Coordination Game)**.** A coordination game is a game with two or more Nash equilibria in which agents choose policies whose support is the same or have some sort of correspondence.

An example of the stag hunt, a coordination game, can be found in Table 2.2. The stag hunt has two DNEs (both agents hunting stags or both agents hunting hares) and a single SNE. **It is with this SNE that we will study coordination among agents using expected value theory and cumulative prospect theory, in Chapter 4 and Chapter 5, respectively.**

2.2.3 Markov Game

A normal-form game, such as the prisoner's dilemma, can be played a number of times. These repeated games[4] capture the idea that an agent will have to take into account the impact of his current action

[4]The famous Axelrod's tournament used the iterated prisoner's dilemma to show that Darwinian evolution of strategies can originate cooperation based on reciprocity [20].

on the future actions of other agents. This dilemma of short-term versus long-term rewards and their associated risk is something we will include in our model, as a CPT value function in a Markov game. We will use Markov chains to model the stochastic nature of real-world environments[5], add decision-making through actions and rewards and derive the Markov decision process and generalize it to multiple agents to create the Markov game.

2.2.3.A Markov Chain

Definition 2.7 (Markov Chain)**.** A sequence of random variables $\{S_t\}_t$ is a (discrete time) Markov chain if it satisfies the first-order Markov property:

$$\mathbb{P}(S_{t+1} = s_{t+1}|S_t = s_t, S_{t-1} = s_{t-1}, ..., S_0 = s_0) = \mathbb{P}(S_{t+1} = s_{t+1}|S_t = s_t). \tag{2.9}$$

In other words, the first-order Markov property states that the state of the future depends only on the present state. A Markov chain can be interpreted as a model of stochastic evolution of a system. A system can be in one of several states which change according to a probability distribution that depends only on the current state. A time-homogeneous, discrete space Markov chain with discrete time and discrete state space $\mathcal{S}$ can be fully specified by the time-independent transition matrix P, a stochastic matrix with elements $P_{ij} = \mathbb{P}(S_{t+1} = j|S_t = i)$, whose rows and columns are sorted in the same fashion as the state space $\mathcal{S}$.

Definition 2.8 (Accessibility, Communication and Irreducibility)**.** State j is **accessible** from state i if there exists an integer $n_{ij} \geq 0$ such that

$$\mathbb{P}(S_{n_{ij}} = j \mid S_0 = i) = p_{ij}^{(n_{ij})} > 0.$$

A state i is said to **communicate** with state j if both i is accessible from j and j is accessible from i.
A **communicating class** is a maximal set of states $C \subseteq \mathcal{S}$ such that for every $i, j \in C$, i communicates with j.
A Markov chain is **irreducible** if its state space $\mathcal{S}$ is a single communicating class.

Definition 2.9 (Recurrence)**.** Let the random variable $T_i = \inf\{t \geq 1 : S_t = i\}$ be the recurrence time to state i. Let $f_{ii}^{(t)} = \mathbb{P}(T_i = n \mid X_0 = i)$ be the probability that the system returns to state i for the first time after t steps. State i is **recurrent** if

$$\mathbb{P}(T_i < \infty \mid X_0 = i) = \sum_{t=1}^{\infty} f_{ii}^{(t)} = 1.$$

[5]When we add risk-sensitive agents we are effectively making agents which are also sensitive to the risk coming not from the other agents but also from the environment.

State i is **positive recurrent** if

$$\mathbb{E}[T_i] = \sum_{t=1}^{\infty} t f_{ii}^{(t)} < \infty.$$

In other words, a Markov chain is irreducible if it is possible to get to any state from any state. Additionally, a state is positive recurrent if it takes a finite amount of time, on average, for the system to return that state.

Theorem 2.1 (Stationary Distribution)**.** *Let $\{S_t\}_t$ be a time-homogeneous, discrete time Markov chain with discrete state space $\mathcal{S}$. If $\{S_t\}_t$ is irreducible and all states in $\mathcal{S}$ are positive recurrent, then it has a* ***stationary distribution*** *$\rho = (\rho_1, ..., \rho_{|\mathcal{S}|})$ such that:*

- $\forall i \in \mathcal{S}, \rho_i \geq 0$*,*
- $\sum_{i \in \mathcal{S}} \rho_i = 1$*, and*
- $\rho = \rho P$*.*

We will now illustrate the use of a Markov chain with a simple example, upon which we will improve on by adding different elements when we discuss Markov decision processes and Markov games.

Example 2.2.1 (Random Walk)**.** A one-dimensional random walk describes the path taken by, for example, a drunkard that moves according to a fixed transition probability distribution on a line. In this example we will model a random walk on a constrained discrete line with a discrete time Markov chain with state space $\mathcal{S} = \{0, ..., 15\}$. We say "line" because, from any state i, it is only to go to states $i-1$ or $i+1$ (if possible due to endpoints) or to remain in state i.

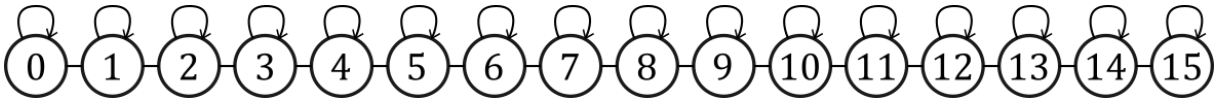

Figure 2.4: The transition diagram of the Markov chain modelling the discrete line with 16 states.

Let us consider that, for all states $i \in \mathcal{S} \setminus \{0, 15\}$, $P_{i,i} = 0.25$, $P_{i,i-1} = 0.5$, $P_{i,i+1} = 0.25$ and that $P_{0,0} = 0.75$, $P_{0,1} = 0.25$, $P_{15,15} = 0.5$, $P_{15,14} = 0.5$. It is easy to see that this Markov chain is irreducible and all its states are positive recurrent and therefore there exists a stationary distribution ρ such that $\rho = \rho P$ (see Figure 2.6).

2.2.3.B Markov Decision Process

If we then add control and motivation to the previously passive agent in a Markov chain we obtain a Markov decision process [21].

Definition 2.10 (Markov Decision Process (MDP))**.** A Markov decision process is a tuple $(\mathcal{S}, \mathcal{A}, P, R)$ with the following elements:

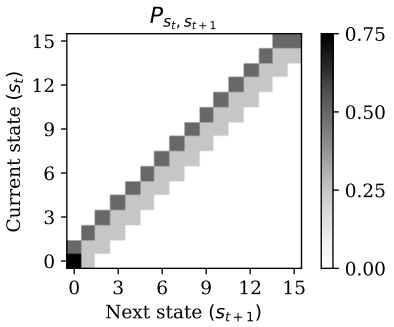

Figure 2.5: The transition matrix of the Markov chain modelling the discrete line with 16 states.

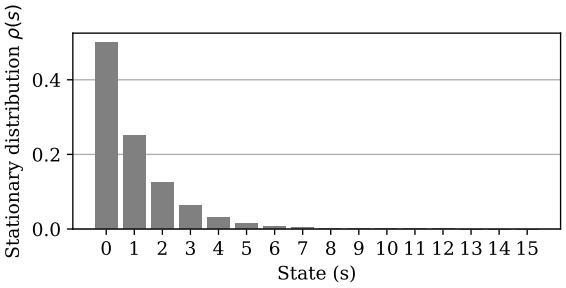

Figure 2.6: Stationary distribution of the drunkard's random walk. It is unsurprisingly skewed to the left, since there is a higher probability of going to the left than staying or going to the right.

- A discrete state space $\mathcal{S}$,
- A finite set of actions $\mathcal{A}$,
- A transition function $P : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0, 1]$, and
- A reward function $R : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$.

A MDP is a model of control in a stochastic environment. A single agent wishes to find, for each state $s \in \mathcal{S}$, a policy $\pi(a|s)$ that maximizes some function of the obtained rewards, which we will call the value V. We will consider only the infinite-horizon expected discounted sum of rewards case where the reward function is solely a function of the current state, in which case the value is defined as

$$V(s, \pi) = \mathbb{E}_{s_{t+1} \sim P(\cdot|\pi, s_t)} \left[\sum_{t=0} \beta^t r(s_t) | s_0 = s \right], \tag{2.10}$$

where the expectation is taken to be on the stochastic dynamics of the system which is fully specified by the decided policy of the agent, and $\beta \in (0, 1)$ is the discount factor, a parameter that both insures the value is finite and controls the agent's sensitivity to long-term rewards. In summary, an agent in a MDP seeks find the optimal policy

$$\pi^* = \underset{\pi}{\text{argmax}}\, V(s, \pi) = \underset{\pi}{\text{argmax}}\, \mathbb{E}_{s_{t+1} \sim P(\cdot|\pi, s_t)} \left[\sum_{t=0} \beta^t r(s_t) | s_0 = s \right], \tag{2.11}$$

which can be solved via a **dynamic programming** scheme such as value iteration [21] (see Appendix B for a brief introduction to dynamic programming in the context of game theory).

Example 2.2.2 (Controlled Random Walk)**.** If our drunkard from Example 2.2.1 is allowed to control its random walk by choosing on which hand he carries his bottle then this decision-making process can be modelled with a MDP, with state space $\mathcal{S} = \{0, ..., 15\}$, action space $\mathcal{A} = \{L, S, R\}$, reward function

r, and transition function P. Let us assume that if he chooses to hold his bottle on the left hand (i.e., chooses action L), then he will move according to the transition matrix P^L presented in Figure 2.5 – similar to Example 2.2.1. If, instead, he chooses to hold his bottle on the right hand (i.e., chooses action R) then he will move according to the transition matrix P^R that, for all states $i \in \mathcal{S} \setminus \{0, 15\}$, $P^R_{i,i} = 0.25$, $P^R_{i,i-1} = 0.25$, $P^R_{i,i+1} = 0.5$, and that $P^R_{0,0} = 0.50$, $P^R_{0,1} = 0.50$, $P^R_{15,15} = 0.75$, $P^R_{15,14} = 0.25$ – see Figure 2.9. He can also choose to hold the bottle with both hands (i.e, choose action S), moving according to the transition matrix P^S that, for all states $i \in \mathcal{S} \setminus \{0, 15\}$, $P^S_{i,i} = 0.50$, $P^S_{i,i-1} = 0.25$, $P^S_{i,i+1} = 0.25$ and that $P^S_{0,0} = 0.75$, $P^S_{0,1} = 0.25$, $P^S_{15,15} = 0.75$, $P^S_{15,14} = 0.25$ – see Figure 2.8.

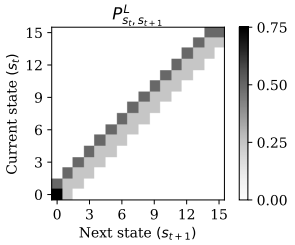

Figure 2.7: The transition diagram of the Markov chain resulting from choosing action L, in the Markov decision process example.

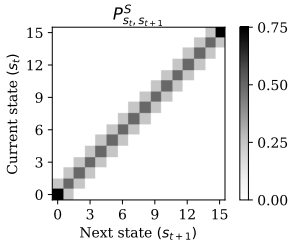

Figure 2.8: The transition diagram of the Markov chain resulting from choosing action S, in the Markov decision process example.

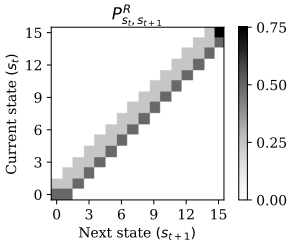

Figure 2.9: The transition diagram of the Markov chain resulting from choosing action R, in the Markov decision process example.

He wants to either get home, at state 3, or go to the next bar, at state 11, but he prefers going to the next bar. Therefore, he is motivated by the reward function r such that $r(3) = +1$, $r(11) = +5$, and $r(s) = 0, \quad \forall s \in \mathcal{S} \setminus \{3, 11\}$.

He wishes to maximize the value in Equation (2.10), with $\beta = 0.9$. Using value iteration (see Appendix B), we obtain the optimal value and policy in Figures 2.10 and 2.11.

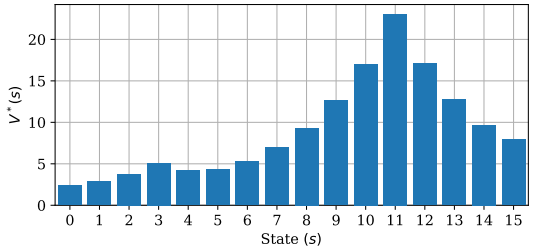

Figure 2.10: The optimal value for the MDP example, with discount factor $\beta = 0.9$.

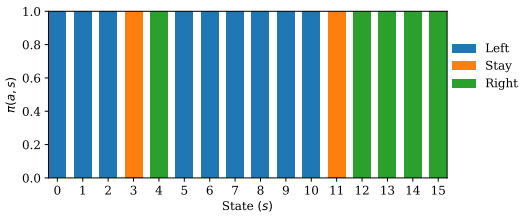

Figure 2.11: The optimal policy for the MDP example, with discount factor $\beta = 0.9$.

2.2.3.C Markov Game

A Markov game [22] is the generalization of MDPs to the multi-agent case.

Definition 2.11 (Markov Game)**.** A Markov game is a tuple (N, S, A, P, R) with the following elements:

- A set of n agents $N = \{1, ..., n\}$,
- A joint state space $\boldsymbol{\mathcal{S}} = \underset{i \in N}{\times} \mathcal{S}_i$, where $\mathcal{S}_i$ is the state space of agent i and $\times$ is the Kronecker product,
- A set of joint actions $\boldsymbol{\mathcal{A}} = \underset{i \in N}{\times} \mathcal{A}_i$, where $\mathcal{A}_i$ is the set of actions of agent i,
- A transition function $P : \boldsymbol{\mathcal{S}} \times \boldsymbol{\mathcal{A}} \times \boldsymbol{\mathcal{S}} \to [0, 1]$, and
- A collection of reward functions $R = \{r_1, ..., r_n\}$.

A Markov game is then a model of multi-agent control in a stochastic environment. Being a game, solving it means finding the joint policy of all agents. A solution concept based on the Nash equilibrium exists, the Markov perfect equilibrium, but we will take a different approach. Regardless, we will assume each agent i has a value they seek to maximize

$$V_i(s, \pi_i, \boldsymbol{\pi_{-i}}) = \mathbb{E}_{s_{t+1} \sim P(\cdot|\pi_i, \boldsymbol{\pi}_{-i}, s_t)} \left[\sum_{t=0}^{\infty} \beta^t r_i(s_t) \right], \tag{2.12}$$

with the dynamics now prescribed by the joint policy $\boldsymbol{\pi} = (\pi_1, ..., \pi_n)$. Each agent will then find their optimal policy π_i^* that maximizes their value, given the joint policy of the other agents $\boldsymbol{\pi}_{-i}$, i.e.,

$$\pi_i^* = \underset{\pi_i}{\text{argmax}}\, V(s, \pi_i, \boldsymbol{\pi_{-i}}) = \underset{\pi}{\text{argmax}}\, \mathbb{E}_{s_{t+1} \sim P(\cdot|\pi_i, \boldsymbol{\pi}_{-i}, s_t)} \left[\sum_{t=0}^{\infty} \beta^t r_i(s_t) \right]. \tag{2.13}$$

By not taking the equilibrium approach, and therefore, foregoing completely rational action, we must have a method through which each agent i obtains the policies of every other agent $\boldsymbol{\pi}_{-i}$.

2.3 Level-k Bounded Rationality - A Theory of Mind Model

In reality, adult people have very well developed prior knowledge about the default behavior of systems they encounter in their every day lives, including of other people. This prior knowledge is a belief system that is learned throughout life to which we call **intuition**. Specifically, when people have beliefs about the behavior of others, it is a form of **intuitive psychology** [23]. A theory of mind model is one mathematical formulation of a theory of intuitive psychology. Since people operate based on their beliefs, the concept of the Nash equilibrium is not a good descriptor of behavior in normal-form games. However, it is possible to obtain a similar equilibrium concept if we model these beliefs, into what is known as **rationalizability** [24, 25], which still assumes agents are rational and common knowledge of rationality. This is because a rational agent maximizes perceived value (which is a function of their beliefs) and not the actual value.

Similar to rationalizability is the **level-k** bounded rationality, a recursive theory of mind model [26]. In the level-k model, all agents make initial assumptions on the behaviors of all agents. These assumptions are level-0 policy beliefs we will call **stereotype policies**. To greatly simplify the formal explanation of the level-k model, we will make the following assumptions:

1. There are only two agents at play, agents 1 and 2,
2. Both agents have the same value function $V_1 = V_2 = V$,
3. They both know the stereotypes policies of one another, denoted $\pi_1^{(0)}$,$\pi_2^{(0)}$, and
4. All the above is common knowledge[6].

With these assumptions, we are greatly simplifying modelling process by eliminating the agent's problem to infer the behavior of others through their beliefs. This is actually what we want because we want to understand how coordination is affected in a theory of mind model, and not how inference mistakes alter that understanding[7].

Now that the agents have a solid basis of assumptions, the level-k model establishes a hierarchy of behaviors (policies) of increasing sophistication level[8]. Agent 1 has a stereotype policy about agent 2 $\pi_2^{(0)}$ and will try to best respond by maximizing the value under the assumption that agent 2 is using $\pi_2^{(0)}$. Agent 1 will then calculate $\pi_1^{(1)} = \underset{\pi_1}{\text{argmax}}\, V(\pi_1, \pi_2^{(0)})$. Agent 2 will go through the same reasoning and calculate $\pi_2^{(1)} = \underset{\pi_2}{\text{argmax}}\, V(\pi_2, \pi_1^{(0)})$. We will refer to the superscript in $\pi_i^{(\cdot)}$ as the sophistication level, such that $\pi_i^{(k)}$ is the **level-k policy**. However, agent 1 can reason agent 2 is not actually using $\pi_2^{(0)}$ but is instead using $\pi_2^{(1)}$ and as such agent 1 will calculate the level-2 policy $\pi_1^{(2)} = \underset{\pi_1}{\text{argmax}}\, V(\pi_1, \pi_2^{(1)})$, and agent 2 will calculate $\pi_2^{(2)} = \underset{\pi_2}{\text{argmax}}\, V(\pi_2, \pi_1^{(1)})$. This tower of policies can be constructed recursively with

$$\begin{aligned} \pi_1^{(k)} &= \underset{\pi_1}{\text{argmax}}\, V(\pi_1, \pi_2^{(k-1)}), \text{ and} \\ \pi_2^{(k)} &= \underset{\pi_2}{\text{argmax}}\, V(\pi_2, \pi_1^{(k-1)}). \end{aligned} \tag{2.14}$$

We will use this scheme to model theory of mind of agents in Markov games, as an alternative to using equilibrium solution concepts.

[6] In other words, they both know that they know this, they both know that they know that they know this, and so on *ad infinitum*.
[7] Which may prove be an interesting avenue of future research.
[8] This structure is what gives the model its name.

3

Related Work

Contents

3.1 Risky Decisions 26
3.2 Risky Games 27
3.3 Intuitive Psychology 28
3.4 Summary of Related Work 29

In this chapter, we will review scientific works which have studied risk-sensitivity and theory of mind.

3.1 Risky Decisions

While there is not yet any one universally satisfactory definition of risk[1], it has been studied at length since its first definition [27]. In most cases, quantitative studies are done using statistical methods to measure the risk of a particular set of outcomes. The analysis of risk is usually done in the domains of business and finance, where there are constant risks that must be taken in to account and controlled [28]. However, we are interested in a different type of risk, the one perceived by humans. We are interested in understanding **risk perception** and how it affects human collective behavior, which is studied in the social sciences as a **cultural theory of risk**[2] [29].

EUT and PT/CPT are risk-sensitive theories of value. However, empirical evidence suggests that PT/CPT are better models of human decision-making than EUT. Private bankers and fund managers behave according to PT and violate EUT [30]. Inexperienced consumers in a well-functioning marketplace behave according to PT while those with more experience behave according to more recent economic theories, showing that learning plays an important role in risk perception [31][3]. A study using a model inspired by PT helped explain properties seen in asset prices in an economy where investors derive direct utility not only from consumption but also from fluctuations in the value of their financial wealth [32]. The presence of reference points was observed in a large database of firms, together with risk seeking behavior for firms below their reference point and risk averse behavior for firms above their reference point, and risk seeking behavior being more intense than risk averse behavior [33]. Prospect theory was used to explain why political actors pursue risky reforms, in spite of political resistance that counteract change [34]. Decision-making models achieved state of the art performance on human judgement datasets by creating neural networks with human-like inference bias by pretraining them with synthetic data generated by CPT [35].

The transformation of cumulative probability rather than individual probability is the crucial idea behind **rank-dependent theory** [36]; it described the choice behavior seen in the Allais paradox [13]. This was the main motivation behind the update done to PT that originated CPT [4]. A discussion of the practical differences between PT and CPT can be found in [37]. However, other models of risk perception exist. **Disappointment aversion** is a model consistent with the Allais paradox that includes EUT as a special case [38]. A modification of PT/CPT endogenizes the reference point based on previous experience of agents [39]. Another modification of PT/CPT, named **third-generation prospect theory**,

[1] Apart from being the collective noun for lobsters.

[2] Cultural theory of risk argues that risk perception is not formed independently of social context, but rather as an emergent property of human systems.

[3] This study also shows agents with intense market experience are more willing to part with their entitlements than lesser-experienced agents, a cognitive bias called **divestiture aversion**, or **endowment effect** in behavioral economics.

makes the same predictions as CPT but it can also account for the **endowment effect** and **preference reversals** [40].

Some criticism on the use of PT/CPT as the default alternative to risk-sensitive models of human choice models have been recently presented [41–43], but we will ignore these under the pretext that there has been no more widely accepted theory of human choice under risk than CPT. Furthermore, our goal is not to use this model to explain human behavior in the most accurate way, but instead, to equip agents with a risk-sensitive model that more closely resembles human risk-sensitivity than standard EUT does.

3.2 Risky Games

In game theory, the equilibrium solution concepts in normal-form games using PT and CPT have been previously studied [44]. A definition of equilibrium in a normal-form game with agents using PT and CPT is proposed and hereafter, introduced for the reader's convenience.

Definition 3.1 (PT- and CPT-equilibrium)**.** In a normal-form game, a joint policy $\boldsymbol{\pi} \in \boldsymbol{\Pi}$ is a **PT-equilibrium** given reference point $\boldsymbol{b} \in \mathbb{R}^n$ if for all $i = 1, ..., n$ and all $\pi_i \in \Pi_i$ we have $V_i^{\mathsf{PT}}(\boldsymbol{\pi}, b_i) \geq V_i^{\mathsf{PT}}(\pi_i, \boldsymbol{\pi}_{-i}, b_i)$. Analogously, a joint policy $\boldsymbol{\pi} \in \boldsymbol{\Pi}$ is a **CPT-equilibrium** given reference point $\boldsymbol{b} \in \mathbb{R}^n$ if for $i = 1, ..., n$ and all $\pi_i \in \Pi_i$, we have $V_i^{\mathsf{CPT}}(\boldsymbol{\pi}, b_i) \geq V_i^{\mathsf{CPT}}(\pi_i, \boldsymbol{\pi}_{-i}, b_i)$.

These are analogous to Nash equilibria, which use EUT as a theory of value. Therefore, it is worthwhile comparing the two equilibria as a way of comparing the two theories of value (i.e., CPT and EUT) in a multi-agent setting.

Prospect theory was used to study the cooperation in the iterated prisoner's dilemma [45].

We suggest the reader to look into [46, 47] that provide an comprehensive description of risk measures on MDPs. In particular, they show that a Bellman equation[4] exists for the CPT-value in both the finite-horizon and discounted infinite-horizon cases. In the discounted infinite-horizon MDP, the recursive solution to the CPT-value is, for a state $s \in \mathcal{S}$ and policy π,

$$\begin{aligned} V^{\mathsf{CPT}}(s, \pi) = &\int_0^\infty w^+ \left(\sum_{a \in A(s)} P_s^a (u^+((r(s) + \beta V^{\mathsf{CPT}}(S, \pi) - b)_+) > \epsilon) \pi(a|s) \right) d\epsilon \\ &- \int_0^\infty w^- \left(\sum_{a \in A(s)} P_s^a (u^-((r(s) + \beta V^{\mathsf{CPT}}(S, \pi) - b)_-) > \epsilon) \pi(a|s) \right) d\epsilon, \end{aligned} \tag{3.1}$$

where b is the reference point and, together with $(\cdot)_+ = \max(0, \cdot)$ and $(\cdot)_- = -\min(0, \cdot)$, they split the

[4] A Bellman equation is a recursive solution of an initially difficult problem which is generally obtained via dynamic programming.

calculation of the value into an integral for gains and another for losses. The discount factor $\beta \in (0, 1)$ controls the importance of long-term rewards over short-term rewards and $r(s)$ is the reward at state $s \in \mathcal{S}$. The random variable S represents the next state and P_s^a is the probability measure conditional to the current state $s \in \mathcal{S}$ and chosen action $a \in A$ (i.e., $P_s^a(\cdot) = \mathbb{P}(\cdot|s_t = s, a_t = a)$). Ultimately, $u^+((r(s) + \beta V^{\text{CPT}}(S, \pi) - b)_+)$ and $u^-((r(s) + \beta V^{\text{CPT}}(S, \pi) - b)_-)$ are random variables representing the utility of a perceived reward (gains or losses). As in the original formulation of CPT, the weighted survival functions of the gains and losses (with respective weighting functions w^+ and w^-) are integrated for all possible utilities, though in this case the probability measure is distorted with the policy π chosen by the agent.

Like the original CPT, the CPT-value in MDPs is a generalization of the EUT-value with utility function u, when $\forall p \in [0, 1] w^+(p) = w^-(p) = p$ and $\forall r \in \mathbb{R} u^+(r) = u^-(r) = u(r)$. For this reason, comparing V^{CPT} and V^{EUT} is straightforward and meaningful. **Part of this thesis consists in extending the CPT-value to the Markov game setting, so we can understand the behavior of CPT-agents and compare to that of EUT-agents**.

3.3 Intuitive Psychology

There has been a recent push in research into the notion of **intuitive theories** and how we can take advantage of these to build agents capable of simulating human intuition [23]. In this thesis, we are interested in **intuitive psychology**, a set of theories agents have upon which causal inference can be done, effectively creating an inference bias that helps behavioral judgements. In multi-agent systems research, these can be seen as theories of agents modelling other agents (see [48] for a recent comprehensive review of autonomous agent modelling papers); agent modelling schemes can be categorized based on their assumptions and what they are trying to model.

In our case, we will be using the level-k bounded rationality model [26] – a recursive theory of mind model. There exists experimental evidence that supports the idea that humans have bounded rationality and that we are rarely in a Nash equilibrium agents [49, 50]. The level-k model has been applied to domains of more than two agents [51], effectively using theory of mind to the behavior of teams and their formation. This is a difficult problem that can be easily be seen with a mathematical analogy. Let $\mathcal{K}_i$ be the operator that, when operated on some knowledge X, gives the knowledge that agent i has about X, that is $\mathcal{K}_i X$. Thus, theory of mind reasoning can be done using these operators. For a given recursion level k, the number of sequences that can be built using n operators $\mathcal{K}_i, i = 1, ..., n$ is $n(n-1)^k$. For the two agent case, this is a constant number with respect to the recursion level and a policy hierarchy is relatively easy to build and operate with. However, for $n > 2$, the number of such sequences is exponential with respect to the recursion depth, because it includes types of reasoning

such as "What 1 thinks that 3 thinks about 2's policy". Because of this, **we will, in this thesis, focus on studying the coordination between two risk-sensitive agents with theory of mind**.

Humans are able to cooperate and coordinate easily due to our evolved intuition [52] and the existence of cultural norms [53] and signals, such as social cues [54] and explicit communication [55]. Some of these mechanisms have been studied and implemented, showing that cooperation of machines with humans and other machines is possible and that, in some cases, it can rival human cooperation using simple learning rules [56].

A major motivation for this thesis was the work in [1]. In it, a formalism known as Linearly-solvable Markov Decision Process (LMDP) (a particular type of MDP) is generalized to the two agent setting and coordination is studied in a scenario similar to the Stag Hunt game. The agents use EUT and are equipped with a level-k model. **They showed coordination increases for increasing sophistication levels of both agents (i.e., higher levels of** k. The LMDP formalism, when extended to the two agent setting, makes the assumption that **agents make repeated decisions in a sequential manner, with agent 1 playing first, in their case**. In this thesis, we use the Markov game framework, a more general framework[5] which is used in learning models [57], to model repeated decisions made in a simultaneous manner. We would also like to note that the rationale in the level-k model should be different in settings of simultaneous and sequential decisions, to account for the added information of the agent who has seen the decision of the other. We will draw inspiration from [1] and study the Markov game version of their stag hunt analogue.

3.4 Summary of Related Work

In this chapter, we have briefly reviewed the literature on risk-sensitive measures, and provided an explanation to why we use CPT as our theory of value, despite there being some evidence that it may not be the perfect descriptor of human behavior. We have seen how CPT is applied to normal-form games and to MDPs in the discounted infinite-horizon case. Based on the reviewed literature, the reader may start to think that descriptive theories of behavior can be used to obtain prescriptive theories in multi-agent settings, such as normal-form games and Markov games.

We draw inspiration from [1], in which was shown that coordination increases with increasing sophistication levels, when two EUT-agents were equipped with level-k models, and played sequentially in an LMDP framework. Appendix A provides a brief explanation of their results and further research concerning the sequential nature of the LMDP and level-k models.

[5] Markov games generalize MDPs while LMDPs are a particular case of them.

4

Model

Contents

4.1 Normal-form Stag Hunt 32

4.2 Markov Game with CPT-Value 33

We will study the coordination of risk-sensitive agents in two settings. First, a normal-form game of stag hunt will give us some idea about how this paradigmatic example changes when agents value actions using CPT rather than the usual EUT. Second, we will create a Markov game analogue of the stag hunt game to gain insight into how two risk-sensitive agents (using CPT to value actions), equipped with a recursive theory of mind (level-k), coordinate in a stochastic environment, where risk comes not only from the actions of the other agent but also from the environment itself.

4.1 Normal-form Stag Hunt

We will look at coordination in a particular instance of the game of stag hunt, with rewards defined as in Table 2.2. We will identify the stochastic Nash equilibrium when both agents calculate the value of actions with EUT and compare it with the CPT-equilibrium, when both agents calculate the value of actions with CPT. Since the stag hunt game is a symmetric game (the rewards are similar for both agents), the optimal policy of agent 1 is equal to the optimal policy of agent 2. The stag hunt game has only two actions, and therefore, a policy can be fully specified by the probability that agent 1 (or agent 2) choosing to hunt stag, which we will denote as $p = \mathbb{P}(\text{Agent } 1 \text{ plays } S)$. **Probability p is then a full descriptor of the equilibria**, which can be obtained for both types of agents. We will refer to this probability as p^{EUT} and p^{CPT} for the Nash equilibrium and the CPT-equilibrium, respectively, and to the value under EUT and under CPT as V^{EUT} and V^{CPT}, respectively. Both equilibria can be obtained by solving these two systems of equations as follows:

$$\begin{aligned} V^{\text{EUT}}(S, p^{\text{EUT}}) &= u(5)p + u(0)(1-p), \\ V^{\text{EUT}}(H, p^{\text{EUT}}) &= u(1), \\ V^{\text{EUT}}(S, p^{\text{EUT}}) &= V^{\text{EUT}}(H, p^{\text{EUT}}), \text{ and} \end{aligned} \tag{4.1}$$

$$\begin{aligned} V^{\text{CPT}}(S, p^{\text{CPT}}) &= u(5)w(p) + u(0)w(1-p), \\ V^{\text{CPT}}(H, p^{\text{CPT}}) &= u(1), \text{ and} \\ V^{\text{CPT}}(S, p^{\text{CPT}}) &= V^{\text{CPT}}(H, p^{\text{CPT}}). \end{aligned} \tag{4.2}$$

In this case, a measure of coordination can be seen as the probability that both agents choose the same action, $\mathbb{P}(a_1 = S, a_2 = S) + \mathbb{P}(a_1 = H, a_2 = H) = p^2 + (1-p)^2$, which is maximal for $p = 0$ and $p = 1$ and minimal for $p = 0.5$.

It is also worth comparing the **expected total reward** given by

$$\mathbb{E}[r_1(p) + r_2(p)|p] = (5+5)p^2 + 2p(1-p) + (1+1)(1-p)^2, \tag{4.3}$$

of both equilibria, p^{EUT} and p^{CPT}, because even when agents coordinate, they may choose a sub-optimal

way of doing so.

In this model, we considered $u(r) = r$, $w(p) = e^{-0.5(-\log(p))^{0.9}}$, and $b_1 = b_2 = 0$ to ensure the theories of value (EUT and CPT) differ only on the perception of probabilities.

4.2 Markov Game with CPT-Value

The CPT-value in an MDP can be recursively calculated with Equation (3.1). Extending this scheme to a Markov game with n agents with joint action space $\boldsymbol{\mathcal{A}} = \mathcal{A}_1 \times ... \times \mathcal{A}_n$ (where $\mathcal{A}_i$ is the action space of agent i) in a joint state space $\boldsymbol{\mathcal{S}} = \mathcal{S}_1 \times ... \times \mathcal{S}_n$ (where $\mathcal{S}_i$ is the state space of agent i) with stochastic dynamics prescribed by the transition function $P_{\boldsymbol{s}}^{\boldsymbol{a}}(\cdot) = \mathbb{P}(\cdot|\boldsymbol{s}_t = \boldsymbol{s}, \boldsymbol{a}_t = \boldsymbol{a}), (s_1, ..., s_n) = \boldsymbol{s} \in \boldsymbol{\mathcal{S}}, (a_1, ..., a_n) = \boldsymbol{a} \in \boldsymbol{\mathcal{A}}$ is relatively straightforward. The CPT-value that agent i places on a joint state $(s_1, ..., s_n) = \boldsymbol{s} \in \boldsymbol{\mathcal{S}}$, given a joint policy $\boldsymbol{\pi} = (\pi_i, \boldsymbol{\pi}_{-i})$ can be obtained via successive iterations of

$$\begin{aligned} V_i^{\boldsymbol{\pi}}(\boldsymbol{s}) = &\int_0^\infty w_i^+ \left(\sum_{\boldsymbol{a} \in A(\boldsymbol{s})} P_{\boldsymbol{s}}^{\boldsymbol{a}} \left(u_i^+ \left((r_i(\boldsymbol{s}) + \beta_i V_i^{\boldsymbol{\pi}}(\boldsymbol{S}) - b_i)_+ \right) > \epsilon \right) \boldsymbol{\pi}(\boldsymbol{a}|\boldsymbol{s}) \right) d\epsilon \\ &- \int_0^\infty w_i^- \left(\sum_{\boldsymbol{a} \in A(\boldsymbol{s})} P_{\boldsymbol{s}}^{\boldsymbol{a}} \left(u_i^- \left((r_i(\boldsymbol{s}) + \beta_i V_i^{\boldsymbol{\pi}}(\boldsymbol{S}) - b_i)_- \right) > \epsilon \right) \boldsymbol{\pi}(\boldsymbol{a}|\boldsymbol{s}) \right) d\epsilon, \end{aligned} \tag{4.4}$$

where $r_i(\boldsymbol{s})$ is the reward given to agent i on joint state $\boldsymbol{s}$, b_i is the reference point of agent i, β_i is the discount factor of agent i, u_i^+ and u_i^- are the utility functions of agent i relative to the gains and losses, respectively, and w_i^+ and w_i^- are the probability weighting functions of agent i relative to the gains and losses. Each agent i will rationally choose a policy π_i that maximizes $V_i^{\boldsymbol{\pi}}$. To make this clearer, we can rewrite Equation (4.4) as

$$\begin{aligned} V_i^{\pi_i, \boldsymbol{\pi}_{-i}}(\boldsymbol{s}) = &\int_0^\infty w_i^+ \left(\sum_{a_i \in \mathcal{A}_i} P_{i,\boldsymbol{s},+}^{a_i, \boldsymbol{\pi}_{-i}}(\epsilon) \pi_i(a_i|\boldsymbol{s}) \right) d\epsilon \\ &- \int_0^\infty w_i^- \left(\sum_{a_i \in \mathcal{A}_i} P_{i,\boldsymbol{s},-}^{a_i, \boldsymbol{\pi}_{-i}}(\epsilon) \pi_i(a_i|\boldsymbol{s}) \right) d\epsilon, \\ \text{where } P_{i,\boldsymbol{s},+}^{a_i, \boldsymbol{\pi}_{-i}}(\epsilon) = &\sum_{\boldsymbol{a}_{-i} \in \boldsymbol{\mathcal{A}}_{-i}(\boldsymbol{s})} P_s^{a_i, \boldsymbol{a}_{-i}}(u_i^+((r_i(\boldsymbol{s}) + \beta_i V_i^{\pi_i, \boldsymbol{\pi}_{-i}}(\boldsymbol{S}) - b_i)_+ > \epsilon) \boldsymbol{\pi}_{-i}(\boldsymbol{a}_{-i}|\boldsymbol{s}) \\ \text{and } P_{i,\boldsymbol{s},-}^{a_i, \boldsymbol{\pi}_{-i}}(\epsilon) = &\sum_{\boldsymbol{a}_{-i} \in \boldsymbol{\mathcal{A}}_{-i}(\boldsymbol{s})} P_s^{a_i, \boldsymbol{a}_{-i}}(u_i^-((r_i(\boldsymbol{s}) + \beta_i V_i^{\pi_i, \boldsymbol{\pi}_{-i}}(\boldsymbol{S}) - b_i)_- > \epsilon) \boldsymbol{\pi}_{-i}(\boldsymbol{a}_{-i}|\boldsymbol{s}). \end{aligned} \tag{4.5}$$

Each agent i will maximize the functional CPT-value $V_i^{\pi_i, \boldsymbol{\pi}_{-i}}$ in every joint state $\boldsymbol{s}$ by choosing the appropriate policy π_i, given the joint policy $\boldsymbol{\pi}_{-i}$. Formally, $\pi_i = \underset{\pi_i'}{\text{argmax}}\, V_i^{\pi_i', \boldsymbol{\pi}_{-i}}(\boldsymbol{s})$ for each joint state $\boldsymbol{s} \in \boldsymbol{\mathcal{S}}$. Optimizing the improper integrals may at first seem a daunting task. To do so, we capitalize on the fact that the survival functions are piece-wise constant, effectively letting us rewrite the improper

integrals as a finite sum. Let $\{\epsilon_k^+\}_{k=0}^{K^+}$ and $\{\epsilon_k^-\}_{k=0}^{K^-}$ be the ordered sets of atoms[1] of the survival functions $P_{i,\boldsymbol{s},+}^{a_i,\boldsymbol{\pi}_{-i}}$ and $P_{i,\boldsymbol{s},-}^{a_i,\boldsymbol{\pi}_{-i}}$, respectively. Replacing the integrals by sums, Equation (4.5) becomes

$$\begin{aligned} V_i^{\pi_i,\boldsymbol{\pi}_{-i}}(s) = &\sum_{k=1}^{K^+} w_i^+ \left(\sum_{a_i \in A_i(s)} P_{i,\boldsymbol{s},+}^{a_i,\boldsymbol{\pi}_{-i}}(\epsilon_k^+)\pi_i(a_i|s) \right) (\epsilon_k^+ - \epsilon_{k-1}^+) \\ &- \sum_{k=1}^{K^-} w_i^- \left(\sum_{a_i \in A_i(s)} P_{i,\boldsymbol{s},-}^{a_i,\boldsymbol{\pi}_{-i}}(\epsilon_k^-)\pi_i(a_i|s) \right) (\epsilon_k^- - \epsilon_{k-1}^-), \end{aligned} \tag{4.6}$$

which simplifies the non-linear optimization problem to,

$$\begin{aligned} &\text{given } \boldsymbol{\pi}_{-i}, u_i^+, u_i^-, w_i^+, w_i^-, \beta_i, b_i, P_{\boldsymbol{s}}^{a_i,\boldsymbol{a}_{-i}}, \\ &\text{find } \pi_i = \underset{\pi_i'}{\text{argmax}}\, V_i^{\pi_i',\boldsymbol{\pi}_{-i}}(\boldsymbol{s}), \forall \boldsymbol{s} \in \boldsymbol{\mathcal{S}}, \\ &\text{subject to } \sum_{a_i \in \mathcal{A}_i} \pi_i(a_i|\boldsymbol{s}) \text{ and } \pi_i(a_i, \boldsymbol{s}) \geq 0, \forall \boldsymbol{s} \in \boldsymbol{\mathcal{S}}, \end{aligned} \tag{4.7}$$

which we can solve using Python's `scipy` implementation of SLSQP (Sequential Least Squares Quadratic Programming).

4.2.1 Markov Stag Hunt

We wish to capture the essence of coordination in a setting where time is relevant. To accomplish this, we will create a Markov game version of the stag hunt in which there are $n = 2$ agents in the joint state space $\boldsymbol{\mathcal{S}} = \mathcal{S}_1 \times \mathcal{S}_2$ (with $\mathcal{S}_1 = \mathcal{S}_2 = \{0, ..., 15\}$) with joint action space $\boldsymbol{\mathcal{A}} = \mathcal{A}_1 \times \mathcal{A}_2$ (with $\mathcal{A}_1 = \mathcal{A}_2 = \{L, S, R\}$), stochastic dynamics from Figure 4.1 (see Figure 2.4 for a similar, single agent state diagram) and reward structure from Figure 4.2. Intuitively, this creates a stochastic environment where two agents, at each time step, choose to move to the left, right or stay in their current state, and receive the reward of that state.

The similarity to the stag hunt can be found in the reward structure; any agent receives a reward of 1 at state 3 (hunting hares), a reward of 5 is given to the agents only if both are in state 11 (hunting stags), and all other states gives a reward of 0 (nothing hunted).

To create a simultaneous decision situation, the dynamics in the joint state space and the individual transition probabilities in Figure 4.1 must be combined in a proper manner, i.e.,

$$P^{a_1,a_2} = \frac{I \otimes P^{a_1} + P^{a_2} \otimes I}{2}, \tag{4.8}$$

where the Kronecker product $\otimes$ ensures an action from agent 1 does not change the state of agent 2

[1] The atoms in both sets are ordered in an increasing manner, i.e. $\epsilon_k > \epsilon_{k-1} \forall k$.

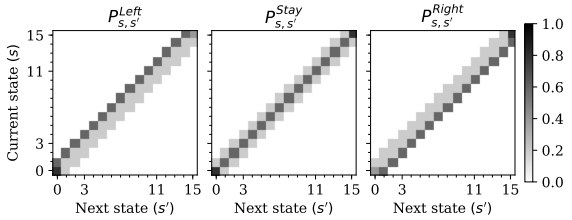

Figure 4.1: The transition matrices of the Markov game, for the three available actions L, S, and R, where L stands for going left, S stands for staying in the same state and R stands for going right. An agent can move to adjacent states or remain by choosing one of these three available actions, but the outcome is not deterministic, i.e. there is a chance of failure. For instance, choosing L does not guarantee that the agent moves to the left and may end up staying in the same place or going the opposite direction.

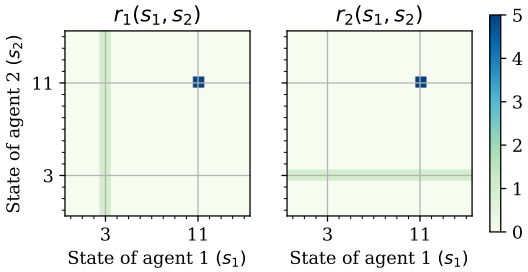

Figure 4.2: The reward functions of both agents. They are symmetrical like the payoff matrix of the stag hunt. State 3 gives reward of $+1$ to an agent in it regardless of the state of the other agent, to represent the hares that can be captured alone but offer little reward. State 11 gives reward of $+5$ to the two agents if they are both in state 11. All other states give a reward of 0.

and vice-versa. The average of this transformed agent transition probability function ensures the joint state space dynamics is independent of who acts first.

Each agent is equipped with a level-k model, which lets them create increasingly sophisticated policies $\pi_1^{(k_1)}$ and $\pi_2^{(k_2)}$, where k_1 and k_2 are the sophistication levels of agent 1 and agent 2, respectively. To simplify the discussion, we will assume that both agents have similar value functionals. Formally, this means that $u_i^+ = u_i^- = u$, $w_i^+ = w_i^- = w$, $\beta_i = \beta$ and $b_i = b$ for $i = 1, 2$, such that $V_1 = V_2 = V$. With this assumption we are saying that both hunters value outcomes in the same exact way. Level-k then dictates that policies of sophistication level k for both agents are obtained via

$$\begin{aligned} \pi_1^{(k)} &= \underset{\pi_1}{\text{argmax}}\, V^{\pi_1, \pi_2^{(k-1)}}(\boldsymbol{s}), \forall \boldsymbol{s} \in \boldsymbol{\mathcal{S}}, \\ \pi_2^{(k)} &= \underset{\pi_2}{\text{argmax}}\, V^{\pi_1^{(k-1)}, \pi_2}(\boldsymbol{s}), \forall \boldsymbol{s} \in \boldsymbol{\mathcal{S}}, \end{aligned} \tag{4.9}$$

and we will assume $\pi_1^{(0)}$ and $\pi_2^{(0)}$ are uniform policies, i.e., policies that choose L,S or R with probability $\frac{1}{3}$, independently of the state.

We will analyze the behavior of both agents as a function of the sophistication order of their policies. To do so, we will make use of the fact that the Markov Stag Hunt game, when conditioned on a joint policy $\boldsymbol{\pi}^{(k_1,k_2)} = (\pi_1^{(k_1)}, \pi_2^{(k_2)})$, is a time-homogeneous, discrete-time, irreducible Markov chain whose

states are all positive recurrent with transition function

$$P_{\boldsymbol{s},\boldsymbol{s}'}^{\pi_1^{(k_1)},\pi_2^{(k_2)}} = \sum_{\substack{a_1 \in \mathcal{A}_1 \\ a_2 \in \mathcal{A}_2}} P_{\boldsymbol{s},\boldsymbol{s}'}^{a_1,a_2} \pi_1^{(k_1)}(a_1|\boldsymbol{s}) \pi_2^{(k_2)}(a_2|\boldsymbol{s}), \forall \boldsymbol{s} \in \boldsymbol{\mathcal{S}}. \tag{4.10}$$

Thus, given $\pi_1^{(k_1)}$ and $\pi_2(k_2)$, there exists a stationary distribution $\rho(k_1, k_2)$ such that

$$\rho(k_1, k_2) = P^{\pi_1^{(k_1)},\pi_2^{(k_2)}} \rho(k_1, k_2). \tag{4.11}$$

The stationary distribution $\rho(k_1, k_2)$ allows us to identify in which states the two agents will most likely be in the long run, regardless of whether the two agents start.

In this model, we considered $u(r) = r$, $w(p) = e^{-0.5(-\log(p))^{0.9}}$, and $b = 0$ to ensure the theories of value (i.e., EUT and CPT) differ only on the perception of probabilities. Furthermore, we assume both agents use a discount factor of $\beta = 0.9$.

5

Results

Contents

5.1 Normal-form Stag Hunt 38

5.2 Markov Stag Hunt 38

5.3 Limitations 42

We evaluate the models defined in the previous chapter. In the first model, we use the stochastic Nash equilibrium and the CPT-equilibrium to understand how coordination differs when agents are using EUT and CPT as theories of value. In the second model, we show the effect of increasingly sophisticated policies of the level-k model on the coordination of EUT- and CPT-agents.

5.1 Normal-form Stag Hunt

Let us assume that utility functions are identical and that the only difference between EUT and CPT is the way agents perceive likelihoods which are captured by the probabilities of an outcome. If, from the perspective of one agent, the other agent will choose S with probability p, then the values of his actions, under EUT and CPT, are given by

$$\begin{aligned} &V^{\mathsf{EUT}}(S) = 5p, V^{\mathsf{EUT}}(H) = 1, \text{ and} \\ &V^{\mathsf{CPT}}(S) = 5w(p), V^{\mathsf{CPT}}(H) = 1. \end{aligned} \tag{5.1}$$

For EUT and CPT, the Nash equilibria correspond to the probabilities p^{EUT} and p^{CPT} that make $V^{\mathsf{M}}(S) = V^{\mathsf{M}}(H)$ for $M \in \{\mathsf{EUT}, \mathsf{CPT}\}$, respectively. Also, we consider $w(x) = \exp\{-0.5(-\log(x))^{0.9}\}$, so we get $p^{\mathsf{EUT}} = 0.2$ and $p^{\mathsf{CPT}} \approx 0.028$.

Our results show that CPT-agents in a normal-form stag hunt game choose to coordinate to hunt hares with higher probability than EUT-agents. Consequently, it readily follows that **CPT increases coordination in the stag hunt game**, since both agents choose the same action more often. In other words, whereas two hunters hunting a stag yields the largest reward, the risk of hunting a stag alone is overshadowed by the safety of hunting a hare.

While hunting hares is sub-optimal, the average sum of rewards does not decrease substantially from the EUT-agents, i.e.,

$$\begin{aligned} &\mathbb{E}_{a_1,a_2}[r_1^{\mathsf{EUT}}(a_1,a_2) + r_2^{\mathsf{EUT}}(a_1,a_2)] = 2, \text{ and} \\ &\mathbb{E}_{a_1,a_2}[r_1^{\mathsf{CPT}}(a_1,a_2) + r_2^{\mathsf{CPT}}(a_1,a_2)] \approx 1.95. \end{aligned} \tag{5.2}$$

This is because an increase in coordination reduces the likelihood of either hunter choosing to hunt stags alone and, consequently, getting a reward of zero. This suggests that CPT-agents are more risk-averse (in a one-shot setting) than EUT-agents, a feature also seen in humans.

5.2 Markov Stag Hunt

The conflict between short- and long-term rewards is of particular interest in domains where time is a relevant factor, and it is also in these domains where a theory of mind may prove useful. To that end,

we studied how EUT- and CPT-agents coordinate in a Markov game version of stag hunt inspired by [1], where both types of agents were equipped with a level-k theory of mind. Hence, predicting several (increasingly sophisticated) behaviors in the form of policies.

The Markov stag hunt game policies – resulting from the optimization problem in Equation (4.7) – can be found in Figure 5.1. There are clear differences between the policies of EUT- and CPT-agents. A good summary of the differences in behavior can be seen by looking at the probability of staying (i.e., choosing S, corresponding to the second and fifth columns of Figure 5.1) – the size of the rightmost green bar indicates the attractiveness of staying at the stag state, which is much larger for CPT agents than it is for EUT agents.

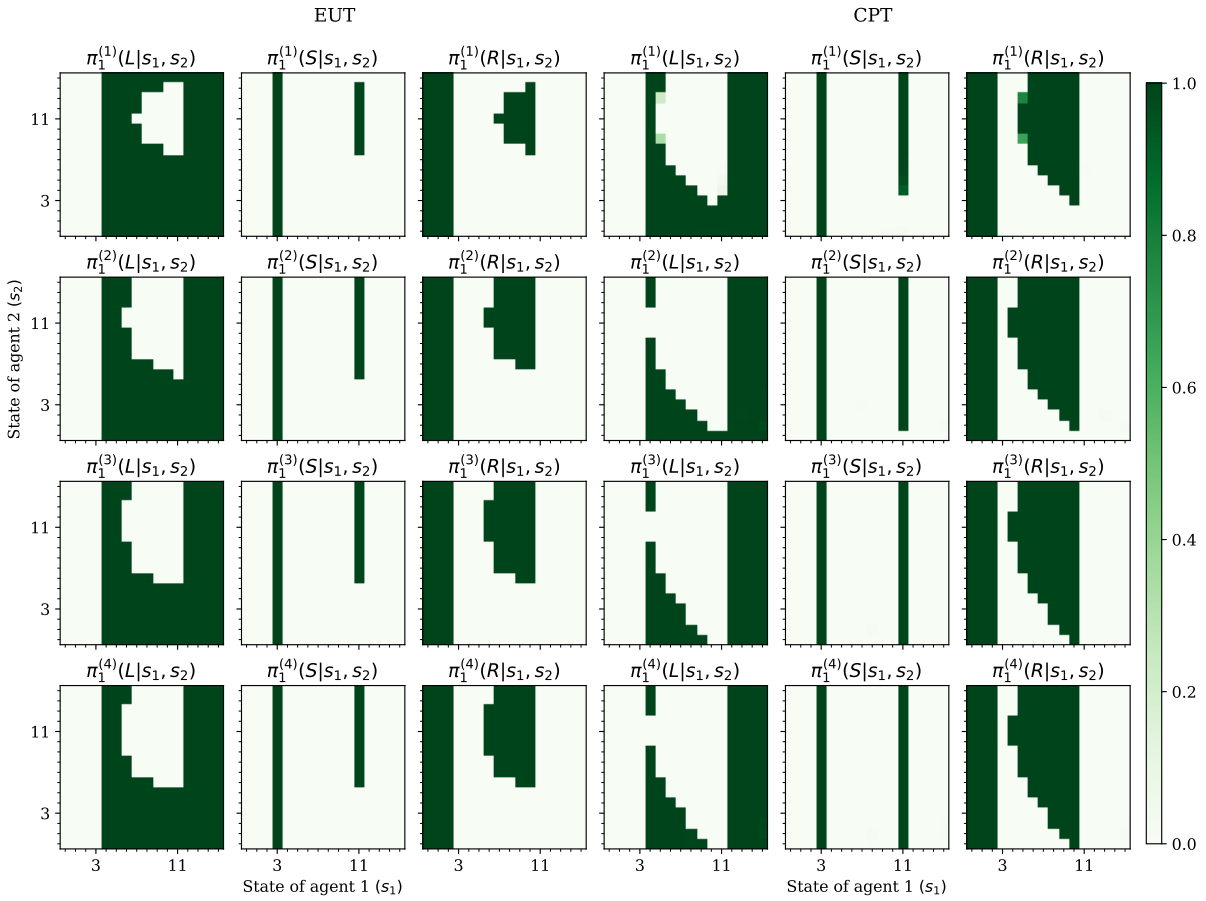

Figure 5.1: Resulting EUT and CPT policies of agent 1 as functions of the agent states, for sophistication levels $k = 1, 2, 3, 4$. Due to the symmetry of the game, the policies for agent 2 are the transpose of these policies.

The results in Figure 5.2 shows that both EUT- and CPT-agents place increasingly more value on the stag state but the latter place substantially more value on the stag state – even in the lowest sophistication levels (i.e., low values of k).

Stationary distributions allow us to easily compare dynamics of EUT- and CPT-agents. Specifically,

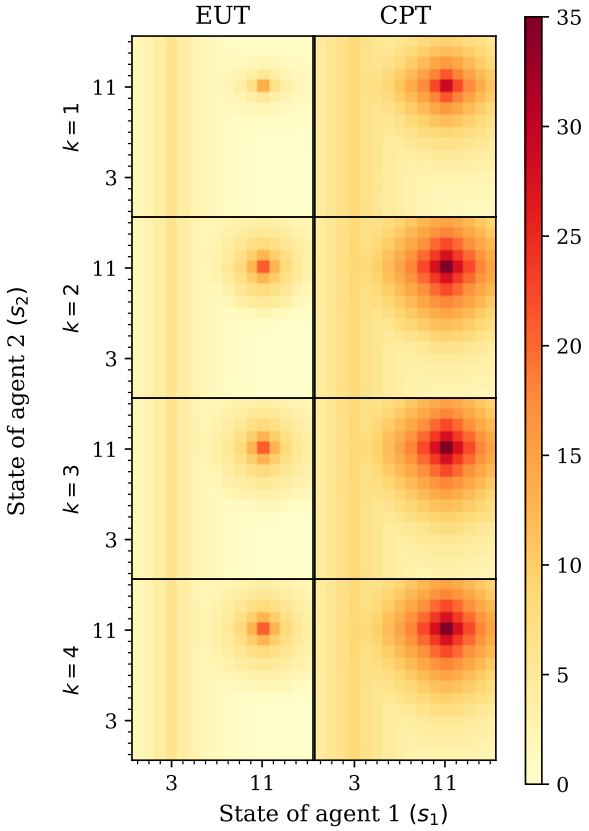

Figure 5.2: EUT- and CPT-values as functions of the agent states, s_1 and s_2, for sophistication levels $k = 1, 2, 3, 4$. We assumed reference points $b_1 = b_2 = 0$, discount factors $\beta_1 = \beta_2 = 0.9$, utility function $u(x) = x$ and weighting function $w(x) = x$ for EUT and $w(x) = e^{-0.5(-\log(x))^{0.9}}$ for CPT

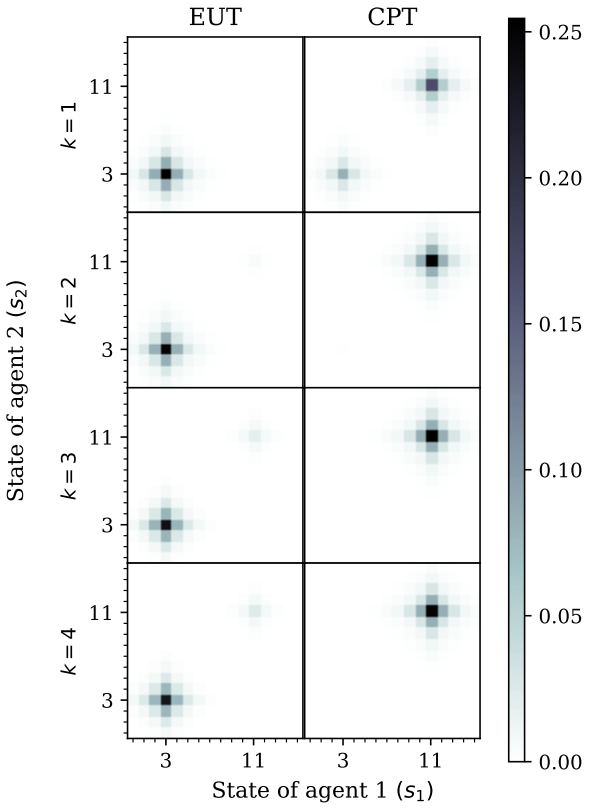

Figure 5.3: Stationary distributions of the resulting Markov chains obtained by conditioning the Markov game to increasingly sophisticated policies, $k = 1, 2, 3, 4$, for EUT- and CPT-agents. We assumed reference points $b_1 = b_2 = 0$, discount factors $\beta_1 = \beta_2 = 0.9$, utility function $u(x) = x$ and weighting function $w(x) = x$ for EUT and $w(x) = e^{-0.5(-\log(x))^{0.9}}$ for CPT.

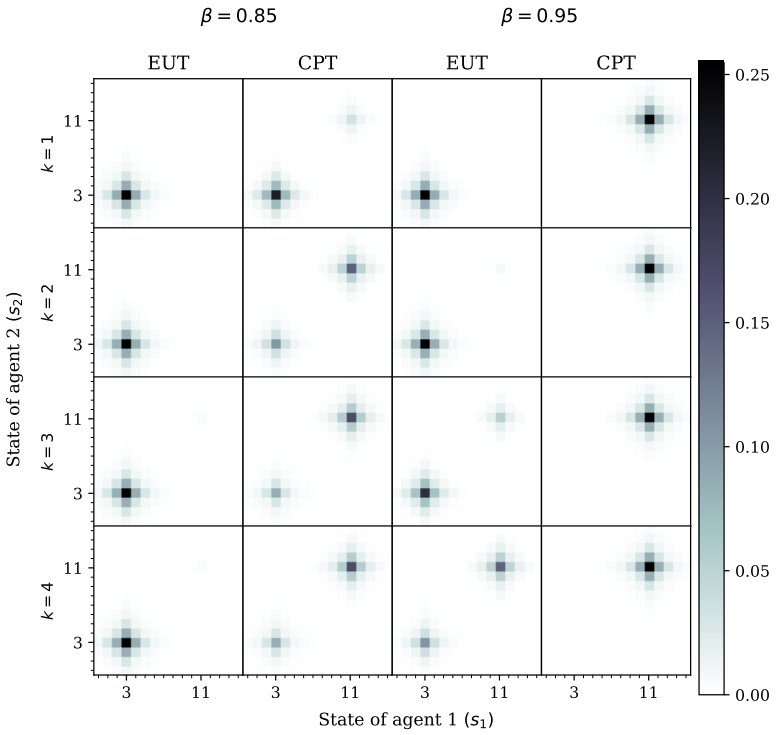

Figure 5.4: Stationary distributions of the resulting Markov chains obtained by conditioning the Markov game to increasingly sophisticated policies, $k = 1, 2, 3$ and 4, for EUT-agents and CPT-agents. We assumed reference points $b_1 = b_2 = 0$, utility function $u(x) = x$ and weighting function $w(x) = x$ for EUT and $w(x) = e^{-0.5(-\log(x))^{0.9}}$ for CPT. (Left) Stationary distribution for EUT- and CPT-agents using discount factor $\beta = 0.85$. (Right) Stationary distribution for EUT- and CPT-agents using discount factor $\beta = 0.85$.

from Figure 5.3, we have that while both EUT- and CPT-agents eventually prefer state 11 (hunting stags) with increasing sophistication levels, CPT-agents dramatically do so. In fact, CPT-agents have a strict preference for the stag state even for sophistication level $k = 1$, where both agents assume the other is using a random policy.

These results suggest that, with increasing k (the sophistication level of policies), CPT-agents coordinate better and choose the optimal stag state, whereas EUT-agents fail to do so.

To analyze the robustness to parametric choices in our setting, we further studied the sensitivity of the coordination to the parameters of the model. We specifically looked at the discount factor β and the reference point b. Figure 5.4 provides evidence that increasing the discount factor (thus increasing the perceived "goodness" of long-term rewards) also increases coordination of both EUT- and CPT-agents, and that the latter still generate more coordination than EUT.

Additionally, we have considered different sophistication levels and reference points. This analysis is captured in Figure 5.5, where we show the stationary distribution of CPT-agents for different sophistication levels and reference points. It readily follows that higher reference points decrease coordination. In other words, hunting stags is perceived as a not-so-good solution when agents have a negative skewed

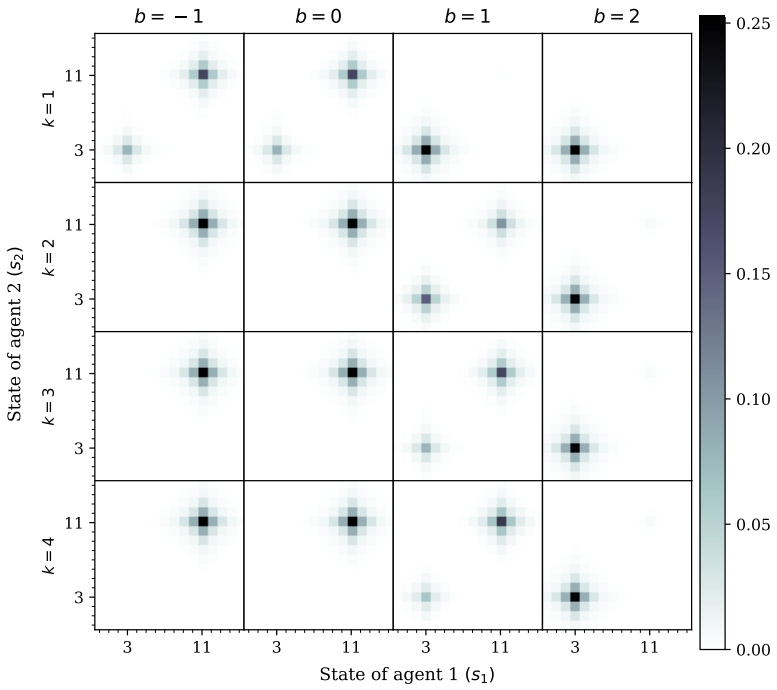

Figure 5.5: Stationary distributions of the resulting Markov chains obtained by conditioning the Markov game to increasingly sophisticated policies, $k = 1, 2, 3$ and 4, for CPT-agents with several reference points $b = -1, 0, 1, 2$. We assumed discount factors $\beta_1 = \beta_2 = 0.9$, utility function $u(x) = x$ and weighting function $w(x) = e^{-0.5(-\log(x))^{0.9}}$.

view of the rewards. Notice that CPT-agents with a higher reference point have a more bleak perception of rewards.

5.3 Limitations

Unfortunately, obtaining the solution to the optimization problem defined in Equation 4.7 is often a daunting task as the numerical approaches suffer from (well known) instability issues for some initial configurations. Consequently, when this occurred, a different but valid initial configuration was selected at random until convergence.

Additionally, the weighting function $w(x) = e^{-0.5(-\log(x))^{0.9}}$ is both computationally expensive and its implementation has to be truncated as $x \to 0$. Therefore, a posynomial approximation $w(x) = 0.00231x^{0.05} + 0.00128x^{0.1} + 0.19578x^{0.35} + 0.59897x^{0.4} + 0.15968x^{0.95} + 0.03318x^{3} + 0.00847x^{23}$ was used, similar to [58].

The reader should also be made aware that the optimization algorithm itself is slow and relatively unstable to some parameter configurations, and therefore, theoretical work on techniques regarding CPT value optimization would prove useful and allow us to readily study agent-based systems with

more than two agents and at more extreme parameter configurations.

6

Conclusion

Contents

6.1 Summary 46

6.2 Suggestions for Future Work 47

6.3 Implications 48

In Chapter 1, we defined two research questions:

- **Q1** - Can cognitive biases concerning risk promote coordination?
- **Q2** - Can increasingly sophisticated levels of theory of mind promote coordination?

In this last chapter, we provide answers to these questions based on the results and discussion in the previous chapters and provide the reader with possible future investigations based on this work.

6.1 Summary

In this thesis, a brief introduction to expected utility theory (EUT), cumulative prospect theory (CPT) and level-k bounded rationality considered in the theory of mind was presented and we stated the relevance of using behavioral models in domains where agents mimic human decisions.

To seek the answer to **Q1**, we studied the normal-form stag hunt game with agents measuring value using CPT and compared the results with agents measuring value using EUT. **Our results suggest CPT-agents in a normal-form stag hunt game coordinate by choosing to hunt hares with higher probability than EUT-agents**. While hunting hares is sub-optimal, the total reward does not decrease substantially from the behavior of the EUT-agents because the probability of hunting stags alone also decreases. This further suggests the risk aversion of human-like agents in a one-shot setting.

However, the conflict between short- and long-term rewards is of particular interest in domains where time is a relevant factor, and it is also in these domains where a theory of mind may prove useful. To answer **Q2**, we studied how EUT- and CPT-agents coordinate in a Markov game version of stag hunt inspired by [1], where both types of agents were equipped with a level-k theory of mind model, that predicts several, increasingly sophisticated behaviors, in the form of policies. **Our results suggest that, with increasing k (the sophistication of policies), CPT-agents coordinate faster and choose the optimal stag state, whereas EUT-agents fail to do so.**

This is a remarkable finding, suggesting that the use of *homo economicus* in multi-agent systems (MAS) may be ruling out on these naturally occurring coordinating behaviors due to their focus on optimality. In fact, most MAS applications use EUT due to the parsimonious mathematical model it provides. Therefore, a shift toward more human-like behavior models may prove useful in this setting.

Cognitive biases and theory of mind are a fundamental part of being human. We have shown that, by including cognitive biases and theory of mind in the dynamics of a coordination game, agents are able to coordinate much more easily.

Furthermore, **increasingly sophisticated policies in the context of bounded rationality (i.e., increasing value of level-k theory of mind) help coordination between both EUT- and CPT-agents**. Additionally, we also provided evidence that higher sophistication levels (i.e., higher than $k = 3$) do not

seem to change the outcome of the two agent setting in the long run. Thus, this latter provides more evidence that **unbounded rationality is not only practically unfeasible, but also unnecessary for coordination**.

Also, we have shown how the consideration of long-term rewards over short-term ones affects the coordination of EUT- and CPT-agents. Specifically, for lower values of the discount factor β, agents will increasingly prefer short-term rewards over long-term rewards. Besides, we have shown that **preferring short-term rewards inhibits the coordination of both EUT- and CPT-agents, while the opposite promotes coordination**. Furthermore, we remarkably observed that **more sophisticated policies in the theory of mind help agents coordinate, even if the long-term reward consideration makes it unlikely at first**.

Lastly, we looked at the sensitivity of the coordination to the reference points of the agents. In particular, we have observed that **higher reference points decrease coordination**, which suggests that the framing of gains and losses plays an important role in the emergence of human coordination.

6.2 Suggestions for Future Work

As we have shown, behavioral agent models provide significantly different system dynamics compared to prescriptive agent models, and therefore, several interesting research directions naturally arise. For instance, multi-agent systems where agents represent people should use a descriptive behavioral model instead of a prescriptive model. Upon realizing this, one can start to develop and study human-based models such as idealized forms of democracy (e.g., liquid democracy [59]), video-game artificial intelligence with human-like behavior (or that is able to understand human-like behavior) and policy-making, or even revisiting already known conflict problems such as the tragedy of the commons and the diffusion of responsibility.

It would also prove interesting to create an inference model to obtain the optimal parameters of this model, similar to [1]. For instance, a Bayesian method to infer the reference point, discount factor, utility and weighting function parameters, and policy sophistication level would enable machines to learn to act in a more personalized manner.

One caveat of the bounded rationality using a level-k model is the assumption that stereotype policies are uniform, which may be rather unrealistic. Therefore, a way of creating more realistic stereotyped policies would be an interesting problem to tackle. One such way is self-play, a reinforcement learning method to train agents by pitting them against themselves and, in an evolutionary manner, preserving winners and discarding losers [60].

In the two-agent level-k model, it is known that humans, in general, do not use more sophistication than level-3 [49]. This creates a finite hypothesis space for the policy levels (i.e., with $k = 0, 1, 2$, and 3).

However, when multiple interacting agents are a part of the environment, it is not enough to specify policy levels as a single number because each agent may have a policy which is a best response against several other policies of different levels. Therefore, there exists a problem of finding a behaviorally plausible hypothesis space for the inferred orders of each agent, which, if solved, would allow inference to be done on a collective level. Specifically, we would like reasoning such as "what you think about what he things that she thinks..." to be described in a simple, yet well-structured manner. The team theory of mind model proposed in [51] is an interesting setting that tackles some of the problems but its solution is computationally costly. At last but not least, experimental verification of the proposed framework could be done via a sociological study, which may also generate interesting data to further validate and expand the proposed model. These and other related research paths may lead to new knowledge of the dynamics of systems comprised of people and, in turn, unlock the knowledge we lack to build artificial entities capable of understanding and simulating human behavior.

6.3 Implications

The field of **affective computing** has become increasingly important to businesses and governments due to the unique ability to use emotional and social intelligence to inform decision-making, tapping into a hidden realm of signals previously only accessible to humans. It was the purpose of this thesis to contribute to this and peripheral fields of research, by studying the effects of cognitive biases and mechanisms such as risk perception and theory of mind on the ability to coordinate.

In important areas of industry, governments and education, machines use cognitive intelligence to solve difficult problems and replace us in time-consuming tasks such as patient diagnosis in healthcare, fraud detection in the financial sector, efficient and automatic anomaly detection in manufacturing, inventory optimization in retail, personalized smarter services in government, demand forecasting in transportation systems like *Uber* and *Lyft*, intrusion detection in IT networks and recommender systems in e-commerce platforms such as Amazon.

Machines capable of emotional and social intelligence have applications to these same fields, but solve problems in which the human element is present. In military applications machines can replace army recruiters by interviewing and selecting human candidates based on emotional cues and more sophisticated machines can psychologically train soldiers before entering a war zone. In stores, machines should automatically identify unhappy shoppers with facial recognition to trigger remedial actions and eventually deal with customer complaints and address concerns of unhappy customers. Current practical applications could also be improved. Recommendation systems using affective data could allow *Spotify* to recommend music and *Netflix* to recommend movies based on your current emotional state.

The main conclusion of this thesis is that, as a consequence of the two conducted experi-

ments, both settings suggest that the reason why humans are good at coordination may stem from the fact that we are cognitively biased to do so. For this reason, machine agents ought to be built to incorporate the cognitive biases of humans if we are to, one day, live among them.

Bibliography

[1] W. Yoshida, R. J. Dolan, and K. J. Friston, "Game theory of mind," *PLoS Computational Biology*, vol. 4, no. 12, 2008.

[2] D. Kahneman and A. Tversky, "On the reality of cognitive illusions." *Psychological review*, vol. 103, no. 3, pp. 582–91, 7 1996.

[3] A. Tversky and D. Kahneman, "The Framing of Decisions and the Psychology of Choice," *Science*, vol. 211, no. 4481, pp. 453–458, 1981.

[4] ——, "Advances in Prospect Theory: Cumulative Representation of Uncertainty," Tech. Rep., 1992.

[5] D. Goleman, *Social Intelligence*. Random House, 2007.

[6] D. Premack and G. Woodruff, "Does the chimpanzee have a theory of mind?" *Behavioral and Brain Sciences*, 1978.

[7] A. Paiva, F. P. Santos, and F. C. Santos, "Engineering Pro-Sociality with Autonomous Agents," *The Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18)*, 2018.

[8] H. R. Sewall, "The Theory of Value before Adam Smith," *Publications of the American Economic Association, 3rd Series*, vol. 2, no. 3, pp. 1–128, 1901.

[9] W. Letwin, *The Origins of Scientific Economics*. Routledge, 2013.

[10] N. Bernoulli, "Correspondence of Nicolas Bernoulli concerning the St. Petersburg Game," 1713. [Online]. Available: https://web.archive.org/web/20150501234331/http://cerebro.xu.edu/math/Sources/NBernoulli/correspondence_petersburg_game.pdf

[11] J. von Neumann and O. Morgenstern, *Theory of Games and Economic Behavior*. Princeton University Press, 1944.

[12] D. Kahneman and A. Tversky, "Prospect Theory: An Analysis of Decision Under Risk," *Econometrica*, vol. 47, no. 2, pp. 263–291, 1979.

[13] M. Allais, “The Foundations of a Positive Theory of Choice Involving Risk and a Criticism of the Postulates and Axioms of the American School (1952),” *Expected utility hypotheses and the Allais paradox*, pp. 27–145, 1979.

[14] D. Ellsberg, “Risk, ambiguity and the Savage Axioms,” *The quarterly journal of economics*, pp. 643–669, 1961.

[15] C. F. Camerer, *Behavioral Game Theory. Experiments in Strategic Interaction*. Princeton University Press, 2011.

[16] J. Hadar and W. R. Russell, “Rules for Ordering Uncertain Prospects,” *The American Economic Review*, vol. 59, no. 1, pp. 25–34, 1969.

[17] R. Aumann and A. Brandenburger, “Epistemic Conditions for Nash Equilibrium,” *Econometrica*, vol. 63, no. 5, pp. 1161–1180, 1995.

[18] A. Rapoport, A. M. Chammah, and C. J. Orwant, *Prisoner’s dilemma: A study in conflict and cooperation*. University of Michigan Press, 1965, vol. 165.

[19] B. Skyrms, *The stag hunt and the evolution of social structure*. Cambridge University Press, 2004.

[20] R. Axelrod, *The Evolution of Cooperation*. Basic Books, 1984.

[21] R. A. Howard, *Dynamic Programming and Markov Processes*, 1960.

[22] L. S. Shapley, “Stochastic Games,” in *Proceedings of the National Academy of Sciences*, 1953, pp. 1095–1100.

[23] T. Gerstenberg and J. B. Tenenbaum, *Intuitive Theories*, M. R. Waldmann, Ed. Oxford University Press, 5 2017, vol. 1.

[24] D. G. Pearce, “Rationalizable Strategic Behavior and the Problem of Perfection,” *Econometrica*, vol. 52, no. 4, pp. 1029–1050, 7 1984.

[25] B. D. Bernheim, “Rationalizable Strategic Behavior,” *Econometrica*, vol. 52, no. 4, pp. 1007–1028, 7 1984.

[26] D. O. Stahl, “Evolution of smartn players,” *Games and Economic Behavior*, vol. 5, pp. 604–617, 1993.

[27] F. H. Knight, *Risk, Uncertainty and Profit*. Courier Corporation, 2012.

[28] Yoe and Charles, *Principles of Risk Analysis Decision Making Under Uncertainty Second Edition*, 2019.

[29] J. Tansey and T. O'riordan, "Cultural theory and risk: a review," Tech. Rep. 1, 1999.

[30] M. Abdellaoui, H. Bleichrodt, and H. Kammoun, "Do financial professionals behave according to prospect theory? An experimental study," *Theory and Decision*, vol. 74, no. 3, pp. 411–429, 2013.

[31] J. A. List, "Neoclassical theory versus prospect theory: Evidence from the marketplace," *Econometrica*, vol. 72, no. 2, pp. 615–625, 2004.

[32] V. Cxvi, N. Barberis, M. Huang, and T. Santos, "Prospect Theory and Asset Prices," *Quarterly Journal of Economics*, no. 1, 2001.

[33] A. Fiegenbaum, "Prospect theory and the risk-return association," *Journal of Economic Behavior & Organization*, vol. 14, no. 2, pp. 187–203, 1990.

[34] B. Vis and K. Van Kersbergen, "Why and how do political actors pursue risky reforms?" *Journal of Theoretical Politics*, vol. 19, no. 2, pp. 153–172, 4 2007.

[35] D. D. Bourgin, J. C. Peterson, D. Reichman, T. L. Griffiths, and S. J. Russell, "Cognitive Model Priors for Predicting Human Decisions," 5 2019. [Online]. Available: http://arxiv.org/abs/1905.09397

[36] J. Quiggin, *Generalized Expected Utility Theory: The Rank Dependent Model*. Springer Netherlands, 1993.

[37] H. Fennema and P. Wakker, "Original and cumulative prospect theory: a discussion of empirical differences," *Journal of Behavioral Decision Making*, vol. 10, no. 1, pp. 53–64, 12 2005.

[38] F. Gul, "A Theory of Disappointment Aversion," *Econometrica*, vol. 59, no. 3, pp. 667–686, 7 1991.

[39] B. Köszegi and M. Rabin, "A Model of Reference-Dependent Preferences," *The Quarterly Journal of Economics*, vol. 121, no. 4, pp. 1133–1165, 2006.

[40] U. Schmidt, C. Starmer, and R. Sugden, "Third-generation prospect theory," *Journal of Risk and Uncertainty*, vol. 36, no. 3, pp. 203–223, 6 2008.

[41] G. W. Harrison and D. Ross, "The empirical adequacy of cumulative prospect theory and its implications for normative assessment," *Journal of Economic Methodology*, vol. 24, no. 2, pp. 150–165, 4 2017.

[42] A. Hofmeyr and H. Kincaid, "Prospect theory in the wild: how good is the nonexperimental evidence for prospect theory?" *Journal of Economic Methodology*, vol. 26, no. 1, pp. 13–31, 1 2019.

[43] A. Kühberger and C. Tanner, "Risky choice framing: Task versions and a comparison of prospect theory and fuzzy-trace theory," pp. 314–329, 2010.

[44] L. P. Metzger and M. O. Rieger, “Non-cooperative games with prospect theory players and dominated strategies,” *Games and Economic Behavior*, vol. 115, pp. 396–409, 5 2019.

[45] W. Zeng, M. Li, and F. Chen, “Cooperation in the evolutionary iterated prisoner’s dilemma game with risk attitude adaptation,” *Applied Soft Computing Journal*, vol. 44, pp. 238–254, 7 2016.

[46] K. Lin and S. I. Marcus, “Dynamic Programming with Non-Convex Risk-Sensitive Measures,” in *American Control Conference*. Washington, DC, USA: IEEE, 2013, pp. 6778–6783.

[47] K. Lin, “Stochastic Systems with Cumulative Prospect Theory,” Ph.D. dissertation, 2013.

[48] S. V. Albrecht and P. Stone, “Autonomous agents modelling other agents: A comprehensive survey and open problems,” 2018.

[49] D. O. Stahl II and P. W. Wilson, “Experimental evidence on players’ models of other players,” Tech. Rep., 1994.

[50] D. O. Stahl and P. W. Wilson, “On Players′ Models of Other Players: Theory and Experimental Evidence,” *Games and Economic Behavior*, vol. 10, no. 1, p. 218–254, 1995.

[51] M. Shum, M. Kleiman-Weiner, M. L. Littman, and J. B. Tenenbaum, “Theory of Minds: Understanding Behavior in Groups Through Inverse Planning,” 1 2019. [Online]. Available: http://arxiv.org/abs/1901.06085

[52] A. Bear, A. Kagan, and D. G. Rand, “Co-evolution of cooperation and cognition: The impact of imperfect deliberation and context-sensitive intuition,” *Proceedings of the Royal Society B: Biological Sciences*, vol. 284, no. 1851, 3 2017.

[53] R. Ghate, S. Ghate, and E. Ostrom, “Cultural norms, cooperation, and communication: taking experiments to the field in indigenous communities,” *International Journal of the Commons*, vol. 7, no. 2, pp. 498–520, 2013.

[54] G. Pezzulo, F. Donnarumma, and H. Dindo, “Human sensorimotor communication: A theory of signaling in online social interactions,” *PLoS ONE*, vol. 8, no. 11, 11 2013.

[55] J. H. Miller, C. T. Butts, and D. Rode, “Communication and cooperation,” *Journal of Economic Behavior & Organization*, vol. 47, pp. 179–195, 2002.

[56] J. W. Crandall, M. Oudah, Tennom, F. Ishowo-Oloko, S. Abdallah, J. F. Bonnefon, M. Cebrian, A. Shariff, M. A. Goodrich, and I. Rahwan, “Cooperating with machines,” *Nature Communications*, vol. 9, no. 1, 12 2018.

[57] M. L. Littman, “Markov games as a framework for multi-agent reinforcement learning,” in *Machine Learning Proceedings*. Elsevier, 1994, pp. 157–163.

[58] M. Cubuktepe and U. Topcu, “Verification of Markov Decision Processes with Risk-Sensitive Measures,” in *Proceedings of the American Control Conference*, vol. 2018-June. Institute of Electrical and Electronics Engineers Inc., 8 2018, pp. 2371–2377.

[59] G. A. O’Donell, “Delegative Democracy,” *Journal of Democracy*, vol. 5, no. 1, pp. 55–69, 1994.

[60] D. Silver, J. Schrittwieser, K. Simonyan, I. Antonoglou, A. Huang, A. Guez, T. Hubert, L. Baker, M. Lai, A. Bolton, Y. Chen, T. Lillicrap, F. Hui, L. Sifre, G. van den Driessche, T. Graepel, and D. Hassabis, “Mastering the game of Go without human knowledge,” *Nature*, vol. 550, no. 7676, pp. 354–359, 2017.

[61] R. D. McKelvey and T. R. Palfrey, “Quantal Response Equilibria for Normal Form Games,” *Games and Economic Behavior*, vol. 10, pp. 6–38, 1994.

[62] ——, “Erratum to: Quantal response equilibria for extensive form games,” *Experimental Economics*, 2015.

[63] E. Todorov, “Linearly-solvable Markov decision problems,” *Advances in Neural Information Processing Systems*, pp. 1369–1376, 2006.

[64] ——, “Efficient computation of optimal actions,” *Proceedings of the national academy of sciences*, vol. 106, no. 28, pp. 11 478–11 483, 2009.

[65] R. J. Aumann, “On the Centipede Game,” *Games and Economic Behavior*, vol. 23, no. 1, pp. 97–105, 4 1998.

[66] R. D. McKelvey and T. R. Palfrey, “An Experimental Study of the Centipede Game,” *Econometrica*, vol. 60, no. 4, pp. 803–836, 1992.

A

Analysis of Game Theory of Mind

This thesis was motivated by a paper called Game Theory of Mind [1]. In this appendix we provide an introduction to the Game Theory of Mind paper and analyze how a sequential version of the level-k model changes the behavior of the agents, something which was not tackled in the original work.

A.1 Description

The Game Theory of Mind paper analyzes the behavior of EUT-agents equipped with level-k, in a setting similar to the second experimental setup in this thesis (see Chapter 4). They make use of a modified LMDP to model the behavior of two agents in the same state space ($\mathcal{S}_1 = \mathcal{S}_2 = \mathcal{S} = \{0, ..., 15\}$) and with a similar reward structure to the Markov Stag Hunt game[1].

The LMDP framework is a special case of the MDP, where a single agent is not presented with a discrete set of actions but is instead presented with some uncontrolled dynamics prescribed by a probability function $\bar{P}$ which can be continuously deformed by assigning to each state in the state space

[1] The reward function is presented as a heatmap plot with no color bar and therefore it was not possible to reproduce the results of the paper with absolute precision.

$\mathcal{S}$ a value such that the controlled dynamics is prescribed by a probability function of the form:

$$P(\boldsymbol{v})_{ij} = \frac{\bar{P}_{ij} e^{\lambda v(i)}}{\sum_{k \in \mathcal{S}} \bar{P}_{kj} e^{\lambda v(k)}}, \tag{A.1}$$

where $P(\boldsymbol{v})_{ij}$ is the controlled probability of going from state j to state i using value $\boldsymbol{v} \in \mathbb{R}^{|\mathcal{S}|}$, $\bar{P}_{ij}$ is the uncontrolled probability of going from state j to state i, and $\lambda \in [0, \infty)$.[2] The reward function $\boldsymbol{r} \in \mathbb{R}^{|\mathcal{S}|}$ represents the motivation of the agent. The agent wishes to maximize the following sum of rewards over an infinite time horizon[3]:

$$\boldsymbol{v} = \mathbb{E}_{P(\boldsymbol{v})} \left[\sum_{t=0}^{\infty} \boldsymbol{r}_t \right]. \tag{A.2}$$

This formulation resembles a form of game-theoretical equilibrium called the Quantal Response Equilibrium (QRE) [61][4], a stochastic version of the Nash equilibrium where agents perceive expected utilities with some random noise. Since agents maximize the perceived expected utilities under this model, the QRE policies are stochastic. An interesting feature of the QRE is that as $\lambda \to \infty$, the QRE tends to the Nash equilibrium in a normal-form game. In other words, as the temperature goes to absolute zero, the agents become rational.

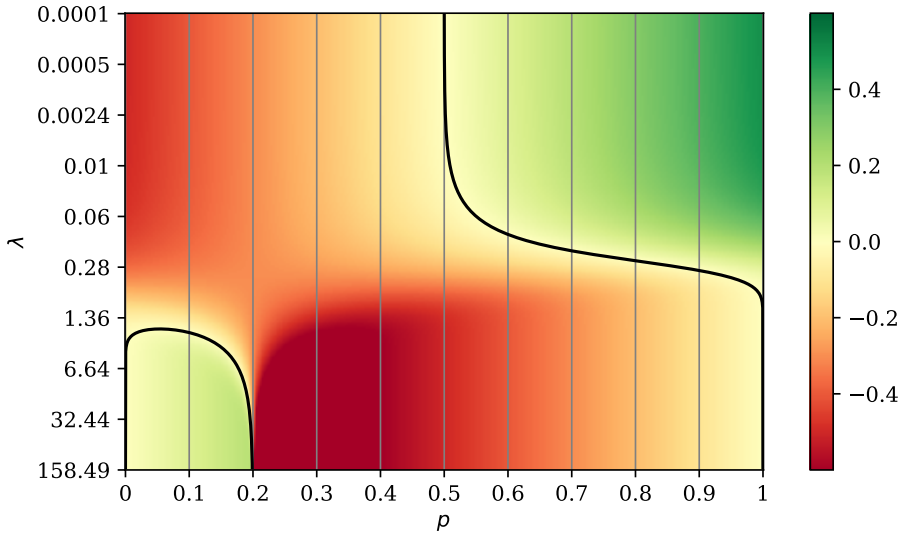

Figure A.1: An illustration of the QRE (black line) in the Stag Hunt game in Table 2.2, as a function of λ. Calculating QREs often involves solving a transcendental equation, which, in this case, we bypass by plotting a surrogate function and observing its zeros.

[2] Those familiar with thermodynamics will recognize this as the Boltzmann distribution. In other words, this deformation creates a stochastic model akin to a thermodynamics system, where the value plays the role of an energy and λ is related to the temperature of a thermal reservoir in a canonical ensemble.

[3] This formulation of value is only possible if the infinite sum is finite, by, for example, assuming that there exists a zero-reward absorbing state or using a discount factor. The original paper does not make any such assumptions and as such, the value does diverge.

[4] The QRE has also been generalized to the extensive-form game setting in [62].

To obtain the optimal value, we can use the following iterative scheme[5] [63, 64]:

$$\boldsymbol{v}_{t+1} = \boldsymbol{r} + \boldsymbol{v}_t P(\boldsymbol{v}_t). \tag{A.3}$$

The LMDP was extended to the two-agent scenario, in the [1]. They accomplish this by assuming that each agent takes turns in a fixed manner, with agent 1 being the first to play. The controlled dynamics $P(\boldsymbol{v}_1, \boldsymbol{v}_2)$ is defined over the Cartesian product $\boldsymbol{\mathcal{S}} = \mathcal{S}_1 \times \mathcal{S}_2$ and now depends on the value of both agents, $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$ and can be written as a function of the controlled probability matrices of both agents as $P(\boldsymbol{v}_1, \boldsymbol{v}_2) = P_2(\boldsymbol{v}_2)P_1(\boldsymbol{v}_1)$, with:

$$\begin{aligned} P_1(\boldsymbol{v}_1)_{ij} &= \frac{\Pi_{1,ij} e^{\lambda v_1(i)}}{\sum_{k \in \boldsymbol{\mathcal{S}}} \Pi_{1,kj} e^{\lambda v_1(k)}}, \\ P_2(\boldsymbol{v}_2)_{ij} &= \frac{\Pi_{2,ij} e^{\lambda v_2(i)}}{\sum_{k \in \boldsymbol{\mathcal{S}}} \Pi_{2,kj} e^{\lambda v_2(k)}}, \\ \Pi_1 &= \mathbb{I} \otimes \bar{P}_1, \\ \Pi_2 &= \bar{P}_2 \otimes \mathbb{I}, \end{aligned} \tag{A.4}$$

where Π_1 and Π_2 represent the uncontrolled transitions of each agent in the joint state space, given the uncontrolled transitions $\bar{P}_1$ and $\bar{P}_2$. The Kronecker product $\otimes$ ensures the transitions of one agent cannot alter the state of the other agent. Each agent has its own reward function, $\boldsymbol{r}_1$ and $\boldsymbol{r}_2$, and it is assumed they are both attempting to maximize the expected sum of rewards, like so:

$$\begin{aligned} \boldsymbol{v}_1 &= \mathbb{E}_{P(\boldsymbol{v}_1, \boldsymbol{v}_2)} \left[\sum_{t=0}^{\infty} \boldsymbol{r}_{1,t} \right], \\ \boldsymbol{v}_2 &= \mathbb{E}_{P(\boldsymbol{v}_1, \boldsymbol{v}_2)} \left[\sum_{t=0}^{\infty} \boldsymbol{r}_{2,t} \right], \end{aligned} \tag{A.5}$$

which each agent can solve in the same manner as the single agent case, given the value vector of the other agent.

A.2 Sequential versus Simultaneous Model

The main difference between the Markov game and the two-agent LMDP is that, in the latter, both agents act in a sequential manner, with agent 1 being the first to play.

The original formulation implements a level-k model by considering the following hierarchy of value

[5] Since it is the relative value between states that is important, the value vector can be normalized by subtracting the maximum of the vector to each entry in order to ensure the.

functions:

$$\begin{aligned}
\boldsymbol{v}_1^{(1)} &= \boldsymbol{r}_1 + \boldsymbol{v}_1^{(1)} P(\boldsymbol{v}_1^{(1)}, \mathbf{0}), \\
\boldsymbol{v}_2^{(1)} &= \boldsymbol{r}_2 + \boldsymbol{v}_2^{(1)} P(\mathbf{0}, \boldsymbol{v}_2^{(1)}), \\
&\vdots \\
\boldsymbol{v}_1^{(k)} &= \boldsymbol{r}_1 + \boldsymbol{v}_1^{(k)} P(\boldsymbol{v}_1^{(k)}, \boldsymbol{v}_2^{(k-1)}), \\
\boldsymbol{v}_2^{(k)} &= \boldsymbol{r}_2 + \boldsymbol{v}_2^{(k)} P(\boldsymbol{v}_1^{(k-1)}, \boldsymbol{v}_2^{(k)}).
\end{aligned} \tag{A.6}$$

This assumes that both agents are equal in every regard. In fact, in the original formulation, the resulting value functions of one agent in the LMDP Stag Hunt game is the transpose of the value functions of the other agent. But this cannot be since the game is inherently sequential and therefore, since agent 2 is the second to play, he should have more information than agent 1.

A repetition of the results for the level-k model implemented in a simultaneous manner, like the original version, can be found in Appendix A.2. The dissimilarity between the value functions for both agents for a given sophistication level is indeed not just the transpose. The resulting stationary distributions, however, still show the switch from the hare state (state 3) to the stag state (state 11).

We investigated a sequential implementation of the level-k model by considering the following hierarchy of value functions:

$$\begin{aligned}
\boldsymbol{v}_1^{(1)} &= \boldsymbol{r}_1 + \boldsymbol{v}_1^{(1)} P(\boldsymbol{v}_1^{(1)}, \mathbf{0}), \\
\boldsymbol{v}_2^{(1)} &= \boldsymbol{r}_2 + \boldsymbol{v}_2^{(1)} P(\boldsymbol{v}_1^{(1)}, \boldsymbol{v}_2^{(1)}), \\
&\vdots \\
\boldsymbol{v}_1^{(k)} &= \boldsymbol{r}_1 + \boldsymbol{v}_1^{(k)} P(\boldsymbol{v}_1^{(k)}, \boldsymbol{v}_2^{(k-1)}), \\
\boldsymbol{v}_2^{(k)} &= \boldsymbol{r}_2 + \boldsymbol{v}_2^{(k)} P(\boldsymbol{v}_1^{(k)}, \boldsymbol{v}_2^{(k)}).
\end{aligned} \tag{A.7}$$

There is now a clear asymmetry in the level-k model, on top of the sequential nature of the two-agent LMDP. The resulting value functions and stationary distributions are displayed in Appendix A.2. With this level-k scheme, the agents are, unsurprisingly, faster at coordinating their efforts to hunt stags, due to the added information of agent 2.

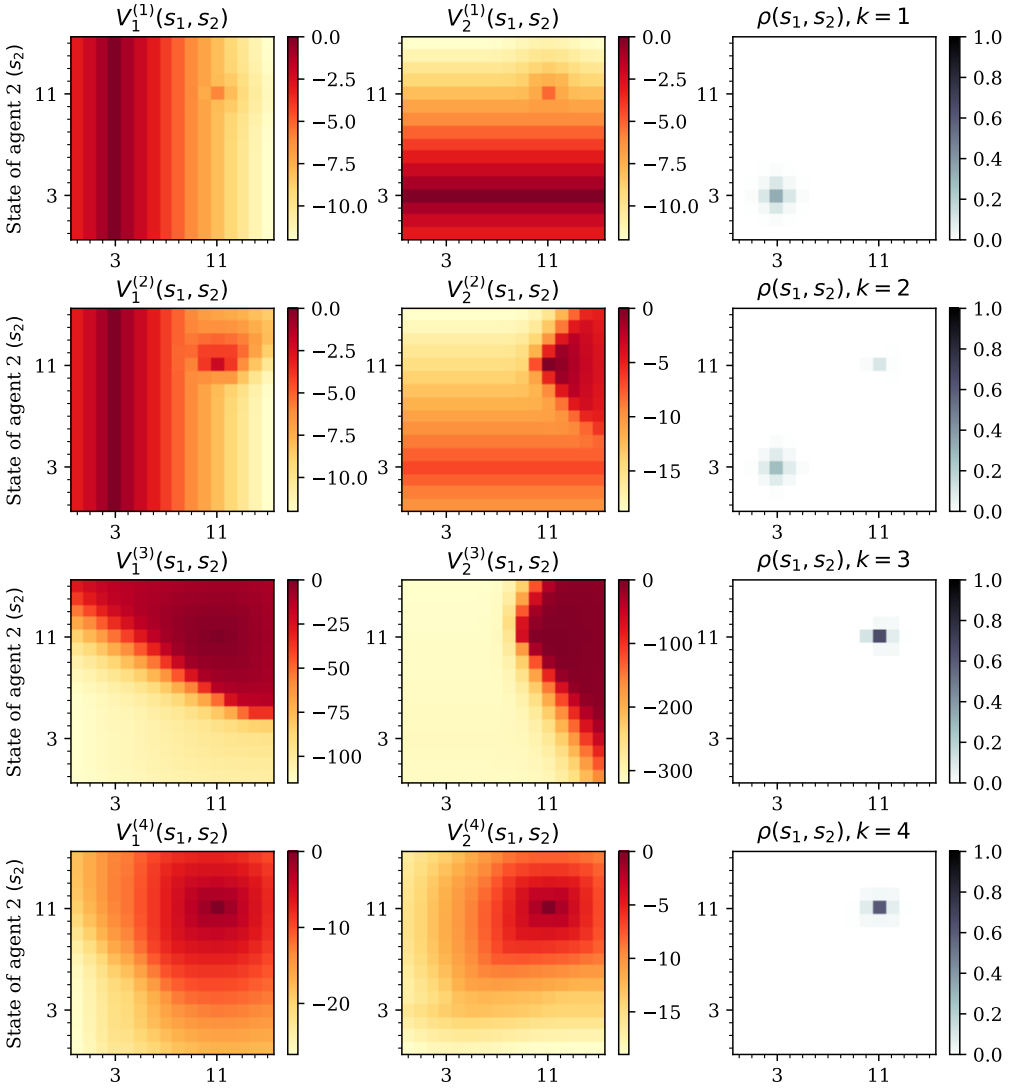

Figure A.2: Results of the simultaneous level-k in [1]. (Left) Value function of agent 1. (Middle) Value function of agent 2. (Right) Stationary distribution of agents. Each row corresponds to the results of a particular level $k = 1, 2, 3, 4$.

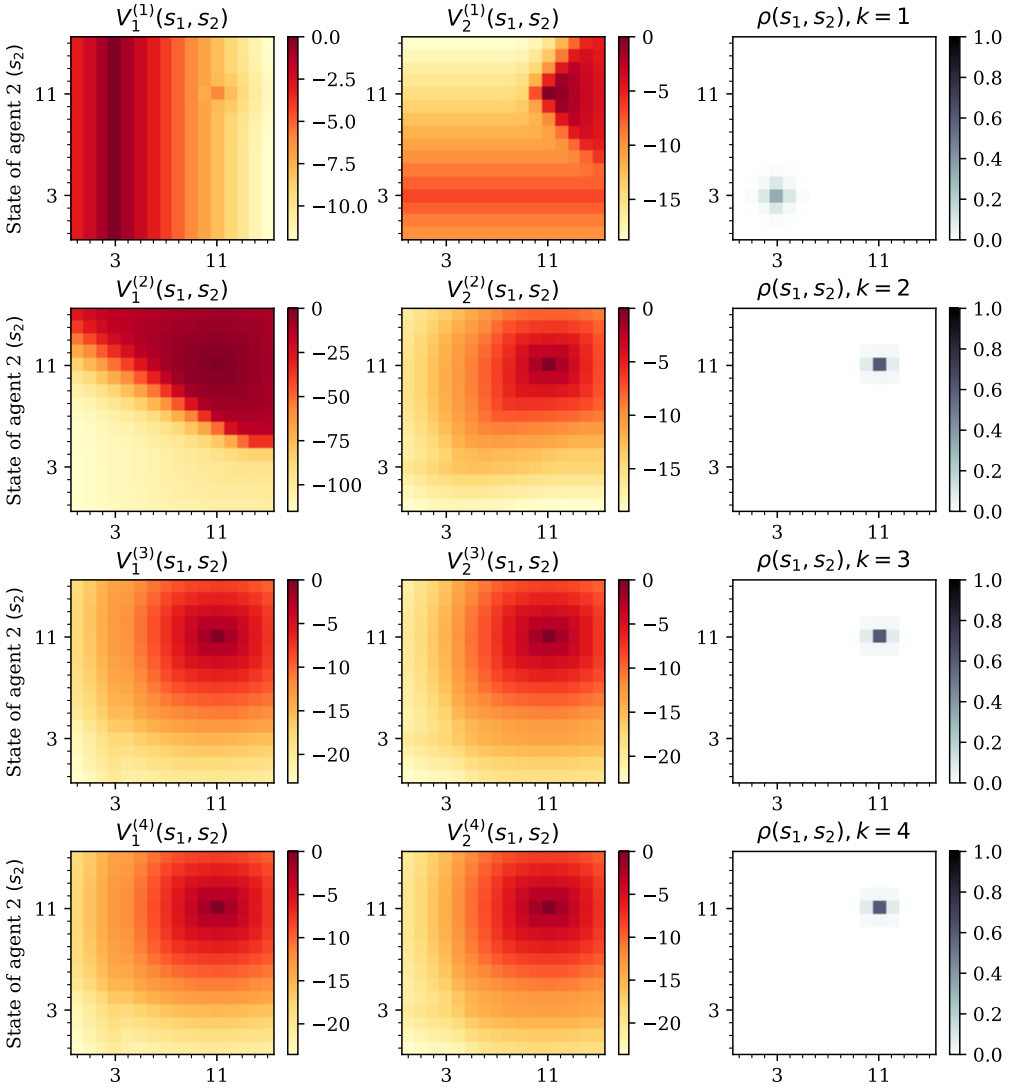

Figure A.3: Results of the proposed sequential level-k. (Left) Value function of agent 1. (Middle) Value function of agent 2. (Right) Stationary distribution of agents. Each row corresponds to the results of a particular level $k = 1, 2, 3, 4$.

B

Dynamic Programming

Dynamic programming is a recursive “divide and conquer” method for solving problems. In essence, it breaks down a difficult problem into smaller, easier to solve problems. Here, we provide an introduction to dynamic programming in the context of games.

B.1 Backwards Induction in the Centipede Game

In game theory, the class of games called **extensive-form games** (sometimes also called **dynamic games**) tries to model interactions between agents over time, in a deterministic fashion. **Repeated games** are a special case of these, wherein a normal-form game is played repeatedly. In an extensive-form game, there may be several normal-form games that can be played depending on the outcomes of previously played normal-form games – in a sort of tree-like structure of normal-form games.

While we will not formally define extensive-form games, since it is not the purpose of this appendix, we will use an example of one, the **centipede game**. In the centipede game, two agents take turns to decide whether to continue or to stop the whole game. By continuing, both agents can increase their potential reward, but the game must either reach the end of the steps or one agent must stop the game

themselves for those rewards to be obtained.

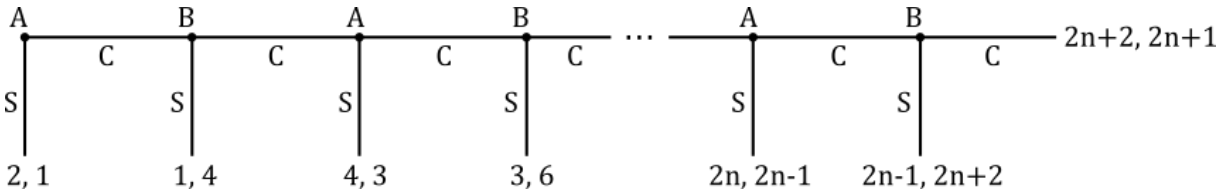

Figure B.1: The tree diagram of the centipede game, an example of an extensive-form game. The name of the game stems from the first appearance where $n = 100$ steps.

Like any game, both agents with to maximize their reward. The usual concept of an equilibrium in extensive-form games is the **subgame perfect equilibrium**, a refinement of the Nash equilibrium.

Definition B.1 (Subgame Perfect Equilibrium)**.** In an extensive-form game, a joint policy $\boldsymbol{\pi}$ is a subgame perfect equilibrium if, for each normal-form game in the extensive-form game, the local joint policy is a Nash equilibrium.

Backward induction can be used to obtain subgame perfect equilibria by reasoning backward in time, from the end of the game to the very start. In the case of the centipede game with n steps, the last agent that plays chooses between continuing C, yielding a reward of $2n + 1$, and stopping S, yielding a reward of $2n + 2$. Being a rational agent (i.e. wishes to maximize his reward), he will choose S, since $2n + 2 > 2n + 1$. The second-to-last agent will then either choose to stop S, yielding a reward of $2n$, or continue C, yielding whatever the next agent will choose. Since we have determined that the last agent will rationally choose S, then the second-to-last agent will only receive $2n - 1$ if he chooses to continue C. Therefore, the second-to-last agent will choose to stop S, since $2n > 2n - 1$. This whole reasoning process can be repeated until the very start of the game, effectively prescribing a joint policy of the two agents whereby they both choose to stop S at every step[1].

B.2 Dynamic Programming in Markov Decision Processes

Backward induction is a dynamic programming method to determine policies. Another instance of dynamic programming can be found in the determination of value in MDPs (see Definition 2.10). Consider the MDP $(\mathcal{S}, \mathcal{A}, P, r)$, where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, $P : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0, 1]$ is the transition probability function, and $r : \mathcal{S} \to \mathbb{R}$ is the reward function. Equation (2.10) represents the infinite-horizon discounted expected sum of rewards value functional that the agent wishes to maximize (with discount factor $\beta \in (0, 1)$). The determination of this value, for a given policy, is done using dynamic programming by rearranging the equation so that it is the sum of the reward at the current state plus the

[1] This result, studied in depth in [65], largely deviates from experimental data [66]

value of the next state which is a random variable:

$$\begin{aligned} V(s,\pi) &= \mathbb{E}_{s_{t+1}\sim p(\cdot|s_t,\pi(s_t))}\left[\sum_{t=0}^{\infty}\beta^t r(s_t)\middle| s_0 = s\right] \\ &= \mathbb{E}_{s_{t+1}\sim p(\cdot|s_t,\pi(s_t))}\left[r(s_0)+\sum_{t=1}^{\infty}\beta^t r(s_t)\middle| s_0 = s\right] \\ &= r(s) + \mathbb{E}_{s_{t+1}\sim p(\cdot|s_t,\pi(s_t))}\left[\sum_{t=1}^{\infty}\beta^t r(s_t)\middle| s_0 = s\right] \\ &= r(s) + \beta\sum_{s'\in\mathcal{S}}\mathbb{E}_{s_{t+1}\sim p(\cdot|s_t,\pi(s_t))}\left[\sum_{t=1}^{\infty}\beta^{t-1} r(s_t)\middle| s_1 = s'\right] p(s'|s_0,\pi(s_0)) \\ &= r(s) + \beta\sum_{s'\in\mathcal{S}} V(s',\pi)p(s'|s,\pi(s)). \end{aligned} \tag{B.1}$$

This is again showing the essence of dynamic programming, the breaking down of a problem by finding a recursive method that we can solve. The optimal value $V^*(s)$, for every state $s \in \mathcal{S}$, is obtained by the following maximization:

$$V^*(s) = \max_{\pi}\left\{r(s,\pi(s)) + \beta\sum_{s'\in S} V(s',\pi)p(s'|s,\pi(s))\right\}. \tag{B.2}$$

One, rather simple algorithm to calculate the optimal value is **value iteration**. It consists in iterating the following recursion:

$$V_{k+1}(s) = \mathcal{T}V_k(s) = \max_{a}\left\{r(s) + \beta\sum_{s'\in S} V_k(s')p(s'|s,a)\right\}, \text{ for all } s \in \mathcal{S}, \tag{B.3}$$

where $V_0(s)$, for all $s \in \mathcal{S}$, is an initial guess of the optimal value function, usually taken to be zero, and $\mathcal{T}$ is the evolution operator. In this case, the policies are deterministic and the maximization is made over the action space $\mathcal{A}$ and not the space of possible policies.

Theorem B.1. *Value iteration converges to a unique value V^*.*

Proof. Let us first rewrite Equation (B.3) in vector form, $\mathcal{T}V_k = \max_a\{R + \beta P^a V_k\}$, with $[P^a]_{i,j} = P(s' = j|s = i, a)$. Value iteration converges if $\lim_{k\to\infty}||V_k - V^*||_\infty = 0$. This can be proven by first showing that the evolution operator $\mathcal{T}$ is a contraction mapping on the metric space $(\mathbb{R}^{|\mathcal{S}|}, ||\cdot||_\infty)$, i.e. that there exists a non-negative real number $c \in [0,1)$ such that, for all value vectors $V, V' \in \mathbb{R}^{|\mathcal{S}|}$,

$$||\mathcal{T}V' - \mathcal{T}V||_\infty \leq c||V' - V||_\infty.$$

This can be proven as follows:

$$
\begin{aligned}
||\mathcal{T}V' - \mathcal{T}V||_\infty &= ||\max_a \{R + \beta P^a V'\} - \max_a \{R + \beta P^a V\}\,||_\infty \qquad \text{(by definition)} \\
&= ||\beta \max_a \{P^a\}\,(V' - V)||_\infty \qquad \text{(simplification)} \\
&\leq \beta ||\max_a \{P^a\}\,||_\infty ||V' - V||_\infty \qquad (||AB|| \leq ||A||||B||) \\
&\leq \beta ||V' - V||_\infty \qquad (\max_a \sum_{s' \in \mathcal{S}} P^a_{s,s'} = 1).
\end{aligned}
\tag{B.4}
$$

Since $\beta \in (0, 1)$, then $\mathcal{T}$ is indeed a contraction mapping on $(\mathbb{R}^{|\mathcal{S}|}, ||\cdot||_\infty)$. Now, with the previous inequality, and the fact that $\mathcal{T}V^* = V^*$, we can show that the limit is zero:

$$
\begin{aligned}
||V_k - V^*||_\infty &= ||\mathcal{T}V_{k-1} - \mathcal{T}V^*||_\infty \\
&\leq \beta ||V_{k-1} - V^*||_\infty \\
&\leq \beta^k ||V_0 - V^*||_\infty
\end{aligned}
\tag{B.5}
$$

Since $||V_k - V^*||_\infty \leq \beta^k ||V_0 - V^*||_\infty$, and $\beta \in (0, 1)$, then $||V_k - V^*||_\infty \xrightarrow{k \to \infty} 0$. □