Thesis for the Degree of Doctor of Philosophy

Models of Cooperation, Learning and Catastrophic Risk

Vilhelm Verendel

Division of Physical Resource Theory
Department of Energy and Environment
Chalmers University of Technology
Gothenburg, Sweden 2016

Models of Cooperation, Learning and Catastrophic Risk
Vilhelm Verendel

ISBN 978-91-7597-349-4

Doktorsavhandlingar vid Chalmers tekniska högskola
Ny serie nr 4030
ISSN 0346-718X

Department of Energy and Environment
Division of Physical Resource Theory

Chalmers University of Technology
SE-412 96 GOTHENBURG, Sweden
Phone: +46 (0)31-772 10 00

Contact Information:
Vilhelm Verendel
Division of Physical Resource Theory
Department of Energy and Environment
Chalmers University of Technology
SE-412 96 GOTHENBURG, Sweden
vive@chalmers.se

Printed by Chalmers Reproservice,
Göteborg, Sweden 2016

Models of Cooperation, Learning and Catastrophic Risk

Vilhelm Verendel
Chalmers University of Technology

Thesis for the degree of Doctor of Philosophy.

ABSTRACT

Our world presents us with dangers and opportunities. Some of these dangers and opportunities are easier to handle if two or more individuals learn to cooperate. This thesis contributes five papers about cooperation, learning and catastrophic risk.

In papers I-II, we consider the Finitely Repeated Prisoners' Dilemma, a model for where cooperation between two players is particularly hard to achieve. We introduce and model strategies that attempt to convince others to cooperate when backward induction can be used to eliminate cooperation for a number of steps from the end. We find that in a population with these strategies, cooperation can become recurrent, and we examine the conditions for this. Recurrent cooperation is possible in an evolutionary model (paper I) as well as in a population of players that are near-perfect Bayesian expected utility-maximizers (paper II).

In paper III, we consider a bargaining model of climate negotiations where players negotiate emissions and sudden catastrophic damage occurs if emissions exceed a threshold amount. We introduce and model a mechanism of strategic reasoning, where players predict the emission bids of others, and consider how this affects the possibility of reaching agreements preventing catastrophic damage. We find that the effect of higher levels of strategic reasoning makes it harder to reach agreements in the model. This effect can be partially mitigated by restricting the range of initial bids in the bargaining process.

In paper IV, we consider the arguments by Hanson and Bostrom about the Great Filter as an attempt to explain the Fermi Paradox. According to these arguments, finding extraterrestrial life on one planet should lower our expectations for humanity's prospects to progress far beyond our current technological capabilities. We model this claim as a Bayesian learning problem and examine the effect a single observation of life has in the model. We find that the conclusion of the argument depends critically on the choice of prior distribution.

In paper V, we consider a model of agricultural markets and land-use competition between food and bioenergy crops. Agents in the model represent farmers who decide which crop to grow depending on predictors that give future price expectations. We model agents who can switch among predictors to make their decisions. We find that some predictor types can be concentrated on key parcels of land, which reduces volatility in crop prices for the system. We also examine several mechanisms that can bring price fluctuations in the system down and closer to a stable state.

Keywords: Cooperation, Finitely Repeated Prisoners' Dilemma, Backward Induction, Climate negotiations, Catastrophic risk, Fermi Paradox, Bayesian analysis, Learning

Acknowledgments

First and foremost I would like to thank my advisers Kristian Lindgren and Daniel Johansson for giving me the opportunity to work with you. Kristian, you have inspired me with your helpful critical thinking and by always being prepared to take a step back and ask what the interesting aspect of a model really is. You showed me the value of being prepared to start over and approach a problem or question from a new angle time after time. Daniel, you together with Kristian formed a useful mix of different complementary viewpoints and you were always ready and willing to discuss new aspects about climate, modeling, and research with me. Thank you!

To all friends and colleagues at the division of Physical Resource Theory and the Department of Energy and Environment at Chalmers, I wish to express my appreciation for the nice working environment which I have much enjoyed. The good seminars and nice atmosphere made the research and work here much more fun.

To my co-authors that I had the good fortune to work with on the papers in this thesis: Daniel, David, Emma, Kristian, Liv and Olle. I can see that our discussions helped me to become a better researcher.

I also want to thank the Swedish Energy Agency and the EU-FP7 project Mathemacs for financial support.

Finally, to my friends and my dear family. My parents, Reet and Vello. Jan, Tomas and Aino. You have often supported me when needed and I am very grateful for this.

To my fiancé Beatrice for being so supportive and being a fantastic mother to our two daughters. Edla and Elvira, thank you too for understanding I had to work late a number of times to finish this thesis. I am grateful for your love and support.

Vilhelm Verendel
Göteborg, February 2016

List of appended papers

- Paper I: Kristian Lindgren and Vilhelm Verendel. *Evolutionary Exploration of the Finitely Repeated Prisoners' Dilemma – The Effect of Out-of-Equilibrium Play.* Published in *GAMES* (2013)

  KL had the idea. KL and VV performed the modeling, simulations and analysis. KL and VV wrote the paper.

- Paper II: Vilhelm Verendel and Kristian Lindgren. *Repeated Prisoners' Dilemma and Bayesian learning of out-of-equilibrium play. Working Paper* (2016).

  VV and KL had the idea. VV and KL performed the modeling, simulations and analysis. VV wrote the paper with input from KL.

- Paper III: Vilhelm Verendel, Daniel J.A. Johansson and Kristian Lindgren. *Strategic reasoning and bargaining in catastrophic climate change games.* Published in *Nature Climate Change* (2015).

  VV had the idea. VV performed the modeling, simulations and analysis with input from DJAJ and KL. VV wrote the paper with input from DJAJ and KL.

- Paper IV: Vilhelm Verendel and Olle Häggström. *Fermi's paradox, extraterrestrial life and the future of humanity: a Bayesian analysis.* Published in *International Journal of Astrobiology* (2016).

  OH had the idea. VV performed the modeling, simulations and analysis with input from OH. OH and VV wrote the paper together.

- Paper V: Liv Lundberg, Emma Jonson, Kristian Lindgren, David Bryngelsson and Vilhelm Verendel. *A cobweb model of land-use competition between food and bioenergy crops.* Published in *Journal of Economic Behaviour and Control* (2015).

  KL had the idea. LL, EJ and KL performed the modeling, simulations and

analysis with input from DB and VV. LL, EJ and KL wrote the paper with input from DB and VV. VV contributed by modeling the agents switching between predictor types and doing part of that analysis.

"Failure is a fascinating thing because most people avoid it. So if you can get to a place where you like it, where you like what you can get out of it, then it's new territory."
- Louie C.K.

Contents

Part I

INTRODUCTION

Chapter 1

Introduction

Our world presents us with dangers and opportunities. Some of these dangers and opportunities are easier to handle when two or more individuals cooperate. One example is climate change: Current climate change is already dangerous, and the danger is increasing [62], but climate negotiations also present opportunities to establish cooperation and mitigate damage (paper III). Collective action problems where cooperation is particularly hard to achieve can be explored using models such as the Finitely Repeated Prisoners' Dilemma, which can be used to reason about how cooperation can be established between two individuals (papers I and II). What individuals believe and learn from observations about whether cooperation will be reciprocated can be crucial for cooperation.

Further, models of learning can also shed light on how to interpret the Great Filter arguments – a response to the Fermi Paradox – about why, so far, we have yet seen no signs of extraterrestrial life. According to these arguments, finding extraterrestrial life on one planet should lower our expectations for humanity's prospects to progress far beyond our current technological capabilities [19, 42]. Bayesian learning makes it possible to analyze whether the conclusion of these arguments hold under different assumptions (paper IV). Learning can also affect the volatility of prices in environments such as commodity markets (paper V).

The aim of this thesis is to improve our knowledge of these and several related questions. The contribution is five papers with five mathematical models. The concepts of cooperation, learning and catastrophic risk connect the five papers. In this introduction, we clarify these three concepts and the relevant methods and research questions.

1.1 Three conceptual pillars

The work in this thesis is supported by three conceptual pillars: *cooperation*, *learning* and *catastrophic risk*, present in each of the papers as follows.

	Cooperation	Learning	Catastrophic risk
Paper I	*X*		
Paper II	*X*	*X*	
Paper III	*X*	*X*	*X*
Paper IV		*X*	*X*
Paper V		*X*	

In the following sections, we take a look at these concepts in more detail.

1.1.1 Cooperation

Some goals that are hard for a single individual to achieve become easier to achieve when several individuals cooperate.

Reaching a goal can sometimes be impossible without cooperation. For example, a table might be too heavy for a single person to move at all, but two people cooperating might be able to do it.[1] Here, a contribution from more than one individual is needed for successful cooperation to come about. Moving up to the level where individual nation-states can cooperate with others, some problems can also require all nation-states to cooperate before a goal can be reached, eradicating global pandemics, such as the smallpox virus,[2] for example.

Smallpox was declared to be eradicated in 1979 after successful international cooperation across the borders during the Cold War. To avoid the pandemic propagating again, starting from a few infected individuals isolated somewhere, it was necessary for each and every country to cooperate to detect and prevent the virus to spread. Had a single country defected from cooperation which required carefully coordinated vaccination efforts [10], the virus could have survived within an individual country, and it could soon have propagated outwards again through our vast social networks. This shows that sometimes *all* individual nation-states must cooperate for the efforts of others not to be in vain.

In other cases, the greater the number of individuals who cooperate, the more progress is gradually made toward the relevant goal. Climate change mitigation to reduce greenhouse gas emissions is a good example. Here, the more countries ("individuals") get involved, the closer we get to a goal of stabilizing greenhouse gas concentrations. In this kind of case, all the benefits from the contributions of others are not necessarily eliminated if an individual country defects.[3]

[1] Example due to Tuomela [72].

[2] Smallpox has led to hundreds of millions of casualities through history, so getting this eradicated has been called "the greatest achievement of international cooperation ever" [9] by Barrett, a leading researcher on international cooperation.

[3] Technically speaking, a single individual (or a single nation) could possibly realize significant climate change mitigation by repeated large-scale injection of sulphur into the stratosphere, which would cool off the atmosphere [13]. A billionaire would be an individual with this power. There is even a term for this: A scenario in which a "greenfinger" takes action. However, this type of intervention poses what might be an unacceptable risk to future generations, because the injection must repeatedly continue to avoid climate damage.

In this thesis, we consider both these forms of cooperation. In the model in papers I and II, both individuals in the Finitely Repeated Prisoners' Dilemma have to contribute for any benefits for a cooperating player to come about. In the model in paper III, not necessarily all individuals have to contribute to reducing greenhouse gas emissions in order to meet the goal of avoiding catastrophe.

We see that the concept of cooperation becomes relevant in many different cases, from everyday situations to global challenges, and the types of individuals that can participate in cooperation among humans can range between individual agents of different forms such as persons, organizations, and the international level of countries. Common is that for cooperation we need contributions from more than one individual. But to what degree is cooperation needed?

Homo sapiens is believed to have flourished on Earth because of the ability to cooperate in large numbers [23]. An early example of large-scale cooperation is human language: It has been argued that sudden improvements in human cognitive capabilities around 70,000 years ago enabled us to share information flexibly and develop shared culture in large groups [43, 59]. Sharing information about the world can be viewed as a form of cooperation, but so can the use of language itself. It has been argued that developing well-established languages also seems to require some form of cooperation and it is unclear which came first [60]. It is reasonable to believe that throughout history, many forms of cooperation have formed a basis for building friendships, families, tribes, villages, cities, organizations and civilizations. In our age, we continue to find new forms of cooperation such as when modern technology enables new forms of large-scale cooperation over vast geographical and cultural distances, as with the Wikipedia.

While it may appear that cooperation can solve many of our problems, it would be naive to expect cooperation necessarily to come about to solve all of them. Our future dangers and opportunities also relate to sustainable development [24] to appropriately control our global natural, social and technological environment for future generations of life. In these and other cases, cooperation can play an important role; if the involved agents were to cooperate, we could all be better off. Evidence suggests that complicated international problems can be solved; for example, countries agreed to protect the ozone layer.[4] However, we know that there are serious barriers to cooperation. Negotiations to reduce greenhouse gas emissions face such barriers. These negotiations only represent one of many global challenges, but climate negotiations may be among the hardest problems to agree on how to solve, because of the many factors involved [10].

We are thus clearly a cooperative species, but cooperation does not always come about. On the one hand, it could be argued that evolution by selection has produced competitive mechanisms that benefit non-cooperating persons at the expense of others who do cooperate. If individuals have the potential for free-riding, there

[4] An effort among countries which culminated in the adoption of the Montreal Protocol in 1987, which Kofi Annan has called "Perhaps the single most successful international environmental agreement to date" (2000) [10].

could in principle be natural, built-in obstacles in any group. On the other hand, we show many signs of altruism and of being predisposed to help others and norms can have a crucial role [23]. This mix of self-interest and altruism is clear; equally clear is the ease with which even trivial circumstances, such as finding a penny in a phone booth, can sometimes influence our willingness to care for others [46]. Our capacity for free-riding means we may need institutions with more or less systematic mechanisms that provide more robust prerequisites for cooperation, such as with the institution of taxation [10].

Thus, while cooperation can be desireable from a collective viewpoint, it need not be for the individual. In any population, an individual can cooperate with or defect from the actions which would make everyone in a group better off. This presents us with the *free-riding problem*, in which an individual gains by not contributing to the effort of the group as whole, i.e., from "defecting".

A number of basic and more advanced cognitive factors have been suggested to facilitate cooperation in humans because they allow detection and avoidance of defectors [28]. These factors include recognizing others, remembering histories with different individuals, the ability to communicate values to others, the ability to understand the values of others and a general cognitive ability to represent costs and benefits in interactions [26]. The factors can have many social aspects as well: Cooperation can be facilitated by the perception of shared fate and of "being in the same boat" [25, 70]. Mechanisms of so-called indirect reciprocity based on reputation and gossip can also shape the willingness of others to engage in non-selfish behaviour [61]. Taken together, this shows a few examples of that our beliefs and learning can have a crucial role for cooperation.

Simulations based on theoretical modeling such as Axelrod's famous computer tournament with different strategies to play the Repeated Prisoners' Dilemma have also led to suggestions for facilitating cooperation [8]. Axelrod's recommentaions include: Emphasizing the repeated nature of interactions so that individuals understand that defecting has consequences for future interactions; to teach reciprocity to make it harder for exploitative free-riders to benefit; enable reputation mechanisms, so that expectation works to shape the willingness of others to engage in cooperation. Understanding the factors that explain cooperation also requires considering individuals other than persons, such as nation-states, for which self-interest can be an important driver. It has been convincingly argued that self-interest is an important factor in the international system of nation-states [55, 10].

Using Elster's definition, "collective action" refers to the course of action by all or most individuals that, when chosen by all or most individuals, leads to the collective best outcome [31]. Collective action is often what is meant by cooperative behaviour. Situations in which there is a conflict between individual and collective interests have been described as "collective action problems". Papers I, II and III contain models of collective action problems, used to study under what assumptions cooperation results. Elster [31] makes a useful distinction between "strong" and "weak" collective action problems as follows.

A *strong collective action problem* is a situation in which (1) each individual derives greater benefits under conditions of universal cooperation than under universal noncooperation, and (2) each individual derives greater benefits when abstaining from cooperation, regardless of what others do.

A *weak collective action problem* is a situation in which (1) is the same as the strong case above, but in which (2) cooperation is individually unstable and individually inaccessible. Individually unstable means that each individual has an incentive to defect from a situation with full cooperation. Individually inaccessible means that an individual has no incentive to take the first step away from a situation of universal noncooperation.

In papers I and II we base the model on two players who are in the strong collective action problem known as the Prisoners' Dilemma (it also counts as a weak collective action problem). The Prisoners' Dilemma is a game-theoretic problem which has been used to reason about hard cooperation problems such as avoiding conflicts and arms races between individuals who can either try to conquer or suffer from not doing so, but the model itself is phrased in general terms about two players, where each chooses between cooperation and defection [8]. For each player, the outcome is better if both choose to cooperate, compared to if both defect. However, given full cooperation, the individual agent can gain an even better outcome by abstaining from cooperation and choosing to defect, so no matter what the other player does, it is always better to defect. Thus, we should expect rational players to choose to defect if they only interact once in the single round version of the game. Papers I and II build on the version of this game in which it gets repeated so that players' responses can have an effect over time.

In paper III we consider a weak collective action problem in a model of climate negotiations where there is an emissions threshold for catastrophe. In the model, negotiators bargain over emissions reductions but also attempt to avoid catastrophic damage. With typical parameters, full defection (full emissions) corresponds to a situation in which no single negotiating party can avoid catastrophic damage on their own. However, if each country cooperates fully (maximal reduction), individual countries would gain from reducing their efforts. Moreover, if other countries contribute a sufficiently large part in avoiding a catastrophe, a single country also gets incentives to contribute and avoid the damage. This makes cooperation individually unstable and individually inaccessible, but it does not satisfy the second requirement for a strong collective action problem; thus, this is a weak collective action problem.

Up to this point in the discussion cooperation is clearly something good for the participants even if it may be hard to bring about. We saw above that a number of circumstances have been found that facilitate higher levels of cooperation among individuals and it could be tempting to claim that we should always seek to make these come about. Maybe there are other forms of cooperation that are less desireable? Cooperation certainly has positive connotations: helping others, acting unselfishly, and overcoming collective action problems, but this is from the perspective of the

participants. Perhaps successful cooperation can bring not only opportunities, but also dangers?

Such examples can be found in many social issues. For example, we might not want particular forms of cooperation to come about in a market that supplies goods, services or technologies efficiently. Companies could see gains from cooperating to form price-setting cartels. It may be desireable for the individual companies to cooperate and keep prices high, compared to engaging in a price competition that lowers the price for customers. This might even have the structure of a Prisoners' Dilemma, where the challenge is for two companies to establish cooperation instead of lowering market prices. Defection (lowering the prices) may quickly attract many of the customers and give one company a greater short-term benefit from more customers, but both companies would fare less well than if cooperating. However, outside observers may want to prevent the companies from finding a solution to this, somewhat stylized, collective action problem of establishing a price cartel. Cooperation between criminals [30] is a phenomenon where we could have clear moral reasons to understand and prevent. These examples remind us that cooperation is a general phenomenon that is neither intrinsically good nor intrinsically bad.

Against this background, do we need to define cooperation? An approach of "we know it when we see it" can be unsatisfactory from a scientific viewpoint. Given that cooperation and the problem of collective action have been discussed for a long time in biology as well as in social science, it is perhaps surprising that the literature exists contains little discussion of the generic concept of cooperation. The first systematic philosophical treatment about the cooperation concept seems to have been written by Tuomela in 2000 [72].

Is there a single definition that can easily capture all forms of cooperation? Tuomela argues that there is not, but suggests two distinct forms: group mode and individual mode cooperation. The first form involves situations where individuals have collective goals (such as two persons both having the same shared collective goal of moving a heavy table). The second form involves situations where there are conflicting interests that can introduce elements of free-riding. Here we have collective action problems such as climate change mitigation and the Prisoners' Dilemma. In these situations, full cooperation can present a temptation for an individual to defect and let others do the work. Tuomela defines such *cooperation in the individual mode* to exist between two agents A and B "if and only if A and B without conflict fit their relevant private goals, preferences, and actions to the other's private goals, preferences, interests and actions (but do not have a collective goal related to the actions in question)". The point of Tuomela's individual mode definition is to describe cooperation that primarily arises out of the individual interests.[5] Tuomela argues that this form of cooperation is the one addressed in non-cooperative game theory discussed below. In this thesis, we do not consider group preferences. All our work falls back on modeling cooperation as ultimately depending on factors and

[5] An individual's private interests are not to be confused with being selfish. A private interest may as well describe a preference for an individual agent which is altruistic.

properties of the invididuals involved, i.e., by methodological individualism [31].

Taken together, studying cooperation can include theoretical, empirical and ethical questions. The theoretical efforts involve building models as described above and can be done using game theory (more in Section 1.2). The empirical efforts involve studying the conditions under which cooperation between individuals actually comes about; the individuals can represent different elements such as persons, organizations and nation-states. Ethical questions are relevant if we seek to determine whether cooperation is desireable or not. This thesis makes an exclusively theoretical contribution. Using quantitative models, we examine some theoretical conditions with assumptions and mechanisms that make cooperation in the models to come about.

1.1.2 Learning

A vast range of our actions are driven not by certain knowledge but mere beliefs. With learning, an individual updates the beliefs it has about the world. Learning can play an essential part of changing a course of action. While we are still learning, we may have to act on the basis of partial or incomplete knowledge. An example: Climate sensitivity is a parameter of core interest in estimating how large temperature to expect from a doubling of carbon dioxide levels in the atmosphere. Climate sensitivity is likely in the range 1.5 to 4.5 degrees Celsius and this range of uncertainty has held for decades [50, 47]. However, with the precise value unknown, the range of this uncertainty can still warrant action as a large part of this interval could have costly damage and effects on current and future persons.

To distinguish learning from action, it will be useful to think that a rational agent can start out by believing something that is wrong or incomplete but still act rationally given that uncertainty. But how can we more precisely distinguish being rational from holding and learning different forms of belief? To distinguish among different forms of belief, Elster [32] presents four different cognitive attitudes that an individual can have towards the surrounding world: *certainty, risk, uncertainty*[6] and *ignorance*.

The attitude of certainty is a particular form of belief that excludes doubts, and beliefs can not be revised. The attitude of risk is a form of belief in which probabilities can be attached to the different possible states of the world, and learning can revise these beliefs. For the attitudes of uncertainty and ignorance, no probabilities come into play. For this thesis, the relevant distinction is between Elster's first two attitudes.

Papers I and II are motivated by the question of whether players should have certainty in their beliefs about backward induction in the Finitely Repeated Prisoners' Dilemma. A backward induction assumption (see Section 1.2.4) means cooperation

[6]It should not be confused with decision-theoretic uncertainty in statistics and economics which can carry meaning of a decision-makes subjective probabilities. Uncertainty in Elster's framework relates to a situation where a decision-maker is unable to form probabilities over the different possible future states of the world.

can not be established in the repeated game, predicting defection throughout the game. However, the Backward Induction Problem [18] and the "Backward Induction Paradox" [64] question the assumption that players hold backward induction beliefs with certainty. The central question of this argument is to ask what players' beliefs about backward induction should be if cooperation gets played in the first round of the repeated game, if they already have certainty in belief about continued defection. The aim of papers I and II is to model different possible reactions when their beliefs are not certain. In paper II, we present an explicit learning model where the extent of backward induction used is observed by players with an belief attitude of risk, where the players gradually learn and revise their beliefs about how much backward induction takes place and what the possible reactions are.

The attitude of risk which allows for revision of beliefs using learning is present in several forms in papers II, III, IV and V. In these papers, learning as a function of the actions of others is used to anticipate future actions of other agents and what effects these have. Learning also becomes relevant in paper IV in analyzing the Great Filter arguments by Hanson and Bostrom, which is based on a statistical claim what we could learn from observing extraterrestial life.

The distinction between rational choice, rational learning and learning in general is important. Learning in general is simply updating beliefs based on observations. Modern theories of rationality typically do not imply any specific beliefs; instead, they allow agents to make prediction errors and act on these because expected utility maximization is a function of beliefs as probabilities. An important special case of learning results from the theory of Bayesian rationality: While it does not dictate which prior beliefs a rational agent should hold, it implies that rational agents should revise their beliefs by Bayes' rule [17]. This is just one out of many possible and suggested learning rules to update beliefs [33], but there are particular justifications for this learning rule, discussed more below, and we use this in papers II and IV. We cover some learning rules in Section 1.4.2 and note that some of them have been argued to be rational.

Rationality and Learning

Rationality commonly has a precise meaning in game theory as expected utility maximization, and this is used to reason about the players' strategies based on their beliefs and preferences [17, 36, 37]. It makes sense to clarify these theories of rationality a bit for two reasons. First, in order to judge whether the requirement of expected utility maximization is a reasonable to hold for rationality, we should know the underlying assumptions. Second, if we use models of bounded rationality that depart from expected utility maximization, we are, mathematically speaking, on less firm ground. This is not necessarily a problem, but it can be good to understand the distinction between rationality and other forms of decision-making.

Modern theories of rationality are rooted in the history and philosophy of games and statistics. The explicit concept of utility was introduced by Bernoulli in the 18th century to resolve the so-called St. Petersburg Paradox [14]. Offering people

the chance to participate in monetary lotteries that had infinite expected outcomes, Bernoulli observed that people were not willing to pay very large sums of money to take part. Bernoulli suggested an explanation for this: People do not primarily care about the monetary value, but the *utility* of money. It would be reasonable for a utility function to show diminishing marginal utility meaning that the first unit of currency would be more worth than the second, which would be worth much more than the millionth unit, and so on. This could make the expected utility of Bernoulli's lottery less than infinite and could form part of the explanation.

Modern theories of rational choice are value-neutral in the sense of not making judgments about what are good or useful outcomes for real people. The primary theoretical concept is the *preference relation* which ranks the different possible options. Rationality is often taken to be described by certain models of individual choice from decision-theory and economics [17, 37]. These show that if a decision-maker has preferences satisfying particular conditions, then the preferences are ranked by expected utility. More precisely, these conditions are axioms about the forms of preference ordering that a decision-maker can hold. The utility concept is then relevant because it comes into play with quantitative maximization of utility corresponding to satisfying such preference orderings. If we hold these axioms to represent rationality, then maximization of expected utility is also rational. Two theories of choice as expected utility maximization are discussed more in Section 1.2.2.

1.1.3 Catastrophic risk

A risk is a possible outcome which is valued negatively, so that there is some form of danger.[7] It makes sense to speak about risk for everything ranging from a person playing roulette to possible effects of some event for human well-being on a global scale. Both cooperation and learning are relevant to negotiating and refining our beliefs about catastrophic risk.

Catastrophic and existential risk

One way to rank risks is by the number of affected individuals and the severity of the outcomes [22]. At some point, when we gradually increase the severity of the consequences so that they have a lasting significant effect for individuals, it becomes relevant to speak of *catastrophic risk*. If we start from the level of catastrophic risk and increase the number of affected individuals, at some point it will make sense to speak of *global catastrophic risks* affecting many individuals severely on a global scale. This group of risks includes severe damage from climate change, pandemics, nuclear war, and secondary events arising from these, such as conflict and serious threats to the social order [22].

If we continue even further, increasing the number of individuals affected, we reach the entire global population. Adding a temporal element to this, some ex-

[7] The term risk is used in many contexts and more technical definitions exists. It should not be confused with the definition of risk in decision-theory or the risky attitude to belief described by Elster above.

treme risks could be severe enough to affect not only the full global population, but all future generations. In Bostrom's terminology, an *existential risk* "is one that threatens the premature extinction of Earth-originating intelligent life or the permanent and drastic destruction of its potential for desirable future development" [20]. Thus, existential risks make up a particularly severe part of global catastrophic risks.

The idea that global catastrophic risks and even existential risks exist has a good scientific basis [22, 39]. Here is one example where there exists very well-justified belief: a few billion years from now, the Sun is expected to no longer provide any good conditions for life on our planet [22]. However, this scenario is of course so vastly remote that it is practically meaningless to care about today. It has however been argued that there are also more near-term existential risks that deserve more of our attention [20, 63, 39]. Existential risk can be attributed to three main sources: a) nature; b) unintentional human acts; and c) intentional human acts [22].

The first group of natural risks are those believed to be most well understood. This group includes major asteroid impacts, supernovas and other astronomical phenomena, supervolcanoes on Earth and severe pandemics [22]. However, these are also the risks judged the least likely to be realized on short time scales such as the coming century. From data compiled about time series of such events [39], it seems that we can expect them to occur hundreds of thousands to millions of years apart.

So what are the more likely existential risks over shorter time spans? The two other groups of existential risk relate to human actions and are believed to mainly involve our increasing technological capabilities [22, 63, 39]. The idea is that increased technological powers to control and modify our environment can be used either for the good or the bad, with some of the dangers being at least catastrophic risks. A well-understood example of this is nuclear technology enabling the potential for large-scale nuclear war, a danger which has existed for over 50 years. However, the idea is also that specific new technologies that can provide great benefits are also believed to grant tremendous destructive power to smaller groups of individuals: These include unintended or intended use of synthetic biology (bioerror or bioterror) [22], powerful artificial superintelligence getting out of our control [21] and other scenarios including destructive use of nanotechnology [39]. All these scenarios come with many factors of uncertainty regarding particular mechanisms and technologies.

Asking experts and the general public to estimate probabilities for existential risks gives the following picture. For the surveyed experts from a Global Catastrophic Risk conference in 2008, the median of the estimated probabilities for extinction during the coming century was 19% [65], which suggests that we can not safely ignore such events during the coming century. For the general public, a more recent poll by authorities in Sweden surveyed over six thousand 18-year olds in 2015, and in this group 3% believed extinction would happen during their lifetime [35]. The median estimated probabilities in the expert survey for extinction during the coming century were weapons technologies (5%), artificial superintelligence going wrong (5%), wars (including nuclear) (4%), pandemies (2%), and nuclear war (1%).

In the public survey, among the 56% believing in an extinction at some future point in time, and when asked about what the cause of extinction would be the most common answer (31%) was climate change.[8] Asking experts and public polls both have their different problems, and these are not directly scientific studies of existential risk. However, the results above suggest the following question: Could climate change be an existential risk?

Climate change: a catastrophic or existential risk?

Climate change is considered a global catastrophic risk because it can have a long-term global effect on the environment, which means dangers for current and future generations [22]. Both an increased average temperature and the extreme ends of the temperature distribution of the weather are important as there can be gradual or sudden damage to critical ecosystems such as ice sheets and global agricultural systems, making areas with human settlements unhabitable. The extent and timing of the damage to be expected for a specific scenario of unmitigated greenhouse gas emissions is hard to quantify precisely, partially because the uncertainty about the equilibrium climate sensitivity[9] spans over several degrees [47, 62].

Most of the increase in the global average surface temperature during the last sixty years is "extremely likely" due to human influence [62]. Whether to mitigate climate change by stabilizing greenhouse gas levels in the atmosphere at a sufficiently low level therefore seems up to us. However, it has been unclear just how severe the consequences of unmitigated climate change can become. Could climate change be also an existential risk? The United Nations Framework Convention on Climate Change, adopted by almost all in 1992, forms the basis for international climate negotiations and involves the term "dangerous":

"The ultimate objective of this convention and any related legal instruments that the Conference of the Parties may adopt is to achieve [...] stabilization of greenhouse gas concentrations in the atmosphere at a level that would prevent *dangerous* anthropogenic interference with the climate system." (italics added)

International climate negotiations have since then specified limiting global warming below 2 or 1.5 degrees Celsius. However, these negotiations do not explicitly define what dangerous means or orient that term in relation to "catastrophic" or "existential", nor do they help determine whether unmitigated climate change poses an existential, or merely a catastrophic, risk. Scenarios have been proposed that could make severe climate change an existential risk, such as the so-called "Venus syndrome", discussed as a possibility by Hansen [40], in which changed atmospheric conditions could set off a runaway effect involving greenhouse gases, temperatures

[8] Followed by world war (16%), resource scarcity (13%), death of the Sun 9% and pandemic (6%) and natural disaster (5%)

[9] Defined as the effect on long-term equilibrium temperature increase from a doubling of CO_2 concentration in the atmosphere.

and boiling up our oceans, but Hansen has later withdrawn from this view [39, 41]. A recent report that surveys existential risks attaches a low probability to the possibility that the effects of climate change on our natural environment could lead to outcomes on the level of existential risk [63]. While a scenario of extreme effects cannot be entirely ruled out because it is believed unlikely, it seems more likely that if climate change were to trigger events on the magnitude of existential risks, this would follow secondary effects such as large-scale competition for access to resources such as freshwater and agricultural land, which could set off a spiral of large-scale conflicts with severe social consequences [39].

Climate negotiations present a challenge to international cooperation because of the sheer number of countries involved as well as complicated questions of historical justice, the moral status of future generations and the collective action problem among nation-states [10]. The normal way to solve collective action problems within nation-states is for the state, as an authority, to establish institutions, in order to prevent free-riding. One example is with the institution of taxation, that ensures that public goods can be achieved [10]. The absence of an international supranational authority certainly has its benefits in allowing national sovereignty. However, this situation makes it more complicated to solve the free-riding problem if the states pursue their own interests [55].

So far, the observed global warming has been gradual, but the question of critical climate thresholds, beyond which catastrophic consequences could suddenly appear, is important as we could possibly learn to detect these tipping points in advance [53, 67]. The effect of increasing emissions beyond some threshold could be to push ecosystems above some limit, setting off an irreversible, catastrophic change. Theories and experiments have been developed to handle such cases involving so-called early warning signals methods [68, 52]. Another question is how the presence of a sharp threshold for catastrophic risk affects climate negotiations. Here, both theory [11] and experiments [12] suggest that a well-known threshold could actually simplify negotiations aimed at keeping below the threshold to avoid catastrophe.

In paper III, we present and analyze a model of climate negotiations under such circumstances with a tipping point: Climate change is assumed to pose a catastrophic risk above a given emissions threshold. In paper IV, we examine the Great Filter argument, a framwork which makes it possible to reason about the level of existential risk in somewhat more general terms.

The Great Filter argument and Existential Risk

Can we reason scientifically about existential risk? While we can reason about existential risks in theory, there are several methodological difficulties in studying many of these risks based on observations. By definition, we cannot observe these events on our planet more than once, as an actual event would mean the end of being an observer. This opens up the door to speculation and arbitrariness in coming up with catastrophic scenarios. It could also be unwarranted to believe that we could produce a final list of existential risks before having observed any existential events.

The former problem has been addressed as part of the discussion about the decision-theoretic problem "Pascal's wager" [39]. The latter problem is acknowledged by using a category of "unknown unknowns" [63], but this does not solve the problem.

The Great Filter argument introduced by Hanson [42] and described further by Bostrom [42], and further quantified by Aldous [19] and Häggström [39], can be viewed as an indirect way of reasoning about generic existential risk. The Great Filter can also be viewed as a proposed answer to a question asked by Fermi in the 1950s, which has come to be known as the Fermi Paradox [74]: How can we explain the absence of scientific observations of extraterrestrial life, so far, in a vast universe much older than Earth? Assume we believe that emergence of life on a planet with progress up to intelligent life to the level of present-day human civilization is very unlikely, but that this happens with a tiny probability p (say, p to be one in a billion) on a randomly chosen planet. Also assume that, given that life on this level has emerged on a given planet, there is further a tiny conditional probability q (for the argument, let q be one in a billion) that this life will progress to a technical level beyond current human capabilities so that it becomes visible across the universe because of technological capabilities such as large-scale energy use and space colonization [2]. To start with, colonizing other planets in the Milky Way would undoubtedly take a very long time, on the scale of millions of years using technology which can allow travel at a few percent of the speed of light, and one could question whether we should be able to see signs of this yet. However, comparing these millions of years to the time scale of billions of years of the universe, this is still a short period, and many planets are estimated to be much older than Earth. A number of other possibilities are available to explain why we should not be able to see other civilizations with advanced technology [74], but we put these aside for the moment and consider the effect of these small assumed values of p and q.

Even with p and q one in a billion, chosen just for the argument, and with the number of planets in the vast visible universe estimated to be around $N = 10^{22}$ [27], we should expect there to be thousands of civilisations (around Npq) out there. Basic evolutionary arguments suggest that if there would be thousands, it would be unlikely to not have seen signs of at least one of them and that not at least one of them could have attempted colonizing space somewhere earlier in time and become observable long ago [44, 42]. Yet, we see no signs. This is the Fermi Paradox. The Great Filter is the hypothesis that there is some kind of filter on the path between the formation of a lifeless planet in the universe and a super-technological civilization that filters out all, or almost all, planets along this path. This great silence that we see suggests that Npq is not large, and if N is very large, then the great silence suggests that pq has to be very small. This means that either p or q, or both of them, have to be small.

Hanson has suggested a number of candidates for the Great Filter and the following examples are from his list [42]. First, it could be necessary for a planet to start out in exactly the same chemical circumstances around the right type of star, and here we could have been very lucky. Second, it could be extremely unlikely that

some form of primitive life emerges such as self-replicating molecules (that could initiate further biological evolution) on a lifeless planet. Currently, how hard this is seems unknown, and there seem to be no established theories that allow estimating this probability based on biological first principles about what is required for life to emerge from a setting like a primordial soup. Third, the Great Filter could relate to various levels of complexity similar to what life on Earth has passed through: as far as is known, for billions of years life on Earth went through long periods of little or no innovation, interrupted by sudden shifts and increase in complexity such as in the Cambrian explosion [42]. Over long time scales, maybe natural existential risks have enough time to wipe out almost all forms of life, and that we have been very lucky. However, we do not know about what mechanisms would strike so often in such varying circumstances so as to let life survive on almost no planets. Finally, the Great Filter may also be located somewhere ahead of a civilization on our technical level. The idea is that somewhere between our current technological capabilities and a more advanced technological civilization, whose actions become visible over longer ranges, there are threats from technology and human actions. It is exactly this that Hanson and Bostrom have argued we could learn more about if we observe extraterrestrial life [42, 19].

Suppose there is a Great Filter: What would be learned from observing extraterrestrial life which has not reached a technological super-civilization? Contrary to common opinion, Hanson and Bostrom have claimed that this observation would be bad news in terms of q. In Bostrom's words: "It would be good news if we find Mars to be completely sterile. Dead rocks and lifeless sands would lift my spirit." [19] Bostrom's argument is roughly this: Since Npq is small, and N large, pq must be small. An observation of life on another planet is evidence of a higher p which could then suggest a smaller q, and this would be bad news for the prospects of a typical civilization on our level to progress on to a super-technological level. More precisely, Bostrom writes that "The effect would be to *shift the probability* more strongly to the hypothesis that the Great Filter is ahead of us, not behind us" (italics added) [19]. The natural interpretation of this claim is that this would lower the expected value of q.

In paper IV, we analyze this argument presented by Bostrom and Hanson with a Bayesian analysis to represent the uncertainty about parameters p and q. We demonstrate that the effect on the posterior for the expected value of q depends on assumptions about the prior distribution and provide an example where the expected value of q does increase. So while the argument holds for many different priors, it does not hold for all, and in paper IV this is examined for a few different scenarios.

1.2 Game theory

Papers I, II, III, and V all make use of methods from game theory and learning. Computational simulation is used to examine properties of the models.

Game theory is the study of strategic interaction between rational agents, introduced in the 1940s by von Neumann and Morgenstern [58]. The interaction between the agents is described by a *game*, which requires defining at least three components. First, a set of two or more *players* representing the decision-making agents. Second, the set of pure *strategies* available to each player, from which a player has to make a choice, specifying the actions of the player in every possible state of the game. Third, players' preferences over the possible outcomes, described by *utility* functions. The use of a utility function is to provide a quantitative description of players' preferences over the possible outcomes. Assuming that players are *rational* corresponds to maximisation using the utility function as will be discussed more below. That a strategy specifies the actions for a player as a function of every possible state of the game means that players can take history into account when deciding what to do.

1.2.1 The Prisoners' Dilemma

A well-known game theoretic model of a strong collective action problem is called the Prisoners' Dilemma (PD), with two players that both have two available strategies [8]. The players make their choices simultaneously, and each player can either choose to "cooperate" (C) or "defect" (D). This is a model for where the outcome (C,C) is particularly hard to achieve and the utility function needs to be defined to describe this.

Figure 1.1(a) shows an example of utilities for a particular PD game. Selecting one row and one column determines utilities for both players as an ordered tuple for player 1 and player 2. Player 1 makes a choice of strategy which can be viewed as choosing one of the rows in the utility matrix. Similarly, player 2 chooses between columns. We can see that the utilities are ordered so that rational choice makes players worse off in full defection compared to the case of cooperation.

(a)

		Player 2 C	Player 2 D
Player 1	C	$3, 3$	$0, 5$
	D	$5, 0$	$1, 1$

(b)

		Player 2 C	Player 2 D
Player 1	C	R, R	S, T
	D	T, S	P, P

Figure 1.1: The single-round Prisoners' Dilemma. Utilities are a function of the strategies chosen by the row and the column players, with left utility to the row player, and right utility to the column player. Panel a) shows a particular instance of the game. Panel b) shows the general form of the utilities and the parameter values for **T**emptation, **R**eward, **P**unishment and **S**ucker ensure the properties of the game if they follow Equations 1.1 and 1.2.

Moreover, for each player choosing D is best no matter of what the other player will do: strategy D is strongly dominant. Thus, no communication or agreements (unless these change the utilities of the game) could result in cooperation between players. Figure 1.1(b) illustrates the notation of the utilities (T, R, P, S) for a generic PD game depending on the players' choices of strategies. For the game to be a PD, it is required that

$$T > R > P > S \tag{1.1}$$

holds for the utilities. To defect gives the "temptation" utility T for a player that defects when the opponent[10] cooperates. Cooperation gives a player the "reward" utility R when the other also cooperates. However, when both players defect a "punishment" utility of P results. Finally, cooperating when the opponent defects gives a "sucker's" utility S.

Rational players will thus end up in the trap of full defection. This is also the only "Pareto inferior" outcome from which both players could improve their utilities (from (P, P) to (R, R)) without making another player worse off. The game is constructed in such a manner that the only outcome which is Pareto inferior is also the only Nash equilibrium, i.e., the pair of strategies from which it is irrational to deviate (see Section 1.2.3 below). This is a strong collective action problem and also shows how individual rationality does not correspond to collective action,[11] because players do not reach any of the other outcomes, of which none is Pareto inferior.

For the repeated Prisoners' Dilemma, where the single game is repeated for a number of rounds, it is also a convention to assume that

$$2R > T + S \tag{1.2}$$

which excludes situations where the two players could find ways to take turns to cooperate and defect. Otherwise, alternating between (C, D) and (D, C) could give better outcome on the average than (C, C) and this would change the structure of the collective action problem. The point of the game is that, by assuming players are rational, we expect the players to be trapped in universal defection.

So is there any room for cooperation at all in the Prisoners' Dilemma? The only cause in the single game would be because of irrationality. However, in papers I and II, we will discuss the Finitely Repeated Prisoners' Dilemma, where the game is repeated a number of rounds and in each round there is a separate Prisoners' Dilemma. There are possibilities for both cooperation and defection in this game, and we return to this in Section 1.2.4 below.

[10] This is simply a convention to help distinguish the player, whose decisions are considered, from the other player.

[11] In the sense of collective action satisfying more of the players' preferences, not in terms of adding up utilities. The utility function assumed in game theory is normally assumed to be defined up to a positive linear tranformation (Section 1.2.2), and so aggregating utilities between the players would not be relevant here. It would be possible to take utilities for only one of the players in Figure 1.1(a) and multiply them by, say, $1,000$. This would represent the same preferences, but make the sum of utilities greater for one of the outcomes involving one player choosing C and one choosing D, unlike the base case.

1.2.2 Rationality

At this point, the reader might be wondering what is so "rational" after all about both players defecting in the Prisoners' Dilemma. If it makes players worse off, perhaps it is not so rational? We now turn to discuss some of the conditions and theory about expected utility maximization which has been termed rational [37, 17].

Rationality in game theory is usually taken to be maximizing expected utility. This is well-defined as long as there is both a utility function and a probability distribution which describes beliefs. But why would expected utility maxmization be rational and not some other decision rule? There are two well-established theories for this that attempt to provide an underlying logic and justification for this decision rule.

First, the theory by von Neumann and Morgenstern [58] shows conditions under which the utility function exist. In game-theoretic models, the starting point is usually to assume the existence of a von Neumann-Morgenstern utility function for each player [34].

Second, the theory by Savage [66] can be used to motivate subjective beliefs and Bayesian learning. Primary to both theories is the technical notion of a *preference relation* $\succcurlyeq$, which is a binary ranking of how a decision-maker ranks pair-wise choices from a set A. To write $a_1 \succcurlyeq a_2$ can be interpreted "a_1 is weakly preferred to a_2" when the decision-maker is presented with two options a_1 and a_2 from a set of available choices A.

von Neumann-Morgenstern utility functions

The decision-theoretic model presented by von Neumann and Morgenstern (vNM) has a decision-maker choosing among a set of "lotteries". Each element a_i in A is a lottery with known probabilities $p_{i,1}, ..., p_{i,m}$ for m distinct outcomes in the set of consequences $C = (c_1, ..., c_m)$. The outcomes are not necessarily monetary, as may come to mind for normal lotteries, but completely general, so that consequences can have elements that are very qualitatively different such as "apple", "death", and "+$100".

Write ΔC for all possible probability distributions over the m. The problem for the decision-maker is to choose among the available options $A \subseteq \Delta C$ to satisfy the preference described by $\succcurlyeq$. The axioms in the vNM theory makes restrictions on how the decision-maker has preferences $\succcurlyeq$ between all pairwise choices in the full set ΔC.

So far the preference relation $\succcurlyeq$ is qualitative in that it is used to make pairwise choices between any pairs a_i and a_j in A with corresponding probabilities $p_{i,1}, ..., p_{i,m}$ and $p_{j,1}, ..., p_{j,m}$. But under what circumstances does there exist a utility function u to quantitatively rank the preferences so that $a_i \succcurlyeq a_j$ if and only if $\sum u(c_k)p_{i,k} \geq \sum u(c_k)p_{j,k}$, for each a_i and a_j? This is expected utility where utilities of outcomes are weighted by the probabilities.

Here is where the vNM result provides an axiomatic justification, if the decision-maker has preferences $\succcurlyeq$ among all the lotteries in ΔC satisfying four axioms: Com-

pleteness, Transitivity, Continuity and Independence. Then, such a utility function u does exist and it is unique up to a positive linear transformation [58].

The following is a discussion to illustrate what these four axioms demand for a preference relation $\succcurlyeq$, and we use examples to illustrate what it can demand from a decision-maker. The axioms have been discussed before many times [51, 36], so the following is meant to present what is alrady well-known facts and discussions about the theory, for the purpose of being able to judge under what conditions to expect utility maximization to hold. Here are the axioms:

Completeness: For each $P, Q \in \Delta C$: $P \succcurlyeq Q$ or $Q \succcurlyeq P$.

This means that all different choices should be comparable to each other, and it seems to be the most straightforward axiom. It demands that the decision-maker should have well-defined preferences and always have well-defined preferences, even among outcomes that could appear to be qualitatively very different. It should not be hard to "compare apples and oranges", at least up to the level of indifference.[12]

Transitivity: For each $P, Q, R \in \Delta C$: if $P \succcurlyeq Q$ and $Q \succcurlyeq R$ then $P \succcurlyeq R$.

Transitive preferences rule out decision-makers being caught in cycles. Example: a cyclical preference could be as follows where $C = \{apple, banana, lemon\}$ corresponding to the outcomes to get a certain fruit. Consider the choice between the three different possible certain choices (certain lotteries) that are all elements from ΔC in the following example which violates the axiom.

Suppose the decision-maker expresses the three preferences as apple $\succcurlyeq$ banana, banana $\succcurlyeq$ lemon and lemon $\succcurlyeq$ apple. This could get into a cycle where for any item there could always be something better to satisfy. Why would this be a problem? In concrete monetary terms, a decision-maker having this structure of preferences acting on a market could end up being used as a "money pump". By handing the decision-maker any given fruit, there could exist an interest in paying a small price to get the preferred option (unless there is full indifference between the choices). But given the preferred option, there is now again another better outcome, again for a small price, and so on. Cyclical preferences can for a number of reasons be judged as irrational because they might never lead to an actual preferred outcome.

For the next axiom, we write $P \succ Q$ for the strict preference of P over Q, which means $Q \not\succcurlyeq P$ and $P \succcurlyeq Q$. Also write $pY + (1-p)Z$ for a lottery which gives $Y \in \Delta C$ with probability p and $Z \in \Delta C$ with probability $1 - p$.

Continuity: For all $P, Q, R \in \Delta C$, if $P \succ Q \succ R$ there always exists $\alpha, \beta \in (0, 1)$ so that $\alpha P + (1 - \alpha)R \succ Q \succ \beta P + (1 - \beta)R$.

[12] Indifference between lotteries a_i and a_j means that both $a_i \succcurlyeq a_j$ and $a_j \succcurlyeq a_i$ hold.

This axiom demands that qualitatively different options can be compared to each other and that the decision-maker is always willing to risk the worst outcome if the probability is just small enough.

Example from Gilboa [36]: consider where $C = \{+\$1, \$0, death\}$ and where the decision-maker expresses $death \succ +\$1 \succ \$0$. Here the axiom demands that there exists a probability α sufficiently high but below 1 so that with $P' = +\$1, Q' = \$0, R' = death$ the agent will have preferences over ΔC that for high $\alpha < 1$: $\alpha P' + (1-\alpha)R' \succ Q'$.

This means the axioms demands there is some lottery where the decision-maker actually prefers to gamble between +$1 and death instead of getting a zero monetary outcome, if death is just very unlikely. A possible response is that this is nonsense, but consider a person having to decide between buying a $1 newspaper and passing a street (with occasional cars) to get a free one. Here, the option to cross the street could represent the lottery between a small gain, with almost certainty, and with a miniscule risk of fatal car accident. Using such examples, it has been argued that real people actually choose such lotteries over certainty in everyday situations and that we take such very small fatal risks might not so easily be ruled out after all.

Independence: For every $\alpha \in (0,1)$ and $P, Q, R \in \Delta C$,
$P \succcurlyeq Q$ if and only if $\alpha P + (1-\alpha)R \succcurlyeq \alpha Q + (1-\alpha)R$.

The independence axiom means that if a lottery P is preferred to Q, then this preference will continue if we add the same possibility with the same probability to both lotteries. It also says that if two lotteries $\alpha P + (1-\alpha)R$ and $\alpha Q + (1-\alpha)R$ are compared then the decision-maker can resolve this preference by comparing only where they differ, between P and Q, and this should hold independently for all α and R.

Example: many experimental results have been found where the independence axiom seems to be systematically violated by people. One example is from Tversky and Kahneman [48] who experimentally examined individual preferences by presenting people with two different problems. Their problems were presented like this:

Problem 1
Choose between the two lotteries:

- Lottery A: $\$2,500$ with probability 0.33, $\$2,400$ with probability 0.66, and $0 with probability 0.01
- Lottery B: $\$2,400$ with probability 1 (certainty)

Problem 2
Choose between the two lotteries:

- Lottery C: $\$2,500$ with probability 0.33, $0 with probability 0.67
- Lottery D: $\$2,400$ with probability 0.34, $0 with probability 0.66

What the experiments showed was that people often prefer B over A but C over D, which violates the axiom. Note that problem 2 can be turned into problem 1 by adding the outcome $\$2,400$ with probability 0.66. If there would exist a utility function u describing this preference, we could normalize it by linear transformation to $u(\$0) = 0$ and from $B \succcurlyeq A$ get that:

$$u(\$2,400) \geq 0.33u(\$2,500) + 0.66u(\$2,400) + 0.01u(\$0)$$
$$\leftrightarrow$$
$$0.34u(\$2,400) \geq 0.33u(\$2,500)$$

But the preference $C \succcurlyeq D$ contradicts this with $0.33u(2,500) \geq 0.34u(2,400)$ showing what can happen if violating the independence assumption [48]. Tversky and Kahneman suggest this is can be explained as a certainty effect: people overweight outcomes that are certain compared to these being merely very probable. A number of models have been suggested to explain such systematic violations of, among other things, the independence axiom [48, 73].

The Savage framework and Bayesian rationality

The theory which has also come to be known as "Bayesian rationality" [17] was introduced by Savage [66] with results similar to the theory of vNM, but under more general conditions. The main conceptual difference is that the preferences are defined over choices where there are no objectively given probabilities and thus no quantified information at all, as in the vNM theory.

The situation is more general involving only qualitatively described "acts", "states" and "consequences". An act maps possible states into consequences. In this general condition, the result from Savage's theory is to present axioms [66], if satisfied for preferences between the acts, that ensure the decision-maker has preferences as if holding a "subjective" probability distribution as well as a utility function. The two then rank preferences in expected utility like described above. Now utilities are instead weighted by the subjective probability distribution.

These additional axioms have also been thoroughly discussed and questioned [36, 17], but what is interesting for this thesis is that Savage's result is consistent with learning using Bayes' rule when the decision-maker makes observations [49, 17]. Thus, this theory of choice under uncertainty has also been used to suggest a particular learning rule to update beliefs given observations. Paper II and IV makes use of models based on a Bayesian analyis.

1.2.3 Solution concepts and Nash equilibrium

A central question in game theory is how one can predict what strategies in a game get chosen by rational players. In order to do this, game theory has different *solution concepts* that describe which strategies will be chosen by the players. Two of the standard solution concepts, Nash equilibrium and backward induction arguments, will be presented below. In the following examples we illustrate the solution concepts with the Prisoners' Dilemma, without trying to generalize as much as possible.

Nash Equilibrium

When there are two players, a *Nash equilibrium* (NE) [57] (s, t) is a pair of strategies such that these strategies are best-response to each other. This means that strategy s is rational against t and t rational against s. No player has any incentive to change in terms of increasing utility, so this is thought of as an equilibrium. With the two players being described by utility functions u_1 and u_2, a pair of strategies (s, t) is a NE when both $u_1(s, t) \geq u_1(s', t)$ and $u_2(t, s) \geq u_2(t', s)$ for all s' and t', where s and s' are taken from strategy set of player 1, and t as well as t' are taken from strategy set of player 2.

In more general circumstances, each strategy in the NE is a best-response to the others, and each finite game has at least one NE [57]. The NE in the general case possibly involves players randomizing among strategies, but in the Prisoners' Dilemma, the only NE is both players defecting. This can be seen in Figure 1.1(a), which shows the utility function of a particular Prisoners' Dilemma game as a matrix. The unique NE here is (D, D), since in other cases at least one rational player could change strategy to increase utility.

1.2.4 The Finitely Repeated Prisoners' Dilemma

We have seen that defection is the only outcome for rational players in the single round Prisoners' Dilemma. In this section, a simple example is used to illustrate why rationality itself does not mean full defection in the FRPD, where the game is repeated in a finite and known number of $N > 1$ rounds. In the PD, rationality independent of belief leads to defection. In the FRPD, predictions about rational players depend on their belief about the opponent in future rounds.

The main difference compared to the one-round game is that repetition allows a player to choose a strategy that reacts to history in previous rounds. This, in turn, makes it relevant for players to consider the effect of their actions on later rounds. As we saw in Section 1.2.2, rational choice in general is determined both by rationality and belief, so that cooperation is possible for rational players if they have a belief that determines cooperation to maximize expected utility.

The role of belief can be illustrated with a simple example of the two-round FRPD. Assume first a player believes with certainty that the opponent will defect in both rounds, no matter the action played out in the first round. In this case, rational choice is to defect in both of the rounds since the maximum possible utility is $2P$. But suppose that a player believes the opponent will cooperate in the first round, and choose a follow-up action in the second to mirror the other player's action in the first round. This strategy is known as "tit-for-tat" [8]. Defecting from the first round would lead to an expected utility of $T + P$, whereas cooperation in the first round followed by defection would lead to expected utility of $R + T$. Now, given the belief we have assumed about the opponent playing tit-for-tat, and following the restrictions on the utilities in PD, it is rational for the player to start by cooperation and then defect in the second round.

This simple example shows that a player's belief can have a crucial role in de-

termining rational choice in the FRPD. Thus, some assumptions on players' beliefs have to be made to unambigously predict the outcome among rational players. It is here that backward induction comes into play: it is motivated by assuming specific beliefs about players and their rationality, but it also seem like it brings players back to the case of full defection in the FRPD.

Backward induction

Backward induction (BI) arguments are often used as solution concepts for repeated games. In finitely repeated games, such as the FRPD, in which the number of rounds is known, the solution of how players choose actions can be guided by the BI procedure as follows [64, 38]. A player can start with considering the last round, in which utility is maximized by defecting. With both players being rational in the sense that they want to maximize utility, the outcome of the last round is clear – mutual defection. But then the next-to-last round turns into the last unresolved round, and the same reasoning applies again resulting in mutual defection also for round $N-1$. The assumption needed is that each player knows that the other one is rational. The procedure then repeats all the way to the first round, reaching the Nash equilibrium with mutual defection from the start of the game [64, 38].

The procedure just described makes a number of assumptions about players' beliefs and their rationality. Under what form of beliefs is it reasonable to assume the BI outcome and can players change from BI play? This is where the basis for BI has been questioned.

1.3 The Backward Induction Problem

While BI predicts full defection in the FRPD, a number of arguments have been raised against using it as a model assumption.

First, it has been found that this prediction is very different compared to what is observed in real-world experiments. These results show that people can cooperate for many rounds in FRPD-like experiments [1] and often start defecting in the final rounds before the end of the game.

Second, the prediction from BI does not match intuition which underlies part of what has been discussed in terms of the "Backward Induction Paradox" [64]. This is not a paradox in a logical sense involving contradictions, but the term reflects that the prediction from BI runs against common intuition. Common intuition says, roughly speaking, that it could very well be rational for a player to cooperate in the first round: Hoping that the other player decides to switch to cooperation for a number of rounds, so that they both fare better than with continued defection [64]. This could perhaps explain the experimental results above.

Against these two criticism, it could be argued that the aim of game theoretic analysis is to make predictions based on assuming rational players and well-justified belief, not necessarily to describe outcomes in the real world. So is there some other critique against the model assumptions?

Third, we have such a critique that Binmore calls the "Backward Induction Problem" [18]. To understand what Binmore means with this, it is possible to start by asking: What would players who have belief about BI for the rest of the game learn if they observed cooperation (or more generally, action which does not follow from BI[13]) in the first round of the game? Two approaches seem possible here: either they change beliefs, those that lead to continued BI, or they do not. And if they could change their beliefs, should a rational player take this into account when choosing whether to cooperate or not?

In the first case, observing cooperation can change the belief about the extent of BI used in the game. We saw in Section 1.2.4 that there are some beliefs that make it rational to cooperate for some rounds, so other beliefs could also change the extent of actual BI being used. This seems to open up for some arbitrariness of choice in that many other beliefs are possible for the player who believed in BI, but got surprised by cooperation. Suppose that the player now switches to cooperation in a number of rounds: It seems like it would be rational to consider cooperating with this player but only up to the next-to-last round of the rounds where the other player. An argument like this could make cooperation unravel again, and in papers I and II we examine some of the conditions where this happens and not. This is done by modeling strategies representing different reactions to cooperation in the first step of the FRPD.

In the second case, a player with certainty about continued BI would expect future play to always continue with defection throughout the game. If players have a certainty in this belief, the belief is self-preserving and the game will continue with defection. There will be no basis for a player to consider a future effect on play leading to something else than continued defection. Using model assumptions where players will believe that beliefs about BI are maintained with certainty throughout the game is what Binmore calls the "Backward Induction Problem" [18]. Binmore refers to a set of beliefs called "Common Knowledge of Rationality" (CKR). CKR describes players who believe that others are rational, and that others believe they are rational, and so on with higher-order beliefs making it common knowledge [4]. The effects of using CKR to justify BI in the first place is what Binmore questions with the Backward Induction Problem. Aumann has shown that CKR implies BI in games of finite information of perfect information [4]. CKR assumptions also lead to BI and full defection in the FRPD [38, 7].

In a sequence of papers [15, 16, 18], Binmore has argued that players should be able to consider play which deviates from what BI predicts, reaching states which Aumann's CKR assumptions imply never happens, and that this would falsify the CKR beliefs, possibly resulting in future actions incompatible with BI. To this, Aumann maintains [6, 5] that CKR implies that players never leave the prediction

[13]More generally, Binmore's critique considers the possibility for out of equilibrium action, and has raised this question for a number of different games. Here we restrict ourselves to the game of Finitely Repeated Prisoners' Dilemma, but Binmore's critique applies to many repeated games and much of the discussion has been around the Centipede, Chain-Store games, etc.

from BI and that this is not an issue with his argument.

Binmore introduces the distinction between knowledge "as belief" and knowledge "as commitment" to explain this difference [18]. This distinction shows the difference between making a model choice where all of a player's beliefs are open to revision and where there are certain forms of beliefs that are always fixed. The question seems to become a philosophical question about whether model assumption are appropriate that prevent players from considering the revision of some beliefs in the model. With Elster's terminology from Section 1.1.2, it appears that Binmore criticises modeling players with the belief attitude of certainty.

Paper I and paper II do not directly contribute to these arguments, but the model assumptions in the paper are compatible with the first case above where some strategies use out-of-equilibrium play in the first round and players can change to cooperation. Following this are possibilities to cooperate for a number of rounds. Thus, we do not assume any CKR beliefs for players. We model the problem of how when different reactions are possible by a population of different strategies. First, in paper I, by choosing a population model from evolutionary game theory where there are no players with explicit beliefs. Second, in paper II, there is a population model of Bayesian players that learn about the strategies that react differently in the population.

1.4 Population models and learning

In four of the papers (I, II, III and V) the research questions are related to whether different forms of equilibria is the natural result of one or more mechanisms. In papers I and II, the question is whether the Nash equilibrium predicted by BI is reached when we allow different reactions and degrees of BI. Therefore, the approach can not only be to start in an equilibrium, but to examine whether equilibrium results from other starting conditions. In papers III and V the equilibria are stable climate negotiation agreements and price levels, respectively.

The approach to avoid starting in equilibrium when making assumptions about the initial conditions of the individuals is typical to that of *complexity economics* [3]. The idea here is to not start with a perfectly ordered system, but a complex system of heterogenous actions, predictions and strategies and see whether equilibrium is an outcome from interactions between the individuals. This is in line with complex systems studies on whether the interaction between many individual elements results in interesting aggregate patterns. By allowing many different starting points, the model describes a wider range of states which the system can be in. This allows to study if and how different ways are possible through which a system reaches the outcome of interest.

Paper I uses a population model based on evolutionary game theory, where the basic idea is biologically inspired: Strategies do not represent the decisions by a rational player, but the outcome of adaptive success involving modifications of genetic representations of strategies. If a strategy performs poorly in terms of utilities

when interacting with others in a population, a selective mechanism filters out this strategy over time at the advantage of others.

Paper II, III and V contain population models with explicit agents that can be heterogeneous in their traits but where learning drives the change of strategy. Similar to above, the papers study the outcome in a population of agents where everyone can interact with everyone.

1.4.1 Evolutionary game theory

Evolutionary game theory is a field that developed from combining ideas about game theory and biology and can be traced back to work by Maynard Smith and Price in the 1970's applying game theory to reason about animal behavior [54]. The approach is, roughly speaking, to have some form of evolutionary selection pressure to pick out strategies based on whether they perform well or not in a population. In evolutionary game theory, it is common to put the emphasis on the rationality of a player aside and to think of each player in a large population as having a fixed strategy. In the population of strategies, each strategy has a *fitness*, depending on the utility of the strategy in an underlying game when considered against the other strategies in the population, and the fitness influences the change in population composition over time. Instead of directly using a solution concept such as Nash equilibrium or backward induction to predict the outcome as in game theory, evolutionary game theory studies properties of some *population dynamic* for the selection among strategies in order to predict what strategies are played as an outcome by interaction within the population. It is possible to think of the dynamic as describing different behaviors in a society, the evolution in an animal population, or more generally in any multi-agent system where the outcome from the behaviour of one individual depends on the behaviour of others. A typical question can then be to consider what different outcomes form an *evolutionarily stable strategy*, which is a strategy that, if it is predominant in the population, cannot be invaded and overtaken by other strategies [54].

Two relevant properties are the size of the population (finite or continuum models) and different forms of interaction between the strategies (this could be at random or something else). A finite population has a smallest given unit of a strategy whereas the continuum models can have an infinitesimal fraction of a strategy remain in the population (it then makes sense to think about an infinite population).

The replicator dynamic is one of the simplest population dynamics which has received a lot of attention [71]. The replicator dynamic assumes that the growth rate of a specific strategy is proportional to how well that strategy performs with respect to average utility compared in the rest of the population. Each Nash equilibrium is a steady state in this dynamic (there are no incentives to change), and each steady state is a Nash equilibrium (otherwise, there would be a drive to change) [33].

An extension that adds a mutation rate $\epsilon \in [0, 1]$ leads to the *replicator-mutation dynamic* [45], where the rate of change for x_i, representing the fraction of strategy

i, in a population is

$$\frac{dx_i}{dt} = x_i \left(s_i - s - \epsilon\right) + \epsilon/n$$

where ϵ is a uniform mutation rate among the n different strategies, s_i describes the average fitness of strategy i in the population, and s describes the average overall fitness in the population.

In paper I, we use the replicator-mutation dynamic to model evolution of strategies in the Finitely Repeated Prisoners' Dilemma motivated by the Backward Induction Problem (see Section 1.3).

1.4.2 Models of learning

Many prescriptive and descriptive models of learning have been suggested in game theory, decision theory, artificial intelligence and economics [33]. A model with a learning agent often has at least three components: a model which describes how the agent makes observations, the parameters which describe the beliefs, and learning rules that update the parameters based on the observations. The model coupled with the parameters is what the agent can use to, e.g., make predictions. A learning rule often forms part of the modeling and analysis and this can be found in papers II, III, IV and V.

Bayesian learning

A theory of Bayesian rationality exists as described in Section 1.2.2, which suggests learning based on Bayesian updating [49]. Using Bayes' rule to let a player update belief, from a prior distribution to a posterior, requires a few different things.

To make Bayesian learning well-defined, two things have be specified. First, a prior distribution. This choice can have some arbitrariness to it, but it can sometimes be argued that the effect of choosing a prior becomes less relevant over time as more observations become available. Second, a model has to be chosen for the likelihood of observations, given the uncertain states of the world. This choice should also be justified and it may not be obvious what the correct model which relates observations to the possible states of the world.

In this thesis, we will use Bayes formula to present a learning rule for repeated learning among players in a population (paper II) and the effect of observations on in a Bayesian analysis (paper IV). In paper II, we will model a player as having uncertainty about the probability to find different strategies in the population within which it interacts to play the FRPD. The aim of this is to study cooperation as an effect of learning when a player interacts with opponents from a population. In paper IV, we study the effect on learning and vary the prior to model and examine Hanson's and Bostrom's Great Filter arguments (Section 1.1.3).

Fictitious play

With fictitious play, players respond to the historical frequency of play by other players. One particular form of fictitious play is that players predict the actions of

others simply as the historical average of their actions.

If players assume they are learning about a stationary environment and do not take into account other players that similarly learn and change in parallell, there is also a Bayesian justification for this in terms of updating from a prior to a posterior. One example of this is using a Dirichlet distribution as a prior together with a multinomial likelihood. The effect of observations in this case is that the probability that next observation will be s_i is simply the historical frequency of observation s_i plus the effect of some initial parameter [33]. The effect of the choice of initial parameters is then gradually diminishing.

In paper II we use a learning rule which is also comparable to fictitious play, based on this Bayesian model. However, we assume that players do not believe they are in a stationary environment and we add a term to discount previous observations. In paper III, one of the learning rules we analyze is a form of fictitious play but here the strategy set is continous for bids of emission levels, where players predict the strategy of others based on their historical bids. This is a learning rule previously used in climate negotiation models [69].

1.4.3 Level-k, strategic reasoning and learning

Level-k models [29] are used to model strategic reasoning in how players anticipate the actions of other players. Level-k models specify players on different levels of strategic reasoning. An important model component is the specification of the level-0 player, which is usually level-1 players' naive model of others. Then, players at level-2 can predict others as level-1 and use this belief as a basis for best response. Including higher levels of k can be done in different ways [29].

The level-0 specification can also represent a learning rule which "naively" predicts other players' strategies based on historical observations. This can be on forms such as simple averagring based on history. In paper III, we use a level-k approach to model strategic reasoning in a model of climate negotiations, where it can be useful to use strategic reasoning. We use the level-k approach to examine the effect of adding strategic reasoning compared to previous work which used a particular level-0 learning rule (paper III, Supplementary Information). We consider other level-0 learning rules that describe discounting of older observations.

Partial best-response dynamic

This is a form of learning which can also be viewed as a form of adaptation, as in each time step of a repeated interaction a fixed part of the population switches from its current action to a best response to the aggregate statistic from the previous period [33].

In paper V, where agents make a predicition of prices in order to maximize their profits, we also have that only a fraction of the producers can change their production decisions in each time step. Thus, our learning rule in paper V can be viewed as a partial best-response dynamic.

Rational expectations

A particular form of belief which has been suggested for models of price movements in markets is known as rational expectations [56]. This represents an idea that agents can perfectly predict the coming prices based on economic theory of equilibrium. Thus, a population of agents with only rational expectations directly end up in equilibrium since the assumed belief justifies playing equilibrium.

The concept of rational expectations is used in economic theory, but the terminology does not make it easier to distinguish rationality from belief. In paper V, we consider the possibility that agents can have beliefs by a form rational expectations that has perfect information about next year's prices.

1.5 Paper I

Motivated by the backward induction problem, we model different reactions to cooperation in the first round of the Finitely Repeated Prisoners' Dilemma. The different reactions for different extent of backward induction are described by different strategies in the strategy set. We make sure that the strategy set always contains a strategy which can defect the round before a cooperative strategy starts to defect. This is made in order to avoid introducing artificial cooperation through cooperative Nash equilibria.

We investigate if the Nash equilibrium solution for two different sets of strategies is reached in an evolutionary context with replicator-mutation dynamics.

The first set consists of conditional cooperators, up to a certain round, while the second set in addition to these contains two strategy types that react differently on the first round action: The "Convincer" strategies insist with two rounds of initial cooperation, trying to establish more cooperative play in the game, while the "Follower" strategies, although being first round defectors, have the capability to respond to an invite in the first round.

The research questions in this paper include:

1. What is the evolutionary outcome with the strategy sets of conditional cooperators, Convincers and Followers?
2. How does the choice of strategy set (the conditional cooperators versus the full strategy set) affect the results?
3. What conditions affect stable cooperation and defection levels in the evolutionary outcomes?
4. What is the effect of mutations?

The findings include that:

1. We show for the conditional cooperators that, as the mutation rate becomes sufficiently small, the cyclic behaviour disappears and the system is attracted to a stable fixed point.
2. The extended strategy set, including Convincers and Followers, allows for cycles in which cooperative players return after a period of defection.
3. For some regions in the parameter space, the evolutionary dynamics does not reach a stable fixed point, but stays in an oscillatory mode.
4. Taken together, this illustrates that the Nash equilibrium play can be unstable at the population level when mutations make explorations off the equilibrium path possible.

1.6 Paper II

In this paper, we follow up on the ideas established by paper I. From the discussion about the backward induction problem we saw that a critique against backward induction as a model assumption rules out belief revision about the extent of backward induction taking place. We model Bayesian players that consider strategies in the Finitely Repeated Prisoners' Dilemma (FRPD) that describe different lengths of applying backward induction. For players the decision between initial defection and cooperation is a strategic choice over the same strategy set as for paper I, but this model considers a finite population with learning players.

The research questions in this paper include:

1. How can we model a player's learning about the strategies in the population using a Bayesian approach?
2. How does the choice of strategy set (only the conditional cooperators versus the full strategy set, with Convincers and Followers) affect the results?
3. Are there some conditions with outcomes of recurrent cooperation in the population?
4. What is the effect of a small rate of ϵ-optimization on cooperation in the population?

The findings include that:

1. One way to model players' beliefs and learning is by a Dirichlet prior and multinomial likelihood of the observed strategies.
2. For both strategy sets, cooperation in the population can be eliminated from backwards as an effect of Bayesian learning, and it can lead to an equilibrium state of full defection.
3. The population can go through recurrent periods of cooperation for sufficiently long FRPD games if there is an arbitrarily small level of ϵ-optimization, meaning that players are indifferent between strategies within distance $\epsilon > 0$ of the optimal strategy in terms of expected utility maximization.
4. In our model, whether cooperation is fully eliminated from the population is highly sensitive to the rationality assumption of expected utility maximization.

1.7 Paper III

Two decades of international negotiations show that agreeing on emission levels for climate change mitigation is a hard challenge. However, if early warning signals would show an upcoming tipping point with catastrophic damage, theory and experiments suggested this could simplify the collective action problem to avoid catastrophe. At the actual threshold, no country would have a free-ride incentive to increase emissions over the tipping point, but it remains for countries to negotiate their emission levels to reach these agreements. We model agents bargaining for emission levels using strategic reasoning to predict emission bids by others and ask how this affects the possibility of reaching agreements that avoid catastrophic damage. It is known that policy elites often use a higher degree of strategic reasoning and in our model this increases the risk for climate catastrophe. Moreover, some forms of higher strategic reasoning make agreements to reduce greenhouse gases unstable. We use empirically informed levels of strategic reasoning when simulating the model.

The research questions in this paper include:

1. How can we extend the previous work in the literature to incorporate and model strategic reasoning?
2. What effects do higher levels of strategic reasoning have on the possibility for agents in the model to agree and avoid climate catastrophe?
3. Does the effect also hold for different level-0 learning specifications?
4. What is the effect of restricting the range of bids for the initial demands on greenhouse gas emissions?

The findings include that:

1. To model strategic reasoning in players, we use a level-k model of players, where heterogeneous levels can represent different players on different levels of strategic reasoning.
2. In our model the inclusion of higher levels of reasoning typically increases the risk for climate catastrophe. Some forms of higher strategic reasoning make agreements to reduce greenhouse gases unstable.
3. The effect also holds for a range of parameters in a different family of level-0 specifications, called exponential smoothing.
4. The measure of restricting bids in the initial round also has some effect on increasing the possibilities of agreement even when including higher levels of strategic reasoning.

1.8 Paper IV

The Great Filter interpretation of Fermi's great silence asserts that Npq is not a very large number, where N is the number of potentially life-supporting planets in the observable universe, p is the probability that a randomly chosen such planet develops intelligent life to the level of present-day human civilization, and q is the conditional probability that it then goes on to develop a technological supercivilization visible all over the observable universe. Evidence suggests that N is huge, which implies that pq is very small. Hanson (1998) and Bostrom (2008) have argued that the discovery of extraterrestrial life would point towards p not being small and therefore a very small q.

The research questions in this paper include:

1. How can we quantitatively model Hanson's and Bostrom's arguments and include the observation of extraterrestrial life?
2. What is the effect on the expected value of q in the posterior distribution by Bayesian updating for a few different priors?
3. Is there some counter-example so that the effect on q contradicts the claims in the previous arguments?
4. What would be the effect of such a counter-example?

The findings include that:

1. Using a Bayesian analysis, we model this as making observations about (1) the great silence from a vast number of planets, and (2) one observation of extraterrestrial life. This is done by modeling the effect of N independent Bernoulli trials and then one additional observation.
2. In our Bayesian analysis, our first two priors (independent uniform, and independent log-uniform) support the previous arguments and give qualitatively similar results. The third one (perfectly correlated log-uniform), however, contradicts it.
3. The example of a prior perfectly correlated on the diagonal contradicts it, and we show that there are priors which are also dense in the parameter space.
4. Taken together, this shows that the effect of one observation on q depends critically on the choice of prior distribution.

1.9 Paper V

We consider land-use competition between food and bioenergy crops when there is limited availability of land. Food price fluctuations have been an issue for centuries and there is need to understand agricultural price dynamics and reason about how mechanisms can contribute to how the prices vary over time. A particular model to address the lag between production and realization is the cobweb model, going back to work in the 1930s.

With an increased demand on biofuels, we can expect the increased demand for bioenergy crops on certain forms of land. An important question is how volatility can be reduced or controlled, and what effect the introduction of demand for bioenergy crops has on the volatility and stability of prices in the system.

The research questions in this paper include:

1. How can we model agents that predict future prices for crops on interacting cobweb markets?
2. What is the effect of having different mixes of predictors on the price volatility?
3. What is the effect of allowing the agents to switch predictors?
4. What are some mechanisms that can reduce price volatility?

The findings include that:

1. We present a model where the agents have heterogeneous production capacities, representing variation in global land quality. The markets are interlinked on the supply side by the limited availability of land.
2. When a sufficient amount of actors with perfect information about next year's prices ("rational" expectations) are introduced to our model the steady state is reached.
3. We find that a more sophisticated (but costly) predictor is concentrated to some key parcels of land, which enables the system to reduce instability significantly. We also find that the adaptive dynamics can cause booms and bust cycles.
4. In our model, the system can also be brought closer to a stable state by introducing costs for changing production type, but it may then be shifted away from the optimum situation predicted by the corresponding equilibrium model.

1.10 Future work

Some of the ideas in this thesis could be extended in several directions.

Paper I raises the question if it could be possible to analytically show that co-operation becomes recurrent for certain parameters of the model for the extended strategy space.

Paper II raises the question what would be the effect of vague observations. Currently, players make perfect observations, i.e., they observe strategies and not actions. A particular sequence of actions may be compatible with several different strategies and this could introduce ambiguity regarding which strategy is observed. It could be possible to let the players use more sophisticated learning rules and examine if players could learn that they are observing learning cyles [33].

Paper III raises the question what other mechanisms that can be relevant from behavioural sciences. Currently the players are myopic in their optimization, but what might be more relevant is looking further to the end result of the negotiations. There seems to be evidence in the behavioral economics literature that more sophisticated negotiation strategies can have an advantage in real negotiations. This could suggest new directions of this work. Other learning mechanisms than fictitious play could also be examined in this framework.

Paper IV raises the question if other consequences could be included to model the judgments from observing extraterrestrial life. It could be possible to apply this type of analysis in a more detailed framework with more factors if we have more knowledge about priors, and perhaps into more philosophical directions.

Paper V raises the question what would happen with different learning rules. One natural thing would be to examine this work with strategic reasoning. Another interesting aspect may be to infer structure and parameters of learning rules starting from real population data.

List of References

[1] James Andreoni and John H Miller. Rational Cooperation in the Finitely Repeated Prisoner's Dilemma: Experimental Evidence. *Economic Journal*, 103(418):570–85, 1993.

[2] Stuart Armstrong and Anders Sandberg. Eternity in six hours: Intergalactic spreading of intelligent life and sharpening the Fermi paradox. *Acta Astronautica*, 89:1–13, 2013.

[3] W. Brian Arthur. *Complexity and the Economy*. Oxford University Press, 2014.

[4] Robert J. Aumann. Backward induction and common knowledge of rationality. *Games and Economic Behavior*, 8(1):6 – 19, 1995.

[5] Robert J. Aumann. Reply to Binmore. *Games and Economic Behavior*, 17(1):138–146, 1996.

[6] Robert J. Aumann. Reply to Binmore and Samuelson. *The Rational Foundations of Economic Behavior. MacMillan, London*, 1996.

[7] Robert J. Aumann. Personal communication, 2014.

[8] Robert Axelrod. *The Evolution of Cooperation: Revised Edition*. Basic books, 2006.

[9] Scott Barrett. The smallpox eradication game. *Public Choice*, 130(1):179–207, 2006.

[10] Scott Barrett. Why Cooperate?: The Incentive to Supply Global Public Goods. *Oxford University Press*, 2010.

[11] Scott Barrett. Climate treaties and approaching catastrophes. *Journal of Environmental Economics and Management*, 66(2):235–250, 2013.

[12] Scott Barrett and Astrid Dannenberg. Sensitivity of collective action to uncertainty about climate tipping points. *Nature Climate Change*, 4(1):36–39, 2014.

[13] Seth D. Baum, Timothy M. Maher Jr, and Jacob Haqq-Misra. Double catastrophe: intermittent stratospheric geoengineering induced by societal collapse. *Environment Systems & Decisions*, 33(1):168–180, 2013.

[14] Daniel Bernoulli. Exposition of a new theory on the measurement of risk. *Econometrica: Journal of the Econometric Society*, pages 23–36, 1954.

[15] Ken Binmore. A Note on Backward Induction. *Games and Economic Behavior*, 17(1):135–137, 1996.

[16] Ken Binmore. Rationality and backward induction. *Journal of Economic Methodology*, 4(1):23–41, 1997.

[17] Ken Binmore. *Rational Decisions*. Princeton University Press, 2008.

[18] Ken Binmore. Interpreting knowledge in the backward induction problem. *Episteme*, 8:248–261, 10 2011.

[19] Nick Bostrom. Where are they? Why I hope the search for extraterrestrial life finds nothing. *MIT Technology Review*, pages 72–78, 2008.

[20] Nick Bostrom. Existential Risk Prevention as Global Priority. *Global Policy*, 4(1):15–31, 2013.

[21] Nick Bostrom. *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press, 2014.

[22] Nick Bostrom and Milan M. Ćirković. *Global Catastrophic Risks*. Oxford University Press, 2008.

[23] Samuel Bowles and Herbert Gintis. *A Cooperative Species: Human Reciprocity and Its Evolution*. Princeton University Press, 2011.

[24] Gru Brundtland, Mansour Khalid, Susanna Agnelli, Sali Al-Athel, Bernard Chidzero, Lamina Fadika, Volker Hauff, Istvan Lang, Ma Shijun, Margarita Morino de Botero, et al. Our Common Future ('Brundtland Report'). 1987.

[25] David Buss. The evolution of happiness. *American Psychologist*, 55(1):15, 2000.

[26] David Buss. *Evolutionary Psychology: The New Science of the Mind*. Psychology Press, 2015.

[27] Arnaud Cassan, D Kubas, J-P Beaulieu, M Dominik, K Horne, J Greenhill, J Wambsganss, J Menzies, A Williams, Uffe Gråe Jørgensen, et al. One or more bound planets per Milky Way star from microlensing observations. *Nature*, 481(7380):167–169, 2012.

[28] Leda Cosmides and John Tooby. Cognitive adaptations for social exchange. *The adapted mind*, pages 163–228, 1992.

[29] Vincent P. Crawford, Miguel A. Costa-Gomes, and Nagore Iriberri. Structural Models of Nonequilibrium Strategic Thinking: Theory, Evidence, and Applications. *Journal of Economic Literature*, 51(1):5–62, March 2013.

[30] Deborah De Felice, Giovanni Giuffrida, Giuseppe Giura, Vilhelm Verendel, and Calogero G Zarba. Information Extraction and Social Network Analysis of Criminal Sentences – A Sociological and Computational Approach. *Informatica e diritto*, 22(1):243–261, 2013.

[31] Jon Elster. Rationality, Morality, and Collective Action. *Ethics*, 96(1):136–155, 1985.

[32] Jon Elster. *Explaining Social Behavior: More Nuts and Bolts for the Social Sciences*. Cambridge University Press, 2015.

[33] Drew Fudenberg and David K Levine. *The Theory of Learning in Games*. MIT press, 1998.

[34] Drew Fudenberg and Jean Tirole. Game Theory. *Cambridge, Massachusetts*, 1991.

[35] Myndigheten för Samhällsskydd och Beredskap. Rapport, 18-årsundersökning, 2015.

[36] Itzhak Gilboa. *Theory of Decision Under Uncertainty*. Econometric Society Monographs. Cambridge University Press, 2009.

[37] Itzhak Gilboa. *Rational Choice*. MIT press, 2010.

[38] Herbert Gintis. *The Bounds of Reason: Game Theory and the Unification of the Behavioral Sciences*. Princeton University Press, 2009.

[39] Olle Häggström. *Here Be Dragons: Science, Technology and the Future of Humanity*. Oxford University Press, 2016.

[40] James Hansen. *Storms of My Grandchildren: The Truth About the Coming Climate Catastrophe and Our Last Chance to Save Humanity*. Bloomsbury Publishing USA, 2010.

[41] James Hansen. Making Things Clearer: Exaggeration, Jumping the Gun, and the Venus Syndrome, 2013. `http://www.columbia.edu/~jeh1/mailings/2013/20130415_Exaggerations.pdf`.

[42] Robin Hanson. The Great Filter – Are We Almost Past It?, 1998. `http://hanson.gmu.edu/greatfilter.html`.

[43] Yuval Noah Harari. *Sapiens: A Brief History of Humankind*. Random House, 2014.

[44] M. Hart. Explanation for the absence of extraterrestrials on Earth. *Quarterly Journal of the Royal Astronomical Society*, 16:128–135, 1975.

[45] Josef Hofbauer. The selection mutation equation. *Journal of Mathematical Biology*, 23:41–53, 1985.

[46] Alice M. Isen and Paula F. Levin. Effect of feeling good on helping: Cookies and kindness. *Journal of Personality and Social Psychology*, 21(3):384, 1972.

[47] Daniel J.A. Johansson, Brian C. O'Neill, Claudia Tebaldi, and Olle Häggström. Equilibrium climate sensitivity in light of observations over the warming hiatus. *Nature Climate Change*, 2015.

[48] Daniel Kahneman and Amos Tversky. Prospect Theory: An Analysis of Decision under Risk. *Econometrica: Journal of the Econometric Society*, pages 263–291, 1979.

[49] Nicholas M. Kiefer and Yaw Nyarko. Savage–Bayesian Models of Economics. In A. Kirman and M. Salmon, editors, *Essays in Learning and Rationality in Economics and Games*. Basil Blackell Press, 1995.

[50] Reto Knutti and Gabriele C Hegerl. The equilibrium sensitivity of the Earth's temperature to radiation changes. *Nature Geoscience*, 1(11):735–743, 2008.

[51] David Kreps. *Notes on the Theory of Choice*. Westview Press, 1988.

[52] Timothy M. Lenton. Early warning of climate tipping points. *Nature Climate Change*, 1(4):201–209, 2011.

[53] Timothy M. Lenton, Hermann Held, Elmar Kriegler, Jim W. Hall, Wolfgang Lucht, Stefan Rahmstorf, and Hans Joachim Schellnhuber. Tipping elements in the Earth's climate system. *Proceedings of the National Academy of Sciences*, 105(6):1786–1793, 2008.

[54] J. Maynard Smith and G. R. Price. The Logic of Animal Conflict. *Nature*, 246(5427):15–18, 1973.

[55] John J. Mearsheimer. The False Promise of International Institutions. *International Security*, 19(3):5–49, 1994.

[56] John F. Muth. Rational Expectations and the Theory of Price Movements. *Econometrica: Journal of the Econometric Society*, pages 315–335, 1961.

[57] John F. Nash. Equilibrium Points in n-Person Games. *Proceedings of the National Academy of Sciences of the United States of America*, 36, 1950.

[58] John Von Neumann and Oskar Morgenstern. *Theory of Games and Economic Behavior*. Princeton University Press, 1944.

[59] Martin A. Nowak, Natalia L. Komarova, and Partha Niyogi. Computational and evolutionary aspects of language. *Nature*, 417(6889):611–617, 2002.

[60] Martin A. Nowak and David C. Krakauer. The evolution of language. *Proceedings of the National Academy of Sciences*, 96(14):8028–8033, 1999.

[61] Martin A. Nowak and Karl Sigmund. Evolution of indirect reciprocity. *Nature*, 437(7063):1291–1298, 2005.

[62] Rajendra K. Pachauri, M.R. Allen, V.R. Barros, J. Broome, W. Cramer, R. Christ, J.A. Church, L. Clarke, Q. Dahe, P. Dasgupta, et al. Climate Change 2014: Synthesis Report. Contribution of Working Groups I, II and III to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change. 2014.

[63] Dennis Pamlin and Stuart Armstrong. Global Challenges: 12 Risks that threaten human civilization. *Global Challenges Foundation, Stockholm*, 2015.

[64] Philip Pettit and Robert Sugden. The Backward Induction Paradox. *The Journal of Philosophy*, pages 169–182, 1989.

[65] Anders Sandberg and Nick Bostrom. Global Catastrophic Risks Survey, Future of Humanity Institute technical report #2008-1, 2008.

[66] Leonard Savage. *The Foundations of Statistics*. Wiley, New York, 1954.

[67] Marten Scheffer. *Critical Transitions in Nature and Society*. Princeton University Press, 2009.

[68] Marten Scheffer, Jordi Bascompte, William A Brock, Victor Brovkin, Stephen R Carpenter, Vasilis Dakos, Hermann Held, Egbert H Van Nes, Max Rietkerk, and George Sugihara. Early-warning signals for critical transitions. *Nature*, 461(7260):53–59, 2009.

[69] Rory Smead, Ronald L Sandler, Patrick Forber, and John Basl. A bargaining game analysis of international climate negotiations. *Nature Climate Change*, 2014.

[70] Elliott Sober and David Sloan Wilson. *Unto Others: The Evolution and Psychology of Unselfish Behavior*. Harvard University Press, 1999.

[71] Peter D. Taylor and Leo B. Jonker. Evolutionary Stable Strategies and Game Dynamics. *Mathematical Biosciences*, 40(1):145–156, 1978.

[72] Raimo Tuomela. *Cooperation: A Philosophical Study*, volume 82. Springer Science & Business Media, 2000.

[73] Amos Tversky and Daniel Kahneman. Advances in Prospect Theory: Cumulative Representation of Uncertainty. *Journal of Risk and Uncertainty*, 5(4):297–323, 1992.

[74] Stephen Webb. *If the Universe Is Teeming with Aliens... Where Is Everybody? Fifty Solutions to the Fermi Paradox and the Problem of Extraterrestrial Life*. Copernicus Books, New York, 2002.