Beyond belief

Doctoral thesis

BEYOND BELIEF: ON REASONING IN PSYCHOLOGICAL GAMES

Niels Mourmans

2020

This book was typeset by the author using LaTeX.

Published by Global Academic Press
Cover based on creations by Zapatosoldador, NartGraphic/Shutterstock.com
ISBN: 978-94-6423-057-4
Printed in The Netherlands by ProefschriftMaken || www.proefschriftmaken.nl

BEYOND BELIEF: ON REASONING IN PSYCHOLOGICAL GAMES

Dissertation

To obtain the degree of Doctor at Maastricht University,
on the authority of the Rector Magnificus, Prof. Dr. Rianne M. Letschert, in accordance with the decision of the Board of Deans,
to be defended in public
on Friday 11 December 2020 at 10.00 hours

by

Niels Jorinus Mourmans

Supervisors
Dr. Ir. Andrés Perea
Dr. Elias Tsakas

Assessment Committee
Prof. Dr. Hans Peters (Chair)
Prof. Dr. Arno Riedl
Prof. Dr. Rineke Verbrugge (University of Groningen, Groningen)
Dr. Christian Bach (University of Liverpool, Liverpool)

Dit onderzoek werd financieel mogelijk gemaakt door de Graduate School of Business and Economics (GSBE).

Acknowledgments

My academic career was kickstarted in the autumn of 2016 when I was approached to start a PhD. Though being a bit unconfident, I did know I had a profound interest in behavioral economics, and a spring course earlier that year on epistemic game theory had instilled a new research passion in me. So eventually I decided to pick up the gauntlet. I look back at a joyful time since then. This does not at all mean it was always a smooth walk though. Finding novel connections requires a lot of insight. To verbalize and write these connections down you need to be as rigorous and precise as possible. And even then many problems take weeks to crack. In those times you often find yourself closed off in a world of your own, desperately trying to grab on to and connect all the dots that are floating around in your head. But on the outside there is meanwhile a lot going on that is helping you tackle all of this. Problems and results to be discussed, minds to be rested and recharged together, inspirations to be found from new shared experiences: I certainly did not walk my PhD-journey alone. My growth as a researcher and teacher as well as the confidence I have gained in myself I have to thank to the interactions with a lot of different people.

First and foremost, I would not have survived my journey without two great supervisors to support me: Andrés and Elias. Thank you for the past three years. I learned a lot from our discussions and from all the feedback I received from you. You are both inspiring in your own way. Andrés, your passion for the field of epistemic game theory is amazing. I admire how many events you participate in or organize in order to inspire the same passion in other people. I especially learned from you that, even though it is not always equally appreciated in the field, to keep confidence in my own work and to keep pursuing my own research passions. Elias, from you I especially appreciate how you always keep an eye on the bigger picture. You had very helpful tips on the best strategy to brand my papers and even how to brand

myself when going on the job market. But your broad view of things was most helpful research-related. Whether it is an idea of my own or someone else presenting a seminar: whatever topic is being discussed, you always show genuine interest and always have insightful comments ready.

Being under the supervision of Andrés and Elias also meant I became a member of the EpiCenter: a research group of nice people that make the rather small field of epistemic game theory feel vibrant and alive. In relation to this, I owe a big thanks to Christian Bach, Shuige Liu and Stephan Jagau, for their great enthusiasm whenever they visited Maastricht.

I would also like to thank the members of the assessment committee for taking the time to thoroughly read this thesis and provide useful feedback. Hans in particular I owe a special word of thanks. Now you are part of my assessment committee towards the end of my PhD-journey. But three and a half years ago your endorsement was the reason I could start that journey to begin with. Furthermore I want to express my gratitude to the KE-secretaries on the third floor: Yolanda, Karin and Vera. Maybe I did not come to your office as often and as long as I could have. But that was mostly because you handled whatever problem I needed help with quickly and very adequately. As long as you are around, the department is in safe hands.

My past three years would not have been half as interesting were it not for all my fellow PhDs at the Quantitative Economics department. I very much enjoyed the closeness of the entire group and I cherish the friendships I made along the way. I want to take this opportunity to highlight a few people in particular. First off, Sean, my first office-mate. If you ever wondered whether your "rambling" about teaching, causal non-causal models, the job market, societal issues or one of your semi-annual injuries ever bothered me: they did not. Where serious, your shared experiences sometimes actually proved to be helpful, even today still. And in the many other cases your laughter provided a nice distraction from work. I could only summarize your presence in one

word: *subarashii*! Qian, my next office mate. Thank you for letting me be your personal guide to Dutch/Phd-life and for listening to whatever stuff I felt I had to share in the office at random moments. Then Shash, whom I am happy to call now my current office-buddy. I may have joked about it sometimes, but your Indian cooking and spices are a delight. Your mother's spices that you always brought along made the Mensa-food much more bearable. Adi, I appreciate your ability to easily take things not too seriously very much. Your small jests and gags managed to recover my sharpness a bit during the day in case my daily can of coke failed on me. Benoit, thank you for sharing so many things: movie nights, homemade limoncello, your dark sense of humor and so many evenings of board games and card games. In particular I appreciate you teaching me Watermelon, the one game I can actually beat you at. Luca, the honest rants you tend to engage in at times are always entertaining: from complaining about the garbage-collection in Maastricht to the contents of blue envelopes we have looked at together one too many times in your office. I certainly learned that my country can be as dysfunctional as yours. Caterina and Li, I was always happy to drop by you two with questions regarding teaching, after-work-dinner plans and what not. A very cheerful office for an always cheerful duo. Farzaneh and Niloofar, *mamnun* for sharing the fun parts of Iranian culture with me. If I had managed to learn a few more words of Farsi I am sure I could have learned much more from all the conversations that I was subject to hearing from across the corridor. And then finally a few words specifically dedicated to KE's very own early bird: Dewi. Thank you Dewi for all the talks we had around eight o'clock in the morning. They were all very energizing for me. From your side I can imagine that some talks left you in confusion instead: as you mentioned many times yourself, I can be 'really vague'. But I like to consider this from the positive side: all the random stuff I mentioned must have been perfect fuel for you to use to innocently poke fun at me (Do not worry, I can easily take it).

I also wish to extend my gratitude to my other colleagues in the KE-department and beyond. This includes all the participants during the

MLSE-seminars for the interesting talks, members of the PhD-football group for the entertaining matches on Wednesdays and my faithful followers of the hiking group to entrust me to lead you guys safely to the onion soup in Noorbeek all those times. But most importantly I want to highlight my (academic) buddy here: Michelle. It started off 4 years ago with regular lunch meetings to escape to the wonderful world of behavioural economics together. We ended up with almost weekly visits to play football or to try and jam on our skates. A paper we have never made together (yet), but at least in the skeelering we have a joint work in progress.

Last but certainly not least I want to dedicate a special word of thanks to my family in helping me reach up to this point in life, particularly to my parents. I know I have not been very clear about what my research actually entails. Well, it took me three years to come up with a good explanation, but I hope this work to be presented to you will provide some much anticipated clarity.

Niels Mourmans
Maastricht
11 December 2020

Contents

1
Introduction

Whenever we make a conscious decision as human beings, we aim for a preferred outcome to take place. Often, the occurrence of such an outcome does not solely depend on our own actions. In many decision-problems that we face in our social world, the presence and actions of others influence the final outcome as well. These other actors may be other people, firms or even nations. We refer to them as *players*. Each of these players may have preferences and motivations of their own. If we are rational human beings, then we have to take such factors into account before making our decision. To decide on our best course of action, we have to form expectations of what other players decide to do and simultaneously be aware that these other players employ similar kinds of thinking processes. That is, we have to engage in strategic reasoning. Game theory is the mathematical framework in which we can model such scenarios of strategic interaction. Each particular scenario we model in the framework, we refer to as a *game*.

The classical approach to non-cooperative game theory is concerned with making predictions about the final outcome of a game in terms of the choices (or behaviour) of players. Ever since the introduction by

Nash (1950, 1951), equilibrium analysis has been the focal point in this. **Epistemic game theory** on the other hand focuses specifically on the underlying strategic reasoning of players in games and the individual behaviour that results from this. The reasoning processes of players are described by so-called *belief hierarchies*: in a game, you face uncertainty about the choice of the other player. To make a reasonable choice, you need to form a belief about the other player's choice, since the outcome of the game is dependent on her choice as well. However, the other player faces a similar uncertainty about your choice. Whatever choice you believe she makes, must also be motivated by a belief of her own about your choice. This gives rise to a new space of uncertainty for you to take into account and thus to form a belief about. Adding on top of the complexity, whatever choice you believe the other player believes you will choose in turn must be believed to be motivated by some belief of yours about the other player's choice. This adds yet another layer of uncertainty we can form beliefs about. Continuing in this manner, we get an infinite chain of beliefs. This chain of beliefs is called a belief hierarchy. A belief hierarchy thus formally captures a line of reasoning steps you as a player in a game can take.

We can impose assumptions on belief hierarchies that we deem intuitively plausible for a rational reasoner to follow and subsequently predict behaviour resulting from such reasoning. This approach is at the very heart of epistemic game theory, as is discussed thoroughly in Perea (2012). Some restrictions bring us back to classical solution concepts such as equilibria and others lead us to entirely novel predictions for games. In a world of strategic uncertainty, belief hierarchies are therefore important objects to consider when analyzing a game. Beyond the strategic uncertainty, belief hierarchies contain information that can be essential for shaping preferences of players in and of itself. This will be clarified next.

Traditional game theory has focused on interactive scenarios where individuals care about material payoffs only. One may think of price-setting by competing firms with the goal to optimize profits or costly effort decisions by individuals a in group project where efforts collec-

tively determine the quality of the final result. However, in many real-life scenarios individuals do not only have preferences that are rooted exclusively in the material outcomes of the game. Rather, they are also often motivated by the beliefs and intentions of themselves and others. These beliefs and intentions follow from the reasoning process of a player. Consider the following example.

Example 1.1 (The dinner session)**.** *You and your friend Ben have plans to cook and then have dinner together this evening. Ben is an excellent chef in his own right and therefore usually takes care of main dishes, whereas you usually prepare side-dishes. Ben is however having a heavy day of teaching, so the roles are reversed this time. Your two options are a low-effort pasta dish and a high-effort Milanese dish of osso bucco and risotto that will require much more attention. Ben's options are to prepare bruschetta or crostini directly when he gets back home. Taste-wise all four possible compositions give you an equal amount of hedonic utility. Ben however is of the strong opinion that the osso buco and bruschetta are by far the best combination, whereas the osso buco and crostini are the worst possible combination.*

There is no way to contact Ben during his teaching. And when he is done teaching you will be riding your bike to his place. Ben's teaching day is demanding, but he will pull through, in part because he is looking forward to having a nice dinner at the end of his day. If his expectations are not met, he will be disappointed. Knowing the intense day he will have had, you would feel guilty for such disappointment.

It would not be a too outlandish assumption to make that if you and Ben are good friends, you will feel guilty if Ben's dinner expectations are not met. You would be motivated to avoid this guilt. Your guilt is dependent on what you believe Ben's expectations are of your choice. If Ben all day expected you to prepare a low-effort pasta dish, you might as well do so as this meets his expectations for the evening. If you believe Ben believes you would prepare the osso buco and that would have been the reason he prepared bruschetta, you believe Ben is expecting an exceptional dinner. To avoid guilt, you will have to

spend more effort and prepare the Milanese osso buco dish. The effort would obviously be like a physical cost for you to bear, but the guilt you experience otherwise would be a *psychological cost* to you.

The motivation in Example 1.1 is called guilt-aversion and is an example of emotions determining behaviour. The idea that emotions and other-regarding preferences influence decision-making is not new and has been taken up long ago by fields such as psychology. Despite Adam Smith ([1759] 2000) himself acknowledging that emotions influence decision-making in crucial ways, modern economics has been silent on this notion for a long time (Elster, 1998). In large part, this was due to a lack of proper machinery to tackle such issues.

Traditional game theory is ill-equipped for dealing with motivations that are belief-dependent, such as guilt-aversion. Namely, it only focuses on motivations defined over the material, observable outcomes of a game, which excludes belief-dependent motivations. As a response, the more general framework of **psychological game theory** was introduced by Geanakoplos et al. (1989) and further developed and refined by Battigalli and Dufwenberg (2009). This framework has allowed for the modelling and experimental testing of a wide range of emotions and other belief-dependent motivations. This includes the aforementioned guilt-aversion (Dufwenberg, 2002; Charness and Dufwenberg, 2006; Battigalli and Dufwenberg, 2007; Attanasi et al., 2013; Attanasi et al., 2016), but also phenomena such as: intention-based reciprocity (Rabin, 1993; Dufwenberg and Kirchsteiger, 2004; Falk and Fischbacher, 2006; Sebald, 2010), frustration and anger in games (Battigalli et al., 2015), deception and lying behaviour (Battigalli et al., 2013; M. Dufwenberg and Dufwenberg, 2018; Gneezy et al., 2018), social norm conformity (Li, 2008; Charness et al., 2019), and also surprise (Khalmetski et al., 2015), which will have a central role to play in CHAPTER 4 of this thesis. The above-mentioned developments have been put to use to explain amongst other the following economic problems: optimal incentive schemes in contract theory by means of reciprocity-concerns (Livio and de Chiara, 2019), individual effort provision in public good settings by means of guilt-averse individuals (Pa-

tel and Smith, 2019) and the level of corruption by public administrators by taking into account guilt-aversion (Balafoutas, 2011).

Psychological game theory can explain both traditional game-theoretic motivations and belief-dependent motivations. In that sense, psychological game theory is indeed a more general framework. This understanding brings about a couple of new challenges. Namely, a more general framework means more cases to consider and therefore possibly more issues that can be encountered. First, game theory is inherently a mathematical framework. More complex problems may require us to enrich mathematics of the framework in order to describe said problems. When moving to the more general framework of psychological game theory, at times we therefore may need to diverge from mathematical tools that are conventionally used. Second is the selection of appropriate solution concepts to analyse games. For instance, the epistemic approach to game theory already puts the use of equilibrium concepts in traditional games under scrutiny. This is because of the correct beliefs assumption that underlies such concepts. Intuitively, this notion means that the belief hierarchy each player has is correctly inferred by all other players, and that this particular event is commonly believed by all players. If the same game is repeated many times among the same players, the reasoning processes of players can be learned in traditional games. But otherwise, there is a lack of an epistemic foundation for correctness of beliefs to occur. In psychological game theory, belief hierarchies are even more prominent in the games via the preferences of players. The question is what further problems can arise from applying particular solution concepts, from an epistemic point of view. Third, being a more general framework implies that results established in traditional game theory do not simply carry over to each psychological game as well. The cases where results fail to carry over can hide meaningful intuitions about the complexities that players are facing in psychological games.

Each of these challenges takes centre-stage in one of the chapters in this thesis.

CHAPTER 2 functions as a preliminary chapter. Here we establish the notation used throughout the thesis and formalize the concepts that are fundamental to all subsequent chapters. We show how to construct belief hierarchies and how to conveniently represent them using types in epistemic models. Moreover, a formal definition of a psychological game is provided and elaborated on. We also discuss the basic rational reasoning concept of common belief in rationality. This concept is the backbone of epistemic game theory.

CHAPTER 3 discusses cautious reasoning in psychological games. Cautious reasoning captures the idea that no choice, even if it is irrational, is ever disregarded as a possibility. In traditional games, rational reasoning processes capturing this notion are modelled using so-called lexicographic belief systems. Such systems include sequences of belief, where each next belief in the sequence is deemed infinitely less important than the previous one. This would usually be sufficient for tie-breaking between two choices. In psychological games however, lexicographic belief systems contain insufficient information to determine tie-breaking under expected utility maximization. Namely, we have to be able to quantify the relation of 'being deemed infinitely less important'. We have to enrich our mathematical framework in order to express such sentences. By employing non-standard belief systems, the issues are essentially resolved.

CHAPTER 4 offsets the notion of common belief in rationality to that of a psychological Nash equilibrium. The contrast between the two notions shows that imposing additional restrictions on reasoning processes of players can be unwanted from a conceptual point of view when we move to the analysis of a psychological game instead of a traditional game. This point is illustrated via the Surprise Exam Paradox. In this relatively old philosophical problem, it appears that rational reasoning leads to unpredictability of an event, even after the event has been announced. In the paradox, the event is a surprising exam announced by a teacher to his student. A resolution to the paradox is provided by modelling it as a psychological game. The few previous game-theoretic approaches applied equilibrium analyses to the prob-

lem. In this chapter we motivate that the restraint to use an equilibrium concept is in fact essential in providing a satisfactory resolution. The goal of the teacher is to instill surprise into the student by his actions. Surprise is an epistemic state: it is a mismatch between a choice made and a belief one formed before that choice is observed. The belief can be derived from the belief hierarchy. At the same time, solution concepts in game theory impose restrictions on the belief hierarchies of players and therefore their possible epistemic states as well. For equilibrium concepts this is the restriction that the belief hierarchy is characterized by correct beliefs. In the case of the Surprise Exam Paradox, we have a game where it is not just the case that there is no epistemic foundation for imposing a correct beliefs assumption. Rather, we show that we have a case where a very strong (epistemic) argument can be made that players should not commonly believe in correct beliefs. By relying on the notions of common belief in rationality and common belief in future rationality instead, we can resolve the paradox in an intuitive way.

CHAPTER 5 deals with the last of the challenges we mentioned. In the seminal work by Geanakoplos et al. (1989) and later on in Jagau and Perea (2017) it was made clear that results that hold in traditional game theory do not always carry over to psychological games. In particular, both papers showed how existence of common belief in rationality need not be guaranteed in a psychological game, though with differing sufficient conditions. Also shown by the latter reference by way of example is that if common belief in rationality as an event can exist in a psychological game, rational choices under this concept are no longer guaranteed to be characterized by iterative procedures as they normally would have. That is, in traditional games we are able to characterize solutions of games by iteratively eliminating choices or strategies. These elimination procedures are intuitive in use and could be perceived as reasoning steps taken by the players themselves (Cubitt and Sugden, 2011; Perea, 2015). Rational choices under common belief in rationality are usually characterized by iterative elimination of strictly dominated choices. In Chapter 5 we find

the families of two-player (expectation-based) psychological games in which this elimination does always work. The games in these families could be viewed as equally easy to reason about as traditional games. We characterize these games based on which orders of beliefs are directly utility-relevant for a player. By introducing the notion of causality diagrams which captures those orders of beliefs that are (indirectly) utility-relevant, we can distinguish between three different cases of families of psychological games. Psychological games that fall outside these three cases bear additional complexities. More specifically, the complexity I found that belief-dependent motivations introduce is that the reasoning processes about two different rationality-events can overlap and interfere with each other. When this happens, iterated elimination of choices will no longer characterize the choices that are possible under common belief in rationality. In such situations decision-makers need to engage in more complex elimination procedures to reason in line with common belief in rationality.

This all challenges the commonly held view that when dealing with emotions, many of which are belief-dependent, irrationality is the norm. In many cases this is not true: to take into consideration motivations such as your own guilt and anger or that of others may in fact require much more complex reasoning. In addition, players need to be extra careful in making assumptions about how the other player reasons. Restrictions on the reasoning process also imply restrictions on the possible epistemic states. When aware of the belief-dependent motivations of others, these motivations and restrictions combined can be illogical in some games. Surprise and correct beliefs are one such example.

2

Basic Definitions and Concepts

Each of the chapters in the main body of this thesis tries to challenge different problems. However, they all find root in the same foundations. We look at all problems with the same fundamental framework of psychological game theory in mind and consider rational reasoning of players that can always be traced back to belief hierarchies that express the basic principle of common belief in rationality. A formal discussion of all these elements is therefore in place.

2.1 Belief hierarchies

In a traditional setting, a player's experienced, ex-post utility depends only on her own choice and her opponents' choices. The player's expected utility of making a particular choice then depends only on her first-order belief of what she expects her opponents to choose. As opposed to this, a player's utility in a psychological game can explicitly, and non-linearly, depend on any order of belief or even the entire belief hierarchy. For instance, in Example 1.1, your utility depends on what you believe Ben expects you to prepare. In order to formally discuss the framework of a psychological game, we should therefore clarify what a belief hierarchy formally is. A belief hierarchy b_i for a player i represents an infinite chain of beliefs. The first element in this chain represents the first-order belief about the opponents' choices. In Example 1.1 this would correspond to what you believe Ben is going to prepare. The second element represents the second-order belief about the opponents' choices combined with the opponents' beliefs about their opponents' choices. In our example, this would for instance be a belief you have about what Ben prepares and a belief about what Ben expects you to prepare. The third element represents the third-order belief about the combination of opponents' choices, opponents' first-order beliefs and the opponents' second-order beliefs. And so on.

Following Brandenburger and Dekel (1993), we formally define beliefs over spaces of uncertainty. These spaces have to be measurable, for each additional layer of uncertainty we add. Consider any Polish space

S. Let $\Delta(S)$ be the set of probability measures on the Borel σ-field over the space of uncertainty S. Finally, endow $\Delta(S)$ with the topology of weak convergence. Then $\Delta(S)$ is a Polish space as well. The primitive space of uncertainty for player i is the set of opponents' choices $\underset{j\neq i}{\times} C_j = C_{-i}$. We can recursively define

$$\begin{aligned} X_i^1 &:= C_{-i} \\ X_i^2 &:= X_i^1 \times \bigtimes_{j\neq i} \Delta(X_j^1) \\ &\vdots \\ X_i^n &:= X_i^{n-1} \times \bigtimes_{j\neq i} \Delta(X_j^{n-1}) \\ &\vdots \end{aligned}$$

Then by $\tilde{B}_i := \times_{n=1}^{\infty} \Delta(X_i^n)$ we denote the set of all possible belief hierarchies for player i. Each belief hierarchy b_i is a vector of (higher-order) beliefs $(b_i^1, b_i^2, ...)$, where the n-th order belief of player i is a probability distribution $b_i^n \in \Delta(X_i^n)$. In the current set-up a belief hierarchy $b_i \in \tilde{B}_i$ may be incoherent in the sense that an n-th order belief may contradict what is stated by the $(n-1)$-th order belief. When defining psychological games, we assume a player's belief hierarchy cannot show such incoherences. More formally, we define coherency as follows.

Definition 2.1. *A belief hierarchy $b_i = (b_i^1, b_i^2, ...)$ expresses **coherency** if for every $n > 1$ we have*

$$marg_{X_i^{n-1}} b_i^n = b_i^{n-1}.$$

Let player i's set of coherent beliefs be denoted by $\tilde{B}_i(1) \subseteq \tilde{B}_i$.

We can make use here of Brandenburger and Dekel (1993)'s Proposition 1. From this proposition we know there exists a homeomorphism $f_i : \tilde{B}_i(1) \to \Delta(C_{-i} \times \tilde{B}_{-i})$. Thus a coherent belief hierarchy can be identified with a probability distribution over the possible combinations of opponents' choices and belief hierarchies. Next to expressing coherency, a player can also believe her opponents express coherency, believe that her opponents believe their opponents express coherency, and so on. This restricts the set of belief hierarchies we will consider further. We can recursively define such sets of belief hierarchies:

$$\tilde{B}_i(k) = \{b_i \in \tilde{B}_i(k-1) | f_i(b_i)(C_{-i} \times \tilde{B}_{-i}(k-1)) = 1\},\ k \geq 2.$$

Consider the set $B_i = \bigcap_{k \geq 0} \tilde{B}_i(k)$. We say a belief hierarchy b_i expresses *coherency and common belief in coherency* if $b_i \in B_i$. Throughout this thesis, whenever we refer to a belief hierarchy b_i, we assume it to be a belief hierarchy in B_i, even if it is not directly stated as such.[1] Moreover, note that $b_i \in B_i$ can be identified by a probability distribution over $C_{-i} \times B_{-i}$ through the homeomorphism f_i as was used in defining $\tilde{B}_i(k)$ for any $k \geq 1$. We will make use of this fact several times throughout the thesis.

2.2 Psychological games

With these elements in place, we can now give a formal definition of a psychological game. All chapters in this thesis concern *static* psychological games, without updating of beliefs. Only in Section 4.4 of CHAPTER 4 we also deal with dynamic psychological games. The necessary tools to analyze such dynamic scenarios we discuss only in that particular chapter. For defining static psychological games we follow the approach taken by Jagau and Perea (2017).

[1] This is the traditional take on belief hierarchies in epistemic game theory. A more recent approach takes coherency and common belief in coherency as a condition on the belief hierarchy that is derived from rational reasoning. See Battigalli et al. (2020).

Definition 2.2. *A **psychological game** is a tuple $G = (C_i, B_i, u_i)_{i \in I}$, where I is the of players, C_i is the finite set of choices for player i[2], B_i denotes the set of belief hierarchies expressing coherency and common belief in coherency and*

$$u_i : C_i \times B_i \to \mathbb{R}$$

is player i's (measurable) utility function.

By this definition, we capture the idea that player i's utility depends explicitly on her full belief hierarchy. Formally speaking a psychological game is a generalisation of a traditional game, since the utility function in a traditional game exclusively depends on first-order beliefs. Moreover, utilities in a traditional game always depend linearly on first-order beliefs. This is not true for psychological games in general, where utilities can depend non-linearly on the full belief hierarchy. Definition 2.2 differs from definitions used in the seminal work by Geanakoplos et al. (1989) and Battigalli and Dufwenberg (2009). Under these two definitions, utility also still explicitly depends on the opponent's choices. In case of the latter approach, utility moreover explicitly depends on the opponents' belief hierarchies as well. These two elements are helpful in visually distinguishing between preferences over outcomes and belief-dependent motivations. However, as Jagau and Perea (2017) point out, all these approaches are essentially equivalent. This can be seen by noting that a belief hierarchy can be identified by a probability distribution over the combinations of the opponents' choices and *their* belief hierarchies $C_{-i} \times B_{-i}$. Hence, a belief hierarchy already includes a conjecture about the opponent's choices and opponent's belief hierarchies. In terms of utility that is deemed relevant at the moment of making a decision, all approaches are thus equivalent in an expected utility framework. The approach we take here will be particularly helpful in the notation of CHAPTER 5, where in the constructive proofs we want to directly couple (rational) choice with belief hierarchies.

[2] C_i may well be a singleton set, indicating a situation where player i does not have any choices to make but where his beliefs matter for the utilities of other players.

Table 2.1: *Example 1.1: Dinner session, traditional game*

		Ben	
		BR	CR
You	$High$	$1, 5$	$1, 0$
	Low	$3, 2$	$3, 3$

Table 2.2: *Example 1.1: Dinner session, psychological game*

Your extreme second-order expectations

	$(BR, High)$	(BR, Low)	$(CR, High)$	(CR, Low)
$High$	1	1	1	-2
Low	0	3	3	3

Your utilities

Ben's extreme first-order beliefs

	$High$	Low
BR	5	2
CR	0	3

Ben's utilities

We can represent psychological games in a similar way as we represent traditional games. All games we consider are so-called belief-finite, expectation-based games. Most games in applications of psychological games belong to this class, and they allow for a matrix-representation of the game (Jagau and Perea, 2018). For an in-depth discussion of this, we refer the reader to CHAPTER 5.

Let us again consider Example 1.1 where you and Ben have a dinner session. If we do not consider the belief-dependent motivations, we can represent the game in matrix form, as is the usual practice for finite static games. This can be seen in Table 2.1. Here, you have the choices $High$ (high-effort osso buco with risotto) and Low (low-effort pasta). Ben can choose to prepare bruschetta (BR) or crostini (CR). As explained in the story, you do not care too much for gastronomy: your choice $High$ always leads to a utility of 1 because of higher effort and your choice Low always leads to a utility of 3 because of lower effort, regardless of what Ben prepares. Ben on the other hand slightly prefers

to prepare CR over BR, but only if you do not prepare the high-effort dish. In the latter case, Ben definitely prefers to prepare BR.

Now let us assume you display guilt-averse motivations. We have two different decision-problems: one for you and one for Ben. Together they constitute the psychological game. This psychological game can be seen in Table 2.2. Ben's utilities are as in a traditional game and can be seen in the lower matrix with utilities. Your utilities change now however, and are dependent on the second-order beliefs. You experience guilt proportional to the utility Ben loses compared to his expectations. This can be seen in the upper matrix of the table. The columns here represent extreme second-order expectations. These are probability-one second-order beliefs. The first column $(BR, High)$ indicates the probability-one second-order belief of you that Ben will prepare BR while he believes you prepare the high-effort dish. If you indeed play $High$ under such a probability-one second-order belief, you believe you will meet Ben's expectations and hence receive a utility of 1, as was the case in the traditional game. If you however play Low, you believe you will fail to meet Ben's expectations for the evening. You believe Ben will fall short 3 units of utility compared to what he expected. For this you will feel guilt, incurred as a psychological cost of 3. Hence, instead of a utility of 3, you will have a utility of 0 when playing Low under this second-order belief. The other columns can be interpreted in exactly the same way (assuming that you are not motivated by 'negative' guilt). Note that in the traditional game the high-effort dish is strictly dominated for you, but in the psychological game it is not.

2.3 Common belief in rationality in psychological games

There are many different ways in which players of a game can come to a decision. A choice may have been made after a very intricate reasoning process, or a choice may be made by a player who is biased in

some way or is limited in her cognitive abilities. It is therefore helpful to have a benchmark to compare all different cases of (limited or biased) reasoning to. This benchmark is that of a rational reasoner. A basic reasoning concept we would expect a perfectly rational reasoner to adhere to is called *common belief in rationality*.

Intuitively, common belief in rationality captures the following idea. A rational reasoner of course wants to make a choice that would yield her the highest utility. However, a player faces uncertainty in a game. So she would decide on the choice that would yield her the highest expected utility given her beliefs about her opponent's choice (in case of a two-player game) and, if belief-dependent motivations are involved, given her belief about her opponent's beliefs, et cetera. However, if all players are assumed to be rational reasoners, she must believe her opponent is trying to do the same thing. So in her belief about her opponent's choice, she must believe that her opponent is trying to maximize his expected utility given his reasoning process. But then along the same argument, she must also believe that her opponent believes she is a rational reasoner. Therefore, she must believe that her opponent believes that the choices of her he assigns positive probability to must maximize expected utility given some line of reasoning. We can continue this reasoning process up to infinity. A belief hierarchy that is in line with such a reasoning process expresses common belief in rationality. Thus, common belief in rationality is an event that a line of reasoning can lead to.

In short, common belief in rationality conveys the idea that nowhere in one's belief hierarchy the rationality of any player is questioned. It has been defined and studied extensively in traditional games. The notion can at least be accredited to Spohn (1982), Bernheim (1984), Pearce (1984) and Tan and Werlang (1988). Common belief in rationality in psychological games captures exactly the same idea as in traditional games. It was first presented in Battigalli and Dufwenberg (2009) by their discussion of common strong belief in rationality in dynamic psychological games (the equivalent of this concept in static games is common belief in rationality). Later it was the focal point in Jagau and

Perea (2017). Our definition of the concept follows that of the latter paper.

Common belief in rationality in a psychological game can be defined recursively. First we consider an *optimal* choice given any belief.

Definition 2.3. *Consider a psychological game G. A choice c_i is **optimal** for a belief hierarchy $b_i \in B_i$ if for all $c_i' \in C_i : u_i(c_i, b_i) \geq u_i(c_i', b_i)$.*

Let $RB_i := \{(c_i, b_i) \in C_i \times B_i | c_i \text{ optimal given } b_i\}$ be the set of combinations of choices and belief hierarchies where the choice is optimal (or: rational) for the linked belief hierarchy. Then we can define what it means to believe in an opponent's rationality. To this end, recall that every belief hierarchy $b_i \in B_i$ is homeomorphic to a probability distribution in $\Delta(\bigtimes_{j \neq i} (C_j \times B_j))$.

Definition 2.4. *Consider a belief hierarchy $b_i \in B_i$ for some player i in G. Player i is said to **believe in the opponent's rationality** with belief hierarchy b_i if $b_i(RB_{-i}) = 1$.*

In line with Spohn (1982), Bernheim (1984), Pearce (1984) and Tan and Werlang (1988) for standard games, we can iterate this argument to get the notion of common belief in rationality in a psychological game (Jagau and Perea, 2017).

Definition 2.5. *Consider a belief hierarchy $b_i \in B_i$ for some player i. Define $B_i(1) = \{b_i \in B_i | b_i(RB_{-i}) = 1\}$. If $b_i \in B_i(1)$, we say b_i expresses **1-fold belief in rationality**.*
*Define $B_i(k) = \{b_i \in B_i(k-1) | b_i \in \Delta(\bigtimes_{j \neq i} (C_j \times B_j(k-1)))\}$, for every $k \geq 1$. We say b_i expresses up to **k-fold belief in rationality** if $b_i \in B_i(k)$. If for every $k \geq 1$, b_i expresses up to k-fold belief in rationality, we say b_i expresses **common belief in rationality**.*

We assume here that the events of expressing k-fold belief in rationality are measurable. Finally, let a rational choice under k-fold belief in rationality and common belief in rationality respectively be denoted as follows.

Definition 2.6. *Consider a choice $c_i \in C_i$. We say c_i is **rational under k-fold belief in rationality** if it is optimal for some belief hierarchy $b_i \in B_i$ that expresses up to $k - fold$ belief in rationality. We say c_i is **rational under common belief in rationality** if it is optimal for some belief hierarchy $b_i \in B_i$ that expresses common belief in rationality.*

This definition of common belief in rationality and rational choice under a belief hierarchy that expresses common belief in rationality is applicable to any class of psychological games. The concept of rationalizability as developed by Bernheim (1984) and Pearce (1984) is very similar to common belief in rationality. Common belief in rationality is a more general concept, allowing for beliefs to be correlated. Rationalizability requires marginal beliefs about opponents' choices to be independent from each other. For two-player games this distinction is irrelevant, as there are no different sets of choices to marginalize over. For two-player games, the two concepts are therefore equivalent (Tan and Werlang, 1988).

We now turn to how belief hierarchies, psychological games and common belief in rationality come together in Example 1.1 of the dinner session. Say you have the specific first-order belief $b_y^1 = BR$. In words this means that with probability-one you believe Ben will prepare bruschetta. A specific second-order belief you can have at the same time is $b_y^2 = (BR, 0.8High + 0.2Low)$. This means that with probability one you believe that Ben will prepare bruschetta while he believes with probability 0.8 that you will prepare the high-effort dish and with probability 0.2 that you will prepare the low-effort dish. A third-order belief you could have is $b_y^3 = (BR, 0.8(High, BR) + 0.2(Low, CR))$. This means that you believe with probability-one that Ben will prepare bruschetta and believe that (i) with probability 0.8 Ben believes

you prepare the high-effort dish and that you believe with probability one that Ben will prepare bruschetta, and (ii) with probability 0.2 Ben believes you prepare the low-effort dish and that you believe with probability one that Ben will prepare crostini. As we move to even higher orders of belief, the objects become even larger and therefore more difficult to express. To determine whether a belief hierarchy expresses common belief in rationality, we would have to continue this process of expressing higher-order beliefs indefinitely, as belief hierarchies involve infinite chains of higher-order beliefs.

Such a process of expressing belief hierarchies can be a very cumbersome endeavour. This is unfortunate, as common belief in rationality is the main building block for all the solution concepts we will employ in the coming chapters. Fortunately, there are methods for modeling belief hierarchies conveniently. The method employed here entails capturing infinite belief hierarchies in an epistemic model. Such an epistemic model relies on assigning types to players, a concept first put forward by Harsanyi (1967-1968). Every type $t_i \in T_i$ holds a belief about the opponents' choice-type combinations. As such, one can derive an infinite chain of beliefs for every type.

Definition 2.7 (Epistemic model in a static psychological game)**.** *Consider a psychological game G. An **epistemic model** $M = (T_i, b_i)_{i \in I}$ for G specifies for every player i a finite set T_i of possible types. Moreover, for every player i and every type $t_i \in T_i$ the epistemic model M specifies a probability distribution $b_i[t_i]$ over the set of opponents' choice-type combinations $C_{-i} \times T_{-i}$. The probability distribution $b_i[t_i]$ represents the belief player i has about the choice-type combinations of her opponents.*

The idea is as follows. Consider the epistemic model in Table 2.3. Take type t_y^{High} for you. If you are of this type, you believe that Ben will play BR and is of type t_b^{BR}. This corresponds with your first-order belief b_y^1 we expressed before in words. Type t_b^{BR} of Ben believes with probability 0.8 that you play $High$ while being of type t_y^{High} and with probability 0.2 that you play Low while being of type t_y^{Low}. Putting t_y^{High} and

t_b^{BR} in a chain, we then retrieve the second-order belief b_y^2 for you that we expressed earlier. Continuing in this fashion, as mentioned before, type t_y^{High} induces the belief that Ben will play BR. Type t_y^{Low} induces the belief that Ben will play CR. Starting at type t_y^{High}: you believe Ben is of type t_b^{BR} and consequently believe Ben believes (i) with probability 0.8 you are of type t_y^{High} and thus believe Ben will play BR and (ii) with probability 0.2 you are of type t_y^{Low} and thus believe Ben will play CR. Putting together with the first-order and second-order beliefs we found, this is exactly the third-order belief b_y^3 we described earlier. Continuing in this fashion, we can retrieve the entire belief hierarchy that the type t_y^{High} represents, and we can do this for every type in the epistemic model.

The model also allows us to check whether a player expresses k-fold belief in rationality for some k by following these chains of types. Let us start with checking if type t_y^{High} expresses 1-fold belief in rationality. Take type t_y^{High} for you. Then you believe Ben is of type t_b^{BR}. Given the belief that type t_b^{BR} induces, we can conclude that indeed BR is an optimal choice for that belief: playing BR gives an expected utility of $0.8 \cdot 5 + 0.2 \cdot 2 = 4.4$, whereas playing CR would only give an expected utility of $0.8 \cdot 0 + 0.2 \cdot 3 = 0.6$. So type t_y^{High} for you, by believing Ben plays BR while being of type t_b^{BR}, believes in Ben's rationality and thus express 1-fold belief in rationality. Next we check if

Table 2.3: Epistemic model Example 1.1: dinner session

Types	$T_y = \{t_y^{High}, t_y^{Low}\}$ $T_2 = \{t_b^{BR}, t_b^{CR}\}$		
Beliefs for You	$b_y[t_y^{High}]$ $b_y[t_y^{Low}]$	= =	(BR, t_b^{BR}) (CR, t_b^{CR})
Beliefs for Ben	$b_b[t_b^{BR}]$ $b_b[t_b^{CR}]$	= =	$0.8(High, t_y^{High}) + 0.2(Low, t_y^{Low})$ (Low, t_y^{Low})

type t_y^{High} for you expresses 2-fold belief in rationality. Following the chain of types from type t_y^{High} we next get to type t_b^{BR}. Type t_b^{BR} of Ben induces the belief $0.8(High, t_y^{High}) + 0.2(Low, t_y^{Low})$. If you are of type t_y^{High}, then you believe Ben plays BR while he believes you play with probability 0.8 $High$ and with probability 0.2 Low. Given this second-order belief, we derive your expected utility from Table 2.2 to be 1 if you play $High$ and $0.8 \cdot 0 + 0.2 \cdot 3 = 0.6$ if you play Low. So given type t_y^{High}, $High$ is an optimal choice. Similarly, if you are of type t_y^{Low}, then you believe Ben plays CR while he believes you play Low. Given type t_y^{Low}, Low is therefore an optimal choice. Then if Ben is of type t_b^{BR}, he believes you make an optimal choice given the belief hierarchy represented by your type. In other words, Ben believes in your rationality and thus expresses 1-fold belief in rationality. Since type t_y^{High} believes with probability one that Ben is of type t_b^{BR} and this type expresses 1-fold belief in rationality, type t_y^{High} expresses 2-fold belief in rationality. And so on. In fact, type t_y^{High} expresses common belief in rationality.

With respect to the event of common belief in rationality, we can make an important analytical remark that we will make use of frequently in the coming chapters.

Remark 2.1. *Consider an epistemic model $M = (T_i, b_i)_{i \in I}$ for some psychological game G. If each type in M expresses 1-fold belief in rationality, then each type in M also expresses common belief in rationality.*

This remark follows directly from the definitions of belief in the opponent's rationality and common belief in rationality: common belief in rationality entails that at no point in the belief hierarchy a player's rationality is questioned. If we can only construct belief hierarchies from a set of types that all express 1-fold belief in rationality, then we cannot construct belief hierarchies that at some point question the rationality of any of the players. In the epistemic model in Table 2.3 it is the case that each type expresses 1-fold belief in rationality. We already

concluded this for the types t_y^{High} and t_b^{BR}. We leave it to the reader to verify this for the remaining two types. The resulting conclusion by applying Remark 2.1 is that each type in the epistemic model expresses common belief in rationality.

A second and final remark with respect to epistemic models is about notation. By means of an epistemic model as defined earlier we can write the utility function as $u_i(c_i, t_i)$, as t_i itself represents a belief hierarchy.

Remark 2.2. *Consider an epistemic model $M = (T_i, b_i)_{i \in I}$ for some psychological game G. We can represent the utility function $u_i(c_i, b_i)$ given some combination of a choice and belief hierarchy (c_i, b_i) by $u_i(c_i, t_i)$, where t_i represents the belief hierarchy b_i.*

In CHAPTER 3 and CHAPTER 4 we will use this notation, as the use of epistemic models in these chapters are directly relevant. In CHAPTER 5 we write utility functions as $u_i(c_i, b_i)$. In this chapter epistemic models are only introduced late for the construction of the proof of the main theorem.

3

Cautious Reasoning in Psychological Games

This chapter is adapted from: "Cautious reasoning in psychological games" (Mourmans, 2018).

3.1 Introduction

Cautious reasoning is an essential component of various concepts of traditional game theory. It encapsulates the idea that a decision-maker has beliefs that do not disregard any of his opponents' options and explains the epistemics behind solution concepts such as elimination of weakly dominated strategies (Luce and Raiffa, 1957; Blume et al., 1991a; Brandenburger et al., 2008), perfect equilibrium (Selten, 1975; Blume et al., 1991b), proper equilibrium (Myerson, 1978; Blume et al., 1991b) and permissibility (Brandenburger, 1992; Börgers, 1994). To give meaning to rational decision-making in scenarios with caution, cautious beliefs are typically modelled using lexicographic probability systems.

A *lexicographic belief* is a finite sequence of beliefs (or "theories") in which the first belief in the sequence is deemed infinitely more important than the second, the second belief infinitely more important than the third, and so on. A player can derive preferences over choices for every belief in the sequence, where preferences based on earlier beliefs in the sequence take precedence over preferences based on beliefs later in the sequence. Under such an interpretation, for a choice to be considered optimal under a lexicographic belief in a traditional game, it must be optimal given the first belief in the sequence. In case of a tie, the choice must then, among those choices with which there was a tie at the previous level, also be considered optimal given the second belief in the sequence. And so on, until the tie is resolved or the sequence ends. Belief in the opponent's rationality can also still be defined in a traditional game with lexicographic beliefs. For instance, we may say a player with lexicographic beliefs believes in his opponent's rationality if in his first belief in the sequence he only considers rational alternatives for his opponents. Then the decision-maker is left free to consider irrational alternatives in beliefs that are 'further' in the sequence.

For similar reasons as in traditional games, one may want to model cautious beliefs for decision-makers in *psychological games*. Psychological games differ from traditional games in the sense that they are

Table 3.1: *Game of King and Queen: Gift*

		Beliefs Queen	
		Games	*Land*
King	*Games*	0	1
	Land	1	0

able to model belief-dependent motivations as well. That is, utilities in psychological games may depend explicitly on the full belief hierarchy instead of just what a player believes his opponents will choose (Geanakoplos et al., 1989; Battigalli and Dufwenberg, 2009).

To illustrate this idea of cautious reasoning in psychological games, consider the game of a King and a Queen in Table 3.1. Being fellow monarchs of neighboring lands, the King finds it a good idea to improve his personal relationship with the Queen by surprising her with a gift during her next visit to the Kingdom. He is thinking of either organizing gladiator games in her honor, lasting for half a year, or transferring a large portion of the coastal lands of his Kingdom that are littered with beaches. The Queen is aware the King has the intention of surprising her with either of these two options. If the King organizes the Games while the Queen expected the transfer of the coastal lands, the King receives a utility of 1. If the King transfers the Land while the Queen expected the Games, the King also receives a utility of 1 by surprising her. All other extreme scenarios result in a utility of 0. If with some probability p he expects the Queen to be incorrect in her prediction of what he will do and with probability $1 - p$ to be correct in her belief, then the King receives a utility of p.

Let us assume that the King knows the Queen is a cautious reasoner and will not disregard either of the two options the King has. Also let us assume that the King with probability $\frac{1}{2}$ thinks that the Queen considers it infinitely more likely that the King will transfer the coastal lands compared to organizing Games and with probability $\frac{1}{2}$ that the Queen considers it infinitely more likely that the King will organize

Games instead of transfer the lands. These orderings can be represented by the King's second-order lexicographic beliefs.

In traditional games, in case of a tie in terms of preferences over choices given a previous belief in a sequence, a player would move further down in his lexicographic belief until the end of the sequence or until the tie is resolved. The case here is that the King's preferences depend on his second-order beliefs, i.e. in particular his beliefs about the Queen's lexicographic beliefs about his choice. In fact, the King deems possible two different first-order lexicographic beliefs for the Queen. Each of these lexicographic beliefs specify the relative importance of her theories within each lexicographic belief. However, we cannot make comparisons between theories from different lexicographic beliefs without imposing extra assumptions [Section 3.2]. For instance, it is well possible that deeming something 'infinitely more likely' means something different probabilistically in the two possible lexicographic beliefs for the Queen that the King considers. In this regard, we can ask ourselves the following question: does the King expect the Queen to be more likely to believe the King will choose Games or Land? At face value, the answer to this is not obvious, yet is relevant for determining the King's preferences.

Preferably, to compare theories from two different sequences of beliefs, one wants to be able to quantify what it means for one theory to be 'infinitely more important' than another. This would involve assigning infinitely small numbers to events that are deemed infinitely less likely to happen than others. This is a feature that non-standard analysis accomplishes by constructing *infinitesimals* (Robinson, 1973) [Section 3.3].

In the remainder of this chapter we will elaborate further on the shortcomings of lexicographic beliefs in modeling cautious reasoning in psychological games. Moreover, the expressive power of non-standard analysis will be contrasted to this. In Section 2, the problem of lexicographic beliefs in psychological games will be discussed. Section 3 illustrates the necessary expressive power of non-standard beliefs in

psychological games and shows how these beliefs can be used to define the cautious reasoning concept of permissibility for such games. Section 4 provides a short conclusion.

3.2 The problem with lexicographic beliefs

Following Blume et al. (1991a), we can define a lexicographic belief as follows.

Definition 3.1. *A **lexicographic belief** b_i for player i on a finite set X is a finite sequence of beliefs $(b_i^1, ..., b_i^n)$, where each element specifies a probability distribution on X. We call b_i^1 the primary theory, b_i^2 the secondary theory, and so on.*

Thus, a lexicographic belief captures an ordering of beliefs, where the $(l-1)^{th}$ belief is deemed infinitely more important than the l^{th} belief. We could impose natural conditions on these sequences of beliefs, such as that $b_i^l \neq b_i^m$ for every $m \neq l$. This ensures no inconsistencies in the ordering occur. That is, we cannot have a scenario where e.g. $b_i^l = b_i^{l+2}$ is deemed infinitely more important than b_i^{l+1} and vice versa. However, as this section will show, regardless of whether lexicographic beliefs meet such conditions, problems may occur in analysing psychological games.

Since we are explicitly dealing with higher-order beliefs or even full belief hierarchies, it is helpful to simplify notation by usage of types (Harsanyi, 1967-1968). To every belief hierarchy of player i, we can assign a type $t_i \in T_i$. Doing this for every player in a psychological game G, and assuming all players can only consider as possible finitely many belief hierarchies, one can construct a finite epistemic model.

Definition 3.2. *Consider a psychological game G with a finite set of choice C_i for each player $i \in I$ where I is a finite set of players. An **epistemic***

Table 3.2: Epistemic model with lexicographic beliefs, Gift game

Type King	$T_1 = \{t_1\}$		
Types Queen	$T_2 = \{t_2, t_2'\}$		
King's beliefs	$b_1[t_1]$	=	$\frac{1}{2}t_2 + \frac{1}{2}t_2'$
Queen's beliefs	$b_2[t_2]$	=	$((Lands, t_1); (Games, t_1))$
	$b_2[t_2']$	=	$((Games, t_1); (Lands, t_1))$

model with lexicographic beliefs *$M = (T_i, b_i)_{i \in I}$ for G specifies for every player i a finite set T_i of possible types. Moreover, for every player i and every type $t_i \in T_i$ the epistemic model specifies a lexicographic belief $b_i[t_i] = (b_i^1[t_i]; b_i^2[t_i]; ...; b_i^n[t_i])$ over the set $C_{-i} \times T_{-i}$ of opponents' choice-type combinations.*

Thus every type specifies a lexicographic belief about the opponents' choice-type pairs, whose types specify a lexicographic belief about their opponents' choice-type pairs. Continuing this process, we can retrieve for each type a hierarchy of lexicographic beliefs.

For now we stay intentionally verbal in what we formally define a psychological game G to be when players have lexicographic beliefs. This is due to measurability issues that may arise as a result of dealing with hierarchies of lexicographic beliefs specifically. For the discussion now, we will assume that a player i makes his decisions in line with subjective expected utility maximization, for some measurable utility function that captures the decision-makers preferences. In a traditional game such a function would be

$$u_i : C_i \times C_{-i} \to \mathbb{R}.$$

Now, denote by $u_i(c_i, b_i^1[t_i])$ the expected utility for player i when choosing c_i and holding the belief $b_i^1[t_i]$, with $b_i^1[t_i]$ being the first theory in the lexicographic belief $b_i[t_i]$. In a *traditional* game (where util-

ities only depend on first-order beliefs) with lexicographic beliefs this implies that, given a type t_i for player i, a choice c_i is *weakly preferred* over another alternative $c_i' \in C_i$ if:

$u_i(c_i, b_i^1[t_i]) > u_i(c_i', b_i^1[t_i])$, or
$u_i(c_i, b_i^1[t_i]) = u_i(c_i', b_i^1[t_i])$ and $u_i(c_i, b_i^2[t_i]) > u_i(c_i', b_i^2[t_i])$, or
...
$u_i(c_i, b_i^1[t_i]) = u_i(c_i', b_i^1[t_i])$ and $u_i(c_i, b_i^2[t_i]) = u_i(c_i', b_i^2[t_i])$ and ... and
$u_i(c_i, b_i^n[t_i]) \geq u_i(c_i', b_i^n[t_i])$.

We say the choice c_i is optimal for $b_i[t_i]$ if the decision maker does not prefer any choice to c_i.

Extending the above intuition of optimality of a choice to psychological games is problematic however. To understand why, let us return to the example of the Queen and the King in Table 3.1. The beliefs of the King and Queen are presented in Table 3.2.

One interpretation of these beliefs could be that organising Games is deemed infinitely more likely *to a higher degree* than Lands for the Queen's belief induced by t_2', than Lands is deemed infinitely more likely than Games for the Queen's belief induced by t_2. Then it is clear that choosing Land is optimal for the King. However, the opposite is possible as well, under which Games is the only optimal choice for the King.

Now, suppose the King had decided to organize the gladiator games for the Queen. However, the spendings during the half-year of the games were so exorbitantly high that the people started a rebellion against the King. The King does not have the required number of faithful soldiers to protect his position. The Queen knows that sending a large force of soldiers of her own to protect her colleague will easily quell the rebellion, whereas a small force will also subdue the rebels, though at the larger cost of losing more King's soldiers. This game is presented in Table 3.3. The Queen has two separate motiva-

Table 3.3: *Game of King and Queen: Aid.*

		Beliefs King	
	Large force	*No aid*	*Small force*
Large force	-2	3	-2
No aid	-2	0	-2
Small force	-1	4	-6

tions here: to lose as few soldiers as possible and to be respected by her fellow ruler the King. Sending a large force will cost her 2 units of utility, whereas a small force will cost her 1 unit of utility. The respect the Queen cares for can come in three forms: if she sends no aid at all whereas the King believes her to send some soldiers to beat the rebellion, she loses 2 units of utility due to an *expected loss* in respect with the King. More important to her is however to see the King be elated with her helping presence. As such, she believes she will be perceived as the 'unexpected savior' by the King. Thus, sending a force (large or small) to subdue the rebels, while she expected the King to believe she would not send any help, will give the Queen a utility of 5. However, sending a small force specifically may also have an exactly opposite effect for the Queen. That is, she certainly does not wish to *reinforce* any existing reservations the King may have about her sending troops *solely* to be perceived as the unexpected savior instead of also caring for his safety and mental well-being. This would occur if the King believes the Queen to only send a small force instead of a large force to quickly quell the rebellion, while the Queen in fact indeed sends such a small force. Such a circumstance would cause her to lose 5 units of utility on top of the 1 unit lost by sending a small force. Overall, the Queen's motivations depend on her second-order beliefs. More specifically, the columns in Table 3.3 refer to the King's probability-one first-order beliefs. The Queen's utility then depends on the combination of her choice and her belief over such probability-one beliefs of the King. These latter objects we call *second-order expectations* (See CHAPTER 5).

Let us make the assumption that the Queen believes the King is a

Table 3.4: Epistemic model with lexicographic beliefs, V1

Type Queen	$T_1 = \{t_1\}$		
Types King	$T_2 = \{t_2, t_2'\}$		
Queen's beliefs	$b_1[t_1]$	=	$\frac{4}{5}t_2 + \frac{1}{5}t_2'$
King's beliefs	$b_2[t_2]$	=	$((LF, t_1); \frac{4}{10}(SF, t_1) + \frac{6}{10}(NA, t_1))$
	$b_2[t_2']$	=	$((SF, t_1); \frac{4}{10}(LF, t_1) + \frac{6}{10}(NA, t_1))$

cautious reasoner. We can model the King's cautious beliefs by lexicographic probabilities over the set of the Queen's choices.[1] Caution implies here that for each type of the Queen considered by the King, positive probability is assigned to each possible choice for the Queen somewhere in the lexicographic belief. The Queen by definition satisfies caution, as there are no choices by the King to be cautious about.

To illustrate the problems regarding lexicographic beliefs in psychological games, a simple epistemic model will suffice. A leading example is presented in Table 3.4. The Queen considers two *types* of the King, each representing one of the King's possible lexicographic beliefs. The King considers a single type for the Queen which induces a non-lexicographic belief. The Queen believes here with probability $\frac{4}{5}$ that the King deems it infinitely more likely that the Queen will send out a large force of soldiers (LF) than a small force (SF) or no aid (NA) at all. Yet, the latter two choices are still considered by the King in the secondary theory of his lexicographic belief with probability $\frac{4}{10}$ and $\frac{6}{10}$ respectively. In a similar manner, the Queen believes with probability $\frac{1}{5}$ that the King deems it infinitely more likely than either a large force or no force at all that a small force will be sent to aid him. However, the Queen also believes with probability $\frac{1}{5}$ that the King still deems a large force or no aid possible in his secondary theory with probability

[1] Caution can also be defined over the strategy-type space, instead of just over the strategy-space (as adopted in this paper). However, this distinction is irrelevant for the problem discussed here.

$\frac{4}{10}$ and $\frac{6}{10}$ respectively.

Were the Queen only to consider the primary theories of each of the King's two possible lexicographic beliefs, then it is clear that the Queen is indifferent between all her options. Namely, we would have that $u_Q(LF) = u_Q(NA) = u_Q(SF) = -2$. Similarly, were the Queen only to look at the secondary theories in the King's lexicographic beliefs, then it is clear that only LF is an optimal choice for her. That is, we would have

$$u_Q(LF) = \frac{4}{5}(\frac{6}{10} \cdot 3 + \frac{4}{10} \cdot (-2)) + \frac{1}{5}(\frac{6}{10} \cdot 3 + \frac{4}{10} \cdot (-2)) = \frac{4}{5} \cdot 1 + \frac{1}{5} \cdot 1 = 1$$

for her choice LF,

$$u_Q(NA) = \frac{4}{5}(\frac{6}{10} \cdot 0 + \frac{4}{10} \cdot (-2)) + \frac{1}{5}(\frac{6}{10} \cdot 0 + \frac{4}{10} \cdot (-2)) = -\frac{4}{5}$$

for her choice NA, and for her choice SF

$$u_Q(SF) = \frac{4}{5}(\frac{6}{10} \cdot 4 + \frac{4}{10} \cdot (-6)) + \frac{1}{5}(\frac{6}{10} \cdot 4 + \frac{4}{10} \cdot (-1)) = \frac{4}{5} \cdot 0 + \frac{1}{5} \cdot 2 = \frac{2}{5}.$$

We cannot make the statement however that this makes the choice to send a large force LF optimal for the Queen by the observations above. By Definition 3.2, we know that $b_2[t_2]$ and $b_2[t_2']$ are both sequences of beliefs on $C_1 \times T_1$ where the beliefs are decreasing in importance. However, we need something even more expressive than this. That is to say, at face value we cannot recover from a lexicographic belief whether 'deeming a choice infinitely less likely' means probabilistically the same thing in $b_2[t_2]$ as in $b_2[t_2']$. It could well be that the secondary theory in $b_2[t_2]$ receives infinitely less weight compared to the secondary theory in $b_2[t_2']$ (or vice versa) with the information we have now. Such information is however relevant for the Queen, since her utility depends on her second-order expectations.

To see this, first consider the secondary *theory* in $b_2[t_2]$ to be deemed infinitely less likely to occur by the Queen in her second-*order* belief

Table 3.5: Epistemic model with lexicographic beliefs, V2

Type Queen	$T_1 = \{t_1\}$		
Types King	$T_2 = \{t_2, t_2'\}$		
Queen's beliefs	$b_1[t_1]$	=	$(\frac{4}{5}t_2 + \frac{1}{5}t_2'; t_2)$
King's beliefs	$b_2[t_2]$	=	$((LF, t_1); \frac{4}{10}(SF, t_1) + \frac{6}{10}(NA, t_1))$
	$b_2[t_2']$	=	$((SF, t_1); \frac{4}{10}(LF, t_1) + \frac{6}{10}(NA, t_1))$

than the secondary theory in $b_2[t_2']$. Say the Queen would fully focus on what she believes would be the King's secondary theories to determine her preferences. This would have as a consequence that the Queen expects the King to believe her choosing SF is infinitely less likely to occur than LF. As a result, the Queen has little concern that the King believes she will only send a small force, leading SF to be the only optimal choice. However, if we reverse the relation between $b_2[t_2]$ and $b_2[t_2']$, it would mean that, as far as secondary theories go, the Queen expects the King to deem SF to be infinitely *more* likely to occur than LF. In such a scenario, sending a large force would be the Queen's only optimal choice. In more general words, lexicographic beliefs carry insufficient information to consistently determine a player's preference over his own choices in a psychological game.

An assumption that one could impose is that 'deeming a belief infinitely more important' means the same thing across lexicographic beliefs of a particular player. However, such a resolution may also not suffice. Consider the scenario in which the Queen also has a lexicographic belief by simply extending her belief $b_1[t_1]$ in Table 3.4 to one as portrayed in Table 3.5. Now, in the Queen's secondary theory she believes with probability-one that the King is of type t_2. Given this theory, she then believes that the King deems it infinitely more likely than anything else that the Queen will choose LF. Under such beliefs, SF would be the only optimal choice for the Queen. It is clear that the Queen's preferences over her choices shaped by her primary the-

ory about the King's primary theories have precedence over the preferences shaped by second-order beliefs relating to all other combinations of theories. It is also obvious that the preferences shaped by the Queen's secondary theory about the King's secondary theories ought to be assigned the least importance. However, with the information provided as of yet it is unclear whether the preferences shaped by the Queen's primary theory about the King's secondary theories should take precedence over those shaped by the Queen's secondary theory about the King's primary theory. This does however matter for determining the Queen's optimal choice: in the former case LF is the optimal choice, but in the latter SF.

This would leave us with two options to go forward from here. First, we could take an axiomatic approach and define an additional *choice rule*. For instance, we could define the ordering of the preferences such that the preferences shaped by the Queen's secondary (and if she had any: her tertiary, quaternary etc.) theory about the King's primary theory (or theories) are deemed infinitely more important than the ones shaped by the Queen's beliefs about non-primary theories of the King. This ordering is however rather random in the sense that we might as well propose a different ordering rule that may be equally intuitive. Instead, we can also look at the primitives of the model of expected utility maximization. We may represent cautious beliefs such that it still allows for deriving optimal choices in psychological games in a non-ambiguous manner and without the need to specify extra choice rules. This implies we need to be able to give a clear description of one event being 'deemed infinitely more likely' than another, a matter elaborated on in the following section.

3.3 Non-standard beliefs as a solution

As the previous discussion has shown, psychological games with cautious beliefs including infinitely small weights on choices will require one to *quantify* what it means for a player to deem one event 'infinitely

more or less likely' than another. Such a quantification can be provided by using *non-standard analysis* (going back to at least Robinson (1973)). The main idea of this form of analysis is that one extends the line of reals $\mathbb{R}$ to a non-Archimedean field of hyperreals $\mathbb{R}^*$, which includes *infinitesimals* yet retains the first-order structure of the line of reals. A strictly positive number $\epsilon \in \mathbb{R}^*$ is called an infinitesimal if $\epsilon \cdot r < 1$ for every $r \in \mathbb{R}$.[2] There is one important property of infinitesimals in a non-Archimedean field that will be highlighted for the purposes of this chapter.

Definition 3.3. *Take two numbers $r, s \in \mathbb{R}^*$ with r and s strictly positive such that $\frac{s}{r}$ is an infinitesimal. That is, $\frac{s}{r}$ is in $\mathbb{R}^*$, but not in $\mathbb{R}$. Then we can say that s is infinitely smaller than r.*

Non-standard analysis allows us to capture the intuition of lexicographic beliefs in the sense that we can still define what it means for one event to be deemed infinitely less likely than another. At the same time, it also allows us to quantify this relation.

The extended field $\mathbb{R}^*$ we will consider is as in Hammond (1994). This is an elementary extension of the line of reals $\mathbb{R}$ by means of any, single infinitesimal number ϵ, which thus does not satisfy the Archimedean property. Using a polynomial construction, we can obtain the ordered field $\mathbb{R}^*$. This works as follows. We want the set $\mathbb{R}^*$ that we are constructing to be a field; it needs to be closed under addition and multiplication. This means that all its members need to be 'rational' numbers in the sense that they can be expressed as a ratio of two, finite polynomial functions. So each member $s \in \mathbb{R}^*$ can be expressed as

$$s = \frac{a_0 + a_1\epsilon + a_2\epsilon^2 + ... + a_n\epsilon^n}{b_0 + b_1\epsilon + b_2\epsilon^2 + ... + b_n\epsilon^n},$$

[2] An infinitesimal may be constructed from a fixed sequence that converges to 0 using ultrafilters or using a polynomial construction as in Robinson (1973). For an in-depth discussion about non-standard analysis in (traditional) game theory, see Hammond (1994) and Halpern (2010).

where ϵ is the single infinitesimal number we started with, $a_k, b_k \in \mathbb{R}$ for all $k \in \{1, ..., n\}$, $b_k \neq 0$ for some $k \in \{1, ..., n\}$, and either $a_0 \neq 0$ or $b_0 \neq 0$.

Using *non-standard probabilities*, we can transform the beliefs in the epistemic model of Table 3.4. Let a non-standard probability distribution p on X assign probabilities $p(x) \in \mathbb{R}^*$, where $p(x) \geq 0$, such that $\sum_{x \in X} p(x) = 1$. We can derive belief hierarchies of non-standard beliefs as we do for standard beliefs, as Rajan (1998) shows. To do so, we need to assume the primitive space of uncertainty is Hausdorff and use the Transfer Principle of non-standard analysis that, roughly speaking, states that any statement that holds about standard objects in a standard space should also hold about non-standard objects in a non-standard space.

Definition 3.4. *A **static psychological game** with nonstandard beliefs is a tuple $G = (C_i, B_i^*, u_i)_{i \in I}$ with I denoting the finite set of players, C_i representing the finite set of choices for player i,[3] B_i^* the set of non-standard belief hierarchies that express coherency and common belief in coherency, and $u_i : C_i \times B_i^* \to \mathbb{R}^*$ representing player i's measurable utility function.*

This definition follows that of Jagau and Perea (2017), but now for psychological games with non-standard beliefs. The belief hierarchies as constructed in Rajan (1998) can be captured in an epistemic model. We construct such an epistemic model as we would do for standard beliefs or did for lexicographic beliefs in Section 3.2 following Brandenburger and Dekel (1993). In Section 3.2 we have shown that in *a* type space related to some psychological game persisting with lexicographic beliefs can already cause problems. We will show in this section that non-standard beliefs provide a solution in such matters.

A finite epistemic model with non-standard beliefs is defined as follows.

[3] C_i may well be a singleton set, indicating a situation where player i does not have any choices to make but where her beliefs matter for the utilities of other players

Definition 3.5. *Consider a psychological game G with non-standard beliefs. An **epistemic model** $M = (T_i, b_i)_{i \in I}$ **with non-standard beliefs** for G specifies for every player i a finite set T_i of possible types. Moreover, for every player i and every type $t_i \in T_i$ the epistemic model specifies a non-standard probability distribution $b_i[t_i]$ on the set $C_{-i} \times T_{-i}$ of opponents' choice-type combinations.*

As long as we assume finite epistemic models, the choice for lexicographic beliefs or non-standard beliefs does not change the nature of underlying game. This means that *if* we can unambiguously define preferences over choices using hierarchies of lexicographic beliefs in a psychological game, then re-defining such hierarchies using non-standard beliefs must result in the same preferences over choices. That is, from Halpern (2010) we know that a lexicographic belief function $\bar{b}_i[t_i]$ and a non-standard belief function $\tilde{b}_i[t_i]$ are equivalent in finite cases. The negative result with regard to lexicographic belief hierarchies for psychological game can thus purely be attributed to the expressive power of such objects and not to any other trade-off.

The result of adapting Table 3.4 to a non-standard beliefs instead of lexicographic beliefs is found in Table 3.6, where $\epsilon > 0$ is an infinitesimal. For both his types the King is cautious, as in both cases he deems possible every choice, be it with a positive real number or a positive infinitesimal number. Note that in this particular example we modelled the King's cautious beliefs such that deeming one event infinitely more likely than another means the same thing for both his types t_2 and t_2', as in both this relation is defined using the same infinitesimal ϵ.

A crucial point here is that non-standard probability measures are defined on a non-Archimedean, *ordered* field. Higher-order, non-standard beliefs then are defined, in a sense, over the set of such probability measures. The resulting probability measure is then still defined on an ordered field. As a result, operations such as multiplication are as usual. This is in contrast to e.g. a second-order lexicographic belief, which is defined over a set of *sequences* of standard probability measures. It is e.g. not clear how to multiply such sequences a priori. This causes

Table 3.6: Epistemic model with non-standard beliefs, V1

Type Queen	$T_1 = \{t_1\}$		
Types King	$T_2 = \{t_2, t_2'\}$		
Queen's beliefs	$b_1[t_1]$	=	$\frac{4}{5}t_2 + \frac{1}{5}t_2'$
King's beliefs	$b_2[t_2]$	=	$(1-\epsilon)(LF, t_1)$
		+	$\epsilon(\frac{4}{10}(SF, t_1) + \frac{6}{10}(NA, t_1))$
	$b_2[t_2']$	=	$(1-\epsilon)(SF, t_1)$
		+	$\epsilon(\frac{4}{10}(LF, t_1) + \frac{6}{10}(NA, t_1))$

Table 3.7: Epistemic model with non-standard beliefs, V2

Type Queen	$T_1 = \{t_1\}$		
Types King	$T_2 = \{t_2, t_2'\}$		
Queen's beliefs	$b_1[t_1]$	=	$\frac{4}{5}t_2 + \frac{1}{5}t_2'$
King's beliefs	$b_2[t_2]$	=	$(1-\epsilon^2)(LF, t_1)$
		+	$\epsilon^2(\frac{4}{10}(SF, t_1) + \frac{6}{10}(NA, t_1))$
	$b_2[t_2']$	=	$(1-\epsilon)(SF, t_1)$
		+	$\epsilon(\frac{4}{10}(LF, t_1) + \frac{6}{10}(NA, t_1))$

difficulties in deriving a unique joint lexicographic probability distribution from two or more separate marginal ones Hammond (1994). This was essentially the root of the problem discussed in Section 3.2 by means of Table 3.4.

The way in which higher-order, non-standard beliefs are constructed, allows us to define the expected utility for a player in a similar manner as how one would define expected utility with standard beliefs. Taking into account an epistemic model, utilities can be defined as a function of choices and types: $u_i(c_i, t_i)$. This utility itself may also involve infinitesimal numbers. An optimal choice for a type with non-standard

beliefs can subsequently be defined as follows.

Definition 3.6. *Consider an epistemic model $M = (T_i, b_i)_{i \in I}$ with **non-standard beliefs** for G and a type t_i for player i in such a model. A **choice** c_i is **optimal** for type t_i of player i if $\forall c_i' \in C_i : u_i(c_i, t_i) \geq u_i(c_i', t_i)$.*

With these tools in hand, we can show that using non-standard beliefs the issue of determining an optimal choice specifically can be resolved. Returning to the game in Table 3.3, we already established the expected utilities for the Queen given the primary theory and given the secondary theory in the previous section. If we take the weighted sum of these, where $(1 - \epsilon)$ is assigned to the primary theory and ϵ in the secondary for both $b_2[t_2]$ and $b_2[t_2']$, we have $u_Q(LF) = (1-\epsilon) \cdot (-2) + \epsilon$ for her choice LF, $u_Q(NA) = (1 - \epsilon) \cdot (-2) + \epsilon \cdot (-\frac{4}{5})$ for her choice NA, and $u_Q(SF) = (1 - \epsilon) \cdot (-2) + \epsilon \cdot (\frac{4}{10})$ for her choice SF. Thus, we have unambiguously derived that under this specific second-order belief sending a large force is optimal for the Queen.

In a similar manner, we can transform the epistemic model in Table 3.4 into another epistemic model with non-standard beliefs that induces the same lexicographic beliefs, but in which sending a small force would be the only optimal choice. This is depicted in Table 3.7. Note however that now the relation of one event being infinitely more likely than another in $b_2[t_2]$ is denoted by ϵ^2 (that of $b_2[t_2']$ is still given by ϵ), where ϵ^2 is infinitely smaller than ϵ. Since the two non-standard beliefs of the King are characterized by two different infinitesimals, we need to take the weighted sum of the utilities given each combination of theories and types to derive expected utilities for the Queen. We then acquire $u_Q(LF) = \frac{4}{5}(-2(1 - \epsilon^2) + \epsilon^2) + \frac{1}{5}(-2(1 - \epsilon) + \epsilon) = -\frac{8}{5} + \frac{8}{5}\epsilon^2 + \frac{4}{5}\epsilon^2 - \frac{2}{5} + \frac{2}{5}\epsilon + \frac{1}{5}\epsilon = -2 + \frac{3}{5}\epsilon + \frac{12}{5}\epsilon^2$. In a similar way, we can get for her choice NA that $u_Q(NA) = -2 - \frac{6}{25}\epsilon + \frac{24}{25}\epsilon^2$ and $u_Q(SF) = -2 + \frac{11}{5}\epsilon + \frac{4}{5}\epsilon^2$ for her choice SF. Clearly then, sending a small force would be optimal. This in part because the 'primary theory' in $b_2[t_2]$ is deemed infinitely more important than its secondary theory to a slightly higher degree than that the primary theory in $b_2[t_2']$

is deemed infinitely more important than its secondary theory. Hence, the Queen expects the King to believe in his primary theories that the Queen is slightly more likely to choose LF than SF. Also important is however that the secondary theory of $b_2[t_2]$ is assigned an infinitely smaller probability than the secondary theory of $b_2[t_2']$. It follows that the Queen expects the King to believe in his secondary theories that LF and NA are infinitely more likely to be chosen than SF. This explains why SF is the only optimal choice.

Once utility (and thus optimality) is defined, we can extend solution concepts from traditional games to psychological games. For notions where caution is an integral part, such as permissibility (Börgers, 1994; Brandenburger, 1992) and common full belief in caution and primary belief in rationality (Perea, 2012), we can use non-standard beliefs to do so.

Definition 3.7. *Consider an epistemic model $M = (T_i, b_i)_{i \in I}$ with non-standard beliefs and a type t_i for player i. Player i has a **cautious type** if, whenever he deems possible an opponent's type t_j for some player j, then for every $c_j \in C_j$ it assigns positive probability in $\mathbb{R}^*$ to (c_j, t_j). This probability may be real or an infinitesimal.*

A type is deemed possible if it is assigned positive probability in $\mathbb{R}^*$ in the belief. Assuming a player to have a full-support belief and believing in the opponent's rationality may be incompatible. However, we can impose the following condition for a choice to be considered rational.

Definition 3.8. *Consider an epistemic model $M = (T_i, b_i)_{i \in I}$ with non-standard beliefs and a type t_i for player i. **Type t_i primarily believes in the opponent's rationality** if for every opponent $j \neq i$ we have that $b_i[t_i](c_j, t_j) \in \mathbb{R}_+$ only if c_j is optimal for t_j, where $\mathbb{R}_+$ is the set of all positive real numbers.*

We can iterate the arguments of believing an opponent is cautious and primarily believing in an opponent's rationality.

Definition 3.9. *Consider an epistemic model $M = (T_i, b_i)_{i \in I}$ with non-standard beliefs and a type t_i for player i. Type t_i expresses 1-fold full belief in caution if it only deems possible opponents' types that are cautious. For every $k > 1$, every player i, and every type $t_i \in T_i$, we say that type t_i expresses k-fold full belief in caution if t_i only deems possible opponents' types that express $(k-1)$-fold full belief in caution.*
*Type t_i expresses **common full belief in caution** if t_i expresses k-fold full belief in caution for every k.*

Definition 3.10. *Consider an epistemic model $M = (T_i, b_i)_{i \in I}$ with non-standard beliefs and a type t_i for player i. Type t_i expresses 1-fold full belief in primary belief in rationality if t_i primarily believes in the opponent's rationality. For every $k > 1$, every player i, and every type $t_i \in T_i$, we say that type t_i expresses k-fold full belief in primary belief in rationality if t_i only deems possible opponents' types that express $(k-1)$-fold full belief in primary belief in rationality.*
*Type t_i expresses **common full belief in primary belief in rationality** if t_i expresses k-fold full belief in primary belief in rationality for every k.*

Then, a *rational* choice under common full belief in caution and primary belief in rationality entails that the choice is optimal for a type t_i that expresses common full belief in caution and primary belief in rationality. Note that the notion of common full belief in caution and primary belief in rationality epistemically justifies the concept of permissibility (in traditional games).

It is clear that there is no type in either the epistemic model of Table 3.6 or the epistemic model of Table 3.7 that expresses common full belief in caution and primary belief in rationality. Namely, the Queen only has a single type, that considers type t_2 and t_2' for the King. Type t_2 primarily believes the Queen will choose LF and type t_2' primarily believes the Queen will choose SF. We already established that in the

Table 3.8: Epistemic model with non-standard beliefs, V3

Type Queen	$T_1 = \{t_1, t_1'\}$		
Types King	$T_2 = \{t_2, t_2'\}$		
Queen's beliefs	$b_1[t_1]$	=	t_2
	$b_1[t_1']$	=	t_2'
King's beliefs	$b_2[t_2]$	=	$(1-\epsilon-\epsilon^2)(LF, t_1')$
		+	$\epsilon(SF, t_1') + \epsilon^2(NA, t_1')$
	$b_2[t_2']$	=	$(1-\epsilon-\epsilon^2)(SF, t_1)$
		+	$\epsilon(NA, t_1) + \epsilon^2(LF, t_1))$

situation of Table 3.6 LF is the only optimal course of action and in the situation of Table 3.7 SF is the only optimal choice. The Queen expects the King however to believe with some positive probability in $\mathbb{R}_+$ that she will choose LF and with some positive probability in $\mathbb{R}_+$ that the she will choose SF. However, by the previous argument LF and SF cannot be optimal choices at the same time in these two situations. Hence, the Queen does not believe the King always primarily believes in her rationality. This does not mean there is no epistemic model in which common full belief in caution and primary belief in rationality is satisfied. One possibility is a scenario in which the Queen believes the King to (partially) believe she is of a different type than she actually is. Namely, one of the motivations of the Queen is to gain respect from the King by surprisingly coming to his aid. Table 3.8 provides an epistemic model in which SF is optimal for type t_1 and LF is optimal for type t_1', where both types express common full belief in caution and primary belief in rationality. The analysis is left to the reader. Note finally that if caution is assumed, NA cannot be optimal, as it is weakly dominated.

3.4 Conclusion

The distinctive feature of a belief modelled by non-standard probabilities compared to a lexicographic belief, is that in case of the latter we know that it is derived from *some* sequence of beliefs, whereas in case of the former we have more information about which *specific* sequence of beliefs it would correspond to. That is, for each lexicographic belief we can find an equivalent non-standard belief that quantifies the probabilistic structure behind one state being 'infinitely more likely' than another. In some cases this extra information evidently is crucial in unambiguously deriving preferences over choices. However, it also stresses that in psychological games, depending on the types of beliefs held by all players, more sorts of information have to be accounted for. If cautious reasoning is involved, the decision-maker needs to be aware of what 'infinitely more likely' means in one considered belief hierarchy of the opponent compared to another.

4

The Surprise Exam Paradox as a Psychological Game

This chapter is adapted from: "Reasoning about the surprise exam paradox: An application of psychological game theory" (Mourmans, 2017).

4.1 Introduction

In psychological games, we model particular belief-dependent motivations by letting preferences of players explicitly depend on the belief hierarchies. From an epistemic point of view, solution concepts used to analyze games are characterized by imposing restrictions on such belief hierarchies. A point of friction is then to what extent meaningful restrictions can be imposed on the belief hierarchies while still retaining the essence of the belief-dependent motivation we are trying to model.

In this chapter I will elaborate on this point by means of analyzing the **Surprise Exam Paradox (SEP)**. This is a well-studied problem, in logic and philosophy alike. The story of the paradox is as follows.

A teacher (male) announces to his student (female) that during the next week she will be given an exam. However, the teacher does not announce on which day of the week the exam will take place. Instead, he announces to the student that, when the exam arrives, the day on which it occurs will have come as a surprise to her. Then, the story goes, the student reasons that the teacher cannot give the exam on Friday. Namely, if Friday has come about and the exam has not been given at that point, the student knows the exam has to be given on Friday and therefore no surprise will be possible. Once Friday is ruled out by the student, only Monday to Thursday are left as viable options for the teacher according to the student. But then by the same reasoning the student cannot think the teacher can choose Thursday any longer: once Thursday has arrived and the exam has not yet been given, the student knows that the exam will be given on Thursday. Following the same line of reasoning, the student will believe that the exam cannot be given on Wednesday, Tuesday or on Monday and thus will conclude that the teacher cannot give a surprise exam. Once Wednesday comes about, the student finds an exam lying on her desk. The student is fully surprised.

Though a seemingly simple problem, the sheer size of the literature on the paradox shows much value has been found in analyzing it, mainly for logicians and philosophers (see Chow (2011) for a comprehensive

overview of the literature on the topic). In particular, logicians have mainly focused on the nature of the teacher's announcement. Surprising the student can be logically defined by the announcement that: (1) the exam will take place next week and that (2) the exact day on which it will take place is not deducible in the day in advance for the student by the preceding statement. This announcement in itself is found to be self-contradictory (Fitch, 1964; Smullyan, 1987). Take for instance the scenario in which the exam has not happened by Thursday. As Smullyan (1987) argues, if statement (1) is known to be true, then the student knows on Thursday evening the exam has to be given, making statement (2) false. If statement (1) were doubted, then the student may not expect an exam at all the next day. In that case statement (2) becomes true. The truth of statement (2) relies on statement (1) being false. This resolution has left some confusion. The argument relies on the premise that the student at all times accepts (or knows) she will receive a surprise exam. But this is not a premise one can logically deduce from the announcement.

Epistemological studies of the problem try to resolve this issue by formulating the problem in such a manner that the student can accept the announcement of the teacher to be either true or false. This is essentially the approach of Quine (1953).[1] Quine argues as follows. When deriving a contradiction in the teacher's announcement, the student can no longer accept his announcement as true. But then the student should have considered this as a possibility from the start. The student can actually discern between four possibilities on Thursday evening. First, (a) an exam will happen tomorrow and the student accepts that as true now, (b) an exam will not happen tomorrow and the student accepts that as true now, (c) an exam will not happen tomorrow but the student does not accept that as true now and (d) an exam will happen tomorrow but the student does not accept that as true now. Possibilities (c) and (d) are deduced after deriving a contradiction, of

[1] Quine (1953), amongst others, technically looked at a different version of the paradox, called the Unexpected Hanging Paradox. However, it represents exactly the same problem.

which possibility (d) actually allows for surprise on Friday. It has been shown that the student cannot accept the announcement as true by the time the last day has occurred. This epistemic state is known as an epistemic blindspot (Sorensen, 1988).

If one wanted to deal with the question of whether the student can *know* the announcement at all times, the resolution offered by Quine (1953) is sufficient. However, if the student were to completely disregard any information disclosed by the announcement, then anything is possible. The student would not *know* what the teacher's intentions are with the exam, or not even *know* an exam will be given. Rather, in a satisfactory solution of the paradox, we at least want the student to learn something from the announcement. Moreover, we want our solution to show why the teacher believes he can vindicate his announcement. This would explain why the teacher made the announcement in the first place. In order to explain why the teacher thinks he can surprise the student, we need to model the teacher's rational reasoning process in the paradox. At the same time, we want our model to explain the student's reasoning process: where she went wrong in her reasoning in the story and whether she could have rationally reasoned to other conclusions instead. Therefore it is useful to address the paradox as a game-theoretic problem.

Surprise is a mental state. It is a mismatch between an observable outcome and a (prior) doxastic or epistemic state of a person. In this particular case the outcome refers to when the teacher *chooses* to give the exam and the doxastic state refers to the student's prior *belief* about this choice of the teacher. The goal of the teacher is to surprise the student. Clearly then, the teacher's decision is driven by his belief about the student's belief about his own decision, i.e. his second-order belief. Therefore, the teacher's preferences are not simply defined over material outcomes of the game played. Instead, the teacher has belief-dependent motivations. As such, the SEP presents itself as a psychological game. By modelling the problem as a psychological game we follow the route taken by Geanakoplos (1996) and therefore do not opt to model it as a traditional game, the route taken by the scarce,

remaining game-theoretic literature on the SEP (Sober, 1998; Ferreira and Bonilla, 2008).[2] However, instead of looking at fixed points as in Geanakoplos, we will formalize and apply the more basic type of iterative reasoning which is used by the student in the story. We will do so by applying notions found in epistemic game theory to this problem.

The only, minimal assumption we wish to impose beforehand in this paradox is that both the teacher and the student are reasoners in line with common belief in rationality. If we include more than two days, we include dynamics to the paradox. Since in the paradox the student appears to employ backward induction reasoning, we will focus on the epistemic game-theoretic analogue of such reasoning for the dynamic setting: common belief in future rationality (Perea, 2014).

In this chapter we provide a game-theoretic resolution of the paradox which also allows us to be specific about the reasoning employed by both the student and the teacher. To this end, we need to be formal about the lines of reasoning used by the players of the game. We consider the concepts of common belief in rationality and common belief in future rationality as basic modes of reasoning. Looking at the paradox from a more fundamental viewpoint in a game-theoretic setting by iteratively defining *rational* reasoning steps gives us a new perspective on the problem. We will describe scenarios in which it makes sense for the teacher to announce a surprise exam by the concept of common belief in rationality.

Common belief in rationality in psychological games is essentially the same as common belief in rationality in traditional games, in the sense that at no point in a belief hierarchy a player's rationality is questioned. There is an important difference to be found in the definition of optimality however, as in psychological games now also belief-dependent

[2]Sober (1998) and Ferreira and Bonilla (2008) model the game as a matching-pennies game. This requires modelling a preference relation for the student as well. It appears hard to plausibly extend the paradox story in such a way that it would lead to clear preferences and choices for the student.

motivations come into play. In the dynamic setting, a comparable distinction is found for common belief in future rationality. Under this concept, at no point in a decision-maker's conditional belief hierarchy future rationality is put into doubt.

We consider two versions of the paradox: one where surprise by giving and not giving the exam are meaningful for the teacher and one where only surprise from giving the exam is meaningful to the teacher. We find that in both cases, full surprise is possible under a belief hierarchy that expresses common belief in rationality. By *full surprise*, as opposed to partial surprise, we mean that the teacher gives the exam on a day of which the student did not believe it would occur on with any positive probability. A crucial element in the version where the teacher only cares for surprise resulting from giving the exam is that the student should deem it possible that the teacher will give the exam on the last day, in order for the teacher to be able to surprise her. This implies that the teacher must believe the student deems it possible the teacher will choose an unsurprising day. This is a belief that seems to go against the announcement he made. The position of the teacher is improved as we include more days, in the sense that he will have more days that can possibly lead to full surprise. Moreover, accepting the announcement as true becomes less problematic. Analyses using psychological Nash equilibrium contrast these results, as the imposed correct beliefs assumption significantly limits the teacher's opportunities to surprise the student.

The remainder of the chapter is organized as follows. In Section 4.2 we define the SEP as a static psychological game. Moreover, we will introduce the concepts of common belief in rationality and psychological Nash equilibrium in relation to the SEP. This will all be applied in Section 4.3, where we analyse several versions of the paradox in a static setting and link the results to previous literature. In Section 4.4 we extend the SEP to a dynamic setting, and discuss the backward induction reasoning concept of common belief in future rationality. This framework is then applied to the two versions of the SEP in a dynamic setting. Finally, we close off with a short conclusion in Section 4.5.

4.2 Preliminaries

We start this section by providing a formal set-up for the analysis of the paradox in a static version. By this we mean we will look at a two-day version where the teacher can only choose to give the exam on Thursday or Friday. We will first introduce the notion of surprise we will be looking at. Subsequently, we define the paradox in the framework of psychological game theory. Finally, common belief in rationality and psychological Nash equilibrium in the paradox will be discussed.

4.2.1 The static SEP as a psychological game

Let us consider the two-day version of the surprise exam paradox (SEP). A teacher makes an announcement to his student that consists of the following two statements: (i) there will be an exam next week on either Thursday or Friday, and (ii) on the evening before the exam will occur, the student will not expect it comes next day. It has already been shown in the logic literature that the student cannot *know* this announcement (Quine, 1953; Kripke, 2011). What we can certainly say is that the teacher's goal is to surprise his student.

Surprise is an epistemic state of a player. It is the mismatch between an observable outcome that a player expected a priori and the actual outcome. The teacher's goal is to surprise the student. The utility from his decision depends on whether he believes he surprised the student with his decision. Therefore, the teacher's utility depends on what he believes the student believes he will do: his *second-order belief*. This makes the teacher's utility belief-dependent. Therefore, we are dealing with a psychological game.

What kind of psychological game does the announcement made by the teacher in the beginning induce? What is clear is that the teacher intends to surprise the student in some way. We define *surprise* as the event that the teacher makes a choice that does not correspond (completely) with the first-order belief of the student. We distinguish between full surprise and partial surprise. *Full surprise* occurs if the

teacher makes a choice that the student in her first-order belief did not assign positive probability to. *Partial surprise* occurs if the teacher makes a choice which the student in her first-order belief did not assign probability-one to.

The student herself knows it is the teacher's intention to surprise her. She herself takes a passive role in the game, yet her beliefs matter for the utility of the teacher. Namely, the teacher derives surprise utility if he gives the exam on a day of which he believes the student did not expect the exam to occur. The utility of the teacher, given a choice, thus depends on his second-order beliefs b_1^2. More specifically, we assume it depends *linearly* on what the teacher *expects* the student to believe he will do. Overall, from the announcement one can learn *at least* the following: statement (i) leads to the interpretation that an exam will inevitably happen.[3] This is also what allows us to model the two-day setting as a static game. Namely, on Friday there is only one choice possible and therefore only one belief over such choice. We decide to capture all the information of what happens on Friday in terms of choices and beliefs in the decision-problem of Thursday. Statement (ii) acts as a revelation of the teacher's preferences as well as a prediction about the state of surprise the student will be in.

This does leave us with the question of what to do with the fact that the student cannot know the literal announcement. This issue most prominently appears on the last day in the SEP. Namely by Friday, if the exam must be given and this is common knowledge, the student would fully anticipate any exam on that day. Therefore there would be no surprise. We will deal with this issue by analyzing two versions of the SEP. In the first version, the teacher may also try to surprise the student by not giving an exam at all at a day. So if the student expected an exam on Thursday, but the teacher on Thursday does not give an exam and instead on Friday, he causes meaningful surprise. In the second version we assume the teacher only cares for surprise by

[3] Some work on the SEP do not make this assumption. E.g., Holliday (2015) distinguishes between the Inevitable and Promised Event.

Table 4.1: *Surprise by giving or not giving exam*

		Beliefs Student	
		Thursday	*Friday*
Teacher	*Thursday*	0	1
	Friday	η	0

Table 4.2: *Only meaningful surprise from giving exam*

		Beliefs Student	
		Thursday	*Friday*
Teacher	*Thursday*	0	1
	Friday	0	0

actually giving an exam. We still model Friday as a possibility, but an exam given on that day will never lead to a meaningful surprise.

Definition 4.1. *A **static surprise exam paradox (SEP)** is a static psychological game with one active player (teacher [1]) and one passive player (student [2]) with the set of choices for the teacher being $C_1 = \{$Thursday, Friday$\}$. We distinguish two versions of the paradox by the preferences of the teacher:*

- *Surprise utility is derived from the probability of a mismatch between any of the teacher's choice and what he expects the student believes he will choose (Table 4.1);*
- *Surprise utility is only derived from the probability between the choice Thursday and what he expects the student believes he will choose (Table 4.2).*

The game in Table 4.1 depicts the version that not only giving the exam on Thursday can come as a surprise to the student and thus give the teacher some utility, but also *not* giving the exam on Thursday can cause a type of surprise that matters for the teacher's utility. This is equivalent to the game considered in Geanakoplos (1996). The rows correspond to the teacher's possible choices, whereas the columns capture the teacher's *extreme second-order expectations*. The teacher will get a utility of 1 when he gives the exam on Thursday while the student believed he would give it on Friday (the ultimate day). The teacher will receive a utility of $0 < \eta \leq 1$ in case he decides not to give the exam on Thursday (and instead on Friday), while the student believed he would give it on Thursday. Any other combination of a choice and

extreme second-order expectation would lead to a utility of 0. We assume the teacher's utility is linear in the second-order beliefs that describe these expectations. For instance, say the teacher believes the student believes with probability 0.5 that the teacher will give the exam on Friday and with probability 0.5 the teacher will give the exam on Thursday. Then choosing Thursday will lead to an expected utility of $0.5 \cdot 1 + 0.5 \cdot 0 = 0.5$. Similarly, if the teacher believes with probability 0.5 that the student thinks he will give the exam on Friday and with probability 0.5 believes the student thinks he will give the exam on Thursday, then choosing Thursday will lead also to an expected utility of $0.5 \cdot 1 + 0.5 \cdot 0 = 0.5$. Namely, both describe the same (second-order) expectation for the teacher: with probability 0.5 it is expected the student believes the teacher will give the exam on Friday and with probability 0.5 it is expected the student believes the teacher will give the exam on Thursday. As such, the extremes of the distribution of the teacher's expectations of what the student believes about his choice are sufficient to represent the teacher's utility in matrix-form. See CHAPTER 5 for a formal discussion on higher-order expectations.

The game in Table 4.2 depicts the second version of the SEP. Here, the teacher only gets utility from surprising the student by giving an exam. It can be argued that this interpretation is closer to the actual crux of the paradox. In the story, the student is trying to figure out when the exam will happen, not when it will not happen.[4] The game should be read in a similar way as in Table 4.1: the rows correspond to the teacher's possible choices, whereas the columns capture the teacher's extreme second-order expectations.

4.2.2 Common belief in rationality

In the SEP, we assume that both players are rational players. To analyse the paradox in the static version, we will therefore apply the reasoning concept of common belief in rationality. We will look at the

[4] For a interesting discussion on this matter, we can refer to Kim and Vadusevan (2017).

concept as defined for psychological games by Jagau and Perea (2017), as explained in CHAPTER 2. The notion of common belief in rationality in psychological games is similar to that of traditional games. Also in psychological games, common belief in rationality for a two-player game entails that a player believes in her opponent's rationality, believes that her opponent believes in the player's rationality, and so on and so forth. The crucial difference, however, can be found in defining optimal choices. The optimality of a choice now depends on (higher)-order beliefs as well or the entire belief hierarchy, instead of just the expectation of the opponent's choice. In the SEP, we define optimality of a choice as follows.

Definition 4.2 (Optimal choice in the SEP)**.**
Consider an epistemic model $M = (T_i, b_i)_{i \in \{1,2\}}$, where 1 = teacher and 2 = student, and a type t_1 for the teacher in such a model. ***A choice*** $c_1 \in \{$*Thursday, Friday*$\}$ ***is optimal*** *for type t_1 of the teacher if for all $c_1' \in \{$Thursday, Friday$\}$: $u_1(c_1, t_1) \geq u_i(c_1', t_1)$.*

In our notation here we make use of Remark 2.2: we represent a utility fucnction $u_i(c_i, b_i)$ by $u_i(c_i, t_i)$, where type t_i captures the belief hierarchy b_i in the relevant epistemic model.

Building on this notion optimality, we can define what it means for a type to express common belief in rationality in the SEP. First we define what it means for a type to believe in an opponent's rationality.

Definition 4.3 (Belief in the opponents' rationality in the SEP)**.**
Consider an epistemic model $M = (T_i, b_i)_{i \in \{1,2\}}$, where 1 = teacher and 2 = student, with a type $t_2 \in T_2$ for the student in that model. Type t_2 of ***the student believes in the teacher's rationality*** *if type t_2 only assigns positive probability to choice-type combinations $(c_1, t_1) \in \{$Thursday, Friday$\} \times T_1$ of the teacher where the choice c_1 is optimal for the type t_1. Every type t_1 of* ***the teacher believes in the student's rationality*** *by default.*

We say the teacher by default believes in the student's rationality as the student makes no choice at all. We can iterate this idea in order to define common belief in rationality in the SEP.

Definition 4.4 (Common belief in rationality in the SEP)**.**
Consider an epistemic model $M = (T_i, b_i)_{i\in\{1,2\}}$, where 1 = teacher and 2 = student, For every player i, and every type $t_i \in T_i$, we say that type t_i expresses 1-fold belief in rationality if t_i believes in the opponent's rationality. For every $k > 1$, every player i, and every type $t_i \in T_i$, we say that type t_i expresses k-fold belief in rationality if t_i only assigns positive probability to opponents' types that express (k-1)-fold belief in rationality.
*Type t_i expresses **common belief in rationality** if it expresses k-fold belief in rationality for every k.*

Finally, we can define a choice that can be rationally made under common belief in rationality as follows.

Definition 4.5 (Rational choice under common belief in rationality in the SEP)**.**
*We say that choice c_1 is a **rational choice for the teacher under common belief in rationality** if there is an epistemic model $M = (T_i, b_i)_{i\in\{1,2\}}$ and a type $t_1 \in T_1$ such that t_1 expresses common belief in rationality, and c_1 is optimal for t_1.*

A rational choice under common belief in rationality only makes sense for the teacher, as the student is inactive in the game.

4.2.3 Psychological Nash Equilibrium

Previous game-theoretic approaches to the SEP employed the concept Nash equilibrium (Sober, 1998; Ferreira and Bonilla, 2008). A Nash equilibrium can be defined as a tuple of first-order beliefs about every player's choices such that they only assign positive probability to choices that are optimal, given the first-order beliefs about the choices

of the other players. There is an equivalent concept for psychological games in psychological Nash equilibrium (Geanakoplos et al., 1989). A psychological Nash equilibrium, however, corresponds to a full belief hierarchy. In line with the notion of a traditional Nash equilibrium, a psychological Nash equilibrium too requires each player to believe that the view of reality is commonly held by all players in the psychological game. That is, if a player i has a certain belief about the choice of opponent j, then i must believe that every other opponent shares that belief. Additionally, if player i has a certain belief about player j's choice, then player i believes that each opponent must believe that player i in fact has this belief. As such, also in a psychological Nash equilibrium, the equilibrium is fully characterized by a player's first-order and second-order beliefs.

These ideas are conceptualized by the notion of a *simple belief hierarchy* (Perea, 2012). In two-player settings such as the SEP, such a simple belief hierarchy is generated by a pair of probabilistic beliefs $\sigma = (\sigma_i)_{i \in \{1,2\}}$ that are independent of each other. In the SEP we have that $\sigma_1 \in \Delta(\{Thursday, Friday\})$. The belief σ_1 is thus a probability measure over the teacher's choice set. In the SEP, the student's choice set is a singleton, therefore σ_2 is too.

The simple belief hierarchy $\beta_i(\sigma)$ that is generated by the combination of beliefs σ states that *(i)* player i has first-order belief σ_j about her opponents' choices. In addition, it states that *(ii)* player i believes that opponent j has belief σ_i about player i's choice, *(iii)* that player i believes that opponent j believes that player i holds belief σ_j about her opponent's choices, *(iv)* et cetera. We are now in a position to define a psychological Nash equilibrium in the SEP.

Definition 4.6 (Psychological Nash equilibrium in the SEP)**.**
The pair of first-order beliefs (σ_1, σ_2) *constitutes a* ***psychological Nash equilibrium*** *if*

$$\sigma_1(c_1) > 0 \Rightarrow \forall c_1' \in \{\textit{Thursday, Friday}\} : u_1(c_1, \beta_1(\sigma)) \geq u_1(c_1', \beta_1(\sigma)).$$

A psychological Nash equilibrium has a natural link to the concept of common belief in rationality. A simple belief hierarchy $\beta_i(\sigma)$ generated by a combination of beliefs σ expresses common belief in rationality, if and only if, σ constitutes a psychological Nash equilibrium. Note that since a Nash equilibrium induces a simple belief hierarchy, it induces a fixed point in reasoning. The same first-order and second-order beliefs are continuously revisited in the belief hierarchy.

4.3 Surprise Exam Paradox (SEP): static version

With these tools at hand, let us turn to analyzing the SEP-game. We shall consider two different versions of the game in order to point out that, irrespective of the scenario at hand, the Surprise Exam Paradox might not be as paradoxical as its name may suggest. We will show that the teacher in both scenarios can believe he can *fully* surprise the student in a way that gives him utility and can thus vindicate his announcement. We will show that in order for full surprise to be possible, the student has to doubt the validity of the announcement made by the teacher. In this sense it fits well in the literature that follows the resolution by Quine (1953). We return to this point in Section 4.3.3.

4.3.1 Static SEP with "surprise from giving and not giving exam"

In Definition 4.1 we described two different games that could be induced by the teacher's announcement. The first version of the SEP is depicted in Table 4.1. This game captures the idea that both giving the exam and not giving the exam can cause a type of surprising event that the teacher cares about.

The main question is whether there is a belief hierarchy for the teacher that satisfies common belief in rationality and such that he can rationally choose to give the exam on Thursday or Friday and surprise the student. Only in such a scenario the teacher can actually believe he can

justify making the announcement in the first place. To this end, consider the belief diagram in Figure 4.1, which visualizes the epistemic model from Table 4.3 and which applies to the scenario of Table 4.1. The types here represent the belief hierarchies of both teacher and student. For instance, type t_1 of the teacher has the belief that the student is of type t_2 and type t_2 holds the belief that the teacher will choose Friday while being of type t_1'. One can easily confirm for the game in Table 4.1 that the belief hierarchy induced by t_1 expresses common belief in rationality: The student with type t_2 believes in the teacher's rationality, as Friday is indeed an optimal choice for a teacher with type t_1'. Namely, type t_1' holds the belief that the student believes the teacher will choose Thursday. Under such a second-order belief, choosing Friday will result in a higher utility than Thursday. Similarly, one can check that type t_2' expresses 1-fold belief in rationality.

Both types t_1 and t_1' automatically express 1-fold belief in rationality. Namely, the student is passive in this game and thus his choice set is a singleton, which then by definition is optimal. As every type in this model expresses 1-fold belief in rationality, every type should also express common belief in rationality, including type t_1. Under type t_1, the teacher expects the student to believe he will choose Friday. Hence when giving the exam on Thursday the teacher would indeed expect to *fully surprise* the surprise. By this we mean that the teacher gives the exam on a day that the student puts no positive probability on at

Table 4.3: Epistemic model static game

Types	$T_1 = \{t_1, t_1'\}$ $T_2 = \{t_2, t_2'\}$
Beliefs for Teacher	$b_1[t_1] = t_2$ $b_1[t_1'] = t_2'$
Beliefs for Student	$b_2[t_2] = (Fr, t_1')$ $b_2[t_2'] = (Th, t_1)$

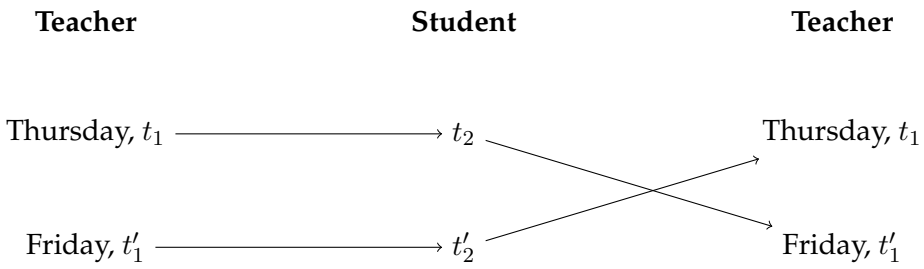


Figure 4.1: Beliefs diagram visualizing epistemic model

all. If in reality the student indeed has the belief hierarchy the teacher believes she has, the student *will* in fact be fully surprised. Type t_1' too expresses common belief in rationality, yet only allows the teacher to catch the student off guard with a surprise worth η by choosing Friday, which gives a utility less or equal to what a full surprise on Thursday would give.

Giving the exam on Thursday or on Friday are both reasonable in terms of common belief in rationality, and both under such reasoning can lead to (full) surprise. The former we might have concluded as well by simply applying an iterative elimination procedure, after which both Thursday and Friday would have survived. Indeed, the SEP belongs to a class of games in which iterative elimination of strictly dominated choices does always characterize exactly all reasonable choices under common belief in rationality.[5] This is because the teacher's utility only depends on his second-order beliefs. In light of this, we can make note here of the fact that Thursday and Friday are both not strictly dominated for the teacher, and hence survive this procedure. We build the epistemic models here for the purpose of being more precise about the belief hierarchies themselves.

The belief hierarchy induced by type t_1 is just one example of a belief hierarchy that satisfies common belief in rationality and under which

[5] See CHAPTER 5.

the teacher can believe to fully surprise the student. All of such belief hierarchies have in common that they induce a second-order expectation that puts probability one on Friday. There are also uncountably many belief hierarchies under which some partial surprise is possible. Only one of such belief hierarchies is induced by the psychological Nash equilibrium (Geanakoplos, 1996), where the teacher *correctly* believes the student has a full-support belief over his options. The equilibrium is given by the belief σ_1 where $\sigma_1(Th) = \frac{1}{\eta+1}$. To see this, first suppose that $\sigma_1(Th) > \frac{1}{\eta+1}$. Then we have if the teacher chooses Thursday

$$\begin{aligned} u_1(Th, \sigma_1(Th)) &= \sigma_1(Th) \cdot 0 + (1 - \sigma_1(Th)) \cdot 1 = 1 - \sigma_1(Th) \\ &< 1 - \frac{1}{\eta + 1} = \frac{\eta}{\eta + 1}. \end{aligned}$$

If the teacher chooses Friday, he receives utility

$$u_1(Fr, \sigma_1(Th)) = \sigma_1(Th) \cdot \eta + (1 - \sigma_1(Th)) \cdot 0 = \sigma_1(Th) \cdot \eta > \frac{\eta}{\eta + 1}.$$

It follows that it would be only optimal for the teacher to choose Friday and $\sigma_1(Th) = 0$, a contradiction. Now suppose that $\sigma_1(Th) < \frac{1}{\eta+1}$. Then the teacher's utility from choosing Thursday will be

$$u_1(Th, \sigma_1(Th)) = 1 - \sigma_1(Th) > \frac{\eta}{\eta + 1}$$

and the utility from choosing Friday will be

$$u_1(Fr, \sigma_1(Th)) = \sigma_1(Th) \cdot \eta < \frac{\eta}{\eta + 1}.$$

It follows that it would be only optimal for the teacher to choose Thursday and $\sigma_1(Th) = 1$, a contradiction. Thus $\sigma_1(Th) = \frac{1}{\eta+1}$ characterizes the unique psychological Nash equilibrium.

In this way, the teacher would never believe to be able to fully surprise the student, as opposed to the example we gave before. The reason

for this discrepancy lies in what it means for the teacher or the student to have a simple belief hierarchy. In the psychological Nash equilibrium of this psychological game we have a combination of beliefs $\sigma = (\sigma_1, \sigma_2)$ where σ_1 is the belief about the teacher's choice and σ_2 is the belief about the student's choice (which is a singleton by definition of the psychological game and thus can be ignored). Let us consider a belief hierarchy $\beta_1(\sigma_1)$, generated by σ_1. Then the teacher must not only believe that the student has belief σ_1 about his own choices, but, because $\beta_1(\sigma_1)$ is a simple belief hierarchy, the teacher must also believe that the student must believe he indeed believes that the student has belief σ_1 about the teacher's choice. And so on. In other words, the teacher must believe the student holds *correct beliefs* throughout. As a result, a simple belief hierarchy, by assuming correct beliefs, takes away much of the power to surprise the student.

4.3.2 Static SEP with "surprise only from giving exam"

Such a fixed-point solution is particularly problematic for exactly the same reason if the teacher exclusively cares for surprise caused by giving an exam. Surprise that the teacher actually cares about can then only derive from one particular type of action, instead of two. This would be the case of Table 4.2. In the game in Table 4.2, the teacher only gets utility from surprising the student by giving an exam. Again, there is only one possible psychological Nash equilibrium here. This equilibrium occurs when the student believes the teacher will choose Thursday, the teacher correctly believes she believes the teacher will choose Thursday and the student correctly believes the teacher indeed holds this second-order belief. Namely, consider a scenario in which $\sigma_1(Th) \neq 1$. This implies that the teacher would think the student believes he will choose Friday with positive probability. It is then only optimal for the teacher to choose Thursday and surprise the student at least a little. However, the student would anticipate this as she is correct about the teacher's second-order beliefs. Consequently, she fully

believes the teacher will choose Thursday. Hence, $\sigma_1(Th) = 1$, a contradiction.

Thus, in the psychological Nash equilibrium there is no room for surprise. This is a very intuitive result as well. Namely, if the teacher believes the student is correct about his second-order beliefs, then the student can easily deduce the teacher's rational choice. Since there is only one way in which the teacher could surprise her, the student deduces the teacher would have to choose Thursday to do so. Any inbound surprise will thus be predicted. However, this defeats the entire notion of surprise.

In contrast, let us now look at the scenario where satisfying common belief in rationality is the minimal requirement. If we again consider the belief diagram in Figure 4.1, one may verify that all belief hierarchies in that model still express common belief in rationality. The reason for this is that it is well possible here for the student to rationally believe the teacher will give the exam on Friday. If the teacher believes the student believes the exam will be given on Thursday, then the teacher expects to receive a utility of 0 in any case. Thus he would be completely indifferent between Thursday and Friday. This is what the student believes if she is of type t_2. Thursday is always an optimal choice. This is what type t_2' of the student believes. Type t_1 and t_1' again by definition express belief in the student's rationality. Hence, it is rational for the student to believe the teacher will give an exam on Friday, only if he expects that the student believes he will give the exam on Thursday. In such a scenario the student believes the teacher must have given up on surprising the student. He is then completely indifferent between Thursday and Friday. If the teacher believes the student has such a mindset, full surprise is still deemed possible.

Thus, there is a mode of thinking possible for the teacher such that he can believe he is able to give the exam on Thursday and fully surprise the student in the process. And if the student in reality has the reasonable belief hierarchy the teacher assigns to her, she will in fact be fully surprised. In the scenario presented in the introduction, the student

makes a valid observation about the teacher's potential reasoning that on Friday he cannot possibly surprise the student. However, it would not logically follow from this that the teacher must therefore believe the student believes the teacher will never give the exam on Friday. Namely, we have given a formal set-up where such reasoning is not the case. By believing the teacher will give the exam on Friday, the student already acknowledges that the announcement made should not be taken so literally. Namely, on Friday surprise will never be possible. Exactly this possibility of doubting the validity of the announcement in a reasonable way is what gives the teacher room to believe he can surprise the student and vindicate his announcement in the first place. This coincides with the argument first made by Quine (1953).

If the student is believed to be a cautious reasoner, there is little the teacher can do to surprise her under a state equivalent to common belief in rationality. To see this, consider some epistemic model with non-standard beliefs $M = (T_i, b_i)_{i \in I}$ as in Definition 3.5 of CHAPTER 3. Take a cautious type t_2^* for the student. Suppose the student with type t_2^*, with some *real*, positive probability $b_2[t_2^*](Fr) \in \mathbb{R}$, $b_2[t_2^*](Fr) > 0$ believes that the teacher will choose Friday and with a positive non-standard or real probability $b_2[t_2^*](Th) \in \mathbb{R}^*$, $b_2[t_2^*](Th) > 0$ believes that the teacher will choose Thursday. Let us assume here reasoning in line with common full belief in caution and primary belief in rationality for the teacher (See Definitions 3.8 and 3.9 in CHAPTER 3). Type t_2^* will only be able to primarily believe in the teacher's rationality, if Friday can indeed be an optimal choice for some teacher's type t_1^*. We should then have $u_1(Fr, t_1^*) \geq u_1(Th, t_1^*)$ for the belief hierarchy that t_1^* induces. We have

$$u_1(Fr, t_1^*) = 0$$

and

$$u_1(Th, t_1^*) = \sum\nolimits_{t_2 \in T_2} b_1[t_1^*](t_2) \cdot b_2[t_2](Fr),$$

where $b_1[t_1^*](t_2)$ is the probability that t_1^* assigns to the student being of type t_2 and $b_2[t_2](Fr)$ the probability that the teacher believes the student's type t_2 assigns to the teacher choosing Friday. Then we have

$u_1(Fr, t_1^*) \geq u_1(Th, t_1^*)$ only if

$$\sum\nolimits_{t_2 \in T_2} b_1[t_1^*](t_2) \cdot b_2[t_2](Fr) = 0.$$

Or equivalently, the teacher's type t_1^* must only believe that the student thinks he will give the exam on Thursday. However, then t_1^* does not believe in the student's caution. But then any arbitrary t_2^* that believes with some real, positive probability that the teacher will give the exam on Friday and primarily believes in the teacher's rationality, does not express 2-fold full belief in caution. Consequently, any type for the teacher that induces a belief hierarchy under which he believes he can surprise the student with some real, positive probability and believes in the student's caution, does not express up to 3-fold full belief in caution and primary belief in rationality. Therefore, he cannot reasonably believe to surprise the student under common full belief in caution and primary belief in rationality. In other words, when the teacher believes the student is cautious, the only choice that is rational under common full in caution and primary belief in rationality is Thursday. Both the student and teacher believe this is the case, making surprise nearly impossible.

4.3.3 Discussion

The analysis of the paradox here differs from logical approaches to the paradox in the sense that we model the paradox interactively. As a result our definition of surprise is different. We define a surprising event as a combination of an outcome (choice) and a first-order *belief* of the student which entails that she did not expect the outcome. In epistemic logic the student is surprised if she could not *know* the exam was coming. The epistemological approach to the paradox is best characterized by the resolution of Quine (1953). Quine identifies a case which can lead to a surprise exam: the student cannot accept the announcement to be true and believes it will come on the unsurprising day of Friday. Then Thursday will be a surprise. This conclusion was derived

by identifying a fallacy in the reasoning of the student in the story. This fallacy relates knowing the announcement or accepting it as true. Namely, to derive the contradiction that the student does in the story, the student must accept the announcement as true throughout the entire week. The contradiction should have made the student realize that she cannot simply do this.[6]

If our analysis is to be adequate, we should be able to link our results to previous intuitions or add to them. In relation to this epistemological approach for the two-day example we can directly confirm one intuition and add further to it. Indirectly we can also model an additional intuition from the epistemological approach.

(1) Corner Case - The analyses in Sections 4.3.1 and 4.3.2 illustrate that the resolution offered by Quine (1953) is a corner-case. By explicitly modelling the teacher's motivations in the paradox, we can allow for the possibility that the teacher also cares for surprising the student by not giving the exam. This corresponds to the game in Table 4.1. We showed that the model in Table 4.3 models belief hierarchies under which the teacher believes he can fully surprise the student. But we also established that many more belief hierarchies that allow for at least partial surprise are possible as well. One of these is the belief hierarchy that is induced by the unique psychological Nash equilibrium if $\eta > 0$.

(2) Quine's solution - The corner case of $\eta = 0$, as in Table 4.2, is indeed the spirit of Quine (1953). In this version, the teacher only cares for surprise from giving an exam. We find here that the teacher can think to surprise in a meaningful way only if he thinks the student believes he will choose Friday. Friday is a choice that can never lead to a surprising exam. Hence, by believing Friday, the student would not accept the announcement made by the teacher as true in a literal sense. This

[6] Other examples that make a similar argument for resolution of the two-day paradox are Kripke (2011), Kim and Vadusevan (2017) and the Inevitable Event in Holliday (2015).

is exactly the case that Quine identifies as the case that would lead to surprise.

(3) Correct beliefs - In our interactive analysis we can recreate the conditions for the student's fallacy as identified by Quine (1953). We do so using the correct beliefs assumption in the game of Table 4.2. Accepting the announcement as true implies that whatever the student deduces from this announcement, any other rational person that heard and accepts that announcement as true should be able to deduce as well. This includes the teacher. From the announcement, the student deduces on the evening of any given day, if the exam has not arrived yet, that it should come the next day. The student accepts the teacher should be able to make these deductions as well. In a game-theoretic setting, it means that if the student concludes something about the surprise-potential of a particular choice, she should believe the teacher concludes the same. This is what the correct beliefs assumption, in conjunction with common belief in rationality, achieves.

To see this, consider the following. Say the student accepts the announcement by the teacher as true and believes the teacher accepts his own announcement as true as well. Let this be commonly believed. The evening before Thursday she logically deduces from the announcement the first-order belief that the exam has to be given on Thursday if the teacher ever intends to surprise the student. Then because she believes the teacher also accepts his announcement as true, she must believe the teacher makes the same deduction for her. So she must believe the teacher believes she indeed believes the exam will be given on Thursday. Because the acceptance of the announcement by both must be commonly believed, it must be commonly believed that the student derives her first-order belief to be such that the teacher will give the exam on Thursday. That is to say, the student believes the teacher is correct about her own beliefs, and this is commonly believed. This is exactly the correct beliefs assumption. In our game-theoretic setting, the student under this assumption derives the same conclusion. Under correct beliefs, common belief in rationality leads

her to conclude that the teacher must give the exam on Thursday and that this is commonly believed.

We can say that accepting the announcement as true is an argument for assuming correct beliefs in the game-theoretic description of the paradox. Our equilibrium analysis shows then that an exam on Thursday will be expected by the student, rendering surprise impossible under the assumptions. This is also the conclusion the student draws, leading eventually to the contradiction with the announcement. According to the resolution of Quine (1953), the student should then no longer have to accept the announcement as true. And in line with this, she should no longer have to believe the teacher himself accepts it as true. But then she has no reason to insist on correct beliefs in the first place.

We concluded from our analyses that, in order for surprise to be deemed possible, in the game of Table 4.2 the student must be able to doubt the validity of the announcement. We can however say something even stronger. The student should be able to believe the teacher cannot surprise her at all. This is what follows from the analysis where it is assumed the student has cautious beliefs : each possible choice for the teacher receives some positive probability in the student's belief. If it is commonly believed that the student is a cautious reasoner, then the teacher can always believe to cause partial surprise, where this partial surprise may be infinitesimally small. We showed that when we add however primary belief in rationality on top of common full belief in caution, Thursday would be the only rational choice for the teacher. Hence, if common full belief in caution and primary belief in rationality is assumed, the student would have to believe with almost probability-one that the teacher will give the exam on Thursday. Then Thursday will be hardly a surprising choice.

Assuming full-support beliefs, while retaining some form of common belief in rationality, thus eliminates almost all possibility of surprise for the teacher if he only cares about surprise from giving an exam. The traditional game-theoretic literature on the SEP has put forward

the idea that the teacher may outsmart the student by making it commonly believed that he is considering a full-support distribution over both Thursday and Friday. The student then would have too guess this distribution (Sober, 1998; Ferreira and Bonilla, 2008). The teacher would then at most expect partial surprise to be possible. We have argued that in a psychological game-setting such an approach will not result in any relevant surprise. Instead, this seems more to be the artefact of artificially including motivations for the student in the game, that are conflicting with the teacher's.

4.4 Dynamic Surprise Exam Paradox

It was found that for both games in Tables 4.1 and 4.2 the teacher can believe he can fully surprise the student. For the game in Table 4.2 this would however require the student to believe the teacher will reasonably choose the always *unsurprising* day of Friday. Only then the teacher can deem it possible that he can fully surprise the student. This may seem like an unsatisfactory explanation to the problem: vindicating the announcement directly depends on the student questioning it.[7] We will show that by adding more days to the paradox-problem, this concern for the version of the SEP in Table 4.2 is alleviated quite a bit. For both versions of the SEP we will show that by extending the SEP further, more opportunities for full surprise can be deemed possible by the teacher. First, however, we will define the formal, dynamic environment in which the SEP takes place.

4.4.1 Description of dynamic SEP

In a dynamic psychological game the utilities also depend on belief-dependent motivations. In a dynamic setting these beliefs are *conditional beliefs*. Namely, as the dynamic game progresses, a player might find out that his opponents are employing strategies he first did not

[7] This is exactly the critique of Kripke (2011) on the resolution by Quine (1953).

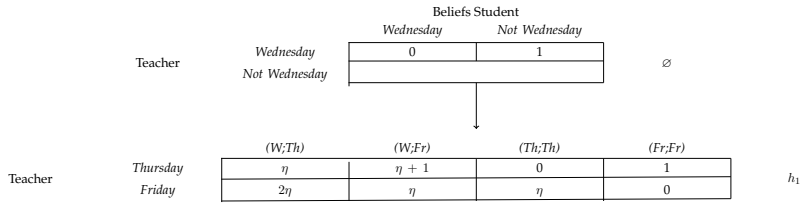


Figure 4.2: Dynamic situation with surprise by giving or not giving exam

expect. The conditioning takes place when a certain information set $h \in H$, where player $i \in I$ makes a choice, is reached in the game. For the set of players in the dynamic surprise exam paradox we have $I = \{Teacher, Student\}$. Moreover, the information sets in a three-day setting correspond to Wednesday (denoted by $\varnothing$) and Thursday (denoted by h_1). See for instance Figure 4.2. Note that in the three-day setting we can distinguish between three pure strategies for the teacher: W (Wednesday), (NW, Th) (not Wednesday but Thursday) and (NW, Fr) (not Wednesday but Friday). We choose to henceforth abbreviate the latter two strategies to Th and Fr respectively.

Like in the static form of the surprise exam game, we will look at two versions of the game, both in which the teacher's utility depends linearly on his conditional second-order expectations. The first version is depicted in Figure 4.2. Here the teacher gets a utility of 1 if he surprises the student by giving the exam, whereas the teacher receives a utility of $0 < \eta \leq 1$ at the end of the game for *each time* that he creates a small surprise event for the student by not giving the exam. The cells in Figure 4.2 illustrate every combination of a choice and an extreme second-order expectation that is relevant for the teacher's utility. In this scenario this means that, when the teacher decides not to give the exam on Wednesday, the subsequent subgame consists of the teacher's possible strategies and vectors of extreme conditional second-order expectations. For instance, $(W; Th)$ indicates that the teacher expects the student to believe at $\varnothing$ that the teacher will be giving the exam on

Wednesday and at h_1, if the exam was not given on Wednesday, believes he will give the exam on Thursday instead. In other words, the teacher's utility at h_1 not only depends on what he believes the student believes at h_1, but also on what he believes the student believed at $\varnothing$. The depicted utilities in Figure 4.2 can then be explained as follows: if the teacher gives the exam on Wednesday while the student believed he would give the exam on a later day, then the teacher gets a utility of 1. If the teacher decides to give the exam not on Wednesday, then the game moves on to the subsequent subgame. However, in the process of moving to the next subgame, the teacher may carry with him a utility of η. This occurs when the teacher expects at h_1 that he has managed to surprise the student at $\varnothing$ by not giving the exam while the student believed he would give one. As a result, if the teacher manages to surprise the student at h_1, the teacher could receive a utility up to $\eta + 1$ in the end. However, if the teacher at h_1 expects not to have surprised the student at $\varnothing$, then we have at h_1 essentially the same game as depicted in Table 4.1 in Section 4.2.1.

It should be mentioned here that according to our description of a dynamic psychological game the extreme conditional second-order beliefs $(Th; Fr)$ and $(Fr; Th)$ should have been included in Figure 4.2 as well. However, there is no particular reason for the student to update his beliefs at h_1 if he already expected the teacher not to give the exam on Wednesday. Because the student is passive in the game, there is no student's action observable for the teacher such that he may reconsider what the student is thinking about him. The beliefs $(Th; Fr)$ and $(Fr; Th)$ do not make sense in that regard. In other words, we assume Bayesian updating here.

Similarly, we can extend the psychological game in which the teacher only cares about surprise caused by giving the exam (see Table 4.2) to include Wednesday as well. The resulting game is depicted in Figure 4.3. In this psychological game, the teacher receives a utility of 1 if he manages to surprise the student by giving the exam. Since on Friday the teacher knows the student knows the exam has to be given if the exam has not been given by that time, choosing Friday as a strategy

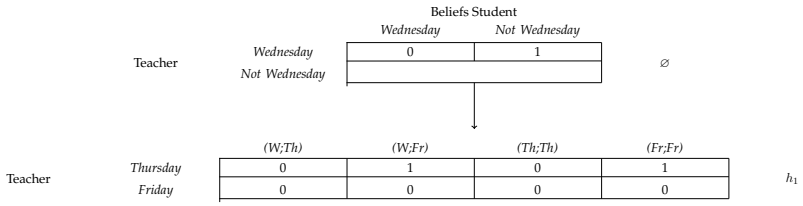


Figure 4.3: Dynamic situation with meaningful surprise only by giving exam

will regardless of the conditional belief hierarchy result in a utility of 0. Here we only defined a dynamic psychological game for the SEP in particular. For a more general definition of dynamic psychological games, the reader is referred to Battigalli and Dufwenberg (2009).

Also in a dynamic setting we can use types to capture belief hierarchies in the SEP. These types form beliefs about the strategy-type combinations of their opponents. This is done for both information sets $\varnothing$ and h_1, resulting in an epistemic model for the three-day dynamic surprise exam game.

Definition 4.7 (Dynamic epistemic model for the Surprise Exam Paradox)**.**
*Consider a dynamic surprise exam game D. A **dynamic epistemic model** $M = (T_i, b_i)_{i \in I}$ for D specifies for both the teacher and the student a finite set of possible types denoted by T_1 and T_2 respectively. For every type $t_1 \in T_1$ for the teacher, we specify at every information set $h \in H = \{\varnothing, h_1\}$ a probability distribution $b_1[t_1, h]$ over the set of the student's types T_2. For every type $t_2 \in T_2$, we specify at every information set $h \in H$ with $H = \{\varnothing, h_1\}$ a probability distribution $b_2[t_2, h]$ over the set of the teacher's strategy-type combinations $S_1(h) \times T_1$ where $S_1(h)$ is the set of the teacher's strategies that lead to h. Hence, $S_1(\varnothing) = \{W, Th, Fr\}$ and $S_1(h_1) = \{Th, Fr\}$.*

4.4.2 Common belief in future rationality

We can extend the notion of common belief in rationality to a dynamic setting as well. As the Surprise Exam Paradox is a problem of backward induction, we will focus on common belief in future rationality (Perea, 2014).[8] Similarly to optimality in a static setting, we say a strategy $s_1 \in \{Wednesday, Thursday, Friday\}$ is optimal for a type $t_1 \in T_1$ at an information set $h \in \{\varnothing, h_1\}$ if $\forall s_1' \in S_1(h)$: $u_1(s_1, \beta_1[t_1, h]) \geq u_1(s_1', \beta_1[t_1, h])$, with $\beta_1[t_1, h]$ being the conditional belief hierarchy represented by type t_1 at information set h. So, in case of the Surprise Exam Paradox, this e.g means that choosing Wednesday for the teacher is only an optimal choice if the expected utility at $\varnothing$ derived from said choice given a conditional belief hierarchy $\beta_1[t_1, \varnothing]$ is higher than what the teacher expects to get from choosing either Thursday or Friday given the same belief hierarchy. Then, belief in the teacher's future rationality can be defined as follows.

Definition 4.8 (Belief in the teacher's future rationality)**.**
Consider a dynamic epistemic model $M = (T_i, b_i)_{i \in I}$ in the Surprise Exam Paradox with a type $t_2 \in T_2$ for the student within that dynamic epistemic model. Moreover, consider an information set $h \in \{\varnothing, h_1\}$ and an information set $h' \in \{\varnothing, h_1\}$ that weakly follows h. Type t_2 believes at h the teacher will choose rationally at h' whenever t_2's conditional belief $b_2[t_2, h]$ only assigns positive probability to strategy-type pairs (s_1, t_1) where s_1 is optimal for t_1 at h' whenever s_1 leads to h'.
Type t_2 ***believes in the teacher's future rationality at h*** *if t_2 believes that the teacher will choose rationally at every h' that weakly follows h.*
We say type t_2 ***believes in the teacher's future rationality*** *if t_2 believes at both $\varnothing$ and h_1 in the teacher's future rationality.*

[8] Dekel et al. (1999) and Asheim and Perea (2005) formally model backward induction reasoning by the notion of sequential rationalizability. Baltag et al. (2009) as well as Penta (2015) propose different concepts as to capture backward induction reasoning, which subtly differ in the restrictions the concepts impose, yet capture the same basic idea. We will however be looking at a direct dynamic counterpart of common belief in rationality: common belief in future rationality (Perea, 2014).

So for the student to believe at $\varnothing$ in the teacher's future rationality, she must believe that the teacher will make an optimal choice at $\varnothing$ and at h_1. Similarly, the student believes in the teacher's future rationality at h_1 if she believes at h_1 that the teacher will choose optimally at h_1. If the student believes in the teacher's future rationality at both $\varnothing$ and h_1 we say she believes in the teacher's future rationality throughout. The teacher always believes in the student's future rationality, as the student has no choices to make. Common belief in future rationality can now be defined as follows for the SEP.

Definition 4.9 (Common belief in future rationality in the Surprise Exam Paradox)**.**
Consider an epistemic model $M = (T_i, b_i)_{i \in I}$ in the dynamic Surprise Exam Paradox. Moreover, let a player i either represent the teacher or the student. For every player i and every type $t_i \in T_i$, we say that type t_i expresses 1-fold belief in future rationality if t_i believes in the opponent's future rationality.
For every $k > 1$, every player i, and every type $t_i \in T_i$, we say that type t_i expresses k-fold belief in future rationality if t_i only assigns positive probability at every information set h to the opponent's types that express (k-1)-fold belief in future rationality.

*Type t_i expresses **common belief in future rationality** if it expresses k-fold belief in future rationality for every k.*

A *rational choice under common belief in future rationality* is then a choice that is optimal given a conditional belief hierarchy that expresses common belief in future rationality.

4.4.3 Psychological subgame perfect equilibrium

The equilibrium concept that is applicable to the dynamic SEPs we are considering is the equivalent of sequential psychological equilibrium with observed past choices by Geanakoplos et al. (1989): *psychological subgame perfection*. Just like in a psychological Nash equilibrium, reasoning from a psychological subgame perfect equilibrium

stems from simple belief hierarchies. This implies that in a dynamic SEP at both information sets $\varnothing$ and h_1, the conditional belief hierarchies are generated by a first-order belief σ_1 about the teacher's strategy. The difference now however is that we have $\sigma_1 = (\sigma_1(h))_{h \in \{\varnothing, h_1\}}$ with $\sigma_1(\varnothing) \in \Delta(S_1(\varnothing))$ and $\sigma_1(h_1) \in \Delta(S_1(h_1))$. Thus σ_1 specifies for both information sets a first-order belief about the available choices at that information set for the teacher. The conditional belief hierarchy $\beta_1(\sigma_1)$ that is generated by σ_1 in the SEP implies that *(i)* the teacher believes at every $h \in \{\varnothing, h_1\}$ that the student has belief $\sigma_1(h')$ at every $h' \in \{\varnothing, h_1\}$, that *(ii)* the teacher believes at every $h \in \{\varnothing, h_1\}$ that the student believes at every $h' \in \{\varnothing, h_1\}$ that the teacher believes at every $h'' \in \{\varnothing, h_1\}$ that the student has belief $\sigma_1(h''')$ at every $h''' \in \{\varnothing, h_1\}$, and so on. Note that also in the dynamic setting, by construction σ_1 implies correctness of beliefs.

We can nove give the following definition for psychological subgame perfection in the SEP.

Definition 4.10 (Psychological subgame perfect equilibrium)**.**
Consider a dynamic surprise exam game portrayed in either Figure 4.2 or in Figure 4.3. Let σ_1 be a first-order belief about the teacher's choice. Let additionally $\beta_1(\sigma_1)$ be the conditional belief hierarchy for the teacher that is generated by σ_1. Then σ_1 constitutes a ***psychological subgame perfect equilibrium*** *if*

$$\forall h \in \{\varnothing, h_1\} : \sigma_1(h)(s_1) > 0 \Rightarrow$$
$$\forall s_1' \in S_1(h) : u_1(s_1, \beta_1(\sigma_1, h)) \geq u_1(s_1', \beta_1(\sigma_1, h)).$$

A psychological subgame perfect equilibrium $\beta_1(\sigma_1)$ is thus such that it assigns in its second-order belief positive probability only to a particular strategy such that this strategy maximizes expected utility given the full conditional belief hierarchy that is generated by σ_1. It should be noted that a psychological subgame perfect equilibrium is not generally equivalent to having a psychological Nash equilibrium at every subgame. Namely, the history of choices made in the past does not

capture all the necessary information for a player to determine his optimal choice, which may also depend on what an opponent believed in the past or what an opponent might have believed given a non-realised history of choices. This is a situation present in the game of Figure 4.2, but not in the game of Figure 4.3.

4.4.4 Dynamic SEP: surprise by giving and not giving the exam

We are now in a position to analyse the nature of the paradox in a dynamic setting. First let us take the dynamic surprise exam game from Figure 4.2. In what ways can the teacher surprise the student in this dynamic setting under common belief in future rationality?

To answer this, let us consider the epistemic model portrayed in Table 4.4. We can confirm that all types for the teacher express common belief in future rationality here by showing that both types of the student believe in the teacher's future rationality. To show this, let us start at type t_2 of the student. The student then believes the teacher is of type

Table 4.4: Dynamic epistemic model for the game in Figure 4.2

	$T_1 = \{t_1, t_1'\}$ $T_2 = \{t_2, t_2'\}$
Types	
Beliefs for Teacher	$b_1[t_1, \varnothing] = t_2$ $b_1[t_1, h_1] = t_2$ $b_1[t_1', \varnothing] = t_2'$ $b_1[t_1', h_1] = t_2'$
Beliefs for Student	$b_2[t_2, \varnothing] = (Fr, t_1')$ $b_2[t_2, h_1] = (Fr, t_1')$ $b_2[t_2', \varnothing] = (W, t_1)$ $b_2[t_2', h_1] = (Th, t_1)$

t_1' at both $\varnothing$ and h_1. If the teacher is of type t_1', he believes that the student believes on Wednesday that he will choose to give the exam on Wednesday and that the student believes on Thursday that he will in fact choose to give the exam on Thursday. Then, on Wednesday it is optimal for the teacher to give the exam at least not on Wednesday, because the student would otherwise anticipate his choice. Since the teacher believes on Wednesday that the student believes on Thursday that the teacher will give the exam on Thursday, the teacher can subsequently only believe to surprise the student by choosing to give the exam on Friday. Hence following the strategy to not give the exam on Wednesday but rather on Friday is optimal for the teacher on Wednesday ($\varnothing$) and Thursday (h_1) if he is of type t_1'. Since these are exactly the beliefs that the student's type t_2 holds while only considering the teacher's type t_1', type t_2 expresses 1-fold belief in future rationality. In case the student is of type t_2', she believes the teacher is of type t_1 at both $\varnothing$ and h_1. The teacher's type t_1 believes the student believes, at both Wednesday and Thursday, that the teacher will give the exam on Friday. Hence, if the teacher is of type t_1, choosing Wednesday is optimal at $\varnothing$ and not choosing Wednesday but Thursday is optimal at h_1 for the teacher. These are exactly the beliefs the student holds if she is of type t_2' on both Wednesday and Thursday. Namely, type t_2' believes at $\varnothing$ that the teacher chooses Wednesday and at h_1 that the teacher will choose Thursday. Hence the student's type t_2' expresses 1-fold belief in future rationality.

As the student is passive, we automatically have that both types of the teacher believe in the student's future rationality at both $\varnothing$ and h_1. Hence all types express common belief in future rationality. Consequently, under common belief in future rationality the teacher can rationally choose to give the exam on Wednesday, Thursday or Friday, since Wednesday and Thursday are optimal for t_1 and Friday is an optimal choice for t_1'.

This epistemic model is a special case in the sense that if the teacher has the belief hierarchy induced by t_1, then he believes he is able to fully surprise the student at every single information set by giving the exam

on Wednesday or Thursday. Namely, at $\varnothing$ he can rationally choose Wednesday while believing that the student fully believes the teacher will not give the exam on Wednesday. If for some reason the teacher chooses not to give the exam on Wednesday while being of type t_1, then still the teacher can fully surprise the student by giving the exam on Thursday, as he believes the student at h_1 believes that the teacher will give the exam on Friday. Additionally, if the teacher has the belief hierarchy induced by t_1', he might expect an even higher expected utility, if $\eta > \frac{1}{2}$. That is, the teacher believes the student believes on Wednesday he will give the exam on Wednesday. By not giving it on Wednesday, the teacher can carry over some utility already from surprising the student by not giving the exam. Overall, the teacher's options for surprising the student have increased, since he has now more days available to surprise the student on.

This result contrasts with the possible strategies and beliefs under the concept of a psychological subgame perfect equilibrium. We know from our discussion at Section 4.3.1 that we must have $\sigma_1(h_1)(Th) = \frac{1}{\eta+1}$. Then, it must be the case that $\sigma_1(\varnothing)(W) = \frac{1}{(\eta+1)^2}$. To see why, consider the contrary. Say $\sigma_1(\varnothing)(W) > \frac{1}{(\eta+1)^2}$. Then it would be always optimal to not choose Wednesday. Namely, we have:

$$u_1((W, \sigma_1), \varnothing) = 1 - \sigma_1(\varnothing)(W) < 1 - \frac{1}{(\eta+1)^2}.$$

By not choosing Wednesday (NW), the expected utility at $\varnothing$ if the teacher expects the student to believe he will give the exam on Wednesday, thus *conditional* on $\sigma_1(\varnothing)(W) = 1$, is

$$u_1((NW, \sigma_1), \varnothing | \sigma_1(\varnothing)(W) = 1) = \frac{\eta}{\eta+1} + \eta = 1 - \frac{1}{\eta+1} + \eta.$$

Namely, in equilibrium the teacher receives $\frac{\eta}{\eta+1}$ at h_1, and he receives an additional η from surprising the student at $\varnothing$ by not giving the exam. Similarly, we have for not choosing Wednesday while the stu-

dent believes the teacher does not choose Wednesday:

$$u_1((NW, \sigma_1), \varnothing | \sigma_1(\varnothing)(W) = 0) = \frac{\eta}{\eta + 1} = 1 - \frac{1}{\eta + 1}.$$

In this case, the student expects the teacher to not choose Wednesday, and hence the teacher does not receive this extra η. As we assumed utility is linear in the second-order expectations, we have an expected utility at $\varnothing$ of

$$\begin{aligned} u_1((NW, \sigma_1), \varnothing) &= \sigma_1(\varnothing)(W)(1 - \frac{1}{\eta + 1} + \eta) + (1 - \sigma_1(\varnothing)(W))(1 - \frac{1}{\eta + 1}) \\ &= 1 + \sigma_1(\varnothing)(W)\eta - \frac{\eta + 1}{(\eta + 1)^2} \\ &> 1 + \frac{1}{(\eta + 1)^2}\eta - \frac{\eta + 1}{(\eta + 1)^2} = 1 - \frac{1}{(\eta + 1)^2}. \end{aligned}$$

Hence, it is optimal for the teacher to not choose Wednesday. However, the student would anticipate that not choosing Wednesday is optimal for teacher and will thus expect him not to choose Wednesday. But then we have $\sigma_1(\varnothing)(W) = 0 < \frac{1}{(\eta+1)^2}$, a contradiction.

Now let us consider the opposite. Let $\sigma_1(\varnothing)(W) < \frac{1}{(\eta+1)^2}$. From the relations above we can infer that $u_1((W, \sigma_1), \varnothing) > u_1((NW, \sigma_1), \varnothing)$. Again, however, the student would be able to anticipate that the teacher would choose to give the exam on Wednesday. This would give us $\sigma_1(\varnothing)(W) = 1 > \frac{1}{(\eta+1)^2}$, a contradiction.

In equilibrium, the expected utility for the teacher at Wednesday from choosing Wednesday is $u_1((W, \sigma_1), \varnothing) = 1 - \frac{1}{(\eta+1)^2}$ and the utility from not choosing Wednesday but either Thursday or Friday is also $u_1((NW, \sigma_1), \varnothing) = 1 - \frac{1}{(\eta+1)^2}$. Similarly, it can also be confirmed that $u_1((Th, \sigma_1), h_1) = u_1((Fr, \sigma_1), h_1) = 1 - \frac{1}{(\eta+1)^2}$. Note that this equilibrium utility is strictly larger than the utility the teacher expects to get in equilibrium in the static game, which was $1 - \frac{1}{(\eta+1)}$. In fact, if we extend the psychological game even further to allow for more days

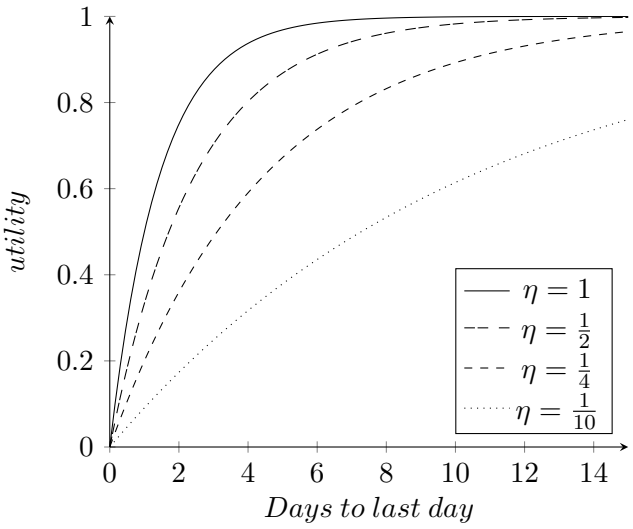


Figure 4.4: Expected utilities in equilibrium for game in Figure 4.2

to potentially give an exam on, the expected utility will increase even further. The intuition behind this is simple: as the number of days between the announcement and the last possible day to give the exam increases, there are more options for the teacher to potentially surprise the student. It thus becomes less likely for the student to anticipate the day of the exam. On the other hand, being able to divert the exam on more occasions, the teacher can accumulate utility from surprising the student by not giving the exam. The effect of this relation on the teacher's expected utility is portrayed in Figure 4.4. For another angle on this result, we can refer the reader to Geanakoplos (1996).

Extending the surprise exam game to more than three days will not have an effect on the possibility of the teacher being able to fully surprise the student under common belief in future rationality or not. That is, we already established that in a two-day example this is already well possible. However, adding more days to the problem will

expand the set of rational strategies under common belief in future rationality. And with that, also the amount of days on which the teacher can reasonably believe he can surprise the student. By slightly modifying the epistemic model in Table 4.4 this can be accommodated for. For instance, in case of a four day example we could simply extend the epistemic model in Table 4.4 such that type t_2 of the student believes at Tuesday the teacher will choose Friday and type t_2' of the student believes on Tuesday the teacher will choose to give the exam on Tuesday. In such a model, the teacher is able to surprise the student under common belief in future rationality on Tuesday, Wednesday, Thursday and Friday, if he is of type t_2.

Thus, much like in the static setting with common belief in rationality, there are in this version of the dynamic surprise exam potentially many belief hierarchies possible that express common belief in future rationality and where the teacher can (partially) surprise the student. These belief hierarchies include at least one where the teacher believes he can fully surprise the student at every information set. In fact, by extending the game by one day compared to the static setting the teacher has gained extra options to fully surprise the student. Under a psychological subgame perfect equilibrium, the options for surprise in this version are limited. Again, the main reason for that observation is the requirement of correct beliefs. Actual full surprise is still not possible under this concept. However, if η is large enough, the partial surprise that is possible in equilibrium gets close to full surprise rather quickly by adding more days to the problem.

4.4.5 Dynamic surprise exam: surprise only by giving the exam

Instead of having $0 < \eta \leq 1$, we could also consider $\eta = 0$. This is the dynamic version of the game discussed in Section 4.3.2, in which the teacher only cares about surprising the student by in fact giving the exam. This is the version of the surprise exam paradox that is most often referred to and best captures the crux of the paradox. The resulting psychological game is presented in Figure 4.3. For the purpose of

analysing this game, we can utilise the epistemic model presented in Section 4.4.4, now repeated in Table 4.5. In a similar fashion as before, it can be verified that all types express common belief in future rationality. Since Wednesday and Thursday are optimal for the teacher's type t_1, and Friday is optimal for his type t_1', the teacher can rationally choose Wednesday, Thursday or Friday under common belief in future rationality.

Type t_1 represents only one example of a conditional belief hierarchy that expresses common belief in future rationality and under which full surprise is possible. That is, we might as well have considered a belief hierarchy encoded by a type t_1^* for the teacher which is similar to his type t_1 except that, at $\varnothing$ and/or h_1, he believes the student also assigns some positive probability to the teacher choosing Thursday instead of only Friday. Then full surprise on Wednesday by choosing Wednesday would still have been possible under common belief in future rationality. Thus, not only the set of rational strategies has become larger when adding an extra day to the problem. Also the set of con-

Table 4.5: Dynamic epistemic model for the game in Figure 4.3, Version A

	$T_1 = \{t_1, t_1'\}$		
Types	$T_2 = \{t_2, t_2'\}$		
Beliefs for Teacher	$b_1[t_1, \varnothing]$	=	t_2
	$b_1[t_1, h_1]$	=	t_2
	$b_1[t_1', \varnothing]$	=	t_2'
	$b_1[t_1', h_1]$	=	t_2'
Beliefs for Student	$b_2[t_2, \varnothing]$	=	(Fr, t_1')
	$b_2[t_2, h_1]$	=	(Fr, t_1')
	$b_2[t_2', \varnothing]$	=	(W, t_1)
	$b_2[t_2', h_1]$	=	(Th, t_1)

ditional belief hierarchies that allow for full surprise has expanded, as the amount of first-order beliefs of the student under which the teacher can expect to fully surprise the student has increased.

Extending the SEP by including more days has another intuitive appeal in explaining the SEP. In the static version, we showed that full surprise on Thursday is *only* deemed possible if the student believes the teacher thinks he cannot surprise the student and thus chooses to give the exam on Friday. This might be considered an unsatisfactory explanation for the paradox, as the teacher announced clearly that the student will be surprised. In the dynamic version however, the teacher can believe he is able fully surprise the student whilst believing the student believes he can do so. Consider for instance the belief hierarchy that is induced by type t_1 in the model of Table 4.6, a belief hierarchy that expresses common belief in future rationality. For the belief hierarchy induced by t_1, Wednesday is still expected to cause a full surprise. However, this time the teacher believes the student believes at both $\varnothing$ and h_1 the teacher will give the exam on Thursday. Concentrating

Table 4.6: Dynamic epistemic model for game in Figure 4.3, Version B

	$T_1 = \{t_1, t_1'\}$		
Types	$T_2 = \{t_2, t_2'\}$		
Beliefs for Teacher	$b_1[t_1, \varnothing]$	=	t_2
	$b_1[t_1, h_1]$	=	t_2
	$b_1[t_1', \varnothing]$	=	t_2'
	$b_1[t_1', h_1]$	=	t_2'
Beliefs for Student	$b_2[t_2, \varnothing]$	=	(Th, t_1')
	$b_2[t_2, h_1]$	=	(Th, t_1')
	$b_2[t_2', \varnothing]$	=	(W, t_1)
	$b_2[t_2', h_1]$	=	(Fr, t_1)

on the belief of the student at h_1, this is a reasonable belief to hold. Namely, the student believes the teacher expects a full surprise on this day, as the student believes the teacher expects she believes he will choose Friday. Friday will always lead to an unsurprising exam. So only in the *fourth-order conditional belief*, belief in an unsurprising exam takes place. Thus, when being of type t_1 in Table 4.6, the teacher can expect to fully surprise the student whilst believing the student also believes he expects to be able to do so. This is because the teacher now has more than one strategy that are reasonable under common belief in (future) rationality and that can possibly lead to surprise: *Wednesday* and *Thursday*. A priori, under common belief in future rationality, the student does not have any reason to find one more reasonable than the other.

Also in a dynamic setting, psychological subgame perfect equilibrium again faces the same pitfall as traditional game theory does. In Section 4.3.1 we already established that $\sigma_1(h_1)(Th) = 1$ needs to be the case. This belief would give the teacher always a utility of 0 at h_1. However, then we know that it must be the case that $\sigma_1(\varnothing)(W) = 1$ too. Namely, if $\sigma_1(\varnothing)(W) < 1$, then the teacher would always be better off by choosing Wednesday. The student would be able to anticipate this and hence believe that $\sigma_1(\varnothing)(W) = 1$, a contradiction.

The discussion above explains where equilibrium concepts tend to go wrong in analyzing the Surprise Exam Paradox. The fact that Thursday is always an optimal choice at h_1 does not mean that Friday is ruled out as a possible choice for the teacher. The teacher only wishes to surprise the student, yet the exam eventually has to be given. If Friday still happens to be ruled out, then surprise on Thursday is no longer possible. However, this again does not mean that it is impossible for the student to consider any day after Wednesday as a choice for the teacher at $\varnothing$. As long as there is a belief hierarchy that expresses common belief in future rationality such that the teacher believes the student believes at any given day that the exam will be given in the future, the teacher can believe surprising the student is possible. In this

particular version of the game this means there is at least another possible belief hierarchy that believes the student believes the teacher will give the exam on the present day. The reasonable doubt of the student in the announcementis what allows the teacher to back up his statements about surprising the student. Much like psychological Nash equilibrium, a psychological subgame perfect equilibrium imposes an additional requirement of correctness in beliefs however, which implies the teacher has a simple belief hierarchy. Whereas extending the Surprise Exam Paradox allowed *more* possibilities for surprise under common belief in future rationality, under psychological subgame perfection there is still a *unique* combination of beliefs, which allows for no surprise. As such, the correct beliefs assumption underlying a psychological subgame perfect equilibrium, like in its static counterpart, significantly reduces the teacher's ability to surprise the student.

4.4.6 Discussion

In a broader discussion, our analyses of the dynamic SEP in Sections 4.4.4 and 4.4.5 establish two things. Both relate to how the approach to the dynamic paradox taken in this section helps in relaxing the conclusion of the epistemological approach, going back to Quine (1953).

(1) Corner Case - Section 4.4.4 proposes a resolution in defining the teacher's preferences in such a way that he also cares for surprise resulting from unexpectedly not giving an exam. Just as we established in Section 4.3.3, the resolution of Quine (1953) for the dynamic scenario is essentially a corner case. First, the part about "a surprise" in the teacher's announcement can always be believed whilst surprise still being possible. This is apparent in the epistemic model of Table 4.5. There the teacher can surprise the student on Wednesday or Thursday, while he believes the student believes he will succeed in surprising her by not giving the exam on Wednesday or Thursday but on Friday. And this can be commonly believed under common belief in future rationality. Second, under this resolution the student does not have to completely disregard the announcement as true, as opposed to what is

argued by Quine (1953). The subgame-perfect equilibrium induces a completely probabilistic conditional belief hierarchy. This implies that at each day the student expect to be partially surprised. As the number of days to the last possible day increases, this partial surprise even approaches full surprise.

(2) Two days against more days - There is a difference between the two-day paradox and the paradox that includes more than two days. This difference occurs in relation to being able to accept the announcement as true. To see this, take the epistemic model of Table 4.6 for the paradox where the teacher only cares for surprise from actually giving the exam. In the static version we established that the teacher has to believe the student does not accept his announcement as true in order to be able to (fully) surprise her under common belief in future rationality. When adding a day to the paradox, we see this is no longer a necessary condition. The teacher can think to surprise the student by choosing Wednesday if it is believed she believed the exam will come on Thursday. As the analysis showed, the student can believe this option Thursday can be expected by the teacher to be surprising as well. The doubt in the validity of the announcement is only required later in the teacher's conditional belief hierarchy, in his fourth-order belief.

When one would extend the paradox with even more days, the same necessary condition will appear: doubt in the validity of the announcement is only required by the fourth-order belief of the teacher in order for surprise to be considered possible. The reason is that in each higher-order belief, also a conjecture needs to be formed conditional on when the game continues until the last Thursday. There we only have two possibilities that the student can believe in: (a) a surprising Thursday because of a teacher's second-order belief that places probability on an unsurprising Friday, or (b) an unsurprising Thursday or Friday.

Letting the paradox run for more than two days allows us to get around the rather unsatisfactory resolution that the teacher can only believe to

surprise the student if he believes the student will not believe his announcement is true. We showed that distinguishing between $n = 2$ days against $n > 2$ days helps in this regard. This conclusion departs from what is mostly argued in literature. Previous resolutions in logic, epistemic logic and game theory argue the paradox can be fully explained by a one-day case (Quine, 1953) or two-day case (Sorensen, 1988; Sober, 1998; Gerbrandy, 2007) only. Notable exceptions include Kripke (2011) and Holliday (2015), who both argue that if the student accepts something as true at the start of the exam week, it does not necessarily follow the student knows her future selves will do this as well. In the $n = 2$ days case, there is no future self.

4.5 Conclusion

The Surprise Exam Paradox has slowly garnered some interest from the field of game theory in recent times. On a surface level, it appears to highlight some potential red flags for backward induction reasoning in games. Common belief in future rationality formally captures where the student's backward induction reasoning goes wrong if he reaches a conclusion that the teacher cannot possibly surprise her. Namely, even though in the actual crux of the paradox the teacher cannot surprise the student on the last day, it does not follow from this that the student cannot believe he will give the exam on the last day.

Common belief in future rationality in a setting of psychological games shows there exists a valid reason for the student to doubt the validity of the teacher's announcement of surprising her by giving the exam, even on the last day. If the student believes all the teacher's routes to surprising the student have been cut off, it can be reasonable for the student to believe the teacher is thinking about giving the exam on the last day. Equilibrium concepts in psychological game theory are inherently incapable of capturing such doubts because of their assumptions

on correctness of beliefs. We argued that imposing such a correct beliefs assumption is strongly linked to the epistemological problem of actually *knowing* the announcement.

In a broader sense, the paradox teaches us to be more careful with the type of assumptions we impose on reasoning of players when dealing with psychological games. In these games, preferences depend not only on outcomes, but also on information captured in belief hierarchies. This includes doxastic states of players at particular points in time, such as first-order beliefs. What the paradox illustrates is that there exist preferences that rely on the *absence* of equilibrium reasoning. Surprise-seeking is one of them.

5

When is Iterated Elimination of Choices Enough?

This chapter is adapted from: "Reasoning in psychological games: when is iterated elimination of choices enough?" (Mourmans, 2019).

5.1 Introduction

From traditional game theory we have become familiar with reasoning about interactive scenarios where individuals care about material payoffs. However, in many real-life scenarios individuals do not only have preferences that are rooted exclusively in the outcomes of the game. Rather, they are also often motivated by the beliefs and intentions of themselves and others. These types of belief-dependent motivations are captured by the framework of psychological game theory (Geanakoplos et al., 1989; Battigalli and Dufwenberg, 2009). While psychological games can introduce very interesting phenomena, they can also be noticeably hard to analyse, certainly compared to traditional games.

A well-established notion that is used to predict behaviour in traditional game theory is the basic concept of common belief in rationality. In traditional games, common belief in rationality is appealing for two reasons. Conceptually, it allows for a one-person perspective on a game, as opposed to Nash equilibrium. This means that when making a choice, a player forms beliefs in her mind about what her opponent will choose. She also forms beliefs about what her opponent believes she will choose. And so on. Based on such individual reasoning, the player reaches her decision. Practically, common belief in rationality is also an intuitive notion to use and straightforward to compute in traditional games due to its characterization in terms of iterated elimination of strictly dominated choices (IESDC). It thus becomes a natural and important question to ask to what extent the IESDC-procedure is able to characterize rational choices under common belief in rationality.

In this chapter, we will explicitly focus on this question. We will consider the question for a particular class of psychological games. As argued in Jagau and Perea (2018), most applications of psychological game theory are *expectation-based psychological games*. In such games, players in a game care only about higher-order expectations. These are sequences of probability distributions that summarize some, but

not all, aspects of a belief hierarchy. One nice property that carries over from traditional games to such expectation-based psychological games is the finite matrix representation of a psychological game. Because of this, such games behave very much like traditional games. Moreover, a finite matrix representation is essential in defining procedures such as IESDC.[1]

Despite the resemblance to traditional games, there are examples of expectation-based psychological games where the IESDC-procedure *fails* to characterize rational choices under common belief in rationality (see for instance Jagau and Perea (2018)). Here we will shed light on the matter of why the IESDC-procedure may fail to characterize common belief in rationality in certain scenarios and why it actually does give an exact characterization in others. We do so by exactly identifying those families of expectation-based psychological games where the IESDC-procedure does give a characterization of rational choices under common belief in rationality. By doing so, we not only identify those families of psychological games that are on a similar level of complexity in terms of reasoning as traditional games. We also point out what can make the other families of psychological games so difficult to reason about, both from the point of view of the player as well as that of the analyst. Our analysis in this chapter focuses on two-player expectation-based psychological games in a static environment without updating of beliefs.

In Theorem 5.1 we show that all rational choices under common belief in rationality must necessarily survive the IESDC-procedure. The other direction however does not need to hold. To briefly illustrate how the IESDC-procedure can fail to characterize rational choices under common belief in rationality, consider the introductory example of an expectation-based psychological game in Table 5.1. Here we have

[1] There are exceptions in psychological game theory that are not expectation-based psychological games. These include modelling preferences regarding anxiety (Caplin and Leahy, 2004) and suspense (Caplin and Leahy, 2001; Ely et al., 2015). To model such preferences, we need more information than just the higher-order expectations.

Table 5.1: *Introductory example*

Player 1's extreme second-order expectations

	(c,a)	(c,b)	(d,a)	(d,b)
a	0	0	0	0
b	1	0	1	1

Player 1's utilities

Player 2's extreme first-order expectations

	a	b
c	0	0
d	0	1

Player 2's utilities

two players: player 1 and player 2. Player 1 has alternatives a and b to choose from, whereas player 2 can choose between options c and d. Player 2's decision problem is as in a traditional game: she cares only about what player 1 will do. This is represented by the lower matrix. Player 1's utility however depends on her full second-order expectation. That is, her expectation about what player 2 is going to do is relevant for her decision, which is her first-order expectation. Additionally however, she cares about what player 2 expects player 1 (herself) to do. These two expectations, one of which is a higher-order expectation, form player 1's second-order expectation. If player 1 chooses a, she always receives a utility of 0. If on the other hand she chooses b, she receives a utility of 0 in case she expects player 2 to choose c while expecting player 2 to believe player 1 will choose b. In all remaining extreme cases of second-order expectations player 1 receives a utility of 1 when choosing b. Player 1's decision problem is depicted by the upper matrix. It is clear here that no choice for player 1 or player 2 is strictly dominated in the relevant decision problem. The IESDC-procedure would therefore not eliminate any choice for any player. However, choice a for player 1 can never be optimal under a belief hierarchy expressing common belief in rationality. Choice a is only optimal under the extreme second-order expectation (c,b), but choice c is never optimal for player 2 given that she expects player 1 to choose b.

The game in Table 5.1 is part of a particular family of games. Namely one in which one player's utility directly depends on her first-order beliefs and the utility of the other depends on her first-order and second-order beliefs. We identify the different families of expectation-based psychological games based on the orders of beliefs that are directly relevant in shaping the belief-dependent motivations of a decision-maker. We call these utility-relevant orders of beliefs or orders of belief in which the utility is variable. For instance, when modelling simple guilt, whatever a player believes about her opponent's choice, her first-order belief, is irrelevant. However, what the player believes about her opponent's first-order beliefs, which is part of her second-order belief, is important. The utility-relevant order of belief for modelling guilt would be the second order of belief. As another example, the game in Table 5.1 then belongs to the family of games where player 1's utility depends on her first and second orders of belief and player 2's utility is variable only in her first order of belief.

In this chapter we characterize those families of expectation-based psychological games where the IESDC-procedure always characterizes exactly the choices that can rationally be made under common belief in rationality. Take the perspective of player 1. The main theorem in the chapter (Theorem 5.2) establishes that the IESDC-procedure *always* characterizes rational choices under common belief in rationality for player 1 if and only if at least one of the following three conditions holds: (i) the utility of player 1 is variable in a single, even order; (ii) the utility of player 1 is variable in a single order of belief and the utility of player 2 is variable in a single order of belief as well; (iii) player 1's utility is only variable in odd orders and player 2's utility is variable in a single, even order of belief z, such that there is no pair x, y of utility-relevant orders for player 1 and no integer n with $x + n \cdot z = y$. The game in Table 5.1 does not belong to any family of games described here. An important observation can be made from this result. That is, if players care about material payoffs, cases $(i) - (iii)$ boil down to traditional games where expected utility only depends on first-order beliefs. In all other cases that involve material payoffs one has to go

beyond the IESDC-procedure to exactly characterize rational choices under common belief in rationality.

In order for a particular choice to be rational, restrictions then need to apply to the orders of beliefs that are utility-relevant. Under strategic reasoning, it makes sense to assume that the players to which these utility-relevant orders pertain play rationally as well. Similarly, these players as well may have belief-dependent motivations, which are rooted in *their* higher-order beliefs. Then, in order for the decision-maker to believe in the players' rationality at her utility-relevant orders, further restrictions need to be imposed on even higher orders of beliefs. And so on. In the end, we obtain a sequence of orders of beliefs that satisfy all aforementioned restrictions. For player 1 in the introductory example of Table 5.1 we can illustrate this via a diagram, as depicted in Figure 5.1. Player 1's utility is variable in her first-order and second-order expectations. These can be directly derived from her first-order and second-order beliefs respectively. For a particular choice C of player 1 to be optimal, restrictions thus need to be imposed on the first-order and second-order beliefs. This is why we have arrows from C to orders 1 and 2 in Figure 5.1. Player 2's utility is variable only in her first-order expectation. In order for player 1 to believe in player 2's rationality at her already restricted first-order belief, further restrictions need to be imposed on the second-order belief. This is why we have an arrow from order 1, which refers to a belief about player 2's choices, to order 2 in Figure 5.1. Order 2 refers to a belief about player 1's choices again. For player 1 not to question her own rationality at the second order of belief given the restrictions that have already been imposed on that order of belief, further restrictions are required on the third and fourth orders of belief. Hence the arrows from order 2 to orders 3 and 4. We can continue establishing such arrows indefinitely. Connected arrows together constitute a path in this diagram.

We refer to a diagram like in Figure 5.1 as a decision-maker's *causality diagram*. Under common belief in rationality, the causality diagram then captures those steps of reasoning of a decision-maker that are directly or indirectly relevant for rationalizing a particular choice. As is

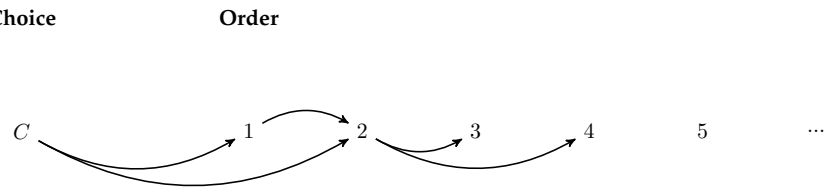


Figure 5.1: *Causality diagram of player* 1 *in Table 5.1*

for instance the case with order 2 in Figure 5.1, in a causality diagram the same order of belief may be reached by multiple paths. In the diagram there, order 2 is reached via the path $(0, 2, ..)$, but also via the path $(0, 1, 2, ...)$. If the same order of belief is reached by multiple paths, it implies that in order to rationalize a particular choice under common belief in rationality, not questioning rationality at different orders of beliefs will require different restrictions on the same higher-order belief. If these restrictions are contradictory, then the choice in question cannot be rational under a belief hierarchy expressing common belief in rationality. Exactly this friction is what the IESDC-procedure cannot pick up on. As a result, the IESDC-procedure may allow for choices that are not rational under common belief in rationality.

If the same order of belief is reached by multiple paths in a causality diagram of a player, we say that two paths in the diagram overlap. If a causality diagram is completely free of any overlapping paths, then also no contradictory restrictions on the same order of belief can occur. The main theorem of this chapter describes exactly these cases by means of families of expectation-based psychological games. Two of the three cases are trivial in the sense that the causality diagram only has a single path. Moreover, none of the cases include scenarios where both players care about materialized outcomes and at least one player has some belief-dependent motivation.

The remainder of this chapter is structured as follows. Section 5.2 discusses the concept of higher-order expectations, expectation-based

psychological games and families of expectation-based psychological games based on utility-variant orders. Section 5.3 discusses the IESDC-procedure and its problems in psychological games. Moreover, we state here the main result of the chapter: Theorem 5.2. In Section 5.4 we introduce the notion of causality diagram to visualize reasoning in expectation-based psychological games. Section 5.5 is fully dedicated to the proof of Theorem 5.2. Some parts of the proof are moved to the Appendix that accompanies this chapter. However, Section 5.5 illustrates these parts of the proof by means of examples. Finally, we end this chapter with some concluding remarks in Section 5.6.

5.2 Higher-order expectations

In CHAPTER 2 we defined static psychological games very generally. In the current chapter we will be looking at a particular broad subclass of psychological games: expectation-based psychological games. In this section we will define such games and provide a general discussion. Furthermore, we will discuss how we will distinguish between different families of psychological games. These families will be defined by the orders of belief that are directly relevant for players' utilities in a given expectation-based psychological game.

5.2.1 Expectation-based psychological games

When modelling emotions or other-regarding preferences, we typically do not use all information contained in higher-order beliefs in the utility functions (Jagau and Perea, 2018). Instead, we use *expectations* about beliefs, which can be derived from (higher-order) beliefs. For instance, in modelling surprise-related preferences we only care about a player's *expectation* about her co-players' first-order beliefs in determining her psychological payoff. To illustrate this, consider the following example.

Example 5.1 (Surprising student by means of exam)**.** *You are Ann's high school teacher in economics. You noticed that Ann's focus during classes has been lacking. Therefore, you wish to give her a wake-up call by surprising her (and the rest of the small class in the process). To this end, at the end of a given school week you vaguely announce to the class that next Monday an exam might be given. You can surprise Ann in two ways. You either give the exam on Monday while Ann does not expect one or you do not give the exam on Monday while she did expect you to give one. Surprising Ann gives you a feeling of psychological satisfaction. Either form of surprise is equally satisfactory to you. Even though you wish to surprise Ann, you also do not want her to fail.*

How do we model your utility as a teacher in the above example? Let us have $I = \{y, a\}$, with $y = you$ and $a = Ann$. For the choice-problem presented to us before Monday we moreover have $C_y = \{exam, no\ exam\}$ for your choice-set as the teacher and $C_a = \{study, not\ study\}$ as the choice set for Ann.

The utility you receive from your decision before Monday depends on two factors: the probability with which you believe Ann will study for the possible exam and the probability with which you expect Ann to believe you will actually give the exam. Let us say that any form of surprise gives you as a teacher one extra unit of utility, Ann failing the exam makes you lose a unit of utility and Ann succeeding makes you gain a unit of utility. If we assume your utility to be additive in these two different components, we can describe your expected utility of giving an exam on Monday by the following relation

$$u_y(exam, b_y) = (1 - \int_{C_a \times B_a} b_a^1(exam) db_y) + (2 \cdot b_y^1(study) - 1).$$

The expected utility function above depends on a summary statistic of the second-order belief induced by your belief hierarchy b_y. The second component in the expected utility function corresponds to your first-order belief that Ann will study. The integral measure on the other hand represents your *expectation* of Ann's first-order belief about your

Table 5.2: *Surprise exam game*

	Extreme Second-order expectations			
	$(study, exam)$	$(study, no\ exam)$	$(not\ study, exam)$	$(not\ study, no\ exam)$
Exam	1	2	-1	0
No exam	1	0	1	0

choice, induced by your belief hierarchy b_y. This summary statistic is called your *second-order expectation* in this setting. For instance, we can define Ann's *first-order expectation* that you will give an exam as $e_a^1[b_a](exam) := b_a^1(exam)$. Then your second-order expectation that Ann will study and that she believes you will give an exam is given by

$$\begin{aligned} e_y^2[b_y](study, exam) &:= \int_{\{study\}\times B_a} e_a^1[b_a](exam) db_y \\ &= \int_{\{study\}\times B_a} \int_{\{exam\}\times B_y} db_a db_y. \end{aligned}$$

Notice that the above expectation is a joint probability measure. In fact, every second-order expectation for you as a teacher is a joint probability measure $e_y^2 \in \Delta(C_a \times C_y)$. We can directly represent your (expected) utility of choosing to give an exam as a function of your second-order expectation that is induced by your belief hierarchy b_y: $u_y(exam, e_y^2[b_y])$. Similarly, the expected utility of not giving an exam can be represented by $u_y(no\ exam, e_y^2[b_y]) = \int_{C_a\times B_a} e_a^1[b_a](exam) db_y$ (if you do not give an exam, you will not see Ann succeed or fail). Since both C_a and C_y are finite sets of choices, we have that $C_a \times C_y$ is finite as well. The set of second-order expectations then has finitely many extreme points. Consequently, the resulting utility for you as a teacher for all possible extreme second-order expectations can be represented in finite-matrix form as in Table 5.2. Following Jagau and Perea (2018), we can define any *higher-order expectation* recursively. To define any particular class of psychological games where utilities depend on higher-order expectations, this is a useful tool.

In a general psychological game G, let us start with the *first-order expectation* $e_i^1[b_i]$ for a player i. This is simply the first-order belief of player i:

$$e_i^1[b_i] := b_i^1 \in \Delta(C_j).$$

The *second-order expectation* of player i that player j will choose c_j while believing that player i will choose c_i is subsequently defined as follows:

$$e_i^2[b_i](c_j, c_i) := \int_{\{c_j\} \times B_j} e_j^1[b_j](c_i) db_i = \int_{\{c_j\} \times B_j} \int_{\{c_i\} \times B_i} db_j db_i.$$

As noted in our surprise exam example (Example 5.1), a second-order expectation is a joint probability measure $e_i^2[b_i] \in \Delta(C_j \times C_i)$. We can recursively define any higher-order expectation following this construction. Note however that second-order expectations are defined over the Cartesian product of two choice sets. As the orders of higher-order expectations increase, this Cartesian product will contain more and more elements as well. For the sake of clarity in our notation, we will therefore define the following sets recursively:

$$W_i^1 = C_j \text{ and } W_i^k = \begin{cases} \underbrace{C_j \times C_i \times ... \times C_i}_{k \text{ times}}, \ k \text{ is even} \\ \underbrace{C_j \times C_i \times ... \times C_j}_{k \text{ times}}, \ k \text{ is odd} \end{cases} \quad \text{for all } k > 1.$$

Each $w_i^k \in W_i^k$ thus has k components and represents a combination of your opponent's and your own choices. Throughout this chapter, we will utilise the following identification of $w_i^k \in W_i^k$ at times as well: $w_i^k = (c_j, w_j^{k-1})$. A k-th order expectation is then defined as a probability measure over W_i^k.

Definition 5.1 (Jagau and Perea (2018))**.** *Consider a two-player psychological game G with a player i and a player j and let b_i be the belief hierarchy for player i. Let $e_i^1[b_i] := b_i^1$ be the first-order expectation for player i given*

b_i. *For* $k \geq 2$, *the* ***k-th order expectation*** $e_i^k[b_i] \in \Delta(W_i^k)$ *of player* i *given belief hierarchy* b_i *is defined as*

$$e_i^k[b_i](w_i^k) := \int_{\{c_j\} \times B_j} e_j^{k-1}[b_j](w_j^{k-1}) db_i, \text{ where } w_i^k = (c_j, w_j^{k-1}).$$

In the integral b_i serves as a probability measure over $C_j \times B_j$, similar to how it was used in Section 2.1.

We capture all possible k-th order expectations in the set $E_i^k := \Delta(W_i^k)$. This allows us to define utilities that depend explicitly on k-th order expectations by $u_i : C_i \times E_i^k \to \mathbb{R}$. Notice here that $e_i^k[b_i]$ for any $k > 1$ given a belief hierarchy b_i also contains the lower-order expectations induced by said belief hierarchy. That is, we have that $marg_{W_i^{k-1}} e_i^k[b_i] = e_i^{k-1}[b_i]$. Much like beliefs, we thus obtain a *hierarchy of expectations* $e_i[b_i] := (e_i^1[b_i], e_i^2[b_i], ...)$ induced by a belief hierarchy b_i.

Two points are worthwhile to elaborate on here leading to the upcoming Definition 5.2. First, the mapping from the set of belief hierarchies to the set E_i^k is surjective but non-injective. For every k-th order expectation, there is a belief hierarchy that induces it. However, a given k-th

Table 5.3: *Illustration of belief hierarchies and second-order expectations, Example 5.1*

Your second-order beliefs	b_y^2	=	$(study, b_a^1)$
	$b_y^{2\prime}$	=	$\frac{1}{2}(study, b_a^{1\prime}) + \frac{1}{2}(study, \hat{b}_a^1)$
Ann's first-order beliefs	b_a^1	=	$\frac{1}{2}(exam) + \frac{1}{2}(no\ exam)$
	$b_a^{1\prime}$	=	$exam$
	$\hat{b}_a^1$	=	$no\ exam$
Your second-order expectations	$e_i^2[b_y]$	=	$\frac{1}{2}(study, exam) + \frac{1}{2}(study, no\ exam)$
	$e_i^2[b_y\prime]$	=	$\frac{1}{2}(study, exam) + \frac{1}{2}(study, no\ exam)$

order expectation may be induced by multiple belief hierarchies. This is illustrated in Table 5.3, where we depict the second-order beliefs of two possible belief hierarchies for you as the teacher. The second-order expectations induced by the two belief hierarchies are equal, whereas the second-order beliefs are not. Indeed, in b_y^2 you are certain about Ann's belief and believe that Ann is uncertain about your choices. In $b_y^{2\prime}$ you are uncertain about Ann's belief but believe that Ann is certain about your choice.

Second, recall again from Section 2.1 that any belief hierarchy b_i can be represented by a probability distribution in $\Delta(C_j \times B_j)$. Take any two belief hierarchies $b_i, b_i' \in B_i$. The **convex combination** $\lambda b_i + (1-\lambda)b_i'$ for any $\lambda \in [0,1]$ is then the belief hierarchy that puts probability $\lambda f_i(b_i)(E) + (1-\lambda) f_i(b_i')(E)$ to every measurable $E \subseteq C_j \times B_j$. With the previous two points in mind, we are now in a position to formally define a psychological game where expectations instead of beliefs matter explicitly for utilities.

Definition 5.2 (Jagau and Perea, 2018)**.** *We call a two-player psychological game $G = (C_i, B_i, u_i)_{i \in I}$ an **expectation-based psychological game** if, for both players i and all choices $c_i \in C_i$,*

(i) $e_i[b_i] = e_i[b_i']$ implies $u_i(c_i, b_i) = u_i(c_i, b_i')$

(ii) utility is linear in the beliefs hierarchies: $u_i(c_i, \lambda b_i + (1-\lambda)b_i') = \lambda u_i(c_i, b_i) + (1-\lambda) u_i(c_i, b_i')$, for all $\lambda \in [0,1]$.

The second condition is that of **belief linearity**. This condition states that the expected utility given c_i and the convex combination of two belief hierarchies b_i and b_i' has to be equal to the convex combination of the two expected utilities induced by the choice $c_i \in C_i$ and by the belief hierarchies b_i and b_i'. Finally, we can impose a last, natural condition.

Definition 5.3. *A psychological game $G = (C_i, B_i, u_i)_{i \in I}$ is **belief-finite** if there is some $n \geq 1$ such that for every choice $c_i \in C_i$, and every two belief hierarchies $b_i, \hat{b}_i \in B_i$ with $b_i^n = \hat{b}_i^n$ we have that $u_i(c_i, b_i) = u_i(c_i, \hat{b}_i)$.*

Table 5.4: *Surprise exam game with a mean teacher*

	Your extreme second-order expectations			
	$(study, exam)$	$(study, no\ exam)$	$(not\ study, exam)$	$(not\ study, no\ exam)$
Exam	0	0	0	1
No exam	1	0	0	0

Teacher's utilities

In words, belief-finiteness means that utility depends only on finite orders of beliefs. Belief-finiteness allows us to have a finite representation of an expectation-based psychological game in matrix form. This is because there are finitely many extreme higher-order expectations under belief-finiteness for a player i to consider. These extreme higher-order expectations are represented in the columns of the matrix (see for instance Table 5.2). The choices, as traditionally is the case, are found in the rows. For the remainder of the chapter, we assume every expectation-based psychological game we will be dealing with is belief-finite.

Finally, note that the example in Table 5.2 assumes that your utility as a teacher by giving the exam is additively separable in wanting to surprise Ann and wanting her to pass the exam. However, by definition of an expectation-based, belief-finite psychological game, utility does not always have to be additive in the different higher-order expectations. To this end, reconsider Example 5.1, but now assume you are a mean teacher. That is, you wish to surprise Ann by giving the exam if she does not study, and you wish to surprise Ann by not giving an exam if she does study. Any other scenario does not interest you. This non-additively separable psychological game is illustrated in Table 5.4.

5.2.2 Order-variable families of psychological games

The problem that a player in any belief-finite, expectation-based psychological game faces can be thought of as a decision problem. Gener-

ally, a *decision problem* can be defined by a triple $D = (C, X, v)$. In this triple, C refers to a finite set of choices, X is a finite set of states and $v : C \times X \to \mathbb{R}$ is a Bernoulli utility function. A choice $c \in C$ is then *optimal* in D if there is a belief $b \in \Delta(X)$ such that

$$\sum_{x \in X} b(x) \cdot v(c, x) \geq \sum_{x \in X} b(x) \cdot v(c', x), \forall c' \in C.$$

In a belief-finite, expectation-based psychological game where utilities only depend up to order k the set of states X would then refer to W_i^k for $k \geq 1$. The utility $v_i(c_i, w_i^k)$ then refers to the utility experienced from choosing c_i while being in state w_i^k.

We can define families of expectation-based psychological games depending on which orders of beliefs are of direct relevance to a player's preferences.

Definition 5.4. *Let $i \in \{1, 2\}$. Take a belief finite, expectation-based psychological game G, where player i's utility function can be summarized by $v_i : C_i \times W_i^n \to \mathbb{R}$. If $v_i(c_i, w_i^n) \neq v_i(c_i, \hat{w}_i^n)$ for some $c_i \in C_i$ and some w_i^n and $\hat{w}_i^n$ that only differ in the m-th order, we say v_i is **variable in the m-th order**.*
*By $\mathcal{G}(N_1, N_2)$ we denote the **family of psychological games** in which player 1's and player 2's utility-variable orders are specified by N_1 and N_2 respectively, with $N_1, N_2 \subseteq \mathbb{N}$.*

When we refer to '(directly) utility-relevant' orders of belief for a player i in the remainder of the chapter, we always mean the orders of belief in which player i's utility is variable.

5.3 Iterated elimination of strictly dominated choices

In this section we will discuss the procedure of iterated elimination of strictly dominated choices in psychological games. Unlike traditional games, this procedure does not always characterize the rational

choices a player can make under common belief in rationality. However, there are some families of games for which this relationship does hold. This leads us to state the main result of this chapter, captured in Theorem 5.2. In the second part of this section we provide intuition on why the elimination procedure may fail in its characterization of rational choices under common belief in rationality.

5.3.1 Iterated elimination of strictly dominated choices in psychological games

In traditional games, choices that are rational under some belief hierarchy expressing common belief in rationality are characterized by iteratively eliminating strictly dominated choices. We say a choice $c \in C$ is *strictly dominated* in a decision problem $D = (C, X, v)$ if there is a randomized choice $r \in \Delta(C)$ such that

$$v(c, x) < \sum_{c' \in C} r(c') \cdot v(c', x), \forall x \in X.$$

Under the set-up presented above, iterative elimination of strictly dominated choices then means that each round of eliminating choices induces a new decision problem for a decision-maker. In each round, those choices are eliminated that are never optimal in the given decision problem. We can use a result by Pearce (1984) for this.

Lemma 5.1 (Pearce's Lemma)**.** *Consider a decision problem $D = (C, X, v)$. Then, $c \in C$ is optimal in D if and only if c is not strictly dominated in D.*

Pearce's original lemma defines the set of states X as the set of choices C_j of the opponent j of a player in a traditional, two-player game. But his proof technique can be used to prove Lemma 5.1 as well. In a belief-finite, expectation-based psychological game where utilities only depend up to order n the set of states X would refer to W_i^n for $n \geq 1$. The utility $v_i(c_i, w_i^n)$ refers to the utility experienced from choosing

c_i while being in state w_i^n. The procedure of iterative elimination of strictly dominated choices (IESDC) is then defined as follows.

Procedure 5.1 (Iterated elimination of strictly dominated choices (IESDC))**.**

Consider a two-player, psychological game $G = (C_i, B_i, u_i)_{i \in I}$ which is expectation-based and belief-finite, and in which utilities depend up to the n-th order expectation. For every player i, consider the full decision problem $(C_i^0, W_i^{n,0}, v_i)$, where $C_i^0 := C_i$, $W_i^{n,0} := W_i^n$ and $v_i : C_i \times W_i^n \to \mathbb{R}$ summarizes the utility function u_i.

Step 1
For each player i, define: $C_i^1 = \{c_i \in C_i |\ c_i$ is not strictly dominated in $(C_i^0, W_i^{n,0}, v_i)\}$.

For each player i, define: $W_i^{n,1} = \begin{cases} C_j^1 \times C_i^1 \times ... \times C_j^1 \text{ if n is odd.} \\ C_j^1 \times C_i^1 \times ... \times C_i^1 \text{ if n is even.} \end{cases}$

***Step* $\mathbf{k \geq 2}$**
For each player i, define: $C_i^k = \{c_i \in C_i^{k-1} |\ c_i$ is not strictly dominated in $(C_i^{k-1}, W_i^{n,k-1}, v_i)\}$.

For each player i, define: $W_i^{n,k} = \begin{cases} C_j^k \times C_i^k \times ... \times C_j^k \text{ if n is odd.} \\ C_j^k \times C_i^k \times ... \times C_i^k \text{ if n is even.} \end{cases}$

For each player i, define $C_i^\infty = \bigcap_{k \geq 1} C_i^k$.

Some explanation is due here. In this procedure we assume that player i only cares for higher-order expectations up to order n (this may or may not include order n). In Step 1, both players i eliminate those choices that are strictly dominated in their respective decision problems. Subsequently we define the resulting sets of combination of choices $W_i^{n,1}$ for each i which are constructed from those sets of choices

that are not strictly dominated in the original decision problems. Using C_i^1 and $W_i^{n,1}$, we then construct a *reduced decision problem* $(C_i^1, W_i^{n,1}, v_i)$, where $v_i : C_i^1 \times W_i^{n,1} \to \mathbb{R}$. Note that for the identification of the utility function v_i we technically abuse notation in this elimination step. Formally we have $v_i : C_i \times W_i^n \to \mathbb{R}$, but we identify it with a restriction on $C_i^1 \times W_i^{n,1}$. After constructing the reduced decision problem, we repeat the process. The procedure ends when no choices can be eliminated any longer for any of the two players.

This procedure leads us to consider the following theorem for this section.

Theorem 5.1 (Rational choice under common belief in rationality requires surviving the procedure)**.** *Consider any belief-finite, expectation-based psychological game G, with two players where utilities depend on up to n-th order expectations. Then every choice that is rational under common belief in rationality must necessarily survive IESDC.*

Proof. Define by R_i^∞ the set of rational choices player i can make under common belief in rationality. We will prove that $R_i^\infty \subseteq C_i^\infty$. We will do this by showing that $R_i^\infty \subseteq C_i^k$ for every $k \geq 1$. This will be done by induction on k. We will start with the case of $k = 1$.

Induction start Take an arbitrary choice $c_i \in R_i^\infty$. This means that c_i is optimal for some belief hierarchy $b_i \in B_i$ that expresses common belief in rationality. By Pearce's Lemma this implies $c_i \in C_i^1$ as well, as c_i being optimal for some belief hierarchy $b_i \in B_i$ among all choices in C_i is exactly the same as c_i not being strictly dominated in the decision problem $(C_i^0, W_i^{n,0}, v_i)$.

Induction step Assume that $R_i^\infty \subseteq C_i^{k-1}$ for some $k \geq 2$ for each player i. Now take some $c_i \in R_i^\infty$. Then we have that c_i is optimal for some belief hierarchy $b_i \in \Delta(C_j \times B_j)$ that expresses common belief in rationality, among all choices in C_i. Common belief in rationality implies that player i believes that her opponent player j makes an optimal choice according to a belief hierarchy that expresses common belief in

rationality. Hence each $(c_j, b_j) \in supp(b_i)$ is such that c_j is optimal for b_j which expresses common belief in rationality. Thus, $c_j \in R_j^\infty$ by definition and by our induction assumption therefore $c_j \in C_j^{k-1}$.

However, if $b_j \in \Delta(C_i \times B_i)$ expresses common belief in rationality, this implies that player i believes that player j believes player i expresses common belief in rationality and makes an optimal choice accordingly. Hence, each $(c_i', b_i') \in supp(b_j)$ is such that c_i' is optimal for b_i' which expresses common belief in rationality. Therefore, $c_i' \in R_i^\infty$ and by the induction assumption, $c_i' \in C_i^{k-1}$. As a result, the choice c_i we started with must be optimal for a belief hierarchy $b_i \in \Delta(C_j^{k-1} \times \Delta(C_i^{k-1} \times B_i))$.

If we continue this line of reasoning that each player i believes rationality is commonly believed, we get that the choice $c_i \in R_i^\infty$ we started with in this induction step must be optimal for a belief hierarchy $b_i \in \Delta(C_j^{k-1} \times \Delta(C_i^{k-1} \times \Delta(C_j^{k-1} \times ...)))$. If we take the n-th order expectation induced by this belief hierarchy, we get

$$e_i^n[b_i] \in \Delta(\underbrace{C_j^{k-1} \times C_i^{k-1} \times ... \times C_j^{k-1}}_{\text{n times}})$$

if n is odd and

$$e_i^n[b_i] \in \Delta(\underbrace{C_j^{k-1} \times C_i^{k-1} \times ... \times C_i^{k-1}}_{\text{n times}})$$

if n is even. Thus choice c_i is optimal for some n-th order expectation $e_i^n[b_i] \in \Delta(W_i^{n,k-1})$ induced by belief hierarchy b_i. By Pearce's Lemma then c_i is not strictly dominated in the decision problem $(C_i, W_i^{n,k-1}, v_i)$. It follows that c_i is then also not strictly dominated in the decision problem $(C_i^{k-1}, W_i^{n,k-1}, v_i)$. Hence, $c_i \in C_i^k$ and since we took an arbitrary $c_i \in R_i^\infty$, we also have that $R_i^\infty \subseteq C_i^k$.

By induction on k, we have that $R_i^\infty \subseteq C_i^k$ for every $k \geq 1$, which

completes the proof.

□

We have thus shown that if a choice is rational under common belief in rationality it must survive the IESDC-procedure. This result holds given a game $G \in \mathcal{G}(N_1, N_2)$ for *any* family of games $\mathcal{G}(N_1, N_2)$, $N_1, N_2 \subseteq \mathbb{N}$. The reverse statement does not always hold. It is not true that those choices that survive the IESDC-procedure are always rational under common belief in rationality (Jagau and Perea, 2018). The example of Table 5.1 illustrates this. There are however specific families of games where, for *all* games in such a family, the IESDC-procedure does exactly characterize rational choices under common belief in rationality.

Theorem 5.2. *Consider any family $\mathcal{G}(N_1, N_2)$ of belief-finite, expectation-based psychological games with two players. For every game in $\mathcal{G}(N_1, N_2)$, each choice for player 1 that survives the IESDC-procedure is also a rational choice under common belief in rationality, if and only if, one of the following conditions is true:*

(i) Player 1's utility and player 2's utility are both variable in a single order of belief;

(ii) Player 1's utility is variable in a single even order of belief;

(iii) Player 1's utility is variable only in odd orders of belief and player 2's utility is variable in a single even order of belief z which is such that there is no pair x, y of player 1's utility-variant orders and no $n \in \mathbb{N}$ with $x + n \cdot z = y$.

An important observation we can make here is that if both players care (among others) about material payoffs, the conditions listed in the theorem above reduce to those that specify a traditional game.

In the remainder of this chapter, Theorem 5.2 will take center-stage.

5.3.2 Illustrating the problem of the IESDC-procedure

Recall the example in Table 5.1 from the Introduction of this chapter. There is an underlying reason why in this example in Table 5.1 the IESDC procedure does not give us exactly the choices one can rationally make under common belief in rationality. In short terms, there is an overlap between the orders in which one's utility is variable on the one hand and the orders of beliefs one needs to consider for expressing 1-fold belief in rationality on the other hand. This is also illustrated in the causality diagram for player 1 in Figure 5.1. We will formalize such diagrams in Section 5.4. In words, we can say the following however. In the diagram in Figure 5.1, the arrow from player 1's choice *C* to order "1" indicates that the optimality of player 1's choice depends on her first-order expectation. The same applies for the arrow from her choice to order "2": the optimality of player 1's choice also depends on her second-order expectation. Player 2's utility only depends on her first-order expectation. This is represented by the arrow from order "1" to order "2". And so on. Clearly then, the rationality of player 1's own choice and believing that player 2 makes a rational choice both directly depend on player 1's second-order expectation. Thus for *c* to be optimal and for player 1 to believe in player 2's rationality, different restrictions need to be imposed on the second-order expectation. In this example these restrictions happen to be in conflict as also explained in the Introduction: in order for choice *a* to be optimal player 1 in her first-order expectation has to believe that player 2 will choose *c* while expecting in her second-order expectation that player 2 expects her to choose *b*. However, in order for choice *c* to be optimal for player 2, she must expect player 1 to choose *a*.

Believing in an opponent's rationality is believing in the event that your opponent makes an optimal choice *given* her belief. Belief in the opponent's rationality thus restricts the combinations of choices and belief hierarchies (c_j, b_j) you can consider for the opponent where c_j is optimal specifically for b_j. The first step of the IESDC-procedure however only eliminates choices for your opponent which are never op-

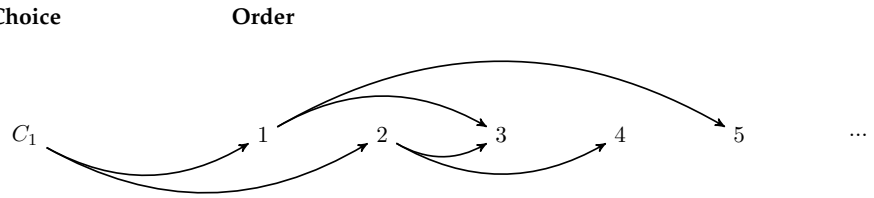


Figure 5.2: *Causality diagram of player* 1 *in game in* $\mathcal{G}(\{1,2\},\{2,4\})$

timal, given *any* belief hierarchy. The second step subsequently only eliminates choices that are never optimal given beliefs that only assign positive probability to choices that are not eliminated in Step 1.

This kind of procedure does not sufficiently restrict the second-order expectations one can consider under belief in the opponent's rationality. One may assign positive probability to an extreme second-order expectation whose two components entail choices for an opponent and oneself that did survive Step 1 of the procedure. However, as illustrated via the example in Table 5.1, we then allow for scenarios in which the player 2's choice in the first component can never be optimal given the conditional probability assigned to player 1's choice that features in the second component. In a traditional game there is no such issue at all, as player 1's utility would only depend on her first-order belief. The rationality of her first-order belief would only depend on her second-order belief. And so on. In this case each step k of eliminating strictly dominated choices corresponds exactly to the reasoning step of expressing up to $(k-1)$-fold belief in rationality. This corresponds also with the observation that traditional games are a special sub-case of case (i).

In the game in Table 5.1 we have that the utility for player 1 is dependent on her first order and second order of belief, which overlap with the orders of belief that matter for expressing 1-fold belief in rationality. That is, 1-fold belief in rationality is determined by a first-order belief of player 1, rationalized by her second-order belief. In general,

this kind of overlap may also occur because of overlap between deeper levels of reasoning. Consider for instance a causality diagram as in Figure 5.2 for player 1. Here player 1's utility depends on her first-order and second-order expectations, whereas player 2's utility depends on her second-order and fourth-order expectations. Here we see that expressing 1-fold belief in rationality and expressing 2-fold belief in rationality both require restrictions on player 1's third-order expectation, which may be in conflict.

First, there is the event of expressing 1-fold belief in rationality, as the first-order expectation is a conjecture about player 2's choice, which is motivated by player 2's second-order and fourth-order expectations. Hence, a rational first-order belief is explained by player 1's third-order expectation and fifth-order expectation[2]. Second, there is the event of expressing 2-fold belief in rationality, which next to the fourth-order expectation, also depends on the third-order expectation. Namely, the second-order belief is a conjecture about player 1's own choice. The utility of player 1 depends on her first-order and second-order expectation. So in order to rationalize the choices in her second-order expectation, player 1 should also consult her third-order and fourth-order beliefs. Hence, there is an overlap in the causality diagram at order 3: one-fold belief in rationality and two-fold belief in rationality both impose restrictions on the third-order belief, and these restrictions may be in conflict.

Memory is a keyword here. The memory of the IESDC-procedure at a particular step only consists of rational choices that survived the procedure up until that step, but not the (higher-order) beliefs used to get there. The perspective of the IESDC-procedure as a reasoning procedure, as put forward by for instance Cubitt and Sugden (2011) and Perea (2015), can be of use here. Take a traditional game. In Step 1 of

[2]We slightly abuse the use of "k-th order expectation" here. Technically, a third-order expectation can be derived from the fifth-order expectation by taking the relevant marginal distribution. With third-order expectation in this context we specifically refer to $marg_{C_2} e_1^3 \in \Delta(C_2)$ where $e_1^3 \in \Delta(W_1^2 \times C_2)$ and by the fifth-order expectation we mean $marg_{C_2} e_1^5 \in \Delta(C_2)$ where $e_1^5 \in \Delta(W_1^4 \times C_2)$.

the procedure you as a player want to check if a choice can be rationalized by a first-order belief. If this is the case, it means it is not strictly dominated. Next, you acknowledge that your opponent has the same thought process. In Step 2 you therefore check if you can rationalize a choice by a first-order belief that takes into account that the opponent also has tried to rationalize her choices. Combining Step 2 for you and Step 1 for the opponent would then lead to a second-order belief, and you expressing 1-fold belief in rationality given a belief hierarchy with such a second-order belief. And so on. Once the procedure stops, you can implicitly form for each surviving choice a belief hierarchy expressing common belief in rationality that rationalizes this choice. In each step of the procedure, we did not need to remember the beliefs used to support a choice in determining whether that choice could be rationalized in the previous step. A similar logic applies to all scenarios covered by Theorem 5.2. In scenarios that are not part of the cases listed in Theorem 5.2, this does not apply. Reasoning about different rationality events, such as 1-fold belief in rationality and 2-fold belief in rationality as in Figure 5.2, can overlap with each other. Concretely for a game associated with a diagram as in Figure 5.2 we have the following: you should recall the higher-order belief you used to support the rationality of the opponent's choice in your first-order belief when reasoning about 2-fold belief in rationality. Note that this is equivalent to saying that you should recall the higher-order belief used to determine whether your opponent's choice in your first-order belief is strictly dominated or not in her full decision-problem. Only by recalling this higher-order belief can you then determine if a particular choice of yours can be rationalized by a belief hierarchy expressing 1-fold *and* 2-fold belief in rationality at the same time. So when overlap in reasoning occurs, we need an elimination procedure that at any step recalls the higher-order beliefs used in previous steps. See Jagau and Perea (2017) for such a procedure.

Theorem 5.2 indicates the cases that describe in which families expectation-based psychological that IESDC-procedure gives an exact characterization of the rational choices one can make under common belief in

rationality. We will formally show this in Section 5.5. We will do so by making use of causality diagrams.

5.4 Causality diagrams

Causality diagrams prove to be a useful analytical tool to think about the problem of overlap in reasoning which appears in psychological games. In order to formally capture the notion of a causality diagram, a discussion on elementary graph theory is in place. A *graph* Σ is a nonempty set of *vertices* V and a (possibly empty) set of *edges* E. In Figure 5.2 the vertices correspond to the choice and the orders of belief, whereas the edges are the arrows between the vertices which indicate in which orders a particular utility is variable. In a *directed graph*, the direction of the edge, also known as an *arc* in a directed graph, matters. In that case we speak of outgoing *arcs* if an arc leaves a vertex and ingoing arcs if an arc goes into a vertex. The amount of outgoing arcs is known as the outdegree, whereas the number of ingoing arcs is the indegree. All the vertices that some vertex x is joined with directly by an outgoing arc is known as the out-neighborhood, whereas all adjacent vertices that x is joined with via ingoing arcs is known as the in-neighborhood. Finally, there is the concept of a path. A *path* in a directed graph is a sequence of vertices that starts at the root, where the k-th element is joined with the $(k-1)$-th element by an ingoing arc. A vertex r is called a *root* if (a) that vertex has only outgoing arcs and (b) for all vertices in the graph that have ingoing arcs, there is a path from r to that vertex. Whenever we refer to a path in the remainder of this chapter, we specifically mean a path that starts at the root. Moreover, we say we have a *divergence point* between two paths p^1 and p^2 in a rooted graph if $p^1 = (p, a, b, ...)$ and $p^2 = (p, a, c, ...)$ with p being a subpath of both p^1 and p^2 and $b \neq c$. The divergence point is then located at vertex a. The root itself can also function as the divergence point.

We can now formally define the concept of a causality diagram.

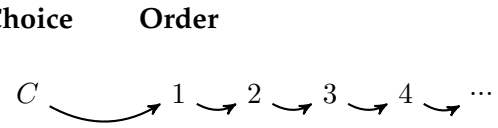


Figure 5.3: *Causality diagram of a player in a traditional game*

Definition 5.5. *Consider a game $G \in \mathcal{G}(N_1, N_2)$. The **causality diagram** $D_1(N_1, N_2)$ for player 1 in the game G is a rooted, directed graph $(\mathbb{N} \cup \{0\}, \mathcal{E})$ with the root being 0. The set of arcs $\mathcal{E}$ is as follows:*

- *For the **root** $r = 0$, establish an arc $(0, a_1)$ for every $a_1 \in N_1$;*
- *Inductively for every $k \geq 2$ do the following:*
  - *For every **even** a_{k-1}, establish an arc (a_{k-1}, a_k) with $a_k = a_{k-1} + b$ for every $b \in N_1$;*
  - *For every **odd** a_{k-1}, establish an arc (a_{k-1}, a_k) with $a_k = a_{k-1} + c$ for every $c \in N_2$.*

It is important to note that each player in a game has her own causality diagram, as players utilities may be variable in different orders. The paths in a causality diagram have a natural interpretation. Each path represents a chain of restrictions. For instance, in Figure 5.2 in order to ensure that some choice c_1 is optimal, the first-order and second-order expectation need to be restricted. If in addition player 1 wants to express 1-fold belief in rationality, restrictions on expectations of an even higher order are necessary. That is, given what player 1 expects player 2 to do, each choice in *that* conjecture can only be made optimal given its own appropriate restrictions on player 1's third-order and fifth-order expectations. And so on. In a traditional game, the causality diagram would look as in Figure 5.3. It is clear that in a traditional game there is only a single path on the causality diagram for each player. Compare this to the causality diagram in Figure 5.2. There, up to three orders, we can already distinguish between four

different paths: $(0, 1, 3, ...)$, $(0, 1, 5, ...)$, $(0, 2, 3, ...)$ and $(0, 2, 4, ...)$. The paths $(0, 1, 3, ...)$ and $(0, 2, 3, ...)$ also clearly display a common vertex after having diverged: vertex order 3. It is this overlap that can cause problems for the characterization of rational choices under common belief in rationality by IESDC.

We say that two paths in a directed graph are *pairwise vertex-disjoint* starting at a particular vertex a if they do not have any vertices in common after this vertex a. This leads us to define the following concept.

Definition 5.6. *A causality diagram is overlap-free if all pairs of paths are pairwise vertex-disjoint after the respective divergence point.*

If we take the interpretation that a path represents a chain of utility-relevant restrictions, then if two paths have a vertex b in common after a divergence point, it follows that two paths lead to two different restrictions on the same set of higher-order beliefs. Of course, these restrictions may clash. If paths are vertex-disjoint however, then the set of higher-order expectations for a particular order is never restricted from multiple angles.

The combinations of variable orders N_1 and N_2 that induce an overlap-free causality diagram are actually identifiable. These correspond to the three cases listed in Theorem 5.2.

Lemma 5.2. *The causality diagram $D_1(N_1, N_2)$ for player 1 in a game $G \in \mathcal{G}(N_1, N_2)$ is overlap-free if and only if one of the following is the case:*

(i) $N_1 = \{x\}$ *and* $N_2 = \{y\}$*;*

(ii) $N_1 = \{x\}$ *with* x *even;*

(iii) N_1 *only consists of odd orders and* $N_2 = \{z\}$ *with* z *even such that there is no pair* $x, y \in N_1$ *and no* $n \in \mathbb{N}$ *where* $x + n \cdot z = y$*.*

Proof. We start off by proving the "if"-direction. We do so for each of the three cases separately. Canonical causality diagrams that represent each of the cases are depicted in Figure 5.4.
$\Leftarrow$:
(i) If $N_1 = \{x\}$ and $N_2 = \{y\}$, then the cardinality of both sets of orders is one. This implies that in the resulting causality diagram every vertex has an outdegree of at most one. Then it follows there is also a unique path in the causality diagram. By definition the causality diagram is then overlap-free, as there is no second path present to have overlap with.

(ii) Clearly, there is a unique path in player 1's causality diagram, containing only even numbers. Hence vertex-disjointness is guaranteed in this case and therefore the causality diagram is overlap-free. This is illustrated in Figure 5.4b.

(iii) The root may be at the start of multiple paths, as $|N_1| \geq 1$. As the out-neighborhood of the root is determined by N_1, the root is only connected to $x \in N_1$, each of which is odd. Each odd vertex's out-neighborhood is determined by $N_2 = \{z\}$. Thus each odd vertex is connected to a different odd vertex, as z is even. Then take two paths in player 1's causality diagram: $(0, x, x+z, ..., x+g\cdot z, ...)$ and $(0, x', x'+z, ..., x'+h\cdot z, ...)$ with $x, x' \in N_1$ and $x \neq x'$. These two paths must be vertex-disjoint. First, note that the divergence point of the two paths is the root. Now assume they do have a vertex in common: let $x + g \cdot z = x' + h \cdot z$. Then $x - x' = (h - g) \cdot z$. Assume without loss of generality that $h > g$. However, this violates the condition that there is no pair $x, x' \in N_1$ such that $x' + n \cdot z = x$ with $n = h - g$. Hence it must be the case that all paths in player 1's causality diagram are pairwise vertex-disjoint and therefore the causality diagram is overlap-free.

For the"only-if" direction, we show that if conditions (i), (ii) and (iii) do not hold, then the causality diagram of player 1 has an overlap. In total there are six scenarios under which none of the three cases listed

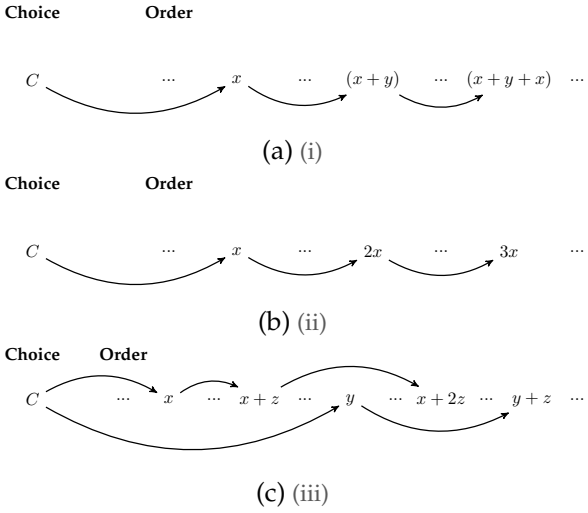


Figure 5.4: *Canonical causality diagram for cases (i), (ii) and (iii) in Lemma 5.2*

in Lemma 5.2 apply: (i) N_1 contains two even orders; (ii) N_1 contains an even and an odd order, whereas N_2 contains an even order; (iii) N_1 contains an even and an odd order, whereas N_2 contains an odd order; (iv) N_1 contains an odd x and an odd y and N_2 contains an even z such that $x = y + n \cdot z$ for some $n \in \mathbb{N}$; (v) N_1 contains two odd orders and N_2 contains also an odd order; and (vi) N_1 contains an odd order and N_2 contains two arbitrary orders. We go by all these cases one-by-one.

$\Rightarrow$:
(i) Consider $\{x, y\} \subseteq N_1$ with both x and y even. Then there exist the following two paths in the causality diagram $D_1(N_1, N_2)$: $(0, x, 2x, ...)$ and $(0, y, 2y, ...)$. These paths share a common vertex in $x \cdot y$. Hence the two paths are not vertex-disjoint after their divergence-point and thus the causality diagram $D_1(N_1, N_2)$ for player 1 is not overlap-free.
(ii) Consider $\{x, y\} \subseteq N_1$ with x even and y odd and an even $z \in N_2$. We have a path $(0, y, y + z, y + 2 \cdot z, ..., y + x \cdot z, ...)$. There is also the path $(0, x, 2x, ..., z \cdot x, z \cdot x + y, ...)$. These two paths have vertex $z \cdot x + y$ in common after the point of divergence, being the root. Hence the

causality diagram is not overlap-free.
(iii) Consider $\{x, y\} \subseteq N_1$ with x even and y odd and an odd $z \in N_2$. We have a path $(0, y, y+z, y+(y+z), (y+z)+(y+z), ..., x(y+z), ...)$. There exists also the path $(0, x, 2x, ..., (y+z)x, ...)$. These two paths share a vertex in $x(y+z)$ after the divergence point, being the root. Hence the causality diagram is not overlap free.
(iv) Consider $\{x, y\} \subseteq N_1$ with both x and y odd and an even $z \in N_2$ such that $x = y + n \cdot z$ for some $n \in \mathbb{N}$. There is a path $(0, y, y+z, y+2 \cdot z..., y+n \cdot z, ...)$. We assumed that $x = y + n \cdot z$. We then have two paths: $(0, x, ...)$ and $(0, y, y+z, y+2 \cdot z, ..., y+n \cdot z, ...)$. These two paths share a vertex in $y + n \cdot z = x$. Hence the causality-diagram for player 1 is not overlap-free.
(v) Consider $\{x, y\} \subseteq N_1$ with both x and y odd and an odd $z \in N_2$. Then the paths $(0, x, x+z, x+z+y, ...)$ and $(0, y, y+z, y+z+x, ...)$ share a vertex after the point of divergence in $x + z + y$. Hence the resulting causality diagram for player 1 contains paths that are not vertex-disjoint after the divergence point and thus is not overlap-free.
(vi) Consider $\{x\} \subseteq N_1$ with x odd and consider two orders $\{y, z\} \subseteq N_2$. The root of the causality diagram of player 1 is then joined to the odd vertex x. This vertex x can be considered as the root of its own subgraph D_2^x. Then, if N_2 contains two even orders, the scenario (i) applies to subgraph D_2^x. Hence, under such a scenario, D_2^x would not be overlap-free. If N_2 contains an even and an odd order, the scenario of (iii) applies to D_2^x. Then also now, D_2^x is not overlap-free. If N_2 contains two odd orders, then the scenario of (v) applies. Also then, D_2^x is not overlap-free. Since the root of player 1's causality diagram is connected to vertex x, the paths that are not pairwise vertex-disjoint in D_2^x are subpaths in player 1's causality diagram. Hence there also exist paths in player 1's causality diagram that are not vertex-disjoint after a point of divergence. Therefore, player 1's causality diagram is also not overlap-free.

Hence, we have shown that if none of the three cases listed in Lemma 5.2 apply, the causality diagram cannot be overlap-free. This ends the proof for Lemma 5.2.

□

In the following section, we will provide and discuss the proof for Theorem 5.2.

5.5 Proof of Theorem 5.2

In order to prove Theorem 5.2, we will split it up into two separate lemmas: Lemmas 5.3 and 5.4. We will prove each of these in turn. For Lemma 5.4 we provide sketches of the proof in this section. The full proof can be found in the Appendix 5.A accompanying this chapter.

Lemma 5.3. *Consider a family of games $\mathcal{G}(N_1, N_2)$. If for every game in $\mathcal{G}(N_1, N_2)$, each choice that survives the IESDC-procedure for player 1 is also a rational choice under common belief in rationality, then the causality diagram of player 1 is overlap-free.*

Proof. Suppose that the causality diagram $D_1(N_1, N_2)$ of player 1 is not overlap-free. We will construct a game G^* in $\mathcal{G}(N_1, N_2)$ such that not every choice for player 1 that survives the IESDC-procedure is rational under common belief in rationality.

We will do this in the following way. First, assume (N_1, N_2) induces a causality diagram for player 1 where the root is already a divergence point for two paths that overlap. Take two of such overlapping paths: a path $(0, a_1, a_2, ..., a_{n-1}, a_n = s, ...)$ and a path $(0, z_1, z_2, ..., z_{m-1}, z_m = s, ...)$. Subsequently we construct a game G^* that has some particular properties. First, we construct G^* such that each choice in C_1 and each choice in C_2 survives the IESDC-procedure. So each choice is optimal for *some* belief hierarchy. Then we construct the game G^* further such that $\bar{c}_1 \in C_1$ is only optimal for a belief hierarchy whose a_1-th order expectation assigns probability one to choice $c[a_1]$ (in C_1^∞ if a_1 is even and in C_2^∞ if a_1 is odd) *and* whose z_1-th order expectation assigns probability one to choice $d[z_1]$ (also in C_1^∞ if z_1 is even and in C_2^∞ if a_1 is

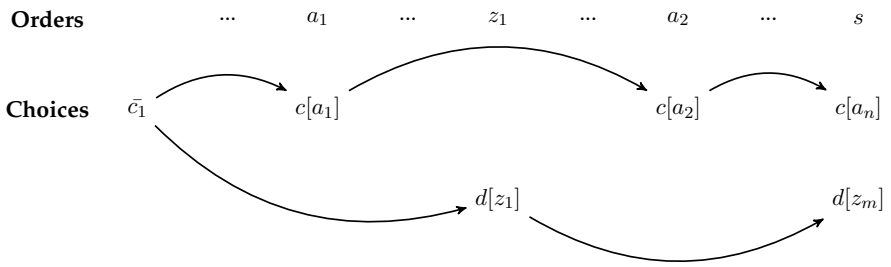


Figure 5.5: Example Proof Lemma 5.3, (part of) beliefs diagram

odd). Choice $c[a_1]$ in turn is only optimal for a $(a_2 - a_1)$-th order expectation assigning probability one to choice $c[a_2]$. And so on, up until we arrive at choice $c[a_n]$ for the s-th order expectation. We can do the same for the second subpath $(0, z_1, z_2, ..., z_m = s)$, which then ends up with a choice $d[z_m]$ for the s-th order expectation. We construct G^* such that $c[a_n] \neq d[z_m]$, and hence $\bar{c}_1$ cannot be optimal while expressing common belief in rationality.

Now, let
$C_1 := \{c[a_k] : a_k \text{ is even and } k \in \{1, ..., n\}\} \cup \{d[z_k] : z_k \text{ is even and } k \in \{1, ..., n\}\} \cup \{\bar{c}_1\} \cup \{x\}$, and
$C_2 := \{c[a_k] : a_k \text{ is odd and } k \in \{1, ..., n\}\} \cup \{d[z_k] : z_k \text{ is odd and } k \in \{1, ..., n\}\} \cup \{y\}$. We assume here that all these choices are different. Note that for order $z_m = a_n = s$, there are two different choices: $c[a_n]$ and $d[z_m]$. Let the choices have the following properties.

1. Choice x always yields a utility of 1 for player 1, under any belief hierarchy b_1;

2. Choice y always yields a utility of 1 for player 2, under any belief hierarchy b_2;

3. Choice $\bar{c}_1$ is only optimal for a higher-order expectation which assigns in the a_1-th component probability one to $c[a_1]$ and in the z_1-th component probability one to $d[z_1]$. Only in that case choice $\bar{c}_1$ leads to a utility of 1. In all other cases, utility is 0;

4. Each $c[a_{k-1}] \in C_1$ for $k \in \{2, ..., n\}$ is such that $u_1(c[a_{k-1}], b_1) = 1$ for any $b_1 \in B_1$ where the $(a_k - a_{k-1})$-th order expectation assigns probability one to $c[a_k]$. If $b_1 \in B_1$ is such that the $(a_k - a_{k-1})$-th order expectation assigns probability one to any choice $c \neq c[a_k]$, then the utility will be 0. Each $d[z_{k-1}] \in C_1$ for $k \in \{2, ..., n\}$ is such that $u_1(d[z_{k-1}], b_1) = 1$ for any $b_1 \in B_1$ where the $(z_k - z_{k-1})$-th order expectation assigns probability one to $d[z_k]$. If $b_1 \in B_1$ is such that the $(z_k - z_{k-1})$-th order expectation assigns probability one to any choice $d \neq d[z_k]$, then the utility will be 0. We do exactly the same for each $c[a_{k-1}], d[z_{k-1}] \in C_2$ for $k \in \{2, ..., n\}$.

5. For simplicity assume $c[a_n]$ and $d[z_m]$ *always* lead to a utility of 1.

Note that *each* choice in C_1 and C_2 leads to a utility of 1 under some extreme higher-order expectation. As a result, no choice is strictly dominated and hence each choice survives the IESDC-procedure.

We will now show that choice $\bar{c}_1$ in the game G^* as constructed before cannot be optimal under a belief hierarchy expressing common belief in rationality, even though it survives the IESDC-procedure. We will do so by arguing that $\bar{c}_1$ cannot be optimal for a belief hierarchy that simultaneously expresses up to a_{n-1}-fold and up z_{m-1}-fold belief in rationality. That is, the latter two events will require conflicting restrictions on the s-th order expectation. Figure 5.5 helps in illustrating this point by depicting (part of) a beliefs diagram of some game G^*.

In game G^*, choice $\bar{c}_1$ is optimal only if the a_1-th order expectation of player 1 assigns probability one to choice $c[a_1]$. Namely, only then the utility for player 1 is equal to 1. For any $k \in \{2, ..., n\}$, choice $c[a_{k-1}]$ is only optimal if the $(a_k - a_{k-1})$-th order expectation (of player 1 if a_{k-1} is even, otherwise of player 2) assigns probability one to choice $c[a_k]$. Thus we obtain a chain of restrictions. In the example of Figure 5.5 this would correspond to the upper path $(0, a_1, a_2, s)$. In order for player 1 to be able to rationally choose $\bar{c}_1$ under a belief hierarchy expressing up to a_{n-1}-fold belief in rationality, the s-th order expectation should thus assign probability one to choice $c[s]$.

Another requirement for choice $\bar{c}_1$ in game G^* to be optimal is that the z_1-th order expectation of player 1 assigns probability one to choice $d[z_1]$. Namely, only under such a condition can the utility of player 1 be equal to 1. For any $k \in \{2, ..., n\}$, choice $d[z_{k-1}]$ is only optimal if the $(z_k - z_{k-1})$-th order expectation (of player 1 if z_{k-1} is even, otherwise of player 2) assigns probability one to choice $d[z_k]$. In Figure 5.5 this chain of restrictions would for instance correspond to the lower path $(0, z_1, s)$. In order for player 1 to be able to rationally choose $\bar{c}_1$ under a belief hierarchy expressing up to z_{m-1}-fold belief in rationality, the s-th order expectation should assign probability one to choice $d[s]$.

We constructed G^* such that $c[s] \neq d[s]$. But then, $\bar{c}_1$ cannot be optimal for a belief hierarchy that expresses both up to a_{n-1}-fold belief in rationality and up to z_{m-1}-fold belief in rationality. Then $\bar{c}_1$ also cannot be optimal for a belief hierarchy expressing common belief in rationality.

Now, we initially assumed that the point of divergence for our two paths was at the root. Not for every causality diagram with overlap this is possible. However, it should certainly be possible to occur within one arc-distance of the root.

Claim 5.1. *Consider a causality diagram $D_1(N_1, N_2)$ for player 1 that is not overlap-free. Then there exist two paths with overlap that either (a) have a point of divergence at the root, or (b) that have a point of divergence at an odd order $a \in N_1$.*

Proof of claim. To prove this, we can point to our proof construction for the "only if"-part of the proof for Lemma 5.2. In this proof, we provided an exhaustive list of six scenarios. In the first five scenarios listed for that proof we were able to construct two paths with overlap that had a divergence point at the root of the causality diagram. So for these scenarios case (a) of the claim is satisfied. For the final scenario we noted that the root had an outgoing arc to an odd vertex $a \in N_1$. This odd vertex was the root of its own subgraph, which always could be categorized under scenario (i), (iii) or (v). As such, in this subgraph there also existed overlapping paths that had their point of divergence

at the root. Hence, for this scenario case (b) is satisfied. This completes the proof of this claim.

So let us now consider the case that the first point of divergence in the causality diagram occurs at the odd order $a \in N_1$. Let us have two paths $(0, a, a_1, a_2, ..., a_{n-1}, a_n = s, ...)$ and $(0, a, z_1, z_2, z_{m-1}, z_m = s, ...)$. Then we can simply construct the game G^* as follows. Let
$C_1 := \{c[a_k] : a_k$ is even and $k \in \{1, ..., n\}\} \cup \{d[z_k] : z_k$ is even and $k \in \{1, ..., n\}\} \cup \{\bar{c}_1\} \cup \{x\}$, and
$C_2 := \{c[a_k] : a_k$ is odd and $k \in \{1, ..., n\}\} \cup \{d[z_k] : z_k$ is odd and $k \in \{1, ..., n\}\} \cup \{\bar{c}_2\} \cup \{y\}$. Again, we assume here that all these choices are different. Let the choices have the following properties.

1. Choice x always yields a utility of 1 for player 1, under any belief hierarchy b_1;

2. Choice y always yields a utility of 1 for player 2, under any belief hierarchy b_2;

3. Choice $\bar{c}_2$ is only optimal for a higher-order expectation which assigns in the $(a_1 - a)$-th component probability one to $c[a_1]$ and in the $(z_1 - a)$-th component probability one to $d[z_1]$. Only in that case choice $\bar{c}_1$ leads to a utility of 1. In all extreme other cases, utility is 0;

4. Each $c[a_{k-1}] \in C_1$ for $k \in \{2, ..., n\}$ is such that $u_1(c[a_{k-1}], b_1) = 1$ for any $b_1 \in B_1$ where the $(a_k - a_{k-1})$-th order expectation assigns probability one to $c[a_k]$. If $b_1 \in B_1$ is such that the $(a_k - a_{k-1})$-th order expectation assigns probability one to any choice $c \neq c[a_k]$, then the utility will be 0. Each $d[z_{k-1}] \in C_1$ for $k \in \{2, ..., n\}$ is such that $u_1(d[z_{k-1}], b_1) = 1$ for any $b_1 \in B_1$ where the $(z_k - z_{k-1})$-th order expectation assigns probability one to $d[z_k]$. If $b_1 \in B_1$ is such that the $(z_k - z_{k-1})$-th order expectation assigns probability one to any choice $d \neq d[z_k]$, then the utility will be 0. We do exactly the same for each $c[a_{k-1}], d[z_{k-1}] \in C_2$ for $k \in \{2, ..., n\}$.

5. For simplicity assume $c[a_n]$ and $d[z_m]$ *always* lead to a utility of 1;

6. Choice $\bar{c}_1$ is only optimal for a higher-order expectation which assigns in the a-th component probability one to $\bar{c}_2$. Only in that case choice $\bar{c}_1$ leads to a utility of 1. In all extreme other cases, utility is 0.

Note that game G^* from the perspective of player 2 is exactly as it was before from the perspective of player 1. It follows then that choice $\bar{c}_2$ cannot be optimal for player 2 under a belief hierarchy expressing common belief in rationality. In the new version of G^* we added an extra choice for player 1: choice $\bar{c}_1$ is only optimal under an a-th order expectation that assigns probability one to choice $\bar{c}_2$ for player 2. Then it follows that choice $\bar{c}_1$ is also never optimal given a belief hierarchy expressing common belief in rationality.

Since we took an arbitrary combinations (N_1, N_2) that leads to a causality diagram for player 1 with overlap, it follows that for each family of games $\mathcal{G}(N_1, N_2)$ we can construct a game G^* as we did here.

□

From Lemma 5.3 we can conclude that if the causality diagram is not overlap-free, there always exist accompanying psychological games in which the IESDC-procedure does not characterize those choices that are rational under common belief in rationality. Next we will show that if the causality diagram *is* overlap-free, then this exact characterization *does* always occur.

Lemma 5.4. *Consider a family of games $\mathcal{G}(N_1, N_2)$. If the causality diagram of player 1 is overlap-free, then for every game in $\mathcal{G}(N_1, N_2)$, each choice that survives the IESDC-procedure for player 1 is also a rational choice under common belief in rationality.*

The actual proof can be found in the appendix. Here we provide an overview of the proof and an explanation by means of examples.

The outline of this proof is as follows. We consider three different scenarios, which together exactly cover all three cases from Theorem 5.2. Scenario (i) corresponds to the scenario that N_1 contains a single even order, scenario (ii) correspond to the case that N_1 and N_2 both consists of a single odd order, and scenario (iii) corresponds to case (iii) of Theorem 5.2. Note here that scenario (iii) covers the subcase (i) of Theorem 5.2 where N_1 contains a single odd order and N_2 contains a single even order.

The proof is constructive. For each of the scenarios listed above, we will take some arbitrary choice $c_1 \in C_1^\infty$ that survives the IESDC-procedure. The goal is to construct a belief hierarchy that expresses common belief in rationality and is such that it optimizes choice c_1. We do so by making use of finite epistemic models where types represent belief hierarchies. These models we introduced in CHAPTER 2. Recall that each type represents a probability distribution over the opponent's choice-type combination. Each types thus already induces a probability distribution over the opponent's choices. Hence we can retrieve first-order beliefs from types in an epistemic model. Since a type also represents a probability distribution over the opponent's types, we can also retrieve a probability distribution over the opponent's first-order beliefs. As such, types also capture second-order beliefs. In a similar fashion we can retrieve from a type in an epistemic model third-order beliefs, fourth-order beliefs, and so on.

The proof of Lemma 5.4 consists of two steps. In Step 1, we first fix an arbitrary choice $c_1 \in C_1^\infty$ for player 1 that survives the IESDC-procedure. For each of the three scenarios, we construct a finite epistemic model with a type $t_1^{c_1}$ that optimizes choice c_1. Moreover, we construct this epistemic model such that for each order of belief k that is on a path in player 1's causality diagram, the type $t_1^{c_1}$ expresses k-fold belief in rationality. We call this on-path belief in rationality.

Definition 5.7. *Consider a game in $\mathcal{G}(N_1, N_2)$ and the causality diagram $D_1(N_1, N_2)$ for player 1. Consider a belief hierarchy b_1 for player 1. We*

*say b_1 expresses **on-path belief in rationality** if it expresses k-fold belief in rationality for every $k \geq 1$ that is on some path in the causality diagram.*

Then, in Step 2, we transform the epistemic model created in Step 1. We ensure that for all the remaining orders of belief l, type $t_1^{c_1}$ expresses also l-fold belief in rationality. Then type $t_1^{c_1}$ will also express common belief in rationality. A formal proof can be found in the appendix. Here, we sketch the proof by means of examples for the three scenarios.

Example Scenario (i): First consider scenario (i). This corresponds to case (ii) in Theorem 5.2. Let $C_1 = \{A, B, C\}$, $C_2 = \{D, E\}$, $N_1 = \{4\}$ and let $N_2 = \{1, 2\}$. Consider the game in Table 5.5. Each choice for player 1 and player 2 in this game is not strictly dominated, and hence $C_1^\infty = \{A, B, C\}$, $C_2^\infty = \{D, E\}$.

First, for A, B and C we will fix fourth-order expectations that will optimize each of these choices in turn. For A, we can take $b_1^A \in \Delta(C_1^\infty)$ with $b_1^A = 0.5A + 0.5B$. For B, we can take $b_1^B = C$ and for C we can take $b_1^C = B$.

Table 5.5: *Illustration of Proof scenario (i)*

Player 1's extreme fourth-order expectations

	A	B	C
A	2	2	0
B	3	0	2
C	0	3	1

Player 1's utilities

Player 2's combinations of extreme first-order and second-order expectations

	(A, D)	(A, E)	(B, D)	(B, E)	(C, D)	(C, E)
D	0	3	0	1	2	0
E	1	1	1	0	1	2

Player 2's utilities

Next, for each choice in $c_1 \in C_1^\infty$, we will construct a type $t_1^{c_1}$ for player 1. Each such type assigns probability one to a type $t_2^{c_1,1}$ for player 2, which on its turn assigns probability one to a type $t_1^{c_1,2}$ for player 1. Type $t_1^{c_1,2}$ on its turn assigns probability one to type $t_2^{c_1,3}$ for player 2. Type $t_2^{A,3}$ is such that $b_2[t_2^{A,3}] = 0.5(A, t_1^A) + 0.5(B, t_1^B)$; type $t_2^{B,3}$ is such that $b_2[t_2^{B,3}] = (C, t_1^C)$; type $t_2^{C,3}$ is such that $b_2[t_2^{C,3}] = (B, t_1^B)$. The probability distributions over choices in the induced beliefs here are thus equal to the fourth-order expectations we fixed before to optimize each choice. We couple the choices assigned positive probability to in these induced beliefs with the respective types we started with. The resulting partial epistemic model is illustrated via the beliefs diagram in Figure 5.6a. We explictely mention that it is partial, since some types induce beliefs that do not yet specify beliefs over choices.

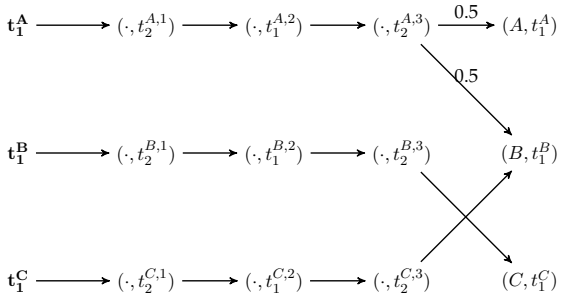
Next, we continue with Step 2 of the proof for this scenario. This involves filling in all blank spaces in our beliefs diagram. We do so in an iterative way. First, fill in random choices in each blank space. This is illustrated in a beliefs diagram in Figure 5.6b, by taking all components that have a superscript 0.

Next, take each right-most matrix. For instance, take the upper-right matrix in the diagram of Figure 5.6b. In this order of belief player 2 expects with probability 0.5 player 1 to choose A while expecting player 1 to believe that player 2 plays E^0. With the remaining probability of 0.5 player 2 expects player 1 to choose B while expecting player 1 to believe that player 2 plays D^0. For such a second-order belief, choice D is optimal. Hence we fill in D^1 in the upper-right matrix. Then, take the upper-middle matrix. Here player 1 in her fourth-order expectation assigns 0.5 to choice C^0 and probability 0.5 to choice B^0. Then choice C is optimal. Hence we fill in C^1 in the upper-middle matrix. Lastly, take the left-upper matrix. Here now player 2 expects player 1 to choose C^1 while expecting player 1 believes player 2 will choose D^1. Thus we get a second-order expectation of (C, D). Given (C, D), choice D is optimal for player 2. Hence, in the left-upper matrix we list choice D^1. We do this for every sequence of matrices in the diagram.

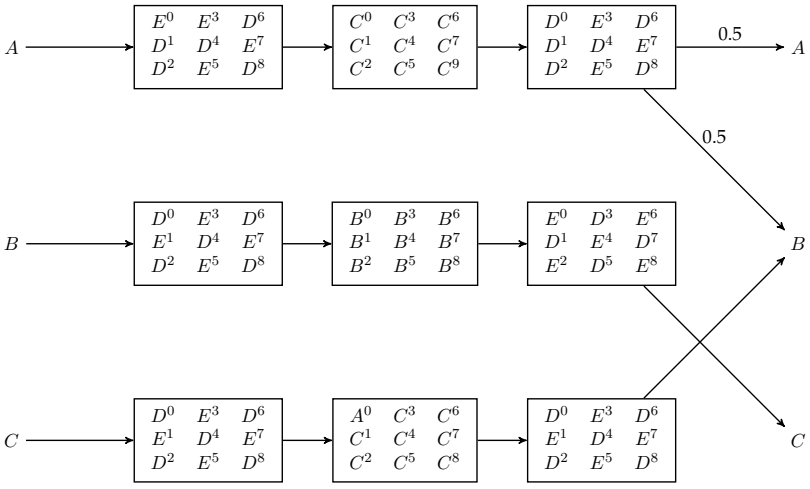
Next, in Iteration 2, we do a similar thing, leading to the choices with superscript 2 in the various matrices.

After a while we see a pattern emerge. We see that iterations 2 and 3 correspond to iterations 4 and 5 respectively. So the pattern repeats itself always after two iterations. Also, we have from iteration 1 onwards that each type expresses belief in the opponent's rationality by construction. Taken together, two recurring iterations from iteration 1 onwards are therefore sufficient to characterize a finite, epistemic model in which each type expresses common belief in rationality. In this case these are Iterations 2 and 3. We loop these iterations indefinitely. In this way, we construct the epistemic model $\mathcal{M}^*$ as is depicted in Figure 5.6c. All beliefs in the second column of the table in this figure correspond to beliefs generated by Iteration 3 of Step 2. All beliefs in the last column of the table correspond to beliefs generated by Iteration 2 of Step 2.

One may verify that in $\mathcal{M}^*$ each type expresses common belief in rationality. This includes types t_1^A, t_1^B and t_1^C . By construction of Step 1 these types optimize choices A, B and C respectively. Hence, for each choice that survives the IESDC-procedure we have managed to construct a type expressing common belief in rationality, such that that choice is also still optimal.



(a) Step 1 scenario (i)



(b) Beliefs diagram Step 2

Figure 5.6: Illustration proof Scenario (i)

Types player 1	$T_1 = \{t_1^A, t_1^B, t_1^C, t_1^{A,2}, t_1^{B,2}, t_1^{C,2}, t_1^{A\prime}, t_1^{B\prime}, t_1^{C\prime}, t_1^{A,2\prime}, t_1^{B,2\prime}, t_1^{C,2\prime}\}$					
Types player 2	$T_2 = \{t_2^{A,1}, t_2^{B,1}, t_2^{C,1}, t_2^{A,3}, t_2^{B,3}, t_2^{C,3}, t_2^{A,1\prime}, t_2^{B,1\prime}, t_2^{C,1\prime}, t_2^{A,3\prime}, t_2^{B,3\prime}, t_2^{C,3\prime}\}$					
Player 1's beliefs	$b_1[t_1^A]$	=	$(E, t_2^{A,1})$	$b_1[t_1^{A\prime}]$	=	$(D, t_2^{A,1\prime})$
	$b_1[t_1^B]$	=	$(E, t_2^{B,1})$	$b_1[t_1^{B\prime}]$	=	$(D, t_2^{B,1\prime})$
	$b_1[t_1^C]$	=	$(E, t_2^{C,1})$	$b_1[t_1^{C\prime}]$	=	$(D, t_2^{C,1\prime})$
	$b_1[t_1^{A,2}]$	=	$(E, t_2^{A,3})$	$b_1[t_1^{A,2\prime}]$	=	$(D, t_2^{A,3\prime})$
	$b_1[t_1^{B,2}]$	=	$(D, t_2^{B,3})$	$b_1[t_1^{B,2\prime}]$	=	$(E, t_2^{B,3\prime})$
	$b_1[t_1^{C,2}]$	=	$(E, t_2^{C,3})$	$b_1[t_1^{C,2\prime}]$	=	$(D, t_2^{C,3\prime})$
Player 2's beliefs	$b_2[t_2^{A,1}]$	=	$(C, t_1^{A,2})$	$b_2[t_2^{A,1\prime}]$	=	$(C, t_1^{A,2\prime})$
	$b_2[t_2^{B,1}]$	=	$(B, t_1^{B,2})$	$b_2[t_2^{B,1\prime}]$	=	$(B, t_1^{B,2\prime})$
	$b_2[t_2^{C,1}]$	=	$(C, t_1^{C,2})$	$b_2[t_2^{C,1\prime}]$	=	$(C, t_1^{C,2\prime})$
	$b_2[t_2^{A,3}]$	=	$0.5(A, t_1^{A\prime}) + 0.5(B, t_1^{B\prime})$	$b_2[t_2^{A,3\prime}]$	=	$0.5(A, t_1^A) + 0.5(B, t_1^B)$
	$b_2[t_2^{B,3}]$	=	$(C, t_1^{C\prime})$	$b_2[t_2^{B,3\prime}]$	=	(C, t_1^C)
	$b_2[t_2^{C,3}]$	=	$(B, t_1^{B\prime})$	$b_2[t_2^{C,3\prime}]$	=	(B, t_1^B)

(c) Epistemic model scenario (i)

Figure 5.6: Illustration proof Scenario (i)

Example Scenario (ii)**:** Consider scenario (ii). This corresponds to case (i) in Theorem 5.2 where for both players the orders are odd. Let, $C_1 = \{A, B, C\}$, $C_2 = \{D, E\}$, $N_1 = \{3\}$ and let $N_2 = \{1\}$. Consider the game in Table 5.6. Note that each choice for player 1 is not strictly dominated. Similarly, for player 2, both choices D and E are not strictly dominated.

First, for A, B and C we will fix third-order expectations that we will optimize each of these choices in turn. For A, we can take $b_1^A \in \Delta(C_1^\infty)$ with $b_1^A = 0.5D + 0.5E$. For B, we can take $b_1^B = E$ and for C we can take $b_1^C = D$. Similary, for choice D of player 2 we can take $b_2^D = 0.6A + 0.4B$ and for choice E we can take $b_2^E = C$.

Next, for each choice $c \in C_1^\infty$, we can construct a type t_1^c. Similarly, for each choice $c \in C_2^\infty$, we can construct a type t_2^c. Each type t_1^c assigns probability one to a type $t_2^{c,1}$, which on its turn induces a belief that assigns probability one to a type $t_1^{c,2}$. Type $t_1^{A,2}$ is such that $b_1[t_1^{A,2}] = 0.5(D, t_2^D) + 0.5(E, t_2^E)$; type $t_1^{B,2}$ is such that $b_1[t_1^{B,2}] = (E, t_2^E)$; type $t_1^{C,2}$ is such that $b_1[t_1^{C,2}] = (D, t_2^D)$. Similarly, we can let type t_2^D be such that $b_2[t_2^D] = 0.6(A, t_1^A) + 0.4(B, t_1^B)$ and type t_2^E be such that $b_2[t_2^E] = (C, t_1^C)$ The resulting partial epistemic model is illustrated via the beliefs dia-

Table 5.6: *Illustration of Proof scenario* (ii)

Player 1's extreme third-order expectations

	D	E
A	2	2
B	0	3
C	3	0

Player 1's utilities

Player 2's extreme first-order expectations

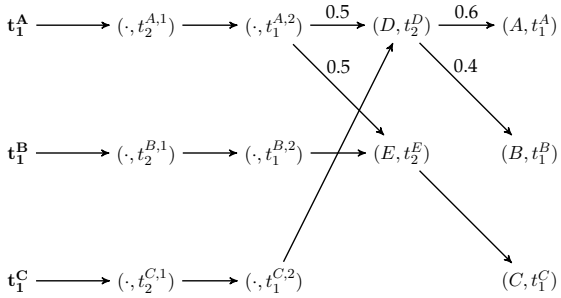
	A	B	C
D	6	6	6
E	8	0	8

Player 2's utilities

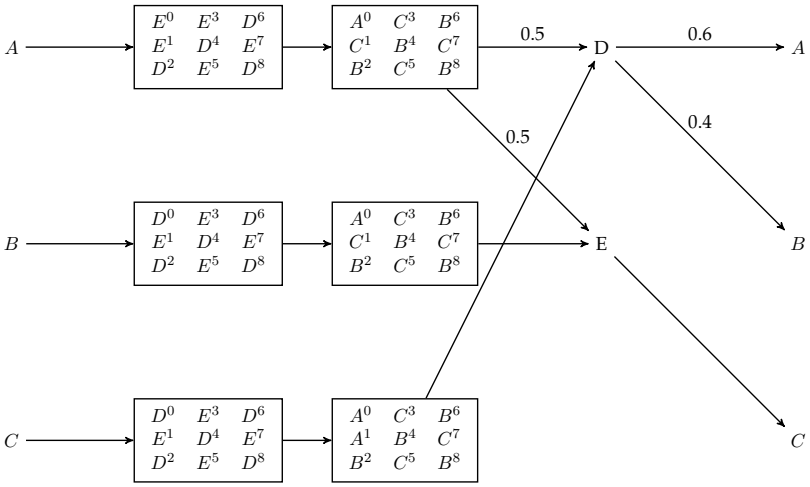
gram in Figure 5.7a.

We then use exactly the same construction method as we did for situation (i) for Step 2. First, for the blank spaces we fill in random choices. These choices have superscript 0 and can be seen in Figure 5.7b. Next, for each sequence of matrices, we will select optimal choices in a backward fashion. For instance, take the right-upper matrix. This order of belief relates to choices of player 1. We know that player 1's utility is variable in order 3. According to the diagram, player 1's third-order expectation reasoned from this matrix is $0.5 \cdot 0.6(D, A, E^0) + 0.5 \cdot 0.4(D, B, D^0) + 0.5(E, C, D^0)$. Summarized, this yields $0.3 \cdot E^0 + 0.7 \cdot D^0$ as a third-order expectation. Given this third-order expectation, we have that choice C is optimal. Hence we list choice C^1 next in this matrix. Similarly, take the left-upper matrix. We know player 2's utility is variable in order 1. According to the diagram, player 2's first-order expectation reasoned from this point is one that assigns probability one to choice C^1. Given this first-order expectation, choice E is optimal. Hence we list E^1. We do this same backward construction for the lower sequence of matrices in the beliefs diagram as well. Now, for the next iteration of Step 2, we do something similar, leading to the choices with superscript 2 in the various matrices.

Just like in scenario (i), we continue this process iteratively. We can note that iterations 4 and 5 yield the same choices as iterations 2 and 3. From Iteration 1 onwards we have that each type constructed expresses belief in the opponent's rationality. Hence, by looping Iteration 2 and 3 we can construct a finite, epistemic model where each type expresses common belief in rationality. Moreover, this would also include types that optimize each choice that survives the IESDC-procedure for player 1 for the same reason as in scenario (i). The resulting epistemic model $\mathcal{M}^*$ is found in Figure 5.7c. All beliefs in the second column of the table in this figure correspond to beliefs generated by Iteration 3 of Step 2. All beliefs in the last column of the table correspond to beliefs generated by Iteration 2 of Step 2.



(a) Step 1 scenario (ii)



(b) Beliefs diagram iteration Step 2

Figure 5.7: Illustration proof Scenario (ii)

Types player 1	$T_1 = \{t_1^A, t_1^B, t_1^C, t_1^{A,2}, t_1^{B,2}, t_1^{C,2}, t_1^{A\prime}, t_1^{B\prime}, t_1^{C\prime}, t_1^{A,2\prime}, t_1^{B,2\prime}, t_1^{C,2\prime}\}$					
Types player 2	$T_2 = \{t_2^{A,1}, t_2^{B,1}, t_2^{C,1}, t_2^D, t_2^E, t_2^{A,1\prime}, t_2^{B,1\prime}, t_2^{C,1\prime}, t_2^{D\prime}, t_2^{E\prime}\}$					
Player 1's beliefs	$b_1[t_1^A]$	=	$(E, t_2^{A,1})$	$b_1[t_1^{A\prime}]$	=	$(D, t_2^{A,1\prime})$
	$b_1[t_1^B]$	=	$(E, t_2^{B,1})$	$b_1[t_1^{B\prime}]$	=	$(D, t_2^{B,1\prime})$
	$b_1[t_1^C]$	=	$(E, t_2^{C,1})$	$b_1[t_1^{C\prime}]$	=	$(D, t_2^{C,1\prime})$
	$b_1[t_1^{A,2}]$	=	$0.5(D, t_2^D) + 0.5(E, t_2^E)$	$b_1[t_1^{A,2\prime}]$	=	$0.5(D, t_2^{D\prime}) + 0.5(E, t_2^{E\prime})$
	$b_1[t_1^{B,2}]$	=	(E, t_2^E)	$b_1[t_1^{B,2\prime}]$	=	$(E, t_2^{E\prime})$
	$b_1[t_1^{C,2}]$	=	(D, t_2^D)	$b_1[t_1^{C,2\prime}]$	=	$(D, t_2^{D\prime})$
Player 2's beliefs	$b_2[t_2^{A,1}]$	=	$(C, t_1^{A,2})$	$b_2[t_2^{A,1\prime}]$	=	$(B, t_1^{A,2\prime})$
	$b_2[t_2^{B,1}]$	=	$(C, t_1^{B,2})$	$b_2[t_2^{B,1\prime}]$	=	$(B, t_1^{B,2\prime})$
	$b_2[t_2^{C,1}]$	=	$(C, t_1^{C,2})$	$b_2[t_2^{C,1\prime}]$	=	$(B, t_1^{C,2\prime})$
	$b_2[t_2^D]$	=	$0.6(A, t_1^{A\prime}) + 0.4(B, t_1^{B\prime})$	$b_2[t_2^{D\prime}]$	=	$0.6(A, t_1^A) + 0.4(B, t_1^B)$
	$b_2[t_2^E]$	=	$(C, t_1^{C\prime})$	$b_2[t_2^{E\prime}]$	=	(C, t_1^C)

(c) Epistemic model scenario (ii)

Figure 5.7: Illustration proof Scenario (ii)

Scenario (iii)

Consider scenario (iii). This corresponds to case (iii) in Theorem 5.2. Take $N_1 = \{1, 7\}$ and $N_2 = \{4\}$. Now, consider the game as depicted in Table 5.7, with $C_1 = \{a, b\}$ and $C_2 = \{c, d, e, f\}$. Note that no choice for player 2 is strictly dominated. No choice for player 1 is strictly dominated either. Thus $C_1 = C_1^\infty$ and $C_2 = C_2^\infty$. Also note that choice a for player 1 is only optimal for the belief $b_1^a = (c^1, d^7)$. The superscripts here in the belief indicate the order of belief the choice corresponds to.

We will focus in this scenario on making choice a optimal for a belief hierarchy expressing k-fold belief in rationality for every order of belief k on a path of player 1's causality diagram. To this end, first fix a type t_1^a for player 1.

Scenario (iii) differs from the previous ones in that we have multiple paths on the causality diagram of player 1. As a result, we cannot use the construction with sequences of probability one beliefs in exactly the same way as we did for scenarios (i) and (ii). To compare more specifically with the previous two scenarios, consider the following example. First, define type t_1^a to be such that it assigns probability one to the choice-type combination $(c, t_2^{a,1})$ for player 2. Let type $t_2^{a,1}$ on its

Table 5.7: *Illustrating game for proof scenario* (iii)

Player 1's combinations of extreme first-order and seventh-order expectations

	(c,c)	(c,d)	(c,e)	(c,f)	(d,c)	(d,d)	...
a	0	1	0	0	0	0	0
b	1	1	1	1	1	1	1

Player 1's utilities

Player 2's extreme fourth-order expectations

	c	d	e	f
c	2	2	0	0
d	0	2	2	0
e	3	0	3	3
f	0	3	0	2

Player 2's utilities

turn assigns probability one to the choice-type combination $(\cdot, t_1^{a,2})$. Let type $t_1^{a,2}$ assign probability one to $(\cdot, t_2^{a,3})$; let type $t_2^{a,3}$ assign probability one to $(\cdot, t_1^{a,4})$; let type $t_1^{a,4}$ assign probability one to $(\cdot, t_2^{a,5})$; let type $(\cdot, t_2^{a,5})$ assign probability one to $(\cdot, t_1^{a,6})$; and let type $t_1^{a,6}$ assign probability one to the choice-type combination (d, t_2^d). Then clearly choice a is optimal given type t_1^a. However, player 1 would not express 1-fold belief in rationality and thus also not on-path belief in rationality. Namely, choice c is only optimal for a probabilistic fourth-order expectation. We have that t_1^a assigns probability one to the choice-type combination $(c, t_2^{a,1})$, from which there follows a sequence of three more probability one beliefs until type $t_1^{a,4}$, which induces a belief that assigns probability one to $(\cdot, t_2^{a,5})$. Whatever choice we fill in to complete the belief $b[t_1^{a,4}]$, choice c assigned probability one to in $b[t_1^a]$ will then never be optimal, as by construction the fourth-order expectation induced by $t_2^{a,1}$ will be non-probabilistic. Another route we could take is to make the belief $b[t_1^{a,4}]$ probabilistic, in such a way that choice c in the support of $b_1[t_1^a]$ becomes optimal. However, then we are not taking into account that the seventh-order expectation induced by type t_1^a should be non-probabilistic. That is, choice a is only optimal for the non-probabilistic belief (or seventh-order expectation) $b_1^a = (c^1, d^7)$. So we cannot fix just any type $t_1^{a,4}$ such that choice c is optimal given type $t_1^{a,1}$.

When constructing types for a partial epistemic model where type t_1^a expresses k-fold belief in rationality for every order k on the causality diagram, we therefore should at all times look at *combinations* of choices. As an (incomplete) illustration of such a partial epistemic model, consider Figure 5.8. In this figure, we have type t_1^a for player 1. All the remaining types have combinations of two choices in their superscripts. The reason for this is that the sequences of types we will construct in this step will be such that they optimize a combination of two choices. These choices appear in different orders of belief. For instance, we define type t_1^a now such that $b_1[t_1^a](c, t_2^{cf}) = 1$. The idea is to

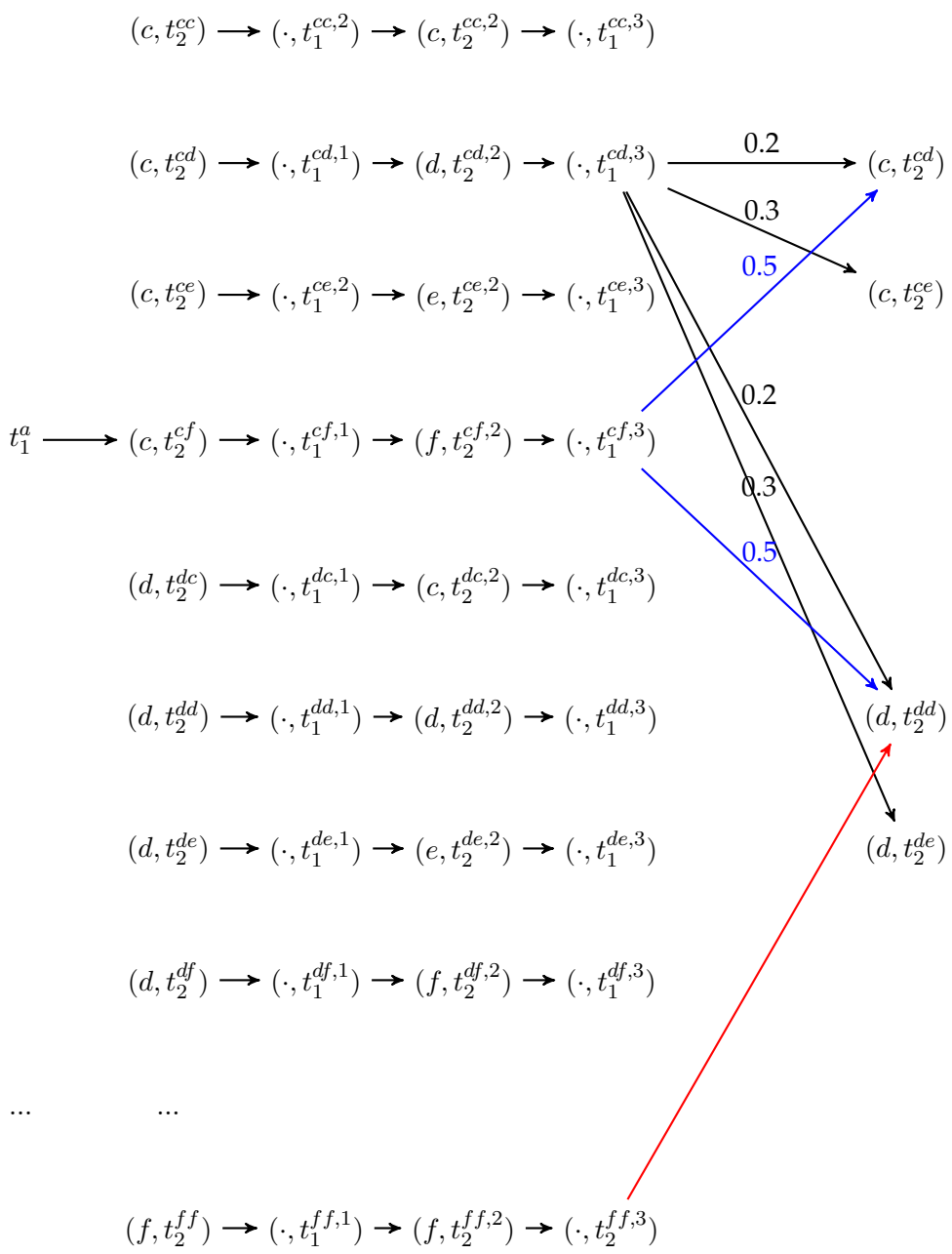


Figure 5.8: Step 1 scenario (iii)

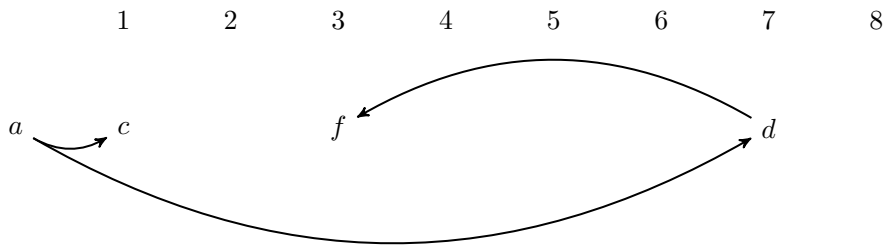


Figure 5.9: Step 1 scenario (iii)

construct t_2^{cf} such that choice c for player 2 is optimal given this type. Then type t_1^a would express 1-fold belief in rationality. We do so by first constructing a sequence of types $(t_2^{cf}, t_1^{cf,1}, t_2^{cf,2}, t_1^{cf,3})$. In this sequence, we have $b_2[t_2^{cf}](\cdot, t_1^{cf,1}) = 1$, $b_1[t_1^{cf,1}](f, t_2^{cf,2}) = 1$ and $b_2[t_2^{cf,2}](\cdot, t_1^{cf,3}) = 1$. Finally, we have $b_1[t_1^{cf,3}] = 0.5(c, t_2^{cd}) + 0.5(d, t_2^{dd})$. The fourth-order expectation that type t_2^{cf} induces is then $b_2^c = 0.5c + 0.5d$. Looking at the utilities for player 2 in Table 5.7, this indeed makes choice c for player 2 optimal.

The second choice listed in the superscript of type t_2^{cf} is the choice f. When player 1 is of type t_1^a, her third-order expectation corresponds to a belief that assigns probability one to choice f. That is, we have that $b_1[t_1^{cf,1}](f, t_2^{cf,2}) = 1$. The reason why we look specifically at the third-order expectation here is illustrated in Figure 5.9. We know that choice a is only optimal given a seventh-order expectation that places probability one on d^7. Directly specifying the seventh-order expectation is problematic since the fifth-order expectation necessarily needs to be probabilistic. Therefore, what we do is the following. For each order of belief in N_1 beyond order $1 + 4 = 5$, we first reason backwards, up until we get in the range of beliefs between orders 1 and 5. In this case, this means we first reason backwards one optimality-relevant step from order 7 to order $7 - 4 = 3$. We fix one choice for player 2 that is optimal given a fourth-order expectation that places probability one on choice d^7. In this game the only optimal choice for

such a fourth-order expectation is f^3. Hence we fix choice f. Note that if we had order 11 instead of 7, we would have to reason two steps backwards first: first from order 11 to order $11 - 4 = 7$ and then from order 7 to order $7 - 4 = 3$.

As we can conclude from Figure 5.8, choice f is optimal for type $t_2^{cf,2}$. That is, we have that $b_2[t_2^{cf,2}](\cdot, t_1^{cf,3}) = 1$ and that $b_1[t_1^{cf,3}] = 0.5(c, t_2^{cd}) + 0.5(d, t_2^{dd})$. In this model, we defined t_2^{cd} such that it induces a second-order expectation that places probability one on choice d. The same applies to type t_2^{dd}. Hence, the fourth-order expectation induced by type $t_2^{cf,2}$ is $b_2^f = 0.5d + 0.5d = d$.

We can construct such types like t_2^{cf} for each combination of choices in $C_2^\infty \times C_2^\infty$. To take another example of a combination of choices, consider (c^1, d^3). Here we define the type t_2^{cd} and the types in the sequence $(t_1^{cd,1}, t_2^{cd,2}, t_1^{cd,3})$ in a similar way as before. This is again illustrated in Figure 5.8. We have that choice c is optimal given a fourth-order expectation $b_2^c = 0.5c + 0.5d$. We have that choice d is optimal given a fourth-order expectation $b_2^d = 0.4d + 0.6e$. The resulting joint probability distribution is $b_2^{cd} = 0.5 \cdot 0.4 \cdot (c, d) + 0.5 \cdot 0.6 \cdot (c, e) + 0.5 \cdot 0.4 \cdot (d, d) + 0.5 \cdot 0.6 \cdot (d, e) = 0.2(c, d) + 0.3(c, e) + 0.2(d, d) + 0.3(d, e)$. Thus we define type $t_1^{cd,3}$ to be such that $b_1[t_1^{cd,3}] = 0.2(c, t_2^{cd}) + 0.3(c, t_2^{ce}) + 0.2(d, t_2^{d,d}) + 0.3(d, t_2^{d,e})$. By construction, type t_2^{cd} then induces a fourth-order expectation b_2^c and type $t_2^{cd,2}$ induces a fourth-order expectation b_2^d. Therefore choice c^1 is optimal given the type t_2^{cd} and choice d^3 is optimal given the type $t_2^{cd,2}$.

For each combination of two choices $(c_2^1, c_2^3) \in C_2^\infty \times C_2^\infty$, we now do the following. Take any such combination. Create a type for this combination: $t_2^{c_2^1 c_2^3}$. Define this type to be such that $b_2[t_2^{c_2^1 c_2^3}]$ assigns probability one to type $t_1^{c_2^1 c_2^3,1}$. Define type $t_1^{c_2^1 c_2^3,1}$ to be such that the belief $b_1[t_1^{c_2^1 c_2^3,1}]$ assigns probability one to $(c_2^3, t_2^{c_2^1 c_2^3,2})$. Define type $t_2^{c_2^1 c_2^3,2}$ such that $b_2[t_2^{c_2^1 c_2^3,2}]$ assigns probability one to $t_1^{c_2^1 c_2^3,3}$. Finally, for type $t_1^{c_2^1 c_2^3,3}$ we need to do the following. For every $c_2 \in C_2$, let $b_2^{c_2} \in \Delta(C_2^\infty)$ be a

fourth-order expectation that makes c_2 optimal. Let $b_2^{c_2^1 c_2^3} \in \Delta(C_2^\infty \times C_2^\infty)$ be the product of $b_2^{c_2^1}$ and $b_2^{c_2^3}$. Now, we define $b_1[t_1^{c_2^1 c_2^3}(3)]$ in the following manner:

$$b_1[t_1^{c_2^1 c_2^3,3}](c_2^{1\prime}, t_2^{c_2^{1\prime} c_2^{3\prime}}) = b_2^{c_2^1 c_2^3}(c_2^{1\prime}, c_2^{3\prime}), \tag{5.1}$$

for each combination of choices $(c_2^{1\prime}, c_2^{3\prime}) \in C_2^\infty \times C_2^\infty$.

Construct such sequences of types for every possible combination of player 2's choices in the product-space $C_2^\infty \times C_2^\infty$. This results in a partial epistemic model as is partially illustrated in Figure 5.8. First, each type $t_2^{c_2^1 c_2^3}$ in this model is such that that choice c_2^1 is optimal given this type. Namely, the fourth-order expectation of type $t_2^{c_2^1 c_2^3}$ is $b_2^{c_2^1}$ by equation (5.1). Also, choice c_2^3 is optimal for type $t_2^{c_2^1 c_2^3,2}$, because the fourth-order expectation of type $t_2^{c_2^1 c_2^3,2}$ is $b_2^{c_2^3}$ by equation (5.1). Second, each type $t_2^{c_2^1 c_2^3}$ in this model expresses 4-fold belief in rationality. Namely, the final type in the sequence of types $t_1^{c_2^1 c_2^3,3}$ is such that it only assigns positive probability to choice-type combinations $(c_2^{1\prime}, t_2^{c_2^{1\prime} c_2^{3\prime}})$. By construction, we have in such choice-type combinations that the choice $c_2^{1\prime}$ is optimal given the type $t_2^{c_2^{1\prime} c_2^{3\prime}}$. Similarly, we also have by construction that each type $t_2^{c_2^1 c_2^3,2}$ expresses 4-fold belief in rationality: from each such type, "following" four arrows in Figure 5.8 always brings us to choice-type combinations $(c_2^{3\prime}, t_2^{c_2^{1\prime} c_2^{3\prime},2})$. As shown before, we have in such choice-type combination that the choice $c_2^{3\prime}$ is optimal given the type $t_2^{c_2^{1\prime} c_2^{3\prime},2}$. Because the types $t_2^{c_2^1 c_2^3}$ and $t_2^{c_2^1 c_2^3,2}$ for each combination of choices (c_2^1, c_2^3) expresses 4-fold belief in rationality and because $N_2 = \{4\}$, we have in fact that each such type for player 2 also expresses k-fold belief in rationality for every k on a path in player 2's causality diagram.

Now, recall that $N_1 = \{1, 7\}$. Looking at Figure 5.8, one arrow away from type t_1^a we have the choice-type combination (c, t_2^{cf}) and seven arrows away from type t_1^a we have the choice-type combinations $(d, t_2^{cd,2})$

and $(d, t_2^{dd,2})$. Hence, the first-order and seventh-order expectation induced by type t_1^a are (c^1, d^7), which means that a is optimal for t_1^a. Also, type t_1^a expresses 1-fold and 7-fold belief in rationality and believes that player 2 expresses k-fold belief in rationality for every k on a path in her causality diagram. It then follows that also player 1 expresses k-fold belief in rationality for every order k on a path in her causality diagram if she is of type t_1^a.

The model illustrated in Figure 5.8 is not completed. We only completed the sequences of types for the following combinations of choices: (c, d), (c, f) and (f, f). We leave the remainder of the illustration of Step 1 to the reader. The backward construction of Step 2 is almost completely analogous to Step 2 for scenarios (i) and (ii). For more details the reader is referred to Appendix 5A.

From Lemma 5.4 we conclude that if the causality diagram of player 1 is overlap-free, then each choice that survives the IESDC-procedure for player 1 is also a rational choice under common belief in rationality. From Lemma 5.3 we concluded the reverse. Together then, Lemma 5.3 and Lemma 5.4 prove Theorem 5.2: the IESDC-procedure exactly characterizes the rational choices under common belief in rationality for player 1 if and only if the player 1's causality diagram is overlap-free. We know the causality diagram is overlap-free if and only if at least one of the cases as is listed in Theorem 5.2 is true.

5.6 Conclusion

Since its introduction by Geanakoplos et al. (1989), psychological game theory has proven to be a competent framework to model many belief-dependent motivations in games. Much of the work in this framework illustrates that, compared to traditional games, reasoning about and in situations with belief-dependent motivations can be rather complex. Some properties of traditional games that add intuition to reasoning in such games do not always carry over to psychological games.

In this chapter we focused on one of such failures: the exact characterization of common belief in rationality by the iterated elimination of strictly dominated choices (IESDC) procedure. The IESDC procedure has proven to be a very useful algorithm to analyse traditional games as it is straightforward to use and an intuitive notion because of its characterization of common belief in rationality. This sparked the question in what kind of psychological games the IESDC-procedure always characterizes rational choices under common belief in rationality. By exactly identifying these cases, we also wished to give intuition as to why the IESDC-procedure may fail in other cases.

The IESDC-procedure takes into account that players may have belief-dependent motivations. The manner in which decision problems are defined clearly lets utilities depend on higher-order expectations. In each elimination round, for those choices that survive the round we can always find *some* belief hierarchy in the (reduced) decision problem such that the relevant choice is optimal. We do this for each (reduced) decision problem independently. The complexity psychological games introduce is that reasoning steps may overlap, as we illustrated via causality diagrams. In order for a belief hierarchy to express k-fold belief in rationality, restrictions need to be imposed on particular higher-order beliefs. Expressing k'-fold belief in rationality may require restrictions on the same higher-order belief. These restrictions can be in conflict. Thus even though a choice (1) may be rational under a belief hierarchy expressing k-fold belief in rationality and (2) that same choice may be rational under a belief hierarchy expressing k'-fold belief in rationality, this choice may not be rational under any belief hierarchy that expresses both k-fold and k'-fold belief in rationality. The IESDC-procedure cannot take into account this friction. To the extent the IESDC-procedure characterizes particular reasoning steps of an individual, it does so for each such reasoning step independently. This completely disregards any overlap in reasoning steps.

When a causality diagram for a player is overlap-free, contradicting restrictions on the same order of belief in order to express common belief in rationality cannot occur. The main result that we have shown in

this chapter is that in precisely such cases IESDC always characterizes the rational choices under common belief in rationality.

In total we identified three cases in which causality diagrams are overlap-free. These include two relatively trivial cases in the sense that the causality diagram has a single path, and one non-trivial case. Though interesting kinds of psychological games can be captured by these three, it can be argued that many types of psychological games that are prominent in practice cannot. Namely, we have that if both players in an expectation-based psychological game care for the material outcome of the game and at least one player has some belief-dependent motivation, already then the IESDC-procedure is not guaranteed to characterize the rational choices under common belief in rationality. In particular in experimental settings this will very often be the case, as subjects need to be incentivized by material pay-offs.

A couple of natural extensions of the questions asked in this chapter arise. First, to keep matters tractable we focused on two-player expectation-based psychological games. We did not venture into the topic of expectation-based psychological games with many players. One interesting additional complexity such settings bring along is the question of whether correlation between the beliefs of opponents matters in the formation of higher-order expectations.

A second natural extension of this research would be to consider similar questions as asked in this chapter for dynamic psychological games. Many instances of belief-dependent motivations arise when players have the opportunity to learn about the beliefs and intentions of their opponents by observing their past behaviour. Such instances can also arise in one-stage games, where the updated belief after play can be utility-relevant as well (Battigalli and Dufwenberg, 2009). In traditional settings, prominent reasoning concepts for dynamic games are for instance common belief in future rationality (which can capture backward induction reasoning) as in Perea (2014) and common strong belief in rationality (which captures forward induction reasoning) as in Battigalli and Siniscalchi (2002). Both concepts are characterized by

procedures that, amongst other things, rely on iteratively eliminating strictly dominated strategies. The natural question then arises to what extent such elimination procedures also succeed in characterizing relevant reasoning concepts in expectation-based psychological games.

Throughout this chapter we assumed common knowledge of players' motivations, including belief-dependent motivations. It is a strong assumption to make that psychological entities such as belief-dependent motivations are completely transparant among all players in a game (Attanasi et al., 2016). Elimination procedures have already been developed for traditional games with incomplete information (see for instance Bach and Perea (2016)). A final extension one therefore could consider is how well such elimination procedures fare in characterizing the relevant rationality concepts in expectation-based psychological games with incomplete information.

5.A Proof of Lemma 5.4

We recall here Lemma 5.4.

Lemma 5.4. *Consider a family of games $\mathcal{G}(N_1, N_2)$. If the causality diagram of player 1 is overlap-free, then for every game in $\mathcal{G}(N_1, N_2)$, each choice that survives the IESDC-procedure for player 1 is also a rational choice under common belief in rationality.*

Proof. We will now prove this lemma for the three scenarios described in Section 5.5. We will do so in two steps. We will take some choice $c_1 \in C_1^\infty$ that survives the IESDC-procedure. Then in Step 1 we will create a partial epistemic model with a type that makes choice c_1 optimal. Moreover, we construct this type such that for each order k on a path in player 1's causality diagram, this type expresses k-fold belief in rationality. Afterwards, in Step 2, we will show that from *any* partial epistemic model including a type as created in Step 1, we can create a full epistemic model with a type that makes choice c_1 optimal and that expresses common belief rationality. We do so by making use of a backward, recursive procedure that in each iteration simultaneously constructs types and choices which are optimal given those types.

Scenario (i)
Step 1
First consider scenario (i) with $N_1 = \{a\}$ and a even. In this Step 1, we will construct a partial epistemic model. By *partial* we mean that we only completely specify the beliefs induced for some particular types.

For each choice $c_1 \in C_1^\infty$, fix an a-th order expectation $b_1^{c_1} \in \Delta(C_1^\infty)$

for which c_1 is optimal.[3] The reason we can do so is as follows. From Lemma 5.1 we know that for each choice that is not strictly dominated in a decision problem, we can find a belief in that decision problem such that the relevant choice is optimal. The final reduced decision problem resulting from the IESDC-procedure leaves the choices in C_1^∞ for player 1 in the decision problem. That is, because order a is even we have as a reduced decision problem after following through with the IESDC-procedure: $(C_1^\infty, C_1^\infty, v_1)$. By Lemma 5.1, we should then have that each choice in C_1^∞ is optimal for some belief in $\Delta(C_1^\infty)$.

Subsequently, for each $c_1 \in C_1^\infty$, construct a type $t_1^{c_1}[c_1]$. Take for each $t_1^{c_1}[c_1]$ a sequence of types $(t_1^{c_1}[c_1], t_2^{c_1,1}[c_1], ..., t_2^{c_1,a-1}[c_1])$. Then, for each $c_1 \in C_1^\infty$ let us have in each such sequence that type $t_1^{c_1}[c_1]$ assigns probability one to type $t_2^{c_1,1}[c_1]$, and that type $t_i^{c_1,n}[c_1]$ with $i \in \{1,2\}$ assigns probability one to type $t_j^{c_1,n+1}$ with $j \neq i$, for each $n \in \{1, ..., a-2\}$.

Next, for each $c_1 \in C_1^\infty$ construct for each $c_1' \in C_1^\infty$ a type $t_1^{c_1,a}[c_1']$. Then define $t_2^{c_1,a-1}[c_1]$ to be such that

$$b_2[t_2^{c_1,a-1}[c_1]](c_1', t_1^{c_1,a}[c_1'']) := \begin{cases} b_1^{c_1}(c_1'), \text{ if } c_1' = c_1'' \\ 0, \text{ otherwise.} \end{cases}$$

Now, we do a similar thing $k^* - 2$ times, where $k^* = max(N_1 \cup N_2)$. For each $p \in \{1, ..., k^* - 2\}$, do the following: For each $c_1, c_1' \in C_1^\infty$ take a sequence of types $(t_1^{c_1,pa}[c_1'], t_2^{c_1,pa+1}[c_1'], ..., t_2^{c_1,(p+1)a-1}[c_1'])$. Then, let us have in each such sequence that type $t_1^{c_1,pa}[c_1']$ assigns probability one to type $t_2^{c_1,pa+1}[c_1']$, and that type $t_i^{c_1,pa+n}[c_1']$ with $i \in \{1,2\}$ assigns probability one to type $t_j^{c_1,pa+n+1}[c_1']$ with $j \neq i$, for each $n \in \{1, ..., a-2\}$.

[3] With a-th order expectation in this context we specifically refer to $marg_{C_1} e_1^a \in \Delta(C_1)$ where $e_1^a \in \Delta(W_1^{a-1} \times C_1)$.

Then, for each $c_1, c_1' \in C_1^\infty$ construct a type $t_1^{c_1,(p+1)a}[c_1']$. Then define $t_2^{c_1,(p+1)a-1}[c_1']$ to be such that

$$b_2[t_2^{c_1,(p+1)a-1}[c_1']](\bar{c}_1, t_1^{c_1,(p+1)a}[c_1'']) := \begin{cases} b_1^{c_1'}(\bar{c}_1), \text{ if } \bar{c}_1 = c_1'' \\ 0, \text{ otherwise.} \end{cases}$$

Finally, consider the case $p = k^* - 1$. For each combination $c_1, c_1' \in C_1^\infty$ take a sequence of types $(t_1^{c_1,(k^*-1)a}[c_1'], t_2^{c_1,(k^*-1)a+1}[c_1'], ..., t_2^{c_1,k^*a-1}[c_1'])$. Then, let us have in each such sequence that type $t_1^{c_1,(k^*-1)a}[c_1']$ assigns probability one to type $t_2^{c_1,(k^*-1)a+1}[c_1']$, and that type $t_i^{c_1,(k^*-1)a+n}[c_1']$ with $i \in \{1,2\}$ assigns probability one to type $t_j^{c_1,(k^*-1)a+n+1}[c_1']$ with $j \neq i$, for each $n \in \{1, ..., a-2\}$.

Finally define $t_2^{c_1,k^*a-1}[c_1']$ to be such that

$$b_2[t_2^{c_1,k^*a-1}[c_1']](\bar{c}_1, t_1^{c_1''}[c_1'']) := \begin{cases} b_1^{c_1'}(\bar{c}_1), \text{ if } \bar{c}_1 = c_1'' \\ 0, \text{ otherwise.} \end{cases}$$

So we have that the distribution over choices induced by type $t_2^{c_1,a-1}[c_1]$ is equal to the distribution over choices represented by the a-th order expectation $b_1^{c_1}$. From type $t_1^{c_1}[c_1]$ there follows a sequence of probability one beliefs up to type $t_2^{c_1,a-1}[c_1]$. The a-th order expectation induced by type $t_1^{c_1}[c_1]$ is thus equal to $b_1^{c_1}$. Taken together, then choice c_1 is optimal given type $t_1^{c_1}$.

A similar line of reasoning holds for each choice c_1' in combination with the type $t_1^{c_1,pa}[c_1']$, for each $c_1 \in C_1^\infty$ and each $p \in \{1, ..., k^* - 1\}$. The distribution over choices induced by type $t_2^{c_1,(p+1)a-1}[c_1']$ is equal to the distribution over choices represented by the a-th order expectation $b_1^{c_1'}$. From type $t_1^{c_1,pa}[c_1']$ there follows a sequence of probability one beliefs up to type $t_2^{c_1,(p+1)a-1}[c_1']$. The a-th order expectation induced by type

$t_1^{c_1,pa}[c_1']$ is thus equal to $b_1^{c_1'}$. Taken together, then choice c_1' is optimal given type $t_1^{c_1,pa}[c_1']$.

We do the above for each $c_1 \in C_1^\infty$. Call the resulting partial epistemic model $\mathcal{M}$. By construction, we have for each $c_1 \in C_1^\infty$ that c_1 is optimal given $t_1^{c_1}[c_1]$. Moreover, each $t_1^{c_1}[c_1]$ also expresses on-path belief in rationality. This is because for each order of belief pa for $p \in \{1, ..., k^* - 1\}$ the type $t_2^{c_1,pa-1}[c_1']$ (with $t_2^{c_1,a-1}[c_1]$ for $p = 1$ specifically) only assigns positive probability to choice-type pairs $(c_1'', t_1^{c_1,(p+1)a}[c_1''])$. In these pairs the choice is optimal for the type by construction. Additionally, for order k^*a, type $t_2^{c_1,k^*a-1}[c_1']$ only assigns positive probability to choice type pairs $(c_1'', t_1^{c_1''}[c_1''])$. Also in these pairs the choice is optimal for the type by construction.

Step 2

In Step 1 we have shown that for every choice $c_1 \in C_1^\infty$ we can always construct a *partial* epistemic model with a type $t_1^{c_1}[c_1]$ for which c_1 is optimal and that expresses on-path belief in rationality. In Step 2 we will now do the following. We will show that if there exists a belief hierarchy expressing on-path belief in rationality for which c_1 is optimal, then there is also a belief hierarchy expressing common belief in rationality for which c_1 is optimal.

Consider a partial epistemic model $\mathcal{M} = (T_i, b_i[t_i])_{i \in \{1,2\}}$ as constructed in Step 1 with a type $t_1^{c_1}[c_1]$ that expresses on-path belief in rationality and for which c_1 is optimal. By means of a backward, recursive procedure we transform this epistemic model such that we get to a new, complete epistemic model that includes a type $t_1^m[c_1, c_1, 0]$ that expresses common belief in rationality and induces the same a-th order expectation as type $t_1^{c_1}[c_1]$ does. The recursive procedure here defines choices *and* types at the same time in each iteration.

The recursive procedure is as follows.

Iteration 0: For each choice $c_1 \in C_1^\infty$, define

$$d^0[c_1, c_1, 0] := c_1.$$

Moreover, for each choice $c_1 \in C_1^\infty$ and each $k \in \{1, 2, ..., a-1\}$, define $d^0[c_1, c_1, k]$ randomly. So

$d^0[c_1, c_1, k] := c'$, for *some* $c' \in C_1^\infty$ if k is even or *some* $c' \in C_2^\infty$ if k is odd.

For each $p \in \{1, ..., k^* - 1\}$ let us have in a similar fashion that

$$d^0[c_1, c_1', pa] := c_1',$$

and for every $k \in \{1, 2, ..., a-1\}$ that

$$d^0[c_1, c_1', pa + k] := c', \text{ for } \textit{some } c' \in C_1^\infty \text{ if } k \text{ is even or } \textit{some } c' \in C_2^\infty \text{ if } k \text{ is odd.}$$

Take a sequence of types $(t_1^0[c_1, c_1, 0], t_2^0[c_1, c_1, 1], ..., t_2^0[c_1, c_1, a-1])$ for every choice $c_1 \in C_1^\infty$. Similarly, for each $p \in \{1, ..., k^* - 1\}$ and each pair of choices $c_1, c_1' \in C_1^\infty$, take a sequence of types $(t_1^0[c_1, c_1', pa], t_2^0[c_1, c_1', pa + 1], ..., t_2^0[c_1, c_1', (p+1)a - 1])$.

Now, for each $c_1 \in C_1^\infty$, define type $t_1^0[c_1, c_1, 0]$ such that

$$b_1[t_1^0[c_1, c_1, 0]] := (d^0[c_1, c_1, 1], t_2^0[c_1, c_1, 1]).$$

Then, define for each $k \in \{1, 2, ..., a-2\}$ type $t_i^0[c_1, c_1, k]$ with $i \in \{1, 2\}$ to be such that

$$b_i[t_i^0[c_1, c_1, k]] := (d^0[c_1, c_1, k+1], t_j^0[c_1, c_1, k+1]).$$

Finally, we define for each choice $c_1 \in C_1^\infty$ type $t_2^0[c_1, c_1, a-1]$ to be such that

$$b_2[t_2^0[c_1, c_1, a-1]](c_1', t_1^0[c_1, c_1', 0]) := b_1^{c_1}(c_1'), \forall c_1' \in C_1^\infty.$$

Similarly, for each $c_1, c_1' \in C_1^\infty$ and each $p \in \{1, ..., k^* - 1\}$ define $t_1^0[c_1, c_1', pa]$ to be such that

$$b_1[t_1^0[c_1, c_1', pa]] := (d^0[c_1, c_1', pa + 1], t_2^0[c_1, c_1', pa + 1]).$$

And define for each $k \in \{1, 2, ..., a - 2\}$ type $t_i^0[c_1, c_1', pa + k]$ to be such that

$$b_i[t_i^0[c_1, c_1', pa + k]] := (d^0[c_1, c_1', pa + k + 1], t_j^0[c_1, c_1', pa + k + 1]).$$

Finally, we define type $t_2^0[c_1, c_1', (p+1)a - 1]$ for $p \in \{1, ..., k^* - 2\}$ to be such that

$$b_2[t_2^0[c_1, c_1', (p + 1)a - 1]](c_1'', t_1^0[c_1, c_1'', (p + 1)a]) := b_1^{c_1'}(c_1''), \forall c_1'' \in C_1^\infty.$$

If $p = k^* - 1$, define type $t_2^0[c_1, c_1', k^*a - 1]$ to be such that

$$b_2[t_2^0[c_1, c_1', k^*a - 1]](c_1'', t_1^0[c_1'', c_1'', 0]) := b_1^{c_1'}(c_1''), \forall c_1'' \in C_1^\infty.$$

Note that by construction of Step 1, we have for each $p \in \{1, ..., k^*\}$ that $b_1^{c_1'}(c_1'') = b_2[t_2^0[c_1, c_1', pa - 1]](c_1'', t_1^0[c_1, c_1'', pa])$ for each $c_1'' \in C_1^\infty$. Moreover, all other types induce a probability one belief. This implies that type $t_1^0[c_1, c_1, 0]$ induces exactly the same a-th order expectation as type $t_1^{c_1}[c_1]$ did in Step 1. Similarly, each type $t_1^0[c_1, c_1', pa]$ induces the same a-th order expectation as type $t_1^{c_1, pa}[c_1']$ did. So for Iteration 0 we essentially take a copy of the epistemic model created in Step 1, but fill in the beliefs that were still incomplete from this step.

Iteration n $\geq$ 1 : For each choice $c_1 \in C_1^\infty$ and each choice $c_1' \in C_1^\infty$ define type $t_2^n[c_1, c_1', k^*a - 1]$ to be such that

$$b_2[t_2^n[c_1, c_1', k^*a - 1]](c_1'', t_1^{n-1}[c_1'', c_1'', 0]) := b_1^{c_1'}(c_1''), \forall c_1'' \in C_1^\infty.$$

For each $c_1, c_1' \in C_1^\infty$, we then also define

$$d^n[c_1, c_1', k^*a{-}1] := c_2', \text{ with } c_2' \text{ optimal given the type } t_2^n[c_1, c_1', k^*a{-}1].$$

Now, for each pair of choices $c_1, c_1' \in C_1^\infty$, define recursively for each *even* $k \in \{2, ..., a-2\}$ starting at $k = a-2$, type $t_1^n[c_1, c_1', (k^*-1)a+k]$ to be such that

$$b_1[t_1^n[c_1, c_1', (k^*{-}1)a{+}k]] := (d^n[c_1, c_1', (k^*{-}1)a{+}k{+}1], t_2^n[c_1, c_1', (k^*{-}1)a{+}k{+}1]).$$

Second, also define

$$d^n[c_1, c_1', (k^*-1)a+k] := \bar{c}_1, \text{ with } \bar{c}_1 \text{ optimal given the type } t_1^n[c_1, c_1', (k^*-1)a+k].$$

Third, define type $t_2^n[c_1, c_1', (k^*-1)a+k-1]$ to be such that

$$b_2[t_2^n[c_1, c_1', (k^*{-}1)a{+}k{-}1]] := (d^n[c_1, c_1', (k^*{-}1)a{+}k], t_1^n[c_1, c_1', (k^*{-}1)a{+}k]).$$

Fourth, also define

$$d^n[c_1, c_1', (k^*-1)a+k-1] := c_2', \text{ with } c_2' \text{ optimal given the type } t_2^n[c_1, c_1', (k^*-1)a+k-1].$$

Finally, for each $c_1, c_1' \in C_1^\infty$ define type $t_1^n[c_1, c_1', (k^*-1)a]$ to be such that

$$b_1[t_1^n[c_1, c_1', (k^*{-}1)a]] := (d^n[c_1, c_1', (k^*{-}1)a{+}1], t_2^n[c_1, c_1', (k^*{-}1)a{+}1]),$$

and define

$$d^n[c_1, c_1', (k^*-1)a], (k^*-1)a] := c_1'.$$

Next, for each $p \in \{0, ..., k^*-2\}$, do the following iteratively, going backwards starting at $p = k^*-2$: for each choice $c_1 \in C_1^\infty$ and each

choice $c_1' \in C_1^\infty$ define type $t_2^n[c_1, c_1', (p+1)a-1]$ to be such that

$$b_2[t_2^n[c_1, c_1', (p+1)a-1]](c_1'', t_1^n[c_1, c_1'', (p+1)a]) := b_1^{c_1'}(c_1''), \forall c_1'' \in C_1^\infty.$$

For each $c_1, c_1' \in C_1^\infty$, we then also define

$$d^n[c_1, c_1', (p+1)a-1] := c_2', \text{ with } c_2' \text{ optimal given the type } t_2^n[c_1, c_1', (p+1)a-1].$$

Now, for each pair of choices $c_1, c_1' \in C_1^\infty$, define recursively for each *even* $k \in \{2, ..., a-2\}$ starting at $k = a-2$, type $t_1^n[c_1, c_1', pa+k]$ to be such that

$$b_1[t_1^n[c_1, c_1', pa+k]] := (d^n[c_1, c_1', pa+k+1], t_2^n[c_1, c_1', pa+k+1]).$$

Second, also define

$$d^n[c_1, c_1', pa+k] := \bar{c}_1, \text{ with } \bar{c}_1 \text{ optimal given the type } t_1^n[c_1, c_1', pa+k].$$

Third, define type $t_2^n[c_1, c_1', pa+k-1]$ to be such that

$$b_2[t_2^n[c_1, c_1', pa+k-1]] := (d^n[c_1, c_1', pa+k], t_1^n[c_1, c_1', pa+k]).$$

Fourth, also define

$$d^n[c_1, c_1', pa+k-1] := c_2', \text{ with } c_2' \text{ optimal given the type } t_2^n[c_1, c_1', pa+k-1].$$

Finally, for each $c_1, c_1' \in C_1^\infty$ define type $t_1^n[c_1, c_1', pa]$ to be such that

$$b_1[t_1^n[c_1, c_1', pa]] := (d^n[c_1, c_1', pa+1], t_2^n[c_1, c_1', pa+1]),$$

and define

$$d^n[c_1, c_1', pa] := c_1'.$$

We do this iteratively for each $p \in \{0, ..., k^* - 2\}$, starting at $p = k^* - 2$.

We have that C_1^∞ and C_2^∞ are finite sets. Moreover, a and k^* are finite orders of belief, and therefore k^*a is as well. Hence, there are iterations m, n with $m > n$ such that:

$$d^m[c_1, c_1, k] = d^n[c_1, c_1, k],\ \forall c_1 \in C_1^\infty, k \in \{0, 1, ..., a-1\},$$

and

$$d^m[c_1, c_1', pa+k] = d^n[c_1, c_1', pa+k],\ \forall c_1, c_1' \in C_1^\infty, k \in \{0, 1, ..., a-1\}, p \in \{1, ..., k^*-1\}.$$

When we find such iterations m and n, we stop the recursive procedure.

Now we create the epistemic model $\mathcal{M}^*$ from the types we have constructed in our recursive procedure. Define $T_1(l) := \{t_1^l[c_1, c_1, k] : c_1 \in C_1^\infty, k \in \{0, ..., a-2\}$ even$\} \cup \{t_1^l[c_1, c_1', pa+k] : c_1, c_1' \in C_1^\infty,\ p \in \{1, ..., k^*-1\},\ k \in \{0, ..., a-2\}$ even$\}$ and $T_2(l) := \{t_2^l[c_1, c_1, k] : c_1 \in C_1^\infty, k \in \{1, ..., a-1\}$ odd$\} \cup \{t_2^l[c_1, c_1', pa+k] : c_1, c_1' \in C_1^\infty,\ p \in \{1, ..., k^*-1\},\ k \in \{1, ..., a-1\}$ odd$\}$. Then, let $T(l) := T_1(l) \cup T_2(l)$. Do this for every $l \in \{n, ..., m\}$.

In $T(n+1)$ specifically, we re-define for each $c_1, c_1' \in C_1^\infty$ the type $t_2^{n+1}[c_1, c_1', k^*a-1]$ to be such that

$$b_2[t_2^{n+1}[c_1, c_1', k^*a-1]](c_1'', t_1^m[c_1'', c_1'', 0]) := b_1^{c_1'}(c_1''), \forall c_1'' \in C_1^\infty.$$

So instead of assigning positive probability to types in $T(n)$, each type $t_2^{n+1}[c_1, c_1', k^*a-1]$ now assigns positive probability to types in $T(m)$. Then define

$$\mathcal{M}^* := (\bigcup\nolimits_{l \in \{n+1, ..., m\}} T_i(l), b[t_i])_{i \in \{1,2\}}.$$

We will show that each type in $\mathcal{M}^*$ expresses common belief in rationality. We will do so in steps.

First, we can note that for each $c_1 \in C_1^\infty$ and each $l \in \{n+1, ..., m\}$ in $\mathcal{M}^*$, choice c_1 is optimal for type $t_1^l[c_1, c_1, 0]$, and that for each $c_1, c_1' \in C_1^\infty$, each $l \in \{n+1, ..., m\}$ and each $p \in \{1, ..., k^* - 1\}$ choice c_1' is optimal for type $t_1^l[c_1, c_1', pa]$.

Namely, from type $t_1^l[c_1, c_1, 0]$ there follows a sequence of probability one beliefs, induced by the sequence of types $(t_1^l[c_1, c_1, 0], t_2^l[c_1, c_1, 1], ..., t_1^l[c_1, c_1, a-2])$. This sequence of probability one beliefs ends at type $t_2^l[c_1, c_1, a-1]$. By construction, we have that

$$marg_{C_1^\infty} b_2[t_2^l[c_1, c_1, a-1]] = b_1^{c_1}.$$

It follows then that type $t_1^l[c_1, c_1, 0]$ induces an a-th order expectation that is equal to $b_1^{c_1}$. We constructed $b_1^{c_1}$ such that c_1 is optimal given $b_1^{c_1}$. Hence c_1 is optimal given type $t_1^l[c_1, c_1, 0]$. This goes for every $l \in \{n+1, ..., m\}$.

Similarly for each $p \in \{1, ..., k^* - 1\}$, from type $t_1^l[c_1, c_1', pa]$ there follows a sequence of probability one beliefs, induced by the sequence of types $(t_1^l[c_1, c_1', pa], t_2^l[c_1, c_1', pa+1], ..., t_1^l[c_1, c_1', (p+1)a-2])$. This sequence of probability one beliefs ends at type $t_2^l[c_1, c_1', (p+1)a-1]$. By construction, we have that

$$marg_{C_1^\infty} b_2[t_2^l[c_1, c_1', (p+1)a-1]] = b_1^{c_1'}.$$

It follows then that type $t_1^l[c_1, c_1', pa]$ induces an a-th order expectation that is equal to $b_1^{c_1'}$. We constructed $b_1^{c_1'}$ such that c_1' is optimal given $b_1^{c_1'}$. Hence c_1' is optimal given type $t_1^l[c_1, c_1', pa]$. This goes for every $l \in \{n+1, ..., m\}$.

Second, we can also show the following is true.

Claim 5.2. *Consider the epistemic model $\mathcal{M}^*$. For each $l \in \{n+1, ..., m\}$, each $k \in \{1, 2, ..., a-1\}$ and each $c_1 \in C_1^\infty$, choice $d^l[c_1, c_1, k]$ is optimal given the type $t_i^l[c_1, c_1, k]$ with $i \in \{1, 2\}$. Moreover, for each $p \in \{1, ..., k^* - 1\}$, each $l \in \{n+1, ..., m\}$, each $k \in \{1, 2, ..., a-1\}$ and each $c_1, c_1' \in C_1^\infty$, choice $d^l[c_1, c_1', pa+k]$ is optimal given the type $t_i^l[c_1, c_1', pa+k]$ with $i \in \{1, 2\}$.*

Proof of claim. We start off with the epistemic model we created when ending the recursive procedure, but *before* $\mathcal{M}^*$ was created.

For each $k \in \{0, 1, ..., a-2\}$ and each $c_1' \in C_1^\infty$ we have by construction that

$$b_i[t_i^n[c_1', c_1', k]](d^n[c_1', c_1', k+1], t_j^n[c_1', c_1', k+1]) = 1 = b_i[t_i^m[c_1', c_1', k]](d^m[c_1', c_1', k+1], t_j^m[c_1', c_1', k+1]),$$

with $d^n[c_1', c_1', k+1] = d^m[c_1', c_1', k+1]$. Note that these were the n and m that determined when to stop our recursive procedure. Moreover, for each $k \in \{0, 1, ..., a-2\}$, each $c_1', \bar{c}_1 \in C_1^\infty$ and each $p \in \{1, ..., k^* - 1\}$ we also have by construction

$$b_i[t_i^n[c_1', \bar{c}_1, pa+k]](d^n[c_1', \bar{c}_1, pa+k+1], t_j^n[c_1', \bar{c}_1, pa+k+1]) = 1 = b_i[t_i^m[c_1', \bar{c}_1, pa+k]](d^m[c_1', \bar{c}_1, pa+k+1], t_j^m[c_1', \bar{c}_1, pa+k+1]),$$

with $d^n[c_1', \bar{c}_1, pa+k+1] = d^m[c_1', \bar{c}_1, pa+k+1]$. Additionally, we have by construction that

$$b_2[t_2^n[c_1', c_1', a-1]](d^n[c_1', c_1'', a], t_1^n[c_1', c_1'', a]) = b_1^{c_1'}[c_1''] = b_2[t_2^m[c_1', c_1', a-1]](d^m[c_1', c_1'', a], t_1^m[c_1', c_1'', a]),$$

for each $c_1'' \in C_1^\infty$. For each $p \in \{1, ..., k^* - 2\}$ we also have that

$$b_2[t_2^n[c_1', \bar{c}_1, pa-1]](d^n[c_1', c_1'', pa], t_1^n[c_1', c_1'', pa]) = b_1^{\bar{c}_1}[c_1''] = b_2[t_2^m[c_1', \bar{c}_1, pa-1]](d^m[c_1', c_1'', pa], t_1^m[c_1', c_1'', pa]),$$

for each $c_1'' \in C_1^\infty$. Finally, we have that

$$b_2[t_2^n[c_1', \bar{c}_1, k^*a - 1]](d^{n-1}[c_1'', c_1'', 0], t_1^{n-1}[c_1'', c_1'', 0]) = b_1^{\bar{c}_1}(c_1'') =$$
$$b_2[t_2^m[c_1', \bar{c}_1, k^*a - 1]](d^{m-1}[c_1'', c_1'', 0], t_1^{m-1}[c_1'', c_1'', 0]),$$

for each $c_1'' \in C_1^\infty$.

Then, for each $c_1' \in C_1^\infty$, the pair of types $t_1^m[c_1', c_1', 0]$ and $t_1^n[c_1', c_1', 0]$ induce the same k^*a-th order *belief*. To see why this is the case, we can employ a recursive argument, for each $p \in \{1, ..., k^* - 1\}$ starting at $p = k^* - 1$.

We can first note that the pair of types
$t_1^m[c_1', \bar{c}_1, (k^* - 1)a]$ and $t_1^n[c_1', \bar{c}_1, (k^* - 1)a]$ for each $c_1', \bar{c}_1 \in C_1^\infty$ induce the same a-th order belief. Namely, from the beginning of the proof of this claim we know that types $t_i^m[c_1', \bar{c}_1, (k^* - 1)a + k]$ and $t_i^n[c_1', \bar{c}_1, (k^* - 1)a + k]$ with $i \in \{1, 2\}$ for each $k \in \{1, ..., a - 2\}$ induce a probability one belief. Moreover, the first-order belief induced by type $t_i^m[c_1', \bar{c}_1, (k^* - 1)a + k]$ for each $k \in \{1, ..., a - 1\}$ is equal to the first-order belief induced by type $t_i^n[c_1', \bar{c}_1, (k^* - 1)a + k]$. As a result, types $t_1^m[c_1', \bar{c}_1, (k^* - 1)a]$ and $t_1^n[c_1', \bar{c}_1, (k^* - 1)a]$ induce the same a-th order *belief*.

Now recall, for each $c_1', \bar{c}_1 \in C_1^\infty$, we have that

$$b_2[t_2^n[c_1', \bar{c}_1, (k^* - 1)a - 1]](d^n[c_1', c_1'', (k^* - 1)a], t_1^n[c_1', c_1'', (k^* - 1)a]) =$$
$$b_1^{\bar{c}_1}(c_1'') =$$
$$b_2[t_2^m[c_1', \bar{c}_1, (k^* - 1)a - 1]](d^m[c_1', c_1'', (k^* - 1)a], t_1^m[c_1', c_1'', (k^* - 1)a])$$

Both types $t_2^m[c_1', \bar{c}_1, (k^* - 1)a - 1]$ and $t_2^n[c_1', \bar{c}_1, (k^* - 1)a - 1]$ thus assign exactly the same probability to choice-type combinations where the choice is equal and the type induces the same a-th order belief. Hence, both types induce the same $(a + 1)$-th order belief.

Now we can employ our recursive argument, starting at $p = k^* - 2$. For $p \in \{1, ..., k^* - 2\}$, assume that types $t_2^m[c_1', \bar{c}_1, (p + 1)a - 1]$ and

$t_2^n[c_1', \bar{c}_1, (p+1)a-1]$ induce the same $((k^*-p-1)a+1)$-th order belief. Then types $t_1^m[c_1', \bar{c}_1, pa]$ and $t_1^n[c_1', \bar{c}_1, pa]$ induce the same $(k^*-p)a$-th order belief. Namely, from the beginning of the proof of this claim we have that types $t_i^m[c_1', \bar{c}_1, pa+k]$ and $t_i^n[c_1', \bar{c}_1, pa+k]$ with $i \in \{1,2\}$ for each $k \in \{1, ..., a-2\}$ induce a probability one belief and moreover induce the same first-order belief. Therefore, types $t_1^m[c_1', \bar{c}_1, pa]$ and $t_1^n[c_1', \bar{c}_1, pa]$ induce the same $(a-1)$-th order belief. Additionally, types $t_1^m[c_1', \bar{c}_1, (p+1)a-2]$ and $t_1^n[c_1', \bar{c}_1, (p+1)a-2]$ assign probability one to types that by assumption induce the same $((k^*-p-1)a+1)$-th order belief. It follows then that types $t_1^m[c_1', \bar{c}_1, pa]$ and $t_1^n[c_1', \bar{c}_1, pa]$ induce the same $(k^*-p)a$-th order belief.

Now recall that for each $c_1', \bar{c}_1 \in C_1^\infty$, we have that

$$b_2[t_2^n[c_1', \bar{c}_1, pa-1]](d^n[c_1', c_1'', pa], t_1^n[c_1', c_1'', pa]) = b_1^{\bar{c}_1}(c_1'') =$$
$$b_2[t_2^m[c_1', \bar{c}_1, pa-1]](d^m[c_1', c_1'', pa], t_1^m[c_1', c_1'', pa])$$

Both types $t_2^m[c_1', \bar{c}_1, pa-1]$ and $t_2^n[c_1', \bar{c}_1, pa-1]$ thus assign exactly the same probability to choice-type combinations where the choice is equal and the type induces the same $(k^*-p)a$-th order belief. Hence, both types induce the same $((k^*-p)a+1)$-th order belief.

Following the same argument, we can establish that types $t_1^m[c_1', c_1', 0]$ and $t_1^n[c_1', c_1', 0]$ induce the same k^*a-th order belief. From the above we know that types $t_2^m[c_1', c_1', a-1]$ and $t_2^n[c_1', c_1', a-1]$ induce the same $((k^*-1)a+1)$-th order belief. From the beginning of the proof of this claim we have that types $t_i^m[c_1', c_1', k]$ and $t_i^n[c_1', c_1', k]$ with $i \in \{1,2\}$ for each $k \in \{1, ..., a-2\}$ induce a probability one belief and moreover induce the same first-order belief. Therefore, they induce the same $(a-1)$-th order belief. Additionally, types $t_1^m[c_1', c_1', a-2]$ and $t_1^n[c_1', c_1', a-2]$ assign probability one to types that by the above recursive argument induce the same $((k^*-1)a+1)$ order belief. It follows then that types $t_1^m[c_1', c_1', 0]$ and $t_1^n[c_1', c_1', 0]$ induce the same k^*a-th order belief. This goes for each $c_1' \in C_1^\infty$.

Denote type $t_2^{n+1}[c_1, \bar{c}_1, k^*a-1]$ that results from our recursive back-

wards procedure but *before* constructing $\mathcal{M}^*$ by $\bar{t}_2^{n+1}[c_1, \bar{c}_1, k^*a - 1]$. In contrast, let the same type that does result from constructing $\mathcal{M}^*$ still be denoted as $t_2^{n+1}[c_1, \bar{c}_1, k^*a - 1]$. Now, we have for each $c_1, \bar{c}_1 \in C_1^\infty$

$$b_2[\bar{t}_2^{n+1}[c_1, \bar{c}_1, k^*a - 1]](c_1', t_2^n[c_1', c_1', 0]) =$$
$$b_2[t_2^{n+1}[c_1, \bar{c}_1, k^*a - 1]](c_1', t_2^m[c_1', c_1', 0]), \forall c_1' \in C_1^\infty.$$

It thus follows that each such type $t_2^{n+1}[c_1, \bar{c}_1, k^*a-1]$ induces the same (k^*a+1)-th order belief in $\mathcal{M}^*$ as it did before $\mathcal{M}^*$ was constructed. All the remaining types in $\bigcup_{l\in\{n+1,...,m\}} T(l)$ remained unchanged when $\mathcal{M}^*$ was constructed: they induce exactly the same belief over choice-type combinations as before. As a result, all types in $\bigcup_{l\in\{n+1,...,m\}} T(l)$ induce at least the same $(k^*a + 1)$-th order belief in $\mathcal{M}^*$ as before $\mathcal{M}^*$ was constructed.

In our backward construction procedure of types and choices, before creating $\mathcal{M}^*$, we constructed each $d^l[c_1, c_1', k]$ for each $l \in \{n + 1, ..., m\}$, $k \in \{1, ..., k^*a - 1\}$ and $c_1, c_1' \in C_1^\infty$ such that it is optimal given type $t_i^l[c_1, c_1', k]$. Now, we have that the maximum directly utility-relevant order of belief for any player is k^* and that each type $t_i^l[c_1, c_1', k]$ at least induces exactly the same $(k^*a + 1)$-th order belief in $\mathcal{M}^*$ as it did before constructing $\mathcal{M}^*$. Hence, we also have in $\mathcal{M}^*$ that $d^l[c_1, c_1', k]$ is optimal given $t_i^l[c_1, c_1', k]$. This completes the proof of this claim.

Since each type in $\mathcal{M}^*$ only assigns positive probability to choice-type combinations $(d^l[c_1, c_1', k], t_i^l[c_1, c_1', k]])$ for $k \in \{0, 1, ..., k^*a - 1\}$, each type only assigns positive probability to choice-type combinations where the choice is optimal given the type. Hence each type in $\mathcal{M}^*$ expresses 1-fold belief in rationality. Therefore also each type in $\mathcal{M}^*$ expresses common belief in rationality.

By our backward, recursive construction, we moreover have that type $t_1^m[c_1, c_1, 0]$ induces an a-th order expectation $b_1^{c_1}$. By construction of Step 1, choice c_1 is optimal given such a higher-order expectation. Hence we have constructed an epistemic model with a type that expresses

common belief in rationality and is such that c_1 is optimal given that type.

In Step 1 we have shown that for every choice $c_1 \in C_1^\infty$ we can construct a partial epistemic model with a type that expresses on-path belief in rationality and that is such that choice c_1 is optimal. In Step 2 we showed that we are then also able to construct a finite, epistemic model with a type that expresses common belief in rationality and that is such that choice c_1 is optimal. This concludes the proof for Scenario (i).

Scenario (ii)
Step 1
Next consider scenario (ii) with $N_1 = \{a\}$ and $N_2 = \{z\}$, a, z odd. In this Step 1, we will construct a partial epistemic model.

For each choice $c_1 \in C_1^\infty$, fix an a-th order expectation $b_1^{c_1} \in \Delta(C_2^\infty)$ for which c_1 is optimal.[4] The reason we can do so is as follows. From Lemma 5.1 we know that for each choice that is not strictly dominated in a decision problem, we can find a belief in that decision problem such that the relevant choice is optimal. The final reduced decision problem resulting from the IESDC-procedure leaves the choices in C_1^∞ for player 1 in the decision problem and the choices in C_2^∞ for player 2. That is, we have as a reduced decision problem after following through with the IESDC-procedure: $(C_1^\infty, C_2^\infty, v_i)$. By Lemma 5.1, we should then have that each choice in C_1^∞ is optimal for some a-th order expectation in $\Delta(C_2^\infty)$. Then, for each $c_1 \in C_1^\infty$, construct a type $t_1^{c_1}$. Similarly, also for each choice $c_2 \in C_2^\infty$, fix a z-th order expectation $b_2^{c_2} \in \Delta(C_1^\infty)$ for which c_2 is optimal. Again, we can do so for the reasons explained above, but then from player 2's perspective. Then, for each $c_2 \in C_2^\infty$, construct a type $t_2^{c_2}$.

[4] With a-th order expectation in this context we specifically refer to $marg_{C_2} e_1^a \in \Delta(C_2)$ where $e_1^a \in \Delta(W_1^{a-1} \times C_2)$.

Take for each $t_1^{c_1}$ a sequence of types $(t_1^{c_1}, t_2^{c_1,1}, ..., t_1^{c_1,a-1})$. Then, let us have in each such sequence that type $t_1^{c_1}$ assigns probability one to type $t_2^{c_1,1}$ if $a > 1$, and type $t_i^{c_1,n}$ probability one to type $t_j^{c_1,n+1}$ for each $n \in \{1, 2, ..., a-2\}$ and with $i \in \{1, 2\}$ and $j \neq i$. Note that if $a = 1$, we treat type $t_1^{c_1}$ such that $t_1^{c_1} = t_1^{c_1,a-1}$. Similarly, take for each $t_2^{c_2}$ a sequence of types $(t_2^{c_2}, t_1^{c_2,1}, ..., t_2^{c_2,z-1})$. Then, for each c_2, let us have in each such sequence that type $t_2^{c_2}$ assigns probability one to type $t_1^{c_2,1}$ if $z > 1$, and type $t_i^{c_2,n}$ probability one to type $t_j^{c_2,n+1}$ for each $n \in \{1, 2, ..., z-2\}$. Again, if $z = 1$, we treat type $t_2^{c_2}$ such that $t_2^{c_2} = t_2^{c_2,z-1}$.

Finally, for each $c_1 \in C_1^\infty$ define type $t_1^{c_1,a-1}$ to be such that, for each $c_2' \in C_2^\infty$,

$$b_1[t_1^{c_1,a-1}](c_2', t_2^{c_2''}) := \begin{cases} b_1^{c_1}(c_2'), \text{ if } c_2' = c_2'' \\ 0, \text{ otherwise.} \end{cases}$$

So we have that the distribution over choices induced by type $t_1^{c_1,a-1}$ is equal to the distribution over choices represented by the expectation $b_1^{c_1}$. From type $t_1^{c_1}$ there follows a sequence of probability one beliefs up to type $t_1^{c_1,a-1}$. The a-th order expectation induced by type $t_1^{c_1}$ is thus equal to $b_1^{c_1}$. Hence, choice c_1 is optimal given type $t_1^{c_1}$.

Similarly, for each $c_2 \in C_2^\infty$ define type $t_2^{c_2,z-1}$ to be such that, for each $c_1' \in C_1^\infty$

$$b_2[t_2^{c_2,z-1}](c_1', t_1^{c_1''}) := \begin{cases} b_2^{c_2}(c_1'), \text{ if } c_1' = c_1'' \\ 0, \text{ otherwise.} \end{cases}$$

So we have that the distribution over choices induced by type $t_2^{c_2,z-1}$ is equal to the distribution over choices represented by the expectation $b_2^{c_2}$. From type $t_2^{c_2}$ there follows a sequence of probability one beliefs up to type $t_2^{c_2,z-1}$. Similarly as before, then type $t_2^{c_2}$ then induces a z-th order expectation that is equal to $b_2^{c_2}$. Then choice c_2 is optimal given type $t_2^{c_2}$.

Call the resulting partial epistemic model $\mathcal{M}$. By construction, we have

for each $c_1 \in C_1^\infty$ that c_1 is optimal given $t_1^{c_1}$, and we have for each $c_2 \in C_2^\infty$ that c_2 is optimal for $t_2^{c_2}$. Moreover, each $t_1^{c_1}$ also expresses on-path belief in rationality. This is because the type $t_1^{c_1,a-1}$ only assigns positive probability to choice type pairs $(c_2', t_2^{c_2'})$. In these pairs the choice is optimal for the type by construction. For similar reasons, each type $t_2^{c_2}$ also expresses on-path belief in rationality.

Step 2
In Step 1 we have shown that for each choice $c_1 \in C_1^\infty$ we can always construct a *partial* epistemic model with a type $t_1^{c_1}$ for which c_1 is optimal and that expresses on-path belief in rationality. In Step 2 we will now do the following. We will show that if there exists a belief hierarchy expressing on-path belief in rationality for which c_1 is optimal, then there is also a belief hierarchy expressing common belief in rationality for which c_1 is optimal.

Consider a partial epistemic model $\mathcal{M}$ as constructed in Step 1 with a type $t_1^{c_1}$ that expresses on-path belief in rationality and for which c_1 is optimal. By means of a backward, recursive procedure we transform this partial epistemic model such that we get to a new, complete epistemic model that now includes a completed type $t_1^m[c_1, 0]$ that expresses common belief in rationality and induces the same a-th order expectation as type $t_1^{c_1}$ does. The recursive procedure in each iteration defines combinations of choices *and* types at the same time.

The recursive procedure is as follows.

Iteration 0: For each choice $c_1 \in C_1^\infty$, define

$$d^0[c_1, 0] := c_1.$$

Moreover, for each choice $c_1 \in C_1^\infty$ and each $k \in \{1, 2, ..., a-1\}$, define $d^0[c_1, k]$ randomly. So

$d^0[c_1, k] := c'$, for *some* $c' \in C_1^\infty$ if k is even or *some* $c' \in C_2^\infty$ if k is odd.

Now, take a sequence of types $(t_1^0[c_1,0], t_2^0[c_1,1], ..., t_1^0[c_1,a-1])$ for every choice $c_1 \in C_1^\infty$. Define type $t_1^0[c_1,0]$ to be such that

$$b_1[t_1^0[c_1,0]] := (d^0[c_1,1], t_2^0[c_1,1]).$$

Then, define for each $k \in \{1,2,...,a-2\}$ type $t_i^0[c_1,k]$ with $i \in \{1,2\}$ to be such that

$$b_i[t_i^0[c_1,k]] := (d^0[c_1,k+1], t_j^0[c_1,k+1]).$$

Finally, we define for each choice $c_1 \in C_1^\infty$ type $t_1^0[c_1,a-1]$ to be such that

$$b_1[t_1^0[c_1,a-1]](c_2', t_2^0[c_2',0]) := b_1^{c_1}(c_2'), \forall c_2' \in C_2^\infty.$$

This implies that type $t_1^0[c_1,0]$ induces exactly the same a-th order expectation as type $t_1^{c_1}$ did in Step 1. So for Iteration 0 we essentially take a copy of the epistemic model created in Step 1, but fill in the beliefs that were still incomplete from this step.

We do a similar thing for each choice c_2. For each choice $c_2 \in C_2^\infty$, define

$$d^0[c_2,0] := c_2.$$

Moreover, for each choice $c_2 \in C_2^\infty$ and each $k \in \{1,2,...,z-1\}$, define $d^0[c_2,k]$ randomly. So

$d^0[c_2,k] := c'$, for *some* $c' \in C_2^\infty$ if k is even or *some* $c' \in C_1^\infty$ if k is odd.

Now, take a sequence of types $(t_2^0[c_2,0], t_1^0[c_2,1], ..., t_2^0[c_2,z-1])$ for every choice $c_2 \in C_2^\infty$. Define type $t_2^0[c_2,0]$ to be such that

$$b_2[t_2^0[c_2,0]] := (d^0[c_2,1], t_1^0[c_2,1]).$$

Then, define for each $k \in \{1,2,...,z-2\}$ type $t_i^0[c_2,k]$ with $i \in \{1,2\}$ to be such that

$$b_i[t_i^0[c_2,k]] := (d^0[c_2,k+1], t_j^0[c_2,k+1]).$$

Finally, we define for each choice $c_2 \in C_2^\infty$ type $t_2^0[c_2, z-1]$ to be such that

$$b_2[t_2^0[c_2, z-1]](c_1', t_1^0[c_1', 0]) := b_2^{c_2}(c_1'), \forall c_1' \in C_1^\infty.$$

Iteration n$\geq$ 1 : We define for each choice $c_2 \in C_2^\infty$ type $t_2^n[c_2, z-1]$ to be such that

$$b_2[t_2^n[c_2, z-1]](c_1', t_1^{n-1}[c_1', 0]) := b_2^{c_2}(c_1'), \forall c_1' \in C_1^\infty.$$

We also define

$$d^n[c_2, z-1] := c_2', \text{ with } c_2' \text{ optimal given the type } t_2^n[c_2, z-1].$$

Define recursively for each *odd* $k \in \{1, 2, ..., z-2\}$ starting at $k = z-2$, type $t_1^n[c_2, k]$ that is such that

$$b_1[t_1^n[c_2, k]] := (d^n[c_2, k+1], t_2^n[c_2, k+1]).$$

Second, also define

$$d^n[c_2, k] := c_1', \text{ with } c_1' \text{ optimal given the a-th order expectation induced by } t_1^n[c_2, k].$$

Third, define type $t_2^n[c_2, k-1]$ to be such that

$$b_2[t_2^n[c_2, k-1]] := (d^n[c_2, k], t_1^n[c_2, k]).$$

Fourth, also define

$$d^n[c_2, k-1] := c_2'', \text{ with } c_2'' \text{ optimal given the type } t_2^n[c_2, k-1].$$

Finally, for each choice $c_2 \in C_2^\infty$ define type $t_2^n[c_2, 0]$ to be such that

$$b_2[t_2^n[c_2, 0]] := (d^n[c_2, 1], t_1^n[c_2, 1]),$$

and define

$$d^n[c_2, 0] := c_2.$$

Next, we do exactly the same thing for choices c_1. We define for each choice $c_1 \in C_1^\infty$ type $t_1^n[c_1, a-1]$ to be such that

$$b_1[t_1^n[c_1, a-1]](c_2', t_2^n[c_2', 0]) := b_1^{c_1}(c_2'), \forall c_2' \in C_2^\infty.$$

For each choice $c_1 \in C_1^\infty$, we then also define

$$d^n[c_1, a-1] := c_1', \text{ with } c_1' \text{ optimal given the type } t_1^n[c_1, a-1].$$

Now, for each choice $c_1 \in C_1^\infty$, define recursively for each *odd* $k \in \{1, 2, ..., a-2\}$ starting at $k = a-2$, type $t_2^n[c_1, k]$ that is such that

$$b_2[t_2^n[c_1, k]] := (d^n[c_1, k+1], t_1^n[c_1, k+1]).$$

Second, also define

$$d^n[c_1, k] := c_2', \text{ with } c_2' \text{ optimal given the a-th order expectation induced by } t_2^n[c_1, k].$$

Third, define type $t_1^n[c_1, k-1]$ to be such that

$$b_1[t_1^n[c_1, k-1]] := (d^n[c_1, k], t_2^n[c_1, k]).$$

Fourth, also define

$$d^n[c_1, k-1] := c_1'', \text{ with } c_1'' \text{ optimal given the type } t_1^n[c_1, k-1].$$

Finally, for each choice $c_1 \in C_1^\infty$ define type $t_1^n[c_1, 0]$ to be such that

$$b_1[t_1^n[c_1, 0]] := (d^n[c_1, 1], t_2^n[c_1, 1]),$$

and define

$$d^n[c_1, 0] := c_1.$$

We have that C_1^∞ and C_2^∞ are finite sets. Moreover, a and z are both finite orders of belief. Hence, there are iterations m and n with $m > n$

such that

$$d^m[c_1, k] = d^n[c_1, k] \text{ and } d^m[c_2, l] = d^n[c_2, l],$$

for each $c_1 \in C_1^\infty$ and $k \in \{0, 1, ..., a-1\}$, and $c_2 \in C_1^\infty$ and $l \in \{0, 1, ..., z-1\}$ respectively. When we find such iterations m and n, we stop the procedure.

Next, we create the epistemic model $\mathcal{M}^*$ from the types we have constructed in our recursive procedure. Define $T_1(l) := \{t_1^l[c_1, k] : c_1 \in C_1^\infty, k \in \{2, ..., a-1\} \text{ even}\} \cup \{t_1^l[c_2, k] : c_2 \in C_2^\infty, k \in \{1, ..., z-2\} \text{ odd}\}$ and $T_2(l) := \{t_2^l[c_2, k] : c_2 \in C_2^\infty, k \in \{2, ..., z-1\} \text{ even}\} \cup \{t_2^l[c_1, k] : c_1 \in C_1^\infty, k \in \{1, ..., a-2\} \text{ odd}\}$. Then, let $T(l) := T_1(l) \cup T_2(l)$. Do this for every $l \in \{n, ..., m\}$.

In $T(n+1)$ specifically, we re-define for each $c_2 \in C_2^\infty$ the type $t_2^{n+1}[c_2, z-1]$. Re-define each such type $t_2^{n+1}[c_2, z-1]$ to be such that

$$b_2[t_2^{n+1}[c_2, z-1]](c_1', t_1^m[c_1', 0]) := b_2^{c_2}(c_1'), \forall c_1' \in C_1^\infty.$$

So instead of assigning positive probability to types in $T(n)$, each type $t_2^{n+1}[c_2, z-1]$ now assigns positive probability to types in $T(m)$. Then define

$$\mathcal{M}^* := (\bigcup\nolimits_{l \in \{n+1, ..., m\}} T_i(l), b[t_i])_{i \in \{1,2\}}.$$

We will show that each type in $\mathcal{M}^*$ expresses common belief in rationality. We will do so in steps.

First we note that choice c_1 is optimal for type $t_1^l[c_1, 0]$, for each $c_1 \in C_1^\infty$ and each $l \in \{n+1, ..., m\}$ in $\mathcal{M}^*$. Namely, from type $t_1^l[c_1, 0]$ there follows a sequence of probability one beliefs, induced by the sequence of types $(t_1^l[c_1, 0], t_2^l[c_1, 1], ..., t_2^l[c_1, a-2])$. This sequence of probability one beliefs ends at type $t_1^l[c_1, a-1]$. Note that if $a = 1$ we treat $t_1^l[c_1, 0]$

as if $t_1^l[c_1, 0] = t_1^l[c_1, a-1]$. By definition, we have that

$$marg_{C_2^\infty} b_1[t_1^l[c_1, a-1]] = b_1^{c_1}.$$

The a-th order expectation induced by type $t_1^l[c_1, 0]$ is thus equal to $b_1^{c_1}$. We constructed $b_1^{c_1}$ such that c_1 is optimal given $b_1^{c_1}$. Hence c_1 is optimal given type $t_1^l[c_1, 0]$. This goes for every $l \in \{n+1, ..., m\}$. For similar reasons, we have that for each $l \in \{n+1, ..., m\}$ and each $c_2 \in C_2^\infty$ that $t_2^l[c_2, 0]$ induces the same z-th order expectation as $b_2^{c_2}$ does. Hence, c_2 is optimal given type $t_2^l[c_2, 0]$.

Second, we can also show the following is true.

Claim 5.3. *Consider the epistemic model $\mathcal{M}^*$. For each $l \in \{n+1, ..., m\}$, each $c_1 \in C_1^\infty$ and each $k \in \{1, 2, ..., a-1\}$, each choice $d^l[c_1, k]$ is optimal given the type $t_i^l[c_1, k]$. Moreover, for each $l \in \{n+1, ..., m\}$, each $c_2 \in C_2^\infty$ and each $k \in \{1, 2, ..., z-1\}$, each choice $d^l[c_2, k]$ is optimal given the type $t_i^l[c_2, k]$.*

Proof of claim. For each $k \in \{0, 2, ..., a-2\}$ and each $c_1' \in C_1^\infty$ we have by construction that

$$b_i[t_i^n[c_1', k]](d^n[c_1', k+1], t_j^n[c_1', k+1]) = 1 =$$
$$b_i[t_i^m[c_1', k]](d^m[c_1', k+1], t_j^m[c_1', k+1])$$

with $d^m[c_1', k+1] = d^n[c_1', k+1]$. We similarly have for each $k \in \{0, 2, ..., z-2\}$ and each $c_2' \in C_2^\infty$ that

$$b_i[t_i^n[c_2', k]](d^n[c_2', k+1], t_j^n[c_2', k+1]) = 1 =$$
$$b_i[t_i^m[c_2', k]](d^m[c_2', k+1], t_j^m[c_2', k+1])$$

with $d^m[c_2', k+1] = d^n[c_2', k+1]$. Additionally, we have that

$$b_1[t_1^n[c_1', a-1]](c_2'', t_2^{n-1}[c_2'', 0]) = b_1^{c_1'}(c_1'') = b_1[t_1^m[c_1', a-1](c_2'', t_2^{m-1}[c_2'', 0]),$$

for each $c_2'' \in C_1^\infty$. And we have that

$$b_2[t_2^n[c_2', z-1]](c_1'', t_1^{n-1}[c_1'', 0]) = b_2^{c_2'}(c_1'') = b_2[t_2^m[c_2', z-1](c_1'', t_1^{m-1}[c_1'', 0]),$$

for each $c_1'' \in C_1^\infty$.

Then for each $c_1' \in C_1^\infty$ the pair of types $t_1^n[c_1', 0]$ and $t_1^m[c_1', 0]$ induce the same $(z+a)$-th order *belief*. To see this, we can first note that for each $c_2'' \in C_2^\infty$ the pair of types $t_2^n[c_2'', 0]$ and $t_2^m[c_2'', 0]$ induce the same z-th order belief. Namely, from the beginning of the proof of this claim we have that types $t_i^m[c_2'', k]$ and $t_i^n[c_2'', k]$ for each $k \in \{0, 1, ..., z-2\}$ induce a probability one belief. Moreover, the first-order belief induced by type $t_i^m[c_2'', k]$ for each $k \in \{0, 1, ..., z-1\}$ is equal to the first-order belief induced by $t_i^n[c_2'', k]$. As a result, types $t_2^m[c_2'', 0]$ and $t_2^n[c_2'', 0]$ in fact induce the same z-th order belief.

Recall that for each $c_1' \in C_1^\infty$ we have that

$$b_1[t_1^n[c_1', a-1]](c_2'', t_2^{n-1}[c_2'', 0]) = b_1^{c_1'}(c_1'') = b_1[t_1^m[c_1', a-1](c_2'', t_2^{m-1}[c_2'', 0]),$$

for each $c_2'' \in C_1^\infty$. Both types $t_1^m[c_1', a-1]$ and $t_1^n[c_1', a-1]$ thus assign exactly the same probability to choice-type combinations where the choice is equal and the type induces the same z-th order belief as established before. It follows that for each $c_1' \in C_1^\infty$ types $t_1^n[c_1', a-1]$ and $t_1^m[c_1', a-1]$ induce the same $(z+1)$-th order belief.

From the beginning of the proof of this claim we have that types $t_i^m[c_1', k]$ and $t_i^n[c_1', k]$ for each $k \in \{0, 1, ..., a-2\}$ induce a probability one belief.. These probability one beliefs end at types $t_1^m[c_1', a-1]$ and $t_1^n[c_1', a-1]$ respectively. We know these types induce the same $(z+1)$-th order expectation. Moreover, the first-order belief induced by type $t_i^m[c_1', k]$ for each $k \in \{0, 1, ..., a-2\}$ is equal to the first-order belief induced by $t_i^n[c_1', k]$. Taken together, types $t_1^m[c_1', 0]$ and $t_1^n[c_1', 0]$ in fact induce the same $(z+a)$-th order belief.

Denote type $t_2^{n+1}[c_2, z-1]$ that results from our recursive, backward procedure but *before* constructing $\mathcal{M}^*$ by $\bar{t}_2^{n+1}[c_2, z-1]$. Now, we have

that for each $c_2 \in C_2^\infty$

$$b_2[\bar{t}_2^{n+1}[c_2, z-1]](c_1', t_1^n[c_1', 0]) = b_2[t_2^{n+1}[c_2, z-1]](c_1', t_1^m[c_1', 0]), \forall c_1' \in C_1^\infty.$$

It follows that each such type $t_2^{n+1}[c_2, z-1]$ induces the same $(z + a + 1)$-th order belief in $\mathcal{M}^*$ as it did before $\mathcal{M}^*$ was constructed. All the remaining types in $\bigcup_{l \in \{n+1,...,m\}} T(l)$ remained unchanged when $\mathcal{M}^*$ was constructed: they induce exactly the same belief over choice-type combinations as before. As a result, all types in $\bigcup_{l \in \{n+1,...,m\}} T(l)$ induce at least the same $(z + a + 1)$-th order belief in $\mathcal{M}^*$ as before $\mathcal{M}^*$ was constructed.

In our backward construction procedure of types and choices, before constructing $\mathcal{M}^*$, we constructed $d^l[c_1, k]$ for each $l \in \{n+1, ..., m\}$, $k \in \{1, 2, ..., a-1\}$ and $c_1 \in C_1^\infty$ such that it is optimal given type $t_i^l[c_1, k]$. Similarly, we constructed $d^l[c_1, k]$ for each $l \in \{n+1, ..., m\}$, $k \in \{1, 2, ..., z-1\}$ and $c_2 \in C_2^\infty$ such that it is optimal given type $t_i^l[c_2, k]$. Now, we have that the maximum order of belief in which either of the players' utility is variable is $max(N_1 \cup N_2)$, which is either a or z. We also have that types $t_i^l[c_1, k]$ and $t_i^l[c_2, k]$ induce exactly the same $(z + a + 1)$-th order belief in $\mathcal{M}^*$ as before $\mathcal{M}^*$ was constructed. Hence, we also have in $\mathcal{M}^*$ that $d^l[c_1, k]$ is optimal given $t_i^l[c_1, k]$ and that $d^l[c_2, k]$ is optimal given $t_i^l[c_2, k]$. This completes the proof of this claim.

Because each type in the epistemic model $\mathcal{M}^*$ only assigns positive probability to choice-type combinations $(d^l[c_1, k], t_i^l[c_1, k])$ for $k \in \{0, 1, ..., a-1\}$ or $(d^l[c_2, k], t_i^l[c_2, k])$ for $k \in \{0, 1, ..., z-1\}$, each type only assigns positive probability to choice-type combinations where the choice is optimal given the type. Hence each type in $\mathcal{M}^*$ expresses 1-fold belief in rationality. Therefore also each type in $\mathcal{M}^*$ expresses common belief in rationality.

By our backward, recursive construction, we also have that type $t_1^m[c_1, 0]$ induces an a-th order expectation $b_1^{c_1}$. By construction of Step 1, choice

c_1 is optimal given such a higher-order expectation. Hence we have constructed an epistemic model with a type that expresses common belief in rationality and is such that c_1 is optimal given that type.

In Step 1 we have shown that for every choice $c_1 \in C_1^\infty$ we can construct a partial epistemic model with a type that expresses on-path belief in rationality and that is such that choice c_1 is optimal. In Step 2 we showed that we are then also able to construct a finite, epistemic model with a type that expresses common belief in rationality and that is such that choice c_1 is optimal. This concludes the proof for Scenario (ii).

Scenario (iii)
Step 1:
Finally, consider scenario (iii). This corresponds to case (iii) of Theorem 5.2. Here we have that $N_1 = \{a, b, ..., x\}$ consists of (possibly multiple) odd orders and that N_2 is of a single, even order such that in the resulting causality diagram for player 1 there are no overlapping paths. In this Step 1, we will construct a partial epistemic model. By *partial* we mean we only completely specify the beliefs for some particular types.

For each choice $c_1 \in C_1^\infty$, fix an expectation $b_1^{c_1}$ which is a probability distribution over the product-space $C_2^{a,\infty} \times C_2^{b,\infty} \times ... \times C_2^{x,\infty}$ and for which c_1 is optimal. The reason we can do so is as follows. From Lemma 5.1 we know that for each choice that is not strictly dominated in a decision problem, we can find a belief in that decision problem such that the relevant choice is optimal. The final reduced problem decision resulting from the IESDC-procedure leaves the choices in C_2^∞ for player 2 in the decision problem. By Lemma 5.1, we should then have that each choice in C_1^∞ is optimal for some belief in the set $\Delta(C_2^{a,\infty} \times C_2^{b,\infty} \times ... \times C_2^{x,\infty})$. Each letter in the superscripts of the

product space $C_2^{a,\infty} \times C_2^{b,\infty} \times ... \times C_2^{x,\infty}$ refers to an order of belief in N_1.

Then, fix a type $t_1^{c_1}[c_1]$ for choice c_1.

The lowest order in N_1 is order a. Let $N_2 = \{z\}$ and take $a + z$. For each remaining order $p \in N_1$, subtract a multiple $n \in \mathbb{N}$ of z from order p such that $p - n \cdot z \in \{a, ..., a + z\}$. Call this order a^p. Note here that, by how case (iii) in Theorem 5.2 is defined, for any orders $b, c \in N_1$ with $b, c \neq a$, we have that $b - n \cdot z \neq c - m \cdot z$, for any combination of n, m. Hence, $a^b \neq a^c$ for any two different orders $b, c \in N_1$.

Take some combination of choices for player 2 $(c^a, c^b, ..., c^x) \in supp(b_1^{c_1})$. For each order $p \in N_1$ we do the following: take choice c^p in $C_2^{p,\infty}$. Let

$$b_2^{c^{p-z}} := c^p$$

be the z-th order expectation for player 2 that puts probability one on c^p.[5] Then there is a choice c^{p-z} in $C_2^{p-z,\infty}$ such that c^{p-z} is optimal given $b_2^{c^{p-z}}$. This follows from the construction of the IESDC-procedure. Namely, every choice $c_2 \in C_2$ that is optimal for some belief $b_2 \in \Delta(C_2^\infty)$ is in C_2^∞.

Next, for each $n \geq 1$, up until we have $p - n \cdot z \in \{a, ..., a + z\}$, we can do the same as we did for choice c^p. Take choice $c^{p-(n-1)z}$. Let

$$b_2^{c^{p-nz}} := c^{p-(n-1)z},$$

be the z-th order expectation for player 2 that puts probability one on $c^{p-(n-1)z}$. Then following the same argument as before there is a choice c^{p-nz} in $C_2^{p-nz,\infty}$ such that c^{p-nz} is optimal given $b_2^{c^{p-nz}}$. We can do this for any $p > a + z$ with $p \in N_1$, up until we have the choice c^{a^p} in $C_2^{a^p,\infty}$.

Each choice c^{p-nz}, given any p and any n, is a choice in C_2^∞. For these

[5] With z-th order expectation in this context we specifically refer to $marg_{C_2} e_2^z \in \Delta(C_2)$ where $e_2^z \in \Delta(W_2^{z-1} \times C_2)$.

choices, we fix the z-th order expectation $b_2^{c^{p-nz}}$ we just constructed before. For all the remaining choices c_2 in C_2^∞, we fix some z-th order expectation b^{c_2} such that c_2 is optimal given $b_2^{c_2}$. Again, we can do so by Lemma 5.1.

The next step is to move on to the construction of types. For *each* combination of choices $\bar{c} = (c^a, c^{a^b}, ..., c^{a^x})$ that results from the construction above, create a type $t_2^{\bar{c}}[\bar{c}]$. We specifically say '*each*', as the support of $b_1^{c_1}$ may include multiple combinations of choices $(c^a, c^b, ..., c^x)$.

For each combination of choices $\bar{c}$, take a sequence of types $(t_2^{\bar{c}}[\bar{c}], t_1^{\bar{c},1}[\bar{c}], ..., t_1^{\bar{c},z-1}[\bar{c}])$. Then in each such sequence let us have that type $t_2^{\bar{c}}[\bar{c}]$ assigns probability one to type $t_1^{\bar{c},1}[\bar{c}]$. Also in each such sequence, let us have, for each $n \in \{1, 2, ..., z-1\}$ that type $t_i^{\bar{c},n}[\bar{c}]$ with $i \in \{i, j\}$ assigns probability one to type $t_j^{\bar{c},n+1}[\bar{c}]$ with $j \neq i$. Additionally, for each $k = a^p - a - 1$, define type $t_1^{\bar{c},k}[\bar{c}]$ such that

$$b_1[t_1^{\bar{c},k}[\bar{c}]] := (c^{a^p}, t_2^{\bar{c},k}[\bar{c}]).$$

Finally we specify the belief that type $t_1^{\bar{c},z-1}[\bar{c}]$ in the sequence induces. First construct for each $\bar{c} \in C_2^{a,\infty} \times ... \times C_2^{a^x,\infty}$ and each $\bar{c}' \in C_2^{a,\infty} \times ... \times C_2^{a^x,\infty}$ a type $t_2^{\bar{c},z}[\bar{c}']$. Then, consider the joint probability distribution $b^{\bar{c}} \in \Delta(C_2^{a,\infty} \times C_2^{a^b,\infty} \times ... \times C_2^{a^x})$. That is,

$$b^{\bar{c}}(\bar{c}') := b_2^{c^a}(c^{a\prime}) \cdot b_2^{c^{a^b}}(c^{a^b\prime}) \cdot ... \cdot b_2^{c^{a^x}}(c^{a^x\prime}),$$
$$\forall \bar{c}' = (c^{a\prime}, c^{a^b\prime}, ..., c^{a^x\prime}) \in C_2^{a,\infty} \times C_2^{a^b,\infty} \times ... \times C_2^{a^x}.$$

Then, define type $t_1^{\bar{c},z-1}[\bar{c}]$ to be such that, for each combination of choices $\bar{c}' = (c^{a\prime}, c^{a^b\prime}, ..., c^{a^x\prime}) \in C_2^{a,\infty} \times C_2^{a^b,\infty} \times ... \times C_2^{a^x}$

$$b_1[t_1^{\bar{c},z-1}[\bar{c}]](c^{a\prime}, t_2^{\bar{c},z}[\bar{c}'']) := \begin{cases} b^{\bar{c}}(\bar{c}'), \text{ if } \bar{c}' = \bar{c}'' \\ 0, \text{ otherwise.} \end{cases}$$

We create such sequences of types for each combination of choices $\bar{c} = (c^a, c^{a^b}, ..., c^{a^x})$.

We follow the same construction another $k^* - 2$ times, where $k^* = max(N_1 \cup N_2)$. For each $y \in \{1, ..., k^* - 2\}$, do the following: for each $\bar{c}, \bar{c}' \in C_2^{a,\infty} \times C_2^{a^b,\infty} \times ... \times C_2^{a^x}$ take a sequence of types $(t_2^{\bar{c},yz}[\bar{c}'], ..., t_1^{\bar{c},(y+1)z-1}[\bar{c}'])$. Then, let us have in each such sequence that type $t_2^{\bar{c},yz}[\bar{c}']$ assigns probability one to type $t_1^{\bar{c},yz+1}[\bar{c}']$, and that type $t_i^{\bar{c},yz+n}[\bar{c}']$ with $i \in \{1, 2\}$ assigns probability one to type $t_j^{\bar{c},yz+n+1}[\bar{c}']$ with $j \neq i$, for each $n \in \{1, ..., z - 2\}$. Additionally, for each $k = a^p - a - 1$, define type $t_1^{\bar{c},yz+k}[\bar{c}']$ to be such that

$$b_1[t_1^{\bar{c},yz+k}[\bar{c}']] := (c^{a^p\prime}, t_2^{\bar{c},yz+k+1}[\bar{c}']),$$

with $\bar{c}' = (c^{a\prime}, ..., c^{a^x\prime})$. Then, for each $\bar{c}, \bar{c}^* \in C_2^{a,\infty} \times C_2^{a^b,\infty} \times ... \times C_2^{a^x}$ construct a type $t_2^{\bar{c},(y+1)z}[\bar{c}^*]$. Then define type $t_1^{\bar{c},(y+1)z-1}[\bar{c}']$ to be such that for each $\bar{c}^* = (c^{a^*}, c^{a^b*}, ..., c^{a^x*}) \in C_2^{a,\infty} \times C_2^{a^b,\infty} \times ... \times C_2^{a^x}$

$$b_1[t_1^{\bar{c},(y+1)z-1}[\bar{c}']](c^{a*}, t_2^{\bar{c},(y+1)z}[\bar{c}'']) := \begin{cases} b^{\bar{c}'}(\bar{c}^*), \text{ if } \bar{c}^* = \bar{c}'' \\ 0, \text{ otherwise.} \end{cases}$$

We do this for every $y \in \{1, ..., k^* - 2\}$.

Finally consider the case $y = k^*-1$. For each $\bar{c}, \bar{c}' \in C_2^{a,\infty} \times C_2^{a^b,\infty} \times ... \times C_2^{a^x}$ take a sequence of types $(t_2^{\bar{c},(k^*-1)z}[\bar{c}'], ..., t_1^{\bar{c},k^*z-1}[\bar{c}'])$. Then, let us have in each such sequence that type $t_2^{\bar{c},(k^*-1)z}[\bar{c}']$ assigns probability one to type $t_1^{\bar{c},(k^*-1)z+1}[\bar{c}']$, and that type $t_i^{\bar{c},(k^*-1)z+n}[\bar{c}']$ with $i \in \{1, 2\}$ assigns probability one to type $t_j^{\bar{c},(k^*-1)z+n+1}[\bar{c}']$ with $j \neq i$, for each $n \in \{1, ..., z - 2\}$. Additionally, for each $k = a^p - a - 1$, define type $t_1^{\bar{c},(k^*-1)z+k}[\bar{c}']$ to be such that

$$b_1[t_1^{\bar{c},(k^*-1)z+k}[\bar{c}']] := (c^{a^p\prime}, t_2^{\bar{c},(k^*-1)z+k+1}[\bar{c}']),$$

with $\bar{c}' = (c^{a\prime}, ..., c^{a^x\prime})$. Then define type $t_1^{\bar{c},k^*z-1}[\bar{c}']$ to be such that for each $\bar{c}^* = (c^{a^*}, c^{a^b*}, ..., c^{a^x*}) \in C_2^{a,\infty} \times C_2^{a^b,\infty} \times ... \times C_2^{a^x}$

$$b_1[t_1^{\bar{c},k^*z-1}[\bar{c}']](c^{a*}, t_2^{\bar{c}''}[\bar{c}'']) := \begin{cases} b^{\bar{c}'}(\bar{c}^*), \text{ if } \bar{c}^* = \bar{c}'' \\ 0, \text{ otherwise.} \end{cases}$$

We create such k^* sequences of z types for each combination of choices $\bar{c} = (c^a, ..., c^{a^x}) \in C_2^{a,\infty} \times C_2^{a^b,\infty} \times ... \times C_2^{a^x}$. All these types together form a partial epistemic model. Call this partial epistemic model $\bar{\mathcal{M}}$.

Extend this partial epistemic model in the following way. Let type $t_1^{c_1}[c_1]$ we fixed at the beginning be at the start of the following sequence of types: $(t_1^{c_1}[c_1], t_2^{c_1,1}[c_1], ..., t_1^{c_1,a-1}[c_1])$. Let type $t_1^{c_1}[c_1]$ be such that it assigns probability one to type $t_2^{c_1,1}[c_1]$ and let type $t_i^{c_1,n}[c_1]$ be such that it assigns probability one to type $t_j^{c_1,n+1}[c_1]$, for each $n \in \{1, 2, ..., a-2\}$.

Now, we have that from each combination of choices $(c^a, c^b, ..., c^x)$ we derive a single combination of choices $\bar{c} = (c^a, c^{a^b}, ..., c^{a^x})$. Then, for each combination of choices $(c^a, c^b, ..., c^x)$ and the combination of choices $\bar{c}$ that is derived from it, define type $t_1^{c_1,a-1}[c_1]$ to be such that

$$b_1[t_1^{c_1,a-1}[c_1]](c^a, t_2^{\bar{c}'}[\bar{c}']) := \begin{cases} b_1^{c_1}(c^a, c^b, ..., c^x), \text{ if } \bar{c}' = \bar{c}, \\ 0, \text{ otherwise.} \end{cases}$$

As will be explained below, then type $t_1^{c_1}[c_1]$ expresses on-path belief in rationality in this partial epistemic model. We do so by first showing that each type $t_2^{\bar{c}}[\bar{c}]$ expresses z-fold belief in rationality, 2z-fold belief in rationality, and so on. Additionally, we show that this type expresses $(a^p - a)$-fold belief in rationality, $z + (a^p - a)$-fold belief in rationality and so on, for every order a^p.

First, it is clear that type $t_2^{\bar{c}}[\bar{c}]$ for any $\bar{c} \in C_2^{a,\infty} \times C_2^{a^b,\infty} \times ... \times C_2^{a^x}$ expresses z-fold belief in rationality by construction. Namely, for any

$\bar{c}' = (c^{a'}, ..., c^{a^x\prime})$, type $t_1^{\bar{c},2z-1}[\bar{c}']$ is such that its distribution over choices in C_2^∞ in the belief it induces is exactly equal to $b^{\bar{c}'}$, which makes choice $c^{a'}$ optimal by construction. Since from type $t_1^{\bar{c},z}[\bar{c}']$ a sequence of $z-1$ probability one beliefs is induced that ends up at type $t_1^{\bar{c},2z-1}[\bar{c}']$, it follows that choice $c^{a'}$ is optimal given type $t_2^{\bar{c},z}[\bar{c}']$. We have in the sequence $(t_2^{\bar{c}}[\bar{c}], ..., t_1^{\bar{c},z-1}[\bar{c}])$ that type $t_1^{\bar{c},z-1}[\bar{c}]$ is constructed such that it only assigns positive probability to choice-type combinations $(c^{a'}, t_2^{\bar{c},z}[\bar{c}']$ where the choice is optimal given such type. Hence type $t_2^{\bar{c}}[\bar{c}]$ expresses z-fold belief in rationality.

Similarly, type $t_2^{\bar{c}}[\bar{c}]$ expresses $(a^p - a)$-fold belief in rationality for each order a^p. To see this, first note that from type $t_2^{\bar{c},a^p-a}[\bar{c}]$ there follows a sequence of probability one beliefs up to type $t_1^{\bar{c},z-1}[\bar{c}]$. Second, type $t_1^{\bar{c},z-1}[\bar{c}]$ induces a belief whose distribution over types is such that it is equal to the distribution that $b^{\bar{c}}$ has over $C_2^{a,\infty} \times ... \times C_2^{a^x,\infty}$. In different terms, the belief $b_1[t_1^{\bar{c},z-1}[\bar{c}]]$ assigns a probability to type $t_2^{\bar{c},z}[\bar{c}']$ that is equal to the probability that $b^{\bar{c}} = b_2^{c^a} \times ... \times b_2^{c^{a^p}} \times ... \times b_2^{c^{a^x}}$ assigns to $\bar{c}' = (c^{a'}, ..., c^{a^p\prime}, ..., c^{a^x\prime})$. Third, by construction, from type $t_2^{\bar{c},z}[\bar{c}']$ there follows a sequence of $(a^p - a)$ probability one beliefs. The $(a^p - a)$-th type in this sequence assigns probability one to specifically the choice-type combination $(c^{a^p\prime}, t_2^{\bar{c},a^p-a}[\bar{c}'])$. Taken together then, type $t_2^{\bar{c},a^p-a}[\bar{c}]$ induces a z-th order expectation that is equal to $b_2^{c^{a^p}}$. By construction choice c^{a^p} is optimal given $b_2^{a^p}$. Hence choice c^{a^p} is optimal given type $t_2^{\bar{c},a^p-a}[\bar{c}]$. Since from type $t_2^{\bar{c}}[\bar{c}]$ there follows a sequence of probability one beliefs up to type $t_1^{\bar{c},a^p-a-1}[\bar{c}]$ and $b_1[t_1^{\bar{c},a^p-a-1}[\bar{c}]](c^{a^p}, t_2^{\bar{c},a^p-a}[\bar{c}]) = 1$, we have that type $t_2^{\bar{c}}[\bar{c}]$ expresses $(a^p - a)$-fold belief in rationality.

Following exactly the same argument, we have that type $t_2^{\bar{c},yz}[\bar{c}']$ for each $\bar{c}, \bar{c}'$ and each $y \in \{1, ..., k^* - 1\}$ expresses both z-fold belief in rationality as well $(a^p - a)$-fold belief in rationality for every order a^p.

Each type $t_2^{\bar{c}}[\bar{c}]$ and each type $t_2^{\bar{c},yz}[\bar{c}']$ in their z-th order beliefs only assign positive probability to choice-type combinations where the types are characterized as before. Hence, each type $t_2^{\bar{c}}[\bar{c}]$ and each type $t_2^{\bar{c},yz}[\bar{c}']$ only assign positive probability in their z-th order beliefs to type that

express both z-fold belief in rationality and $(a^p - a)$-fold belief in rationality for every order a^p. It then follows that each type $t_2^{\bar{c}}[\bar{c}]$ and each type $t_2^{\bar{c},yz}[\bar{c}']$ then also expresses $(a^p - a + z)$-fold belief in rationality for every order a^p, $2z$-fold belief in rationality, $(a^p - a + 2z)$-fold belief in rationality for every order a^p $3z$-fold belief in rationality, and so on.

We started off with type $t_1^{c_1}[c_1]$. In the a-th order belief, the belief hierarchy induced by $t_1^{c_1}[c_1]$ exclusively assigns positive probability to the choice-type combinations $(c^a, t_2^{\bar{c}}[\bar{c}])$ where $\bar{c}$ starts with c^a. By construction choice c^a is optimal given type $t_2^{\bar{c}}[\bar{c}]$, hence type $t_1^{c_1}[c_1]$ expresses a-fold belief in rationality. Moreover, each such type $t_2^{\bar{c}}[\bar{c}]$ expresses $(a^p - a + z)$-fold belief in rationality for every order a^p, $2z$-fold belief in rationality, $(a^p - a + 2z)$-fold belief in rationality for every order a^p $3z$-fold belief in rationality, and so on. It follows that type $t_1^{c_1}[c_1]$ then also expresses $(a^p + z)$-fold belief in rationality for every order a^p, $a + z$-fold belief in rationality, $(a^p + 2z)$-fold belief in rationality for every order a^p, $a + 2z$-fold belief in rationality, and so on. Then, type $t_1^{c_1}[c_1]$ expresses on-path belief in rationality. Additionally, $t_1^{c_1}[c_1]$ was constructed such that c_1 was optimal given the type.

Hence, we have constructed an epistemic model in which c_1 is optimal given a type that expresses on-path belief in rationality.

Step 2
We will now develop a similar recursive, backward construction for scenario (iii) as we did earlier for scenarios (i) and (ii).

In Step 1 we have shown that for choice $c_1 \in C_1^\infty$ we can always construct a *partial* epistemic model with a type $t_1^{c_1}[c_1]$ for which c_1 is optimal and that expresses on-path belief in rationality. In Step 2 we will now do the following. We will show that if there exists a belief hierarchy expressing on-path belief in rationality for which c_1 is optimal, then there is also a belief hierarchy expressing common belief in rationality for which c_1 is optimal.

Consider a partial epistemic model $\bar{\mathcal{M}} = (T_i, b_i[t_i])_{i \in \{1,2\}}$ as constructed in Step 1. We had here for each combination of choices $\bar{c} = (c^a, c^{a^b}, ..., c^{a^x})$ a type $t_2^{\bar{c}}[\bar{c}]$ that expressed $(a^p - a + yz)$-fold belief in rationality and $(y+1)z$-fold belief in rationality for any $y \in \mathbb{N}$. By means of a backward, recursive procedure we transform this epistemic model such that we get to a new, complete epistemic model that includes a type $t_2^m[\bar{c}, \bar{c}, 0]$ that expresses common belief in rationality and induces the same yz-th order expectation and $(a^p - a + yz)$-th order expectation for every $y \in \mathbb{N}$ as type $t_2^{\bar{c}}[\bar{c}]$ does. The recursive procedure here defines choices *and* types at the same time in each iteration.

The recursive procedure is as follows.

Iteration 0:
For each combination of choices
$\bar{c} = (c^a, c^{a^b}, ..., c^{a^x}) \in C_2^{a,\infty} \times C_2^{a^b,\infty} \times ... \times C_2^{a^x,\infty}$, define

$$d^0[\bar{c}, \bar{c}, 0] := c^a.$$

Also define for each order a^p

$$d^0[\bar{c}, \bar{c}, a^p - a] := c^{a^p}.$$

Moreover, for each $k \in \{1, 2, ..., z-1\}$ with $k \neq a^p - a$ for any order a^p, define $d^0[\bar{c}, \bar{c}, k]$ randomly:

$d^0[\bar{c}, \bar{c}, k] := c'$, for *some* $c' \in C_2^\infty$ if k is even or *some* $c' \in C_1^\infty$ if k is odd.

For each $y \in \{1, ..., k^* - 1\}$, define for each $\bar{c}, \bar{c}' \in C_2^{a,\infty} \times C_2^{a^b,\infty} \times ... \times C_2^{a^x,\infty}$ with $\bar{c}' = (c^{a\prime}, ..., c^{a^x\prime})$

$$d^0[\bar{c}, \bar{c}', yz] := c^{a\prime},$$

and for each order a^p

$$d^0[\bar{c}, \bar{c}', a^p - a + yz] := c^{a^p\prime},$$

and finally for each $k \in \{1, 2, ..., z-1\}$ with $k \neq a^p - a$ for any order a^p

$$d^0[\bar{c}, \bar{c}', yz + k] := c', \text{ for } \textit{some } c' \in C_2^\infty \text{ if } k \text{ is even or}$$
$$\textit{some } c' \in C_1^\infty \text{ if } k \text{ is odd.}$$

Take a sequence of types $(t_2^0[\bar{c}, \bar{c}, 0], ..., t_1^0[\bar{c}, \bar{c}, z-1])$ for every combination of choices $\bar{c}$. Similarly, for each $y \in \{1, ..., k^* - 1\}$ and each pair $\bar{c}, \bar{c}' \in C_2^{a,\infty} \times ... \times C_2^{a^x,\infty}$ take a sequence of types $(t_2^0[\bar{c}, \bar{c}', yz], ..., t_1^0[\bar{c}, \bar{c}', (y+1)z - 1])$.

Now, for each combination of choices $\bar{c}$, define type $t_2^0[\bar{c}, \bar{c}, 0]$ such that

$$b_2[t_2^0[\bar{c}, \bar{c}, 0]] := (d^0[\bar{c}, \bar{c}, 1], t_1^0[\bar{c}, \bar{c}, 1]).$$

Then, define for each $k \in \{1, 2, ..., z-2\}$ type $t_i^0[\bar{c}, \bar{c}, k]$ with $i \in \{1, 2\}$ to be such that

$$b_i[t_i^0[\bar{c}, \bar{c}, k]] := (d^0[\bar{c}, \bar{c}, k+1], t_j^0[\bar{c}, \bar{c}, k+1]),$$

where $j \neq i$. Finally, we define type $t_1^0[\bar{c}, \bar{c}, z-1]$ to be such that

$$b_1[t_1^0[\bar{c}, \bar{c}, z-1]](c^{a\prime}, t_2^0[\bar{c}, \bar{c}', z]) := b^{\bar{c}}(\bar{c}'),$$
$$\forall \bar{c}' = (c^{a\prime}, c^{a^b\prime}, ..., c^{a^x\prime}) \in C_2^{a,\infty} \times C_2^{a^b,\infty} \times ... \times C_2^{a^x,\infty}.$$

In a similar manner, for each pair $\bar{c}, \bar{c}' \in C_2^{a,\infty} \times ... \times C_2^{a^x,\infty}$ and each $y \in \{1, ..., k^* - 1\}$, define type $t_2^0[\bar{c}, \bar{c}', yz]$ to be such that

$$b_2[t_2^0[\bar{c}, \bar{c}', yz]] := (d^0[\bar{c}, \bar{c}', yz+1], t_1^0[\bar{c}, \bar{c}', yz+1]).$$

And define for each $k \in \{1, 2, ..., z-2\}$ type $t_i^0[\bar{c}, \bar{c}', yz + k]$ with $i \in \{1, 2\}$ to be such that

$$b_i[t_i^0[\bar{c}, \bar{c}', yz + k]] := (d^0[\bar{c}, \bar{c}', yz + k + 1], t_j^0[\bar{c}, \bar{c}', yz + k + 1]),$$

where $j \neq i$. Finally, we define type $t_1^0[\bar{c}, \bar{c}', (y+1)z-1]$ for $y \in \{1, ..., k^*-1\}$ to be such that

$$b_1[t_1^0[\bar{c}, \bar{c}', (y+1)z-1]](c^{a''}, t_2^0[\bar{c}, \bar{c}'', (y+1)z]) := b^{\bar{c}'}(\bar{c}''),$$
$$\forall \bar{c}'' = (c^{a''}, ..., c^{a^x''}) \in C_2^{a,\infty} \times C_2^{a^b,\infty} \times ... \times C_2^{a^x,\infty}.$$

If $y = k^* - 1$, define type $t_1^0[\bar{c}, \bar{c}', k^*z - 1]$ to be such that

$$b_1[t_1^0[\bar{c}, \bar{c}', k^*z-1]](c^{a''}, t_2^0[\bar{c}'', \bar{c}'', 0]) := b^{\bar{c}'}(\bar{c}''),$$
$$\forall \bar{c}'' = (c^{a''}, ..., c^{a^x''}) \in C_2^{a,\infty} \times C_2^{a^b,\infty} \times ... \times C_2^{a^x,\infty}.$$

Note that by construction of Step 1, we have for each $y \in 1, ..., k^*$ that $b^{\bar{c}'}(\bar{c}'') = b_1[t_1^0[\bar{c}, \bar{c}', yz-1]](\bar{c}'', t_2^0[\bar{c}, \bar{c}', yz])$ for each $\bar{c}''$. And for each order a^p we have that $b_1[t_1^0[\bar{c}, \bar{c}', a^p - a - 1 + (y-1)z]] = (d^0[\bar{c}, \bar{c}', a^p - a + (y-1)z], t_2^0[\bar{c}, \bar{c}' a^p - a + (y-1)z])$. Moreover, all other types induce probability one beliefs. So for Iteration 0 we essentially take a copy of the epistemic model $\bar{\mathcal{M}}$ created in Step 1, but fill in the beliefs that were still incomplete from this step.

Iteration n $\geq$ 1 :
For each pair of combinations of choices $\bar{c}, \bar{c}' \in C_2^{a,\infty} \times ... \times C_2^{a^x,\infty}$, take a sequence of types $(t_2^n[\bar{c}, \bar{c}', (k^*-1)z], ..., t_1^n[\bar{c}, \bar{c}', k^*z-1])$. Then, define type $t_1^n[\bar{c}, \bar{c}', k^*z-1]$ to be such that

$$b_1[t_1^n[\bar{c}, \bar{c}', k^*z-1]](c^{a''}, t_2^{n-1}[\bar{c}'', \bar{c}'', 0]) := b^{\bar{c}'}(\bar{c}''),$$
$$\forall \bar{c}'' = (c^{a''}, ..., c^{a^x''}) \in C_2^{a,\infty} \times ... \times C_2^{a^x,\infty}.$$

We then also define

$$d^n[\bar{c}, \bar{c}', k^*z-1] := c_1', \text{ with } c_1' \text{ optimal given the type } t_1^n[\bar{c}, \bar{c}', k^*z-1].$$

We also define for each order a^p

$$d^n[\bar{c}, \bar{c}', (k^*-1)z + (a^p - a)] := c^{a^p\prime}.$$

Now, for each pair $\bar{c}, \bar{c}' \in C_2^{a,\infty} \times ... \times C_2^{a^x,\infty}$, define recursively for each *even* $k \in \{2, ..., z-2\}$ starting at $k = z-2$, type $t_2^n[\bar{c}, \bar{c}', (k^*-1)z+k]$ to be such that

$$b_2[t_2^n[\bar{c}, \bar{c}', (k^*-1)z+k]] := (d^n[\bar{c}, \bar{c}', (k^*-1)z+k+1], t_1^n[\bar{c}, \bar{c}', (k^*-1)z+k+1]).$$

Second, if $k \neq a^p - a$ also define

$$d^n[\bar{c}, \bar{c}', (k^*-1)z+k] := c_2^*, \text{ with } c_2^* \text{ optimal given the type } t_2^n[\bar{c}, \bar{c}', (k^*-1)z+k].$$

Third, define type $t_1^n[\bar{c}, \bar{c}', (k^*-1)z+k-1]$ to be such that

$$b_1[t_1^n[\bar{c}, \bar{c}', (k^*-1)z+k-1]] := (d^n[\bar{c}, \bar{c}', (k^*-1)z+k], t_1^n[\bar{c}, \bar{c}', (k^*-1)z+k-1].$$

Fourth, also define

$$d^n[\bar{c}, \bar{c}', (k^*-1)z+k-1] := c_1^*, \text{ with } c_1^* \text{ optimal given the type } t_1^n[\bar{c}, \bar{c}', (k^*-1)z+k-1]].$$

Finally, define type $t_2^n[\bar{c}, \bar{c}', (k^*-1)z]$ to be such that

$$b_2[t_2^n[\bar{c}, \bar{c}', (k^*-1)z]] := (d^n[\bar{c}, \bar{c}', (k^*-1)z+1], t_1^n[\bar{c}, \bar{c}', (k^*-1)z+1]),$$

and define

$$d^n[\bar{c}, \bar{c}', (k^*-1)z]] := c^{a\prime}.$$

Next, for each $y \in \{1, ..., k^*-2\}$, do the following iteratively, going backwards starting at $y = k^*-2$: For each pair $\bar{c}, \bar{c}' \in C_2^{a,\infty} \times ... \times C_2^{a^x,\infty}$, take a sequence of types $(t_2^n[\bar{c}, \bar{c}', yz], ..., t_1^n[\bar{c}, \bar{c}', (y+1)z-1])$. Then, define type $t_1^n[\bar{c}, \bar{c}', (y+1)z-1]$ to be such that

$$b_1[t_1^n[\bar{c}, \bar{c}', (y+1)z-1]](c^{a\prime\prime}, t_2^n[\bar{c}, \bar{c}'', (y+1)z]) := b^{\bar{c}'}(\bar{c}''),$$
$$\forall \bar{c}'' = (c^{a\prime\prime}, ..., c^{a^x\prime\prime}) \in C_2^{a,\infty} \times ... \times C_2^{a^x,\infty}.$$

We then also define

$$d^n[\bar{c}, \bar{c}', (y+1)z - 1] := c_1', \text{ with } c_1' \text{ optimal given the type } t_1^n[\bar{c}, \bar{c}', (y+1)z - 1].$$

We also define for each order a^p

$$d^n[\bar{c}, \bar{c}', yz + (a^p - a)] := c^{a^p\prime}.$$

Now, for each pair $\bar{c}, \bar{c}' \in C_2^{a,\infty} \times ... \times C_2^{a^x,\infty}$, define recursively for each *even* $k \in \{2, ..., z-2\}$ starting at $k = z - 2$, type $t_2^n[\bar{c}, \bar{c}', yz + k]$ to be such that

$$b_2[t_2^n[\bar{c}, \bar{c}', yz + k]] := (d^n[\bar{c}, \bar{c}', yz + k + 1], t_1^n[\bar{c}, \bar{c}', yz + k + 1]).$$

Second, if $k \neq a^p - a$ also define

$$d^n[\bar{c}, \bar{c}', yz + k] := c_2^*, \text{ with } c_2^* \text{ optimal given the type } t_2^n[\bar{c}, \bar{c}', yz + k].$$

Third, define type $t_1^n[\bar{c}, \bar{c}', yz + k - 1]$ to be such that

$$b_1[t_1^n[\bar{c}, \bar{c}', yz + k - 1]] := (d^n[\bar{c}, \bar{c}', yz + k], t_1^n[\bar{c}, \bar{c}', yz + k - 1].$$

Fourth, also define

$$d^n[\bar{c}, \bar{c}', yz + k - 1] := c_1^*, \text{ with } c_1^* \text{ optimal given the type } t_1^n[\bar{c}, \bar{c}', yz + k - 1]].$$

Finally, define type $t_2^n[\bar{c}, \bar{c}', yz]$ to be such that

$$b_2[t_2^n[\bar{c}, \bar{c}', yz]] := (d^n[\bar{c}, \bar{c}', yz + 1], t_1^n[\bar{c}, \bar{c}', yz + 1]),$$

and define

$$d^n[\bar{c}, \bar{c}', yz]] := c^{a\prime}.$$

We do this iteratively for each $y \in \{1, ..., k^* - 2\}$, starting at $y = k^* - 2$.

Finally, for $y = 1$, we again do the following: For each pair $\bar{c}, \bar{c}' \in C_2^{a,\infty} \times ... \times C_2^{a^x,\infty}$, take a sequence of types $(t_2^n[\bar{c}, \bar{c}, 0], ..., t_1^n[\bar{c}, \bar{c}', z - 1])$. Then, define type $t_1^n[\bar{c}, \bar{c}, z - 1]$ to be such that

$$b_1[t_1^n[\bar{c}, \bar{c}, z - 1]](c^{a\prime\prime}, t_2^n[\bar{c}, \bar{c}'', z]) := b^{\bar{c}}(\bar{c}''),$$
$$\forall \bar{c}'' = (c^{a\prime\prime}, ..., c^{a^x\prime\prime}) \in C_2^{a,\infty} \times ... \times C_2^{a^x,\infty}.$$

We then also define

$$d^n[\bar{c}, \bar{c}, z - 1] := c_1', \text{ with } c_1' \text{ optimal given the type } t_1^n[\bar{c}, \bar{c}', z - 1].$$

We also define for each order a^p

$$d^n[\bar{c}, \bar{c}, (a^p - a)] := c^{a^p}.$$

Now, for each $\bar{c} \in C_2^{a,\infty} \times ... \times C_2^{a^x,\infty}$, define recursively for each *even* $k \in \{2, ..., z - 2\}$ starting at $k = z - 2$, type $t_2^n[\bar{c}, \bar{c}, k]$ to be such that

$$b_2[t_2^n[\bar{c}, \bar{c}, k]] := (d^n[\bar{c}, \bar{c}, k + 1], t_1^n[\bar{c}, \bar{c}, k + 1]).$$

Second, if $k \neq a^p - a$ also define

$$d^n[\bar{c}, \bar{c}, k] := c_2^*, \text{ with } c_2^* \text{ optimal given the type } t_2^n[\bar{c}, \bar{c}, k].$$

Third, define type $t_1^n[\bar{c}, \bar{c}, k - 1]$ to be such that

$$b_1[t_1^n[\bar{c}, \bar{c}, k - 1]] := (d^n[\bar{c}, \bar{c}, k], t_1^n[\bar{c}, \bar{c}, k - 1].$$

Fourth, also define

$$d^n[\bar{c}, \bar{c}, k - 1] := c_1^*, \text{ with } c_1^* \text{ optimal given the type } t_1^n[\bar{c}, \bar{c}, k - 1]].$$

Finally, define type $t_2^n[\bar{c}, \bar{c}, 0]$ to be such that

$$b_2[t_2^n[\bar{c}, \bar{c}, 0]] := (d^n[\bar{c}, \bar{c}, 1], t_1^n[\bar{c}, \bar{c}, 1]),$$

and define

$$d^n[\bar{c}, \bar{c}, 0]] := c^a.$$

We have that C_1^∞ and C_2^∞ are finite sets. Additionally, z and k^* are finite orders of belief and thus k^*z is a finite order of belief as well. Finally, we have that $C_2^{a,\infty} \times ... \times C_2^{a^x,\infty}$ is a finite set as well. Together, we then have that there are iterations m, n with $m > n$ such that:

$$d^m[\bar{c}, \bar{c}, k] = d^n[\bar{c}, \bar{c}, k],\ \forall \bar{c} \in C_2^{a,\infty} \times ... \times C_2^{a^x,\infty}, k \in \{0, 1, ..., z-1\},$$

and

$$d^m[\bar{c}, \bar{c}', yz + k] = d^n[\bar{c}, \bar{c}', yz + k],\ \forall \bar{c}, \bar{c}' \in C_2^{a,\infty} \times ... \times C_2^{a^x,\infty}, \\ y \in \{1, ..., k^* - 1\}.$$

When we find such iterations m and n, we stop the recursive procedure.

Now we create the epistemic model $\mathcal{M}^*$ from the types we have constructed in our recursive procedure. Define

$$\begin{aligned} T_2(l) :=& \{t_2^l[\bar{c}, \bar{c}, k] : \bar{c} \in C_2^{a,\infty} \times ... \times C_2^{a^x,\infty}, k \in \{0, ..., z-2\} \text{ even}\} \cup \\ & \{t_1^l[\bar{c}, \bar{c}', yz + k] : \bar{c}, \bar{c}' \in C_2^{a,\infty} \times ... \times C_2^{a^x,\infty},\ y \in \{1, ..., k^* - 1\}, \\ & k \in \{0, ..., z-2\} \text{ even}\} \end{aligned}$$

and

$$\begin{aligned} T_1(l) :=& \{t_1^l[\bar{c}, \bar{c}, k] : \bar{c} \in C_2^{a,\infty} \times ... \times C_2^{a^x,\infty}, k \in \{0, 1, ..., z-1\} \text{ odd}\} \cup \\ & \{t_1^l[\bar{c}, \bar{c}', yz + k] : \bar{c}, \bar{c}' \in C_2^{a,\infty} \times ... \times C_2^{a^x,\infty},\ y \in \{1, ..., k^* - 1\}, \\ & k \in \{0, ..., z-1\} \text{ odd.}\} \end{aligned}$$

Then, let $T(l) := T_1(l) \cup T_2(l)$. Do this for every $l \in \{n, ..., m\}$.

In $T(n+1)$ specifically, we re-define for each $\bar{c}, \bar{c}' \in C_2^{a,\infty} \times ... \times C_2^{a^x,\infty}$

the type $t_1^{n+1}[\bar{c}, \bar{c}', k^*z - 1]$ to be such that

$$b_1[t_1^{n+1}[\bar{c}, \bar{c}', k^*z - 1]](c^{a\prime\prime}, t_2^m[\bar{c}'', \bar{c}'', 0]) := b^{\bar{c}'}(\bar{c}'')$$
$$\forall \bar{c}'' = (c^{a\prime\prime}, ..., c^{a^x\prime\prime}) \in C_2^{a,\infty} \times ... \times C_2^{a^x,\infty}.$$

So instead of assigning positive probability to types in $T(n)$, each type $t_1^{n+1}[\bar{c}, \bar{c}', k^*z - 1]$ now assigns positive probability to types in $T(m)$. Then define

$$\mathcal{M}^* := (\bigcup_{l \in \{n+1,...,m\}} T_i(l), b[t_i])_{i \in \{1,2\}}.$$

We will show that each type in $\mathcal{M}^*$ expresses common belief in rationality. We will do so in steps.

First, for the epistemic model $\mathcal{M}^*$, we can note that for each combination of choices $\bar{c} = (c^a, c^{a^b}, ..., c^{a^x})$ and each $l \in \{n+1, ..., m\}$, choice c^a is optimal for type $t_2^l[\bar{c}, \bar{c}, 0]$. We can also say that for each $\bar{c} = (c^a, c^{a^b}, ..., c^{a^x})$, $\bar{c}' = (c^{a\prime}, c^{a^b\prime}, ..., c^{a^x\prime})$, each $l \in \{n+1, ..., m\}$ and each $y \in \{1, ..., k^* - 1\}$ that choice $c^{a\prime}$ is optimal for type $t_2^l[\bar{c}, \bar{c}', yz]$.

Namely, from type $t_2^l[\bar{c}, \bar{c}, 0]$ there follows a sequence of probability one beliefs, induced by the sequence of types $(t_2^l[\bar{c}, \bar{c}, 0], ..., t_2^l[\bar{c}, \bar{c}, z - 2])$. This sequence of probability one beliefs ends at type $t_1^l[\bar{c}, \bar{c}, z - 1]$. By our recursive backwards construction and the way we defined $b^{\bar{c}}$ in Step 1, we have that

$$marg_{C_2^\infty} b_1[t_1^l[\bar{c}, \bar{c}, z - 1]] = b_2^{c^a}.$$

It follows then that type $t_2^l[\bar{c}, \bar{c}, 0]$ induces a z-th order expectation that is equal to $b_2^{c^a}$. And we constructed $b_2^{c^a}$ such that c^a is optimal given $b_2^{c^a}$. Hence c^a is optimal given type $t_2^l[\bar{c}, \bar{c}, 0]$. This goes for every $l \in \{n+1, ..., m\}$.

Similarly, for each $y \in \{1, ..., k^* - 1\}$, from type $t_2^l[\bar{c}, \bar{c}', yz]$ there follows a sequence of probability one beliefs, induced by the sequence of

types $(t_2^l[\bar{c}, \bar{c}', yz], ..., t_2^l[\bar{c}, \bar{c}', (y+1)z-2])$. This sequence ends at type $t_1^l[\bar{c}, \bar{c}', (y+1)z-1]$. By construction, we have that

$$marg_{C_2^\infty} b_1[t_1^l[\bar{c}, \bar{c}', (y+1)z-1]] = b_2^{c^{a'}}.$$

It follows then that type $t_2^l[\bar{c}, \bar{c}', yz]$ induces a z-th order expectation that is equal to $b_2^{c^{a'}}$. We constructed $b_2^{c^{a'}}$ such that $c^{a'}$ is optimal given $b_2^{c^{a'}}$. Hence $c^{a'}$ is optimal given type $t_2^l[\bar{c}, \bar{c}', yz]$. This goes for every $l \in \{n+1, ..., m\}$.

Second, we can also note that for each combination of choices $\bar{c}$, each order a^p and each $l \in \{n+1, ..., m\}$ we have that choice c^{a^p} is optimal for type $t_2^l[\bar{c}, \bar{c}, a^p - a]$. Moreover, for each pair $\bar{c}, \bar{c}'$, each order a^p, each $l \in \{n+1, ..., m\}$ and each $y \in \{1, ..., k^* - 1\}$ we have that choice $c^{a^p\prime}$ is optimal for the type $t_2^l[\bar{c}, \bar{c}', a^p - a + yz]$.

Namely, from type $t_2^l[\bar{c}, \bar{c}, 0]$ there follows a sequence of probability one beliefs, induced by the sequence of types $(t_2^l[\bar{c}, \bar{c}, 0], ..., t_2^l[\bar{c}, \bar{c}, z-2])$. This sequence of probability one beliefs ends at type $t_1^l[\bar{c}, \bar{c}, z-1]$. By our recursive backwards construction, we have that

$$b_1[t_1^l[\bar{c}, \bar{c}, z-1](d^l[\bar{c}, \bar{c}', z], t_2^{l-1}[\bar{c}, \bar{c}', z]) = b^{\bar{c}}(\bar{c}'),$$
$$\forall \bar{c}' = (c^{a'}, c^{a^b\prime}, ..., c^{a^x\prime})' \in C_2^{a,\infty} \times C_2^{a^b,\infty} \times ... \times C_2^{a^x,\infty}.$$

By Step 1 we defined $b^{\bar{c}}$ as the joint probability distribution of the z-th order expectations $b_2^{c^{a^p}}$ for all orders a^p. Now, from each type $t_2^{l-1}[\bar{c}, \bar{c}', 0]$ there again follows a sequence of probability one beliefs up to at least type $t_2^{l-1}[\bar{c}, \bar{c}', a^p - a - 1 + z]$. And type $t_2^{l-1}[\bar{c}, \bar{c}', a^p - a - 1 + z]$ assigns by construction of our recursive procedure probability one to choice $c^{a^p\prime}$. Taken together, it follows that type $t_2^l[\bar{c}, \bar{c}, a^p - a]$ induces a z-th order expectation that is equal to $b_2^{c^{a^p}}$. By construction of Step 1, we have that choice c^{a^p} is optimal given $b_2^{c^{a^p}}$. Hence, choice c^{a^p} is also optimal given type $t_2^l[\bar{c}, \bar{c}, a^p - a]$.

Similarly, for each $y \in \{1, ..., k^* - 1\}$, from type $t_2^l[\bar{c}, \bar{c}', yz]$ there follows

a sequence of probability one beliefs, induced by the sequence of types $(t_2^l[\bar{c}, \bar{c}', yz], ..., t_2^l[\bar{c}, \bar{c}', (y+1)z-2])$. This sequence of probability one beliefs ends at type $t_1^l[\bar{c}, \bar{c}', (y+1)z-1]$. By our recursive backwards construction, we have that

$$b_1[t_1^l[\bar{c}, \bar{c}', (y+1)z-1](d^l[\bar{c}, \bar{c}'', (y+1)z], t_2^{l-1}[\bar{c}, \bar{c}'', (y+1)z]) = b^{\bar{c}'}(\bar{c}''),$$
$$\forall \bar{c}'' = (c^{a''}, c^{a^b}{}'', ..., c^{a^x}{}'')' \in C_2^{a,\infty} \times C_2^{a^b,\infty} \times ... \times C_2^{a^x,\infty}.$$

By Step 1 we defined $b^{\bar{c}'}$ as the joint probability distribution of the z-th order expectations $b_2^{c^{a^p}{}'}$ for all orders a^p. Now, from each type $t_2^{l-1}[\bar{c}, \bar{c}'', (y+1)z]$ there again follows a sequence of probability one beliefs up to at least type $t_2^{l-1}[\bar{c}, \bar{c}'', a^p - a - 1 + (y+1)z]$. And type $t_2^{l-1}[\bar{c}, \bar{c}'', a^p - a - 1 + (y+1)z]$ assigns by construction of our recursive procedure probability one to choice $c^{a^p}{}''$. Taken together, it follows that type $t_2^l[\bar{c}, \bar{c}', a^p - a + yz]$ induces a z-th order expectation that is equal to $b_2^{c^{a^p}{}'}$. By construction of Step 1, we have that choice $c^{a^p}{}'$ is optimal given $b_2^{c^{a^p}{}'}$. Hence, choice $c^{a^p}{}'$ is also optimal given type $t_2^l[\bar{c}, \bar{c}', a^p - a + yz]$.

Third, we can also show the following is true.

Claim 5.4. *Consider the epistemic model $\mathcal{M}^*$. For each $l \in \{n+1, ..., m\}$, each $k \in \{1, 2, ..., z-1\}$ for $k \neq a^p - a$ for any order a^p and each combination of choices $\bar{c} \in C_2^{a,\infty} \times C_2^{a^b,\infty} \times ... \times C_2^{a^x,\infty}$, each choice $d^l[\bar{c}, \bar{c}, k]$ is optimal given the type $t_i^l[\bar{c}, \bar{c}, k]$ with $i \in \{1, 2\}$. Moreover, for each $y \in \{1, ..., k^* - 1\}$, for each $l \in \{n+1, ..., m\}$, each $k \in \{1, 2, ..., z-1\}$ for $k \neq a^p - a$ for any order a^p and each pair $\bar{c}, \bar{c}' \in C_2^{a,\infty} \times C_2^{a^b,\infty} \times ... \times C_2^{a^x,\infty}$, each choice $d^l[\bar{c}, \bar{c}', yz + k]$ is optimal given the type $t_i^l[\bar{c}, \bar{c}', yz + k]$ with $i \in \{1, 2\}$.*

Proof of claim. We start of with the epistemic model we created when ending the recursive procedure, but *before* $\mathcal{M}^*$ was created.

For each $k \in \{0, 1, ..., z-2\}$ and each $\bar{c}'$ we have by construction that

$$b_i[t_i^n[\bar{c}', \bar{c}', k](d^n[\bar{c}', \bar{c}', k+1], t_j^n[\bar{c}', \bar{c}', k+1]) = 1 =$$
$$b_i[t_i^m[\bar{c}', \bar{c}', k](d^m[\bar{c}', \bar{c}', k+1], t_j^m[\bar{c}', \bar{c}', k+1]),$$

with $d^n[\bar{c}', \bar{c}', k+1] = d^m[\bar{c}', \bar{c}', k+1]$. Note that these were the n and m that determined when to stop our recursive procedure. Moreover, for each $k \in \{0, 1, ..., a-2\}$, each $\bar{c}', \bar{c}^* \in C_2^{a,\infty} \times ... \times C_2^{a^x,\infty}$ and each $y \in \{1, ..., k^* - 1\}$ we also have by construction

$$b_i[t_i^n[\bar{c}', \bar{c}^*, yz+k](d^n[\bar{c}', \bar{c}^*, yz+k+1], t_j^n[\bar{c}', \bar{c}^*, yz+k+1]) = 1 =$$
$$b_i[t_i^m[\bar{c}', \bar{c}^*, yz+k](d^m[\bar{c}', \bar{c}^*, yz+k+1], t_j^m[\bar{c}', \bar{c}^*, yz+k+1]),$$

with $d^n[\bar{c}', \bar{c}^*, yz+k+1] = d^m[\bar{c}', \bar{c}^*, yz+k+1]$. Additionally, we have by construction that

$$b_1[t_1^n[\bar{c}', \bar{c}', z-1](d^n[\bar{c}', \bar{c}'', z], t_2^n[\bar{c}', \bar{c}'', z]) = b^{\bar{c}'}(\bar{c}'') =$$
$$b_1[t_1^m[\bar{c}', \bar{c}', z-1](d^m[\bar{c}', \bar{c}'', z], t_2^m[\bar{c}', \bar{c}'', z]),$$

for each $\bar{c}'' \in C_2^{a,\infty} \times ... \times C_2^{a^x,\infty}$. For each $y \in \{1, ..., k^* - 2\}$ we also have that

$$b_1[t_1^n[\bar{c}', \bar{c}^*, yz-1](d^n[\bar{c}', \bar{c}'', yz], t_2^n[\bar{c}', \bar{c}'', yz]) = b^{\bar{c}^*}(\bar{c}'') =$$
$$b_1[t_1^m[\bar{c}', \bar{c}^*, yz-1](d^m[\bar{c}', \bar{c}'', yz], t_2^m[\bar{c}', \bar{c}'', yz]),$$

for each $\bar{c}'' \in C_2^{a,\infty} \times ... \times C_2^{a^x,\infty}$. Finally, we have that

$$b_1[t_1^n[\bar{c}', \bar{c}^*, k^*z-1](d^{n-1}[\bar{c}'', c_1'', 0], t_2^{n-1}[\bar{c}'', \bar{c}'', 0]) = b^{\bar{c}^*}(\bar{c}'') =$$
$$b_1[t_1^m[\bar{c}', \bar{c}^*, k^*z-1](d^{m-1}[\bar{c}'', \bar{c}'', 0], t_2^{m-1}[\bar{c}'', \bar{c}'', 0]),$$

for each $\bar{c}'' \in C_2^{a,\infty} \times ... \times C_2^{a^x,\infty}$.

Then, for each $\bar{c}' \in C_2^{a,\infty} \times ... \times C_2^{a^x,\infty}$, the pair of types $t_2^m[\bar{c}', \bar{c}', 0]$ and $t_2^n[\bar{c}', \bar{c}', 0]$ induce the same k^*z-th order *belief*. To see why this is the

case, we can employ a recursive argument, for each $y \in \{1, ..., k^* - 1\}$ starting at $y = k^* - 1$.

We can first note that the pair of types $t_2^m[\bar{c}', \bar{c}^*, (k^*-1)z]$ and $t_2^n[\bar{c}', \bar{c}^*, (k^*-1)z]$ for each $\bar{c}', \bar{c}^*$ induce the same z-th order belief. Namely, from the beginning of the proof of this claim we know that types $t_i^m[\bar{c}', \bar{c}^*, (k^* - 1)z+k]$ and $t_i^n[\bar{c}', \bar{c}^*, (k^*-1)z+k]$ with $i \in \{1, 2\}$ for each $k \in \{1, ..., z-2\}$ induce a probability one belief. Moreover, the first-order belief induced by type $t_i^m[\bar{c}', \bar{c}^*, (k^* - 1)z + k]$ for each $k \in \{1, ..., z-1\}$ is equal to the first-order belief induced by type $t_i^n[\bar{c}', \bar{c}^*, (k^* - 1)z + k]$. As a result, types $t_i^m[\bar{c}', \bar{c}^*, (k^* - 1)z + k]$ and $t_i^n[\bar{c}', \bar{c}^*, (k^* - 1)z + k]$ induce the same z-th order *belief*.

Now recall, for each $\bar{c}', \bar{c}^* \in C_2^{a,\infty} \times ... \times C_2^{a^x,\infty}$, we have that

$b_1[t_1^n[\bar{c}', \bar{c}^*, (k^* - 1)z - 1](d^n[\bar{c}', \bar{c}'', (k^* - 1)z], t_2^n[\bar{c}', \bar{c}'', (k^* - 1)z]) = b_1^{\bar{c}^*}(\bar{c}'') = b_1[t_1^m[\bar{c}', \bar{c}^*, (k^* - 1)z - 1](d^m[\bar{c}', \bar{c}'', (k^* - 1)z], t_2^m[\bar{c}', \bar{c}'', (k^* - 1)z])$.

Both types $t_1^m[\bar{c}', \bar{c}^*, (k^* - 1)z - 1]$ and $t_1^n[\bar{c}', \bar{c}^*, (k^* - 1)z - 1]$ thus assign exactly the same probability to choice-type combinations where the choice is equal and the type induces the same z-th order belief. Hence, both types induce the same $(z + 1)$-th order belief.
Now we can employ our recursive argument, starting at $y = k^* - 2$. For $y \in \{1, ..., k^* - 2\}$, assume that types $t_1^m[\bar{c}', \bar{c}^*, (y + 1)z - 1]$ and $t_1^n[\bar{c}', \bar{c}^*, (y+1)z-1]$ induce the same $((k^* - y - 1)z + 1)$-th order belief. Then types $t_2^m[\bar{c}', \bar{c}^*, yz]$ and $t_2^n[\bar{c}', \bar{c}^*, yz]$ induce the same $(k^* - p)z$-th order belief. Namely, from the beginning of the proof of this claim we have that types $t_i^m[\bar{c}', \bar{c}^*, yz + k]$ and $t_i^n[\bar{c}', \bar{c}^*, yz + k]$ with $i \in \{1, 2\}$ for each $k \in \{1, ..., z-2\}$ induce a probability one belief and moreover induce the same first-order belief. Therefore, they induce the same $(z - 1)$-th order belief. Additionally, types $t_2^m[\bar{c}', \bar{c}^*, (y + 1)z - 2]$ and $t_2^n[\bar{c}', \bar{c}^*, (y+1)z-2]$ assign probability one to types that by assumption induce the same $((k^*-y)z+1)$-th order belief. It follows then that types $t_2^m[\bar{c}', \bar{c}^*, yz]$ and $t_2^n[\bar{c}', \bar{c}^*, yz]$ induce the same $(k^* - y)z$-th order belief.

Now recall that for each $\bar{c}', \bar{c}^* \in C_2^{a,\infty} \times ... \times C_2^{a^x,\infty}$, we have that

$$b_1[t_1^n[\bar{c}', \bar{c}^*, yz-1]](d^n[\bar{c}', \bar{c}'', yz], t_2^n[\bar{c}', \bar{c}'', yz]) = b_1^{\bar{c}_1}(c_1'') =$$
$$b_1[t_1^m[\bar{c}', \bar{c}^*, yz-1]](d^m[\bar{c}', \bar{c}'', yz], t_2^m[\bar{c}', \bar{c}'', yz])$$

Both types $t_1^m[\bar{c}', \bar{c}^*, yz-1]$ and $t_1^n[\bar{c}', \bar{c}^*, yz-1]$ thus assign exactly the same probability to choice-type combinations where the choice is equal and the type induces the same $(k^*-y)z$-th order belief. Hence, both types induce the same $((k^*-y)z+1)$-th order belief.

Following the same argument, we can establish that types $t_2^m[\bar{c}', \bar{c}', 0]$ and $t_2^n[\bar{c}', \bar{c}', 0]$ induce the same k^*z-th order belief. From the above we know that types $t_1^m[\bar{c}', \bar{c}', z-1]$ and $t_1^n[\bar{c}', \bar{c}', z-1]$ induce the same $((k^*-1)y+1)$-th order belief. From the beginning of the proof of this claim we have that types $t_i^m[\bar{c}', \bar{c}', k]$ and $t_i^n[\bar{c}', \bar{c}', k]$ with $i \in \{1,2\}$ for each $k \in \{1, ..., z-2\}$ induce a probability one belief and moreover induce the same first-order belief. Therefore, they induce the same $(z-1)$-th order belief. Additionally, types $t_2^m[\bar{c}', \bar{c}', z-2]$ and $t_2^n[\bar{c}', \bar{c}', z-2]$ assign probability one to types that by the above recursive argument induce the same $((k^*-1)z+1)$ order belief. It follows then that types $t_2^m[\bar{c}', \bar{c}', 0]$ and $t_2^n[\bar{c}', \bar{c}', 0]$ induce the same k^*z-th order belief. This goes for each $\bar{c}' \in C_2^{a,\infty} \times ... \times C_2^{a^x,\infty}$.

Denote type $t_1^{n+1}[\bar{c}, \bar{c}^*, k^*z-1]$ that results from our recursive backwards procedure but *before* constructing $\mathcal{M}^*$ by $\bar{t}_1^{n+1}[\bar{c}, \bar{c}^*, k^*z-1]$. In contrast, let the same type that does result from constructing $\mathcal{M}^*$ still be denoted as $t_1^{n+1}[\bar{c}, \bar{c}^*, k^*z-1]$. Now, we have for each combination of choices $\bar{c} \in C_2^{a,\infty} \times ... \times C_2^{a^x,\infty}$

$b_1[\bar{t}_1^{n+1}[\bar{c}, \bar{c}^*, k^*z-1]](c^{a\prime}, t_2^n[\bar{c}', \bar{c}', 0]) = b_1[t_1^{n+1}[\bar{c}, \bar{c}^*, k^*z-1]](c^{a\prime}, t_1^m[\bar{c}', \bar{c}', 0])$,
$\forall \bar{c}' = (c^{a\prime}, ..., c^{x\prime}) \in C_2^{a,\infty} \times ... \times C_2^{a^x,\infty}$.

It thus follows that each such type $t_1^{n+1}[\bar{c}, \bar{c}^*, k^*z - 1]$ induces the same (k^*z+1)-th order belief in $\mathcal{M}^*$ as it did before $\mathcal{M}^*$ was constructed. All the remaining types in $\bigcup_{l\in\{n+1,...,m\}} T(l)$ remained unchanged when $\mathcal{M}^*$ was constructed: they induce exactly the same belief over choice-type combinations as before. As a result, all types in $\bigcup_{l\in\{n+1,...,m\}} T(l)$ induce at least the same $(k^*z + 1)$-th order belief in $\mathcal{M}^*$ as before $\mathcal{M}^*$ was constructed.

In our backward construction procedure of types and choices, before creating $\mathcal{M}^*$, we constructed each $d^l[\bar{c}, \bar{c}', k]$ for each $l \in \{n + 1, ..., m\}$, $k \in \{1, ..., k^*z - 1\}$ and $\bar{c}, \bar{c}' \in C_2^{a,\infty} \times ... \times C_2^{a^x,\infty}$ such that it is optimal given type $t_i^l[\bar{c}, \bar{c}', k]$. Now, we have that the maximum directly utility-relevant order of belief for any player is k^* and that each type $t_i^l[\bar{c}, \bar{c}', k]$ at least induces exactly the same $(k^*z + 1)$-th order belief in $\mathcal{M}^*$ as it did before constructing $\mathcal{M}^*$. Hence, we also have in $\mathcal{M}^*$ that $d^l[\bar{c}, \bar{c}', k]$ is optimal given $t_i^l[\bar{c}, \bar{c}', k]$. This completes the proof of this claim.

Since each type in $\mathcal{M}^*$ only assigns positive probability to choice-type combinations of the likes of $(d^l[\bar{c}, \bar{c}', k], t_i^l[\bar{c}, \bar{c}', k]])$ for $k \in \{0, 1, ..., k^*a - 1\}$, each type only assigns positive probability to choice-type combinations where the choice is optimal given the type. Hence each type in $\mathcal{M}^*$ expresses 1-fold belief in rationality. Therefore also each type in $\mathcal{M}^*$ expresses common belief in rationality.

We have now the following result. For each combination of choices $\bar{c} = (c^a, c^{a^b}, ..., c^{a^x})$, we have for choice c^a constructed type $t_2^m[\bar{c}, \bar{c}, 0]$ such that c^a is optimal given that type. This type also expresses common belief in rationality in $\mathcal{M}^*$ Additionally, for each pair $\bar{c}, \bar{c}'$ with $\bar{c}' = (c^{a\prime}, c^{a^b\prime}, ..., c^{a^x\prime})$, the type $t_1^m[\bar{c}, \bar{c}', a^p - a - 1 + yz]$ always assigns probability one to choice $c^{a^p\prime}$.

As a final step, we extend this finite epistemic model in the following way. Consider again the type $t_1^{c_1}[c_1]$ we fixed in Step 1 of this proof. Choice c_1 is optimal given some higher-order expectation $b_1^{c_1} \in \Delta(C_2^{a,\infty} \times C_2^{b,\infty} \times ... \times C_2^{x,\infty})$.

We have from Step 1 that $t_1^{c_1}[c_1]$ is at the start of the following sequence of types: $(t_1^{c_1}[c_1], ..., t_1^{c_1,a-1}[c_1])$. Type $t_1^{c_1}[c_1]$ is such that it assigns probability one to type $t_2^{c_1,1}[c_1]$ and for each $n \in \{1, 2, ..., a-2\}$ we have that type $t_i^{c_1,n}[c_1]$ is such that it assigns probability one to type $t_j^{c_1,n+1}[c_1]$.

Second, recall that each combination of choices $\bar{c} = (c^a, c^{a^b}, ..., c^{a^x})$ was derived from a different combination of choices in $C_2^{a,\infty} \times C_2^{b,\infty} \times ... \times C_2^{x,\infty}$. Then, by construction of Step 1 we had for each $(c^a, c^b, ..., c^x)$ and each combination of choices $\bar{c}$ that is derived from it, that type $t_1^{c_1,a-1}[c_1]$ is such that

$$b_1[t_1^{c_1,a-1}[c_1]](c^a, t_2^m[\bar{c}', \bar{c}', 0]) := \begin{cases} b_1^{c_1}(c^a, c^b, ..., c^x), \text{ if } \bar{c}' = \bar{c}, \\ 0, \text{ otherwise,} \end{cases}$$

where type $t_2^m[\bar{c}', \bar{c}', 0]$ now replaces type $t_2^{\bar{c}}[\bar{c}']$ from Step 1. Taken together, we have that type $t_1^{c_1,a-1}[c_1]$ is constructed such that choice c_1 is optimal given type $t_1^{c_1}[c_1]$. Moreover, type $t_1^{c_1,a-1}[c_1]$, by construction of $\mathcal{M}^*$ only assigns positive probability to choice-type combinations where the choice is optimal given the type and the type expresses common belief in rationality.

Finally, for each $n \in \{1, 2, ..., a-2\}$, do the following in a stepwise manner, starting at $n = a-2$. Take type $t_i^{c_1,n}[c_1]$. Let $b_i[t_i^{c_1,n}[c_1]] := (c', t_j^{c_1,n+1}[c_1])$, with c' being optimal given type $t_i^{c_1,n+1}[c_1]$. Likewise, let $b_1[t_1^{c_1}[c_1]] := (c', t_2^{c_1,1}[c_1])$, with c' such that it is optimal given $t_2^{c_1,1}[c_1]$. Then, type $t_i^{c_1,n}[c_1]$ for each n and type $t_1^{c_1}[c_1]$ express belief in the opponent's rationality. As such, we have iteratively connected type $t_1^{c_1}[c_1]$ exclusively to types that express common belief in rationality. Hence, type $t_1^{c_1}[c_1]$ expresses common belief in rationality. Call the resulting epistemic model $\mathcal{M}^{**}$.

Thus, we constructed a finite epistemic model with a type that expresses common belief in rationality for which choice c_1 is optimal.

In Step 1 we have shown that for every choice $c_1 \in C_1^\infty$ we can construct a partial epistemic model with a type that expresses on-path belief in rationality and that is such that choice c_1 is optimal. In Step 2 we showed that we are then also able to construct a finite, epistemic model with a type that expresses common belief in rationality and that is such that choice c_1 is optimal. This concludes the proof for Scenario (iii). This also concludes the proof for Lemma 5.4 as a whole.

□

6

Conclusion

This chapter provides a general conclusion from an academic perspective. More detailed comments can be found separately in the previous chapters. The focus of this thesis has been on the intersection between psychological game theory and epistemic game theory. Both of these fields developed in order to provide a better understanding of particular aspects of strategic decision-making processes. Psychological game theory provides a framework that allows for the modeling and study of belief-dependent motivations. These motivations include very real and regularly visible phenomena such as guilt, reciprocity, and the leading example in this thesis: surprise. The aim of epistemic game theory has been to model the reasoning processes that are implicitly assumed when applying traditional game-theoretic solution concepts. The machinery that developed from these endeavours moreover has been used to describe novel reasoning concepts. However, the lessons learned in the past from epistemic game theory regarding traditional games do not directly translate to the more general class of psychological games. With this in mind, I identified three challenges at the start of thesis. Each of them has been tackled in a separate chapter.

CHAPTER 3 considered issues in mathematical modeling that can arise when turning to psychological games. To model cautious reasoning, lexicographic probability systems are conventionally used. In psychological games, capturing belief hierarchies in such probability systems can cause essential information to go lost. We are in fact in need of a way to quantify an 'infinitely less important than'-relationship. Belief hierarchies modeled using non-standard belief accomplish this.

In CHAPTER 4 the Surprise Exam Paradox (SEP) was considered from a game-theoretic perspective and a resolution was provided. In a broader sense, the SEP points to a new argument to be critical on the use of equilibria as solution concepts to psychological games. In psychological games, the common rationale of learning equilibria by repeated interaction is already difficult to apply, as beliefs are not necessarily ex-post observable. The SEP however points out that there can exist belief-dependent motivations, such as surprise, that can be conceptually incompatible with the correct beliefs assumption underlying prominent equilibrium concepts.

CHAPTER 5 focused on the procedure of iterated elimination of strictly dominated choices (IESDC). In traditional games this procedure exactly characterizes rational choices under belief hierarchies expressing common belief in rationality. This is a nice result, as IESDC is simple in its use and intuitively appealing. In psychological games the result does not hold in general. We classified psychological games based on the orders of belief matter for the utility of players. If none of three specific conditions is satisfied by the family a particular game is part of, then this can spell trouble for the characterizing power of the IESDC-procedure. I found that an additional complexity that belief-dependent motivations introduce is that the reasoning about two different rationality-events can overlap and interfere with each other. When this happens, iterated elimination of choices breaks down in its characterization. In such situations decision-makers need to engage in more complex elimination procedures to reason in line with common belief in rationality.

In a broader view, this thesis provides some cautionary lessons for doing analysis in strategic settings with belief-dependent motivations. Most belief-dependent motivations of interest are emotions. Many emotions are known to inhibit rational processes and responses in some way or another. At the same time, psychological games tend to be rather complex. For a player in a psychological game to thus behave rationally in line with even the basic notion of common belief in rationality, will require significant cognitive effort, as CHAPTER 5 elaborated on. One thus needs to be cautious in drawing conclusions about the cognitive abilities of emotional types in strategic settings. In case interactions between non-emotional types behave more in line with a certain solution concept, it may well be that including a player *with* belief-dependent motivations makes the rational reasoning more difficult than without such a player. Additionally, it is possible that some types of belief-dependent motivations are conceptually incompatible with with certain solution concepts, depending on the context. Surprise and equilibrium concepts with the underlying correct beliefs assumption can be seen as one such pair. On the other side of the spectrum, there may be belief-dependent motivations that could pair up more naturally with reasoning assumptions such as the correct beliefs assumption. For now however, I remain speculative about this. Future experimental research in the lab can provide much clearer insight on these matters.

Bibliography

Asheim, G. and Perea, A. (2005). Sequential and quasi-perfect rationalizability in extensive games. *Games and Economic Behavior*, *53*, 15–42.

Attanasi, G., Battigalli, P. and Manzoni, E. (2016). Incomplete-information models of guilt aversion in the trust game. *Management Science*, *62*(3), 648–667.

Attanasi, G., Battigalli, P. and Nagel, R. (2013). Disclosure of belief-dependent preferences in the trust game. *IGIER Working Paper no. 506*.

Bach, C. and Perea, A. (2016). Incomplete information and generalized iterative strict dominance. *Epicenter Working Paper No. 7*.

Balafoutas, L. (2011). Public beliefs and corruption in a repeated psychological game. *Journal of Economic Behavior and Organization*, *78*, 51–59.

Baltag, A., Smets, S. and Zvesper, J. (2009). Keep 'hoping' for rationality: A solution to the backward induction paradox. *Synthese*, *169*, 301–333.

Battigalli, P., Charness, G. and Dufwenberg, M. (2013). Deception: The role of guilt. *Journal of Economic Behaviour and Organization*, *93*, 227–232.

Battigalli, P., Corrao, R. and Sanna, F. (2020). Epistemic game theory without types structures: An application to psychological games. *Games and Economic Behavior*, *120*, 28–57.

Battigalli, P. and Dufwenberg, M. (2007). Guilt in games. *American Economic Review*, *97*(2), 170–176.

Battigalli, P. and Dufwenberg, M. (2009). Dynamic psychological games. *Journal of Economic Theory*, *144*, 1–35.

Battigalli, P., Dufwenberg, M. and Smith, A. (2019). Frustration, aggression, and anger in leader-follower games. *Games and Economic Behavior*, *117*, 15–39.

Battigalli, P. and Siniscalchi, M. (2002). Strong belief and forward induction reasoning. *Journal of Economic Theory*, *106*(2), 356–391.

Bernheim, B. D. (1984). Rationalizable strategic behavior. *Econometrica*, *52*(4), 1007–1028.

Blume, L., Brandenburger, A. and Dekel, E. (1991a). Lexicographic probabilities and choice under uncertainty. *Econometrica*, *59*, 61–79.

Blume, L., Brandenburger, A. and Dekel, E. (1991b). Lexicographic probabilities and equilibrium refinements. *Econometrica*, *59*, 81–98.

Börgers, T. (1994). Weak dominance and approximate common knowledge. *Journal of Economic Theory*, *64*, 265–276.

Brandenburger, A. and Dekel, E. (1993). Hierarchies of beliefs and common knowledge. *Journal of Economic Theory*.

Brandenburger, A., Friedenberg, A. and Keisler, H. (2008). Admissibility in games. *Econometrica*, *76*(2), 307–352.

Brandenburger, A. (1992). Lexicographic probabilities and iterated admissibility. In P. Dasgupta (Ed.), *Economic analysis of markets and games* (pp. 282–290). Cambridge, MA: MIT Press.

Caplin, A. and Leahy, J. (2001). Psychological expected utility theory and anticipatory feelings. *Quarterly Journal of Economics*, *116*, 55–79.

Caplin, A. and Leahy, J. (2004). The supply of information by a concerned expert. *Economic Journal*, *114*, 487–505.

Charness, G. and Dufwenberg, M. (2006). Promises and partnership. *Econometrica*, *74*(6), 1579–1601.

Charness, G., Naef, M. and Sontuoso, A. (2019). Opportunistic conformism. *Journal of Economic Theory*, *180*, 100–134.

Chow, T. Y. (2011). The surprise examination or unexpected hanging paradox. *Working paper*.

Cubitt, R. P. and Sugden, R. (2011). The reasoning-based expected utility procedure. *Games and Economic Behavior*, *71*(2), 328–338.

Dekel, E., Fudenberg, D. and Levine, D. K. (1999). Payoff information and self-confirming equilibrium. *Journal of Economic Theory*, *89*, 165–185.

Dufwenberg, M. and Kirchsteiger, G. (2004). A theory of sequential reciprocity. *Games and Economic Behavior*, *47*(2), 269–298.

Dufwenberg, M. (2002). Marital investments, time consistency and emotions. *Journal of Economic Behaviour and Organization*, *48*, 57–69.
Dufwenberg, M., Jr. and Dufwenberg, M. (2018). Lies in disguise - a theoretical analysis of cheating. *Journal of Economic Theory*, *175*, 248–264.
Elster, J. (1998). Emotions and economic theory. *Journal of Economic Literature*, *36*, 47–74.
Ely, J., Frankel, A. and Kamenica, E. (2015). Suspense and surprise. *Journal of Political Economy*, *123*, 215–260.
Falk, A. and Fischbacher, U. (2006). A theory of reciprocity. *Games and Economic Behavior*, *54*(2), 293–315.
Ferreira, J. L. and Bonilla, J. Z. (2008). The surprise exam paradox, rationality, and pragmatics: A simple game theoretic analysis. *Journal of Economic Methodology*, *15*(3), 285–299.
Fitch, F. (1964). A Goedelized formulation of the prediction paradox. *American Philosophical Quarterly*, *1*(2), 161–164.
Geanakoplos, J. (1996). The Hangman's Paradox and Newcomb's Paradox as psychological games. *Cowles Foundation Discussion Paper*, *No. 1128*.
Geanakoplos, J., Pearce, D. and Stacchetti, E. (1989). Psychological games and sequential rationality. *Games and Economic Behavior*, *1*(1), 60–79.
Gerbrandy, J. (2007). The surprise examination in dynamic epistemic logic. *Synthese*, *155*, 21–33.
Gneezy, U., Kajackaite, A. and Sobel, J. (2018). Lying aversion and the size of a lie. *American Economic Review*, *108*(2), 419–453.
Halpern, J. Y. (2010). Lexicographic probability, conditional probability, and nonstandard probability. *Games and Economic Behavior*, *68*, 155–179.
Hammond, P. J. (1994). Scientific philosopher. In P. Humphreys and P. Suppes (Eds.). Dordrecht: Kluwer.
Harsanyi, J. (1967-1968). Games with imcomplete information played by "Bayesian players". *Management Science*, *14*, 159-182 320-334 486–502.
Holliday, W. H. (2015). Simplifying the surprise exam. *Working Paper*.

Jagau, S. and Perea, A. (2018). *Expectation-based psychological games and psychological expected utility* [Work in progress].

Jagau, S. and Perea, A. (2017). Common belief in rationality in psychological games. *Epicenter Working Paper No. 10*.

Khalmetski, K., Ockenfels, A. and Werner, P. (2015). Surprising gifts: Theory and laboratory evidence. *Journal of Economic Theory*, *159*, 163–208.

Kim, B. and Vadusevan, A. (2017). How to expect a surprising exam. *Synthese*, *194*(8), 3101–3133.

Kripke, S. (2011). On two paradoxes of knowledge. In S. Kripke (Ed.), *Philosophical troubles: Collected papers, vol 1* (pp. 27–51). New York: Oxford University Press.

Li, J. (2008). The power of conventions: A theory of social preferences. *Journal of Economic Behaviour and Organizaiton*, *65*(3), 489–505.

Livio, L. and de Chiara, A. (2019). Friends or foes? optimal incentives for reciprocal agents. *Journal of Economic Behavior and Organization*, *167*, 235–278.

Luce, R. and Raiffa, H. (1957). *Games and decisions: Introduction and critical survey*. New York: Wiley.

Mourmans, N. J. (2017). Reasoning about the surprise exam paradox: An application of psychological game theory. *Working Paper*.

Mourmans, N. J. (2018). Cautious reasoning in psychological games. *Working Paper*.

Mourmans, N. J. (2019). Reasoning in psychological games: When is iterated elimination of choices enough? *Working Paper*.

Myerson, R. (1978). Refinements of the nash equilibrium concept. *International Journal of Game Theory*, *7*, 73–80.

Nash, J. (1950). Equilibrium points in n-person games. *Proc. Nat. Ac. Sc. USA*, *36*(1), 48–49.

Nash, J. (1951). Non-cooperative games. *The Annals of Mathematics*, *54*(2), 286–295.

Patel, A. and Smith, A. (2019). Guilt and participation. *Journal of Economic Behavior and Organization*, *167*, 267–295.

Pearce, D. (1984). Rationalizable strategic behavior and the problem of perfection. *Econometrica*, *52*(4), 1029–1050.

Penta, A. (2015). Robust dynamic implementation. *Journal of Economic Theory*, *160*, 280–316.
Perea, A. (2012). *Epistemic game theory: Reasoning and choice*. Cambridge: Cambridge University Press.
Perea, A. (2014). Belief in the opponents' future rationality. *Games and Economic Behavior*, *83*, 231–254.
Perea, A. (2015). Finite reasoning procedures for dynamic games. In J. van Benthem, S. Ghosh and R. Verbrugge (Eds.), *Models of strategic reasoning: Logics, games and communities* (pp. 63–90). Heidelberg: Springer.
Quine, W. (1953). On a so called paradox. *Mind*, *67*, 382–384.
Rabin, M. (1993). Incorporating fairness into game theory and economics. *American Economic Review*, *83*, 1281–1302.
Rajan, U. (1998). Trembles in the Bayesian foundations of solution concepts of games. *Journal of Economic Theory*, *82*, 248–266.
Robinson, A. (1973). Function theory on some nonarchimedean fields. *American Mathematical Monthly:Papers in the Foundations of Mathematics*, *80*(6), S87–S109.
Sebald, A. (2010). Attribution and reciprocity. *Games and Economic Behavior*, *68*(1), 339–352.
Selten, R. (1975). Reexamination of the perfectness concept for equilibrium points in extensive games. *International Journal of Game Theory*, *4*, 25–55.
Smith, A. ([1759] 2000). *The theory of moral sentiments*. New York: Prometheus Books.
Smullyan, R. M. (1987). *Forever undecided: A puzzle guide to Goedel*. New York: Knopf.
Sober, E. (1998). To give a Surprise Exam, use Game Theory. *Synthese*, *115*(3), 355–373.
Sorensen, R. A. (1988). *Blindspots*. Oxford: Oxford University Press.
Spohn, W. (1982). How to make sense of game theory. In W. Stegmuller, W. Balzer and W. Spohn (Eds.), *Philosophy of economics* (pp. 239–270). Heidelberg; New York: Springer Verlag.

Tan, T. and Werlang, S. R. C. (1988). The Bayesian foundations of solution concepts of games. *Journal of Economic Theory*, *45*, 370–391.

Tolkien, J. (1954). *The fellowship of the ring: Being the first part of the lord of the rings*. London: Allen & Unwin.

Valorisation

> *"Many that live deserve death. And some that die deserve life. Can you give it to them? Then do not be so eager to deal out death in judgement. For even the wise cannot see all ends....My heart tells me he [Gollum] has some part to play yet, for good or ill, before the end; and when that comes the pity of Bilbo may rule the fate of many - yours not least."*
>
> - Tolkien (1954, p.78)

In the above passage from J.R.R. Tolkien's *The Lord of the Rings*, Gandalf lectures Frodo Baggins about the value of life. More in particular about the value of Gollum's life, a wicked creature that poses a direct threat to Frodo and the fate of Middle-Earth, as Frodo holds possession of the powerful One Ring. With Gandalf's words in mind, Frodo chooses to show pity and mercy to Gollum in future encounters with the creature. This choice of kindness sets up the eventual eucatastrophe that saves Frodo and the entirety of Middle-Earth from evil, by having Gollum be alive at the right place and right time to fall with the One Ring into the fiery depths of Mount Doom.

There is an important lesson about **value** in this passage. To be very clear: I am certainly not that humble that I would compare this thesis to the abhorrent creature that is Gollum. Nor am I conceited enough to be saying that the fruits of my still young academic career will ever save Earth at a certain point in time. In fact, assuming such levels of certainty and finality with regard to creating and diffusing value is often unfounded. Frodo's decision to show kindness is one made when faced with heavy uncertainty and a very long path ahead. He supports his own decision however by his acquired understanding that any life has the potential to add some material value over time, if given the opportunity.

Also when attributing societal value to fundamental research, uncertainty and a long time horizon are inherent challenges. In fundamental research one seeks to create knowledge for the sake of knowledge itself. However, by providing a large and well-structured knowledge base to support on, it serves applied researchers in conducting their research. It is applied research that can provide direct societal impact. For economics, one can think here of policy recommendations, market design and patents.

The process of translating academic knowledge to practice and attributing societal value to it is called **valorisation**. This thesis looks at fundamental aspects of psychological game theory and therefore certainly falls into the category of fundamental research. As such, the process of valorisation for this thesis is not straightforward. Covering fundamental research, we must assume that the future societal value that can be attributed to the knowledge created in this thesis will accrue in indirect ways and over an extended period of time. The main avenue for this value creation for practical matters will be experimental research, as will be discussed in what follows.

Both psychological game theory as well as epistemic game theory are grounded in a practical view. The field of psychological game theory is motivated by the observation that in real life many people do not only have self-regarding preferences in the way that traditional game theory models them. Instead, people also exhibit belief-dependent motivations in economic situations, such as guilt-aversion or reciprocity-concerns. Epistemic game theory is motivated by explicitly modeling the reasoning processes of players in a game, instead of using common solution concepts as a black box. An important point of note here is that these black boxes, such as the classical Nash equilibrium, work well in traditional settings that are repeated often, such as in auctions or competition between businesses in markets. The repetition in such traditional game-theoretic settings allows for the learning of beliefs and thus also for eventually achieving correct beliefs, the reasoning assumption behind equilibrium concepts. However, many economic settings are not that often repeated or are even one-shot in-

teractions between multiple players. Equilibrium concepts are known to not perform well in describing decisions made by people under such conditions, because learning cannot take place. Examples include higher than expected effort provision in contests or more specifically overbidding in auctions. With psychological games these problems are amplified, as the variable of interest are the beliefs themselves, which may never be observed or deduced. Then there is not even a relevant basis to learn from. Additionally, as this thesis shows, the assumptions placed on reasoning processes by equilibrium concepts may be incompatible with belief-dependent motivations in certain settings. And even if all necessary assumptions for an equilibrium are met, a common issue in psychological games is that there are often multiple equilibria possible. A priori it is then still not unambiguously clear how players in fact will behave.

All of the previous points stress the necessity for a permissive view on *how* players in a game can reason towards particular decisions. Epistemic game theory provides the appropriate machinery to formalize testable hypotheses for experimental researchers regarding such reasoning processes. The results and discussion in each of the chapters in this thesis can help experimental researchers in doing so for psychological game theory in particular. The range of topics addressed in experimental research that have direct relevance for society is very diverse. Examples include understanding charitable giving when players can experience guilt or shame and tax morale with concerns for reciprocity or again aversion to guilt.

Understanding and describing real-life behaviour is important. However, the ultimate goal of game theory is to be able to provide well-founded predictions for economic scenarios. The accuracy of these predictions depend on many factors, but an important one is the complexity of the decision-problem at hand. Cognitive abilities are limited and so is time available for reasoning in many decision-problems. This is called bounded rationality. The more complex a decision-problem is to reason about, the less accurate the predictions may be. In this thesis we have argued that psychological games can be noticeably hard to

reason about. In CHAPTER 5 in particular we have argued that the simplicity of a psychological game can be partially assessed by whether it is solvable by an iterative elimination procedure that only eliminates (finitely many) choices. We provided conditions for when this is always possible and showed that if it is not solvable by such a procedure it is because of conflict in the underlying reasoning processes that can occur. Certainly more research is needed to investigate whether this intuition of the problem also extends to dynamic games. It will be the task for experimental research to find out whether in practice decision-makers on average do act more in line with predictions in such "simpler" settings than in other settings.

There is a lot of practical relevance to be found in such questions about bounded rationality. Namely, in many economic settings the game is designed with an intention. For instance, we can think of a manager designing for a group of workers with belief-dependent motivations incentive schemes that specifies per output level a certain reward. The manager may for instance have to choose between joint incentives and individual incentives for the workers, or between a scheme where workers have to simultaneously commit to a certain output level and a scheme where they do so sequentially. Given the predictions, the manager can set up an optimal incentive scheme that maximizes her organisation's output. But if the game is hard to reason about, predictions based on a particular solution concept may get inaccurate and potentially lead to a waste of resources. To the extent these resources are government-funded or to the extent the organisation has a strong societal merit, this of course can trickle down to society. The previous is an example in the labour market, but one can also think of relevant examples in public choice settings with local (government) representatives that have belief-dependent motivations or tax audit designs that foster tax morale.

Finally, a few words on possible direct value for practitioners. It would be too much of a stretch to say that knowing how to surprise a student with an announced exam has much direct societal merit. However, the classroom is perhaps the one place where results of the type of

knowledge created in this thesis can create some direct societal impact. Many undergraduate and graduate degrees offer game theory courses. And although the exact results discussed in this thesis are not directly applicable to them, the intuitions may be of use to alumni in the future when they set up or manage their own organisations. For psychological game theory in particular, the intuitions will mostly be of use in small-scale organisation where one can distinguish on an individual basis in some way. Namely, business, corporations or other organisations cannot have psychological motivations. The individual human beings who make up these organisations however can have such motivations. Taking lessons from psychological game theory and acknowledging people may exhibit a wide array of motivations will create better people's managers, which is practically relevant.

In sum, the discussions and results in this thesis are mostly useful to experimental researchers. In general it helps in better formalizing testable hypotheses in strategic settings with belief-dependent motivations. More specifically there may be lessons found with regard to bounded rationality and accuracy of predictions. For the latter point more research is needed, also on a fundamental level. The fruits of such research can be relevant for e.g. labour markets and public choice settings.

Nederlandse Samenvatting

Dit proefschrift onderzoekt fundamentele aspecten van **psychologische spelen**. Met een spel bedoelen we een beslissingsituatie waarin een individu (speler) moet redeneren over de keuze en redenatie voor die keuze van een ander individu om zelf tot een rationele keuze te komen. Hierbij kunt u denken aan de hoogte van een bod tijdens een blinde veiling of de te leveren inspanning voor een groepsproject op de werkvloer. Vooraf weet u niet hoe hoog de anderen gaan bieden, of hoeveel inspanning uw collega's gaan leveren. Derhalve zult u daarover moeten redeneren en verwachtigen over maken. Gebaseerd op deze verwachtingen probeert u dan een uitkomst van voorkeur te bewerkstelligen met uw eigen keuze. spelen waarbij de uiteindelijke relevante uitkomst objectief waarneembaar is, oftewel materialiseert, kunnen als traditionele spelen bestempeld worden. De relevante uitkomst kan gewaardeerd worden in geldseenheden, hoeveelheden van een bepaald object, maar ook hedonisch nut zoals geluk. We noemen dit ook wel het 'nut' van een uitkomst. Als standaardaannames hebben we dat (1) een speler een rationele keuze maakt en (2) een speler een rationeel redenatie-proces heeft begaan om tot haar keuze te komen. Een rationele keuze in een traditioneel spel is een keuze waarvan u verwacht dat het leidt tot het hoogste nut, gegeven uw verwachting over wat de andere speler gaat kiezen. Een dergelijke *verwachting* of *geloof* noemen we een *belief*. In een rationeel redenatie-proces moet u als speler de verwachting hebben dat de tegenstander ook een rationele keuze maakt. Maar tegelijkertijd moet u ook de verwachting hebben dat de andere speler een rationele keuze maakt gestoeld op de verwachting dat u een rationele keuze maakt, enzovoorts. Een oneindige hiërarchie van verwachting (beliefs) noemen we een *belief hierarchy*. En als een belief hierarchy voldoet aan het type restricties hierboven uitgelegd zeggen we dat de belief hierarchy voldoet aan het redenatie-concept van **common belief in rationality**.

Een psychologische spel onderscheidt zich van een traditioneel spel in die zin dat het nut van een speler nu ook direct kan afhangen van de verwachtingen. Of het nut kan direct afhangen van hogere-orde verwachtingen. Een voorbeeld hiervan is uw verwachting van andermans verwachting van uw keuze. De motivaties van mensen voor hun keuzes in het dagelijks leven zijn vaak te herleiden tot zulke hogere-orde verwachtingen. Hierbij kunt u denken aan het ontwijken van schuldgevoel: u voelt zich schuldig wanneer u verwacht dat u andermans verwachtingen niet zult hebt kunnen waarmaken, waar die verwachtigen weer afhangen van de verwachting van die andere speler van wat u gaat kiezen. Een ander voorbeeld, en leidraad in dit proefschrift, is het concept van verrassing: u probeert een ander persoon te verrassen wanneer u een keuze maakt waarvan u verwacht dat het exact de tegenovergestelde keuze is van wat die andere persoon verwacht zal hebben dat u gaat kiezen. Derhalve hangt het verwachte nut van de keuze af van u gelooft dat de andere speler verwacht van uw keuze.

Door de motivaties af te laten hangen van een extra, nieuwe dimensie, neemt de complexiteit van de te analyseren spelen toe. Speltheoretische intuïties die golden voor traditionele spelen zijn niet direct over te brengen naar de meer generieke klasse van psychologische spelen. Bovendien is het niet op voorhand duidelijk welke nuttige (theoretische) resultaten uit de traditionele speltheorie één voor één toepasselijk zijn voor psychologische speltheorie. In dit proefschrift werpen we een epistemische blik op psychologische spelen om enkele van deze aspecten te verklaren.

Hoofdstuk 2 geeft een algemene introductie tot de meest basale concepten en definities die regelmatig terugkeren in het proefschrift. We geven een constructie van belief hierarchies, definiëren het concept van common belief in rationality en geven een definitie voor een (statisch) psychologisch spel.

Hoofdstuk 3 werpt een blik op het modeleren van redenatie-processen in psychologische spelen. In het bijzonder kijken we naar zogenoemd

'voorzichtig redeneren', waar iedere mogelijke keuze van de tegenstander een positieve kans toegekend moet krijgen in de verwachting van een speler. Een belief hierarchy met uitsluitend zulke verwachting kan niet als normaal gemodeleerd worden. Voor traditionele spelen is de meest gangbare optie om lexicografische kanssystemen te gebruik voor de representatie van zulke belief hierarchies met voorzichtige verwachtingen. We tonen in dit hoofdstuk aan dat bij een keuze voor lexicografische kanssystemen er echter relevante informatie verloren kan gaan in psychologische spelen. Verwachtingen gepresenteerd door middel van zogeheten 'non-standard beliefs' bieden een oplossing hier.

In **hoofdstuk 4** weerleggen we het redenatie-concept van common belief in rationality in psychologische spelen met dat van een psychologisch Nash evenwicht. In het hoofdstuk wordt beargumenteert waarom de extra restricties die een psychologisch Nash evenwicht introduceert ten opzichte van common belief in rationality in het bijzonder voor psychologische spelen problematisch kunnen zijn. We doen dit aan de hand van een beschouwing van een paradox uit de logica: de paradox van het verrassende examen (SEP). Een oplossing voor deze paradox wordt gegeven door het probleem te modeleren als een psychologisch spel. Door common belief in rationality toe te passen kan een logische oplossing voor de paradox gegeven worden. Een toepassing van een psychologisch Nash evenwicht leidt echter niet tot een gewenst resultaat. Dit is vanwege de aanname van juiste verwachtingen die impliciet wordt opgelegd op de belief hierarchies bij het concept van een psychologisch Nash evenwicht. Problematisch hier is dat de motivatie van verrassing direct afhangt van de belief hierarchy. In een traditioneel spel is dit niet aan de orde. De beperkingen opgelegd aan de mogelijke belief hierarchies door de aanname van juiste verwachtingen staan haaks op het concept van verrassing in de paradox.

Hoofdstuk 5 beschouwt de procedure van iteratieve eliminatie van strikt gedomineerde keuzes (IESDC). In traditionele spelen is er het resultaat dat stelt dat alle rationele keuze die gemaakt kunnen worden onder common belief in rationality exact gekarakterizeerd worden door IESDC. Dit is een zeer nuttig resultaat, daar IESDC vrij een-

voudig is in gebruik. Bovendien is het een intuïtieve procedure, waar iedere stap in de procedure zelfs gezien kan worden als een redenatie-stap in een redenatie-proces. Dit resultaat is echter niet direct toepasbaar op psychologische spelen. In dit hoofdstuk tonen we aan dat verwachtings-afhankelijke motivaties ervoor kunnen zorgen dat er conflicten in een redenatie-proces kunnen optreden. Het kan bijvoorbeeld zo zijn dat om een bepaalde keuze van u te rationaliseren, er restricties nodig zijn op zowel de verwachting van de keuze van uw tegenstander alsmede de verwachting van de verwachting die de tegenstander zal hebben gehad van uw keuze, de tweede-orde verwachting. Onder common belief in rationality moet u tegelijkertijd verwachten dat uw tegenstander een rationiële keuze maakt, waar die keuze gestoeld kan zijn op de tegenstander's verwachting van uw keuze. Om een keuze te rationaliseren onder common belief in rationality kan het dus in een psychologische spel zo zijn dat er tegelijkertijd verschillende restricties geplaatst moeten worden op uw tweede-orde verwachting. Of op zelfs hogere ordes van verwachtingen. Deze restricties kunnen in conflict met elkaar zijn. Dit soort overlap in restricties kan echter niet worden opgepakt door de IESDC-procedure. We kunnen psychologische spelen karakteriseren door middel van welke ordes van verwachting nutsbepalend zijn. We bewijzen op een constructieve manier de voorwaarden onder welke overlap in restricties in het redenatie-proces nooit kunnen voorkomen. In dergelijke psychologische spelen werkt de IESDC-procedure altijd 'naar behoren'.

De hoop is dat dit proefschrift toekomstig onderzoek verder op weg kan helpen. Voornamelijk is er de hoop dat er voor toegepast onderzoek praktische lessen te vinden zijn met betrekking tot keuzes voor uitkomst concepten en met betrekking tot inzicht in mogelijke redenatie-processen van individuën in situaties met emoties of andere verwachtings-afhankelijke motivaties.

Curriculum Vitae

Niels Mourmans was born on June 4, 1994 in Berg en Terblijt, The Netherlands. He attended high school between 2006 and 2012 at Sint-Maartenscollege in Maastricht, where he received his Atheneum diploma. Afterwards he studied the bachelor Fiscal Economics at Maastricht University. He completed his bachelor in 2015 as an honours student. Subsequently, Niels engaged in the two-year Economic and Financial Research Master. He obtained his M.Sc. degree in July 2017.

After graduation, Niels joined the Department of Quantitative Economics (KE) at Maastricht University in September 2017 as a Ph.D. Candidate, under the supervision of Dr. Andrés Perea and Dr. Elias Tsakas. The results of his research are presented in this thesis. Niels presented his work at various international conferences, such as the the International Conference for Logic and the Foundations of Game and Decision Theory (LOFT) in Milan (2018) and the Lisbon Meetings in Game Theory and Applications in Lisbon (2019).