Reasoning about the Surprise Exam Paradox:

An application of psychological game theory*

Niels J. Mourmans†

EPICENTER Working Paper No. 12 (2017)

Abstract

In many real-life scenarios, decision-makers do not exclusively care for materialized outcomes from decisions they and their co-players make but also display other-regarding preferences such as reciprocity and surprise. Psychological game theory is able to model such belief-dependent motivations. In this paper we discuss the reasoning concepts of common belief in rationality and common belief in future rationality in a psychological game-theoretic setting and use them to provide an explanation for the puzzle of the Surprise Exam Paradox. We consider two versions of the surprise exam game, both in a static and dynamic scenario. In the version that best captures the actual crux of the paradox, we show that, as long as no cautious reasoning is imposed, full surprise is always possible. This contrasts the previous game-theoretic literature on the Surprise Exam Paradox, which relied on equilibrium concepts for traditional and psychological games alike and showed that at most partial surprise is possible under these concepts.

JEL Classification: C72, D03, D83, D84

Keywords: Psychological games, Surprise Exam Paradox, Epistemic game theory, Common belief in rationality, Non-standard beliefs

*I wish to thank my supervisors Andrés Perea and Elias Tsakas for their many useful comments and support throughout this research project.

†Department of Quantitative Economics, School of Business and Economics, Maastricht University, 6200 MD Maastricht, THE NETHERLANDS; EPICENTER, School of Business and Economics, Maastricht University, 6200 MD Maastricht, THE NETHERLANDS. Email: n.mourmans@maastrichtuniversity.nl

1 Introduction

Traditional concepts in game theory assume decision-makers only care about outcomes that result from decisions made by themselves and others. Many real-life decisions do not exclusively depend on such materialized outcomes however. In interactions between individuals, decision-makers often display motivations that are fuelled by altruism, feelings of reciprocity or further other-regarding preferences. Many of such psychological payoffs rely on what others expect the decision-maker to think and do. Traditional game theory is inapt to truly capture such motivations: it assumes that decision-makers solely care for the *decisions* of others when deciding upon their optimal course of action.

The field of psychological game theory is a response to these considerations and studies the interaction of individuals with belief-dependent motivations. It has allowed for modeling many different intention-based emotions in the framework of game theory, such as reciprocity (Rabin, 1993; Dufwenberg and Kirchsteiger, 2004), anger (Battigalli et al., 2015) and aversion to perceived cheating (Dufwenberg and Dufwenberg, 2016). In comparison, research on the theoretical foundations of psychological games has been quite limited thus far. Whereas the theory of psychological games has mostly been focusing on psychological Nash equilibrium (Geanakoplos et al., 1989) and psychological sequential equilibrium (Battigalli and Dufwenberg, 2009), little is known about more basic notions of iterative reasoning such as common belief in rationality (Brandenburger and Dekel, 1987; Tan and Werlang, 1988) and common belief in future rationality (Perea, 2014) in psychological games. Though steps have been made already (Battigalli and Dufwenberg, 2009; Bjorndahl et al., 2016; Sanna, 2016; Jagau and Perea 2017), much still remains to be explored, especially in dynamic settings.

A better understanding of the reasoning processes underlying psychological games could help in shedding light on questions that the current theoretical literature has not yet been able to provide a satisfactory answer to. An interesting case we will consider in this regard is *The Surprise Exam Paradox*. Paradoxes have had a central role in studying human reasoning. They highlight flaws or limits in the understanding of a whole range of different problems. Among those, the Surprise Exam Paradox in particular is a puzzle that has garnered much interest, from multiple academic fields. It could be described as follows:

A geography teacher announces to his student that during the next week he will be given an exam. However, the teacher does not announce on which day of the week the exam will take place. That is, he lets the student know that he wishes to surprise him. Reasoning by backward induction will lead the student to believe 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 the exam. Once Wednesday comes about, the student finds an exam lying on his desk. He did not expect this, by the discussion above. It is a paradoxical outcome in his eyes.

Though a seemingly simple problem, the sheer size of the literature on the paradox shows its value not only to game-theorists, but also to philosophers and logicians (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 (Shaw, 1958). That is, surprising the student can be defined by the announcement that: (1) the exam will take place next week and (2) that the exact day on which it will take place is not deducible in advance for the student by the preceding statement. This announcement in itself is found to be self-contradictory (Smullyan, 1987). The philosophical school of epistemology has tried to resolve this self-contradiction 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 the approach of Quine (1953).[1] He found that an exam can at least come as a surprise to the student on Friday if the student does not accept the teacher's announcement to be true[2]More specifically, Quine (1953) showed that the student cannot be surprised if he *knows* the announcement. This assumes some level of caution by the student about the announcement, implying a role for cautious reasoning within the paradox. Even though it is realistic for the student to doubt the announcement, it does not coherently explain how the student can justifiably believe the announcement to be false. Moreover, it does not give insight into why in the end the teacher is still able to surprise the student. These are issues neither the epistemic school nor the logicians properly address. This all directly highlights the importance of studying how the teacher can possibly believe to surprise the student within this debate, instead of just looking at the student's reasoning process. A more interactive framework is needed to analyse such a matter. Game theory is able to provide this. In this regard, two questions are important to ask ourselves: how do we model the Surprise Exam Paradox in the language of game theory and what solution concept(s) do we use to analyse it?

Only recently have game-theorists tackled the puzzle, with the literature on the topic remaining scarce. Sober (1998) models the Surprise Exam Paradox as an iterated matching-pennies game, in which the student chooses what he anticipates. Using an equilibrium analysis, he finds that with some positive probability the student can be surprised in a unique, mixed-strategy equilibrium. Sober argues that because the teacher adopts a distribution of choices, the student cannot always correctly guess what the teacher is going to do. Ferreira and Bonilla (2008) try to reconcile the results found in Sober with pragmatic logic. They argue that many of the knowledge and reasoning concepts introduced to the problem by logicians are not needed to understand the paradox in a game-theoretic setting. They mostly confirm the findings of Sober. Also modeling beliefs as actions, they find for multiple game forms that the teacher can indeed at best partially (with positive probability less than one) surprise the student in a subgame perfect equilibrium, by adopting a random distribution of choices. That is to say, there is only a mixed-strategy equilibrium in each game form. From the teacher's reasoning perspective, this implies that full surprise (surprise with probability one) is not possible. However, from the paradox scenario described earlier, it is clear that there exist events for the teacher to believe in, in which fully surprising the student is possible. An additional matter of consideration with this traditional game-theoretic approach is that a distinction is made between beliefs the student may hold and what the student may choose to anticipate about what the teacher will choose. This allows for many conceivable scenarios in which the student may believe one thing, but anticipates something different, whereas conceptually beliefs and anticipations are the same. This inevitably leads one to seek for prudential reasons that help in explaining the possibility of the student being willfully blind with respect to his own beliefs, in order to choose to anticipate something else (Sober, 1998). However, such a discussion distracts from the actual core of the paradox. The teacher is merely interested in surprising the student by doing exactly the opposite of what he believes the student is thinking. Any motivations the student may have in convincing himself what he chooses to anticipate are thus irrelevant for the teacher.

[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.

[2].

Geanakoplos (1996) takes a different perspective on the matter which circumvents the previous issue. He uses psychological game theory to model the teacher's utility as a function of his second-order beliefs. This crucially includes the belief the teacher has about what the student believes the teacher is going to choose. Employing the concept of psychological Nash equilibrium (Geanakoplos et al., 1989), in one of two versions considered it is shown that under said concept, common knowledge about the teacher's belief-dependent motivations allow the student to completely predict when the exam will happen. Though a manner is found in which the psychological game can be transformed to allow for near full surprise under a psychological subgame perfect equilibrium, also the analysis of Geanakoplos shows us that equilibrium concepts such as the (psychological) Nash equilibrium do not provide us with all the right tools to predict full surprise and thus to, in a sense, resolve the Surprise Exam Paradox. In a game-theoretic framework, a broader perspective is thus warranted.

In light of the limited work on the epistemic foundations of psychological games, the Surprise Exam Paradox presents itself as an interesting thought experiment. Being a game that fully revolves around a teacher that wishes to surprise a student by giving an unexpected exam, it neatly captures the idea of belief-dependent motivations. Moreover, it is part of a class of games that allows for a straightforward transformation from static scenarios to dynamic scenarios. In both static and dynamic scenarios we will consider two variants of the game in order to provide a complete picture of the game-theoretic reasoning behind the supposed paradox. At the same time, the approach from epistemic game theory can provide a fresh take on the much debated mechanisms behind the paradox on itself as well. Namely, it is able to formalize how intuitions from logic about reasoning and beliefs are inherently present in a game-theoretic discussion of the paradox.

The purpose of this paper is twofold. First, to gain a deeper understanding of the reasoning processes of decision-makers in psychological games, we discuss and also expand upon the epistemics of psychological game theory. We consider the concepts of common belief in rationality and common belief in future rationality as basic modes of reasoning and introduce the notion of caution to the setting of psychological games. Second, using the theoretical foundations discussed and introduced in this paper, our goal is to add to the scarce game-theoretic literature that tries to resolve the Surprise Exam Paradox. Overall, we wish to answer the following question in this paper: *Can the concepts of common belief in rationality and common belief in future rationality resolve the Surprise Exam Paradox, and if so, how?*

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 his belief hierarchy a player's rationality is questioned (Battigalli and Dufwenberg, 2009; Bjorndahl et al., 2016; Sanna, 2016; Jagau and Perea, 2017). There is an important difference to be found in the definition of optimality however, as in psychological games now also belief-dependent motivations come into play. If we extend the paradox game to the dynamic scenario, a comparable distinction is found for the concept of common belief in future rationality, where at no point in a decision-makers belief conditional hierarchy rationality pertaining to now and in the future is put into doubt.

We consider two versions of the surprise exam paradox to apply these concepts: one where surprise by giving and not giving the exam is possible and one where only surprise from giving the exam is possible. We find that in both cases, full surprise is possible under a belief hierarchy that expresses common belief in rationality. A crucial element in the version where surprise is only possible from giving the exam is that the student should deem it possible that the teacher cannot simultaneously give the exam and surprise the student. We further elaborate on this by introducing the notion of caution to psychological games, modeled by non-standard probabilities. We show that if the student is a cautious reasoner, the teacher cannot hope to fully surprise the student in any way. These findings translate to the dynamic scenario, when considering common belief in future

rationality. 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 paper is organized as follows. In Section 2 we formally define the concept of a static psychological game. Moreover, the static reasoning concept of common belief in rationality in psychological games will be discussed, as well as its link to the equilibrium concept of psychological Nash equilibrium. This will all be applied in Section 3, where we analyse several variants of the Surprise Exam Paradox in a static scenario. In Section 4, cautious reasoning in psychological games will be introduced and applied to the paradox. In Section 5 we extend the Surprise Exam Paradox to a dynamic scenario, and discuss the backward induction reasoning concept of common belief in future rationality and its link to psychological subgame perfection. This framework is then applied to two versions of the Surprise Exam Paradox in a dynamic scenario. Finally, we conclude with some closing remarks in Section 6.

2 Preliminaries

We start this section by giving a formal definition of a psychological game. Subsequently, common belief in rationality in psychological games and psychological Nash equilibrium are discussed. The discussion will be general and applies to any static psychological game.

2.1 Static psychological games

Psychological games have been developed to model decision-problems where the utility of a player is allowed to explicitly depend on his higher-order beliefs. We will first concentrate on psychological games in a static scenario. Following Jagau and Perea (2017), we can formally define such a static psychological game as follows.

Definition 2.1. *A **static psychological game** 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 belief hierarchies for player i that express coherency and common belief in coherency, and

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

representing player i's utility function.

A belief hierarchy $b_i \in B_i$ for a player i consists of a belief about the opponent's set of choices, a belief about the opponents' choices and the opponents' beliefs about their opponents' choices, and so on. Hence, a belief hierarchy is a chain of beliefs, where each component of the chain represents a certain *order* of belief. For instance, b_i^1 represents the first-order belief about the opponents' choices and b_i^2 represents the second-order belief about the opponents' choices combined with the opponents' beliefs about their opponents' choices. Note that in a psychological game, the utility of a player may depend on any order of belief. In the Surprise Exam game we are considering in this paper, the utility function depends on the second-order belief b_i^2 specifically. Although utility functions in psychological games can depend on any higher-order belief, second-order beliefs will therefore be the key focus of this paper. In addition to this, the condition of coherency ensures that

[3] 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.

any k-th order belief does not contradict the $(k-1)$-th order of belief (Brandenburger and Dekel, 1993). Though not a direction taken here, Sanna (2016) shows one may also abstain from imposing the assumption of coherency and common belief in coherency a priori on the belief hierarchies when defining a psychological game.

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 may depend non-linearly on the full belief hierarchy.

As belief hierarchies involve infinite chains, writing them down explicitly can be a very cumbersome endeavor. Fortunately, there are methods for modeling such infinite chains of beliefs 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 2.2 (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 specifies a probability distribution $b_i(t_i)$ over the 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 his opponents.*

The coherency and common belief in coherency assumption assures here that any belief hierarchy can in fact be represented by a type in an epistemic model. By means of an epistemic model as defined above we can furthermore write the utility function as $u_i(c_i, \beta_i(t_i))$, where $\beta_i(t_i)$ represents the entire belief hierarchy that is generated by type t_i. Finally, whenever t_i and t_i' induce the same belief hierarchy, we should have $u_i(c_i, \beta_i(t_i)) = u_i(c_i, \beta_i(t_i'))$.

2.2 Common belief in rationality

In order to analyse basic reasoning in a static psychological game like the static Surprise Exam Paradox, we will first look at the concept of common belief in rationality in psychological games, as defined in Jagau and Perea (2017). It should be mentioned here that Bjorndahl et al. (2016) define rationalizability in language-based games, which is an even larger class of games, in a similar vein. Moreover, the static version of the common strong belief in rationality concept of Battigalli and Dufwenberg (2009) is equivalent to common belief in rationality in psychological games as well.

The concept of common belief in rationality in psychological games is similar to that of traditional games. Also in psychological games, common belief in rationality entails that every player i believes in his opponents' rationality, believes that his opponents believe in their opponents' rationality, and so on and so forth. A crucial difference, however, can be found in defining optimal choices.

Definition 2.3 (Optimal choice in a static psychological game)**.**
*Consider an epistemic model $M = (T_i, b_i)_{i \in I}$ 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 $\forall c_i' \in C_i : u_i(c_i, \beta_i(t_i)) \geq u_i(c_i', \beta_i(t_i))$.*

So optimality of a particular choice in a psychological game is defined as that choice being optimal given a belief hierarchy instead of just the first-order belief. Building on this notion, the concept of common belief in rationality remains similar to that of common belief in rationality in traditional games (Bernheim, 1984; Pearce, 1984; Brandenburger and Dekel, 1987; Tan and

Werlang, 1988). That is, we can first define what it means for a type to believe in an opponent's rationality.

Definition 2.4 (Belief in the opponents' rationality)**.**
Consider an epistemic model $M = (T_i, b_i)_{i \in I}$ with a type $t_i \in T_i$ for player i within that epistemic model. Type t_i of player i ***believes in the opponents' rationality*** *if type t_i only assigns positive probability to opponents' choice-type combinations $(c_j, t_j) \in C_j \times T_j$ where the choice c_j is optimal for the type t_j, for every $j \neq i$.*

Analogously to Tan and Werlang (1992), we can subsequently iterate this argument in order to define what common belief in rationality in a psychological game entails.

Definition 2.5 (Common belief in rationality)**.**
Consider an epistemic model $M = (T_i, b_i)_{i \in I}$. 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. [4]

Definition 2.6 (Rational choice under common belief in rationality)**.**
We say that choice c_i can be rationally made by player i under common belief in rationality if there is an epistemic model $M = (T_i, b_i)_{i \in I}$ and a type $t_i \in T_i$ such that t_i expresses common belief in rationality, and c_i is optimal for t_i.

2.3 Psychological Nash Equilibrium

Previous research on belief-dependent motivations in game-theoretic settings often revolved around the idea of a psychological Nash equilibrium (Geanakoplos et al., 1989). This concept provides a generalisation of the traditional solution concept of a Nash equilibrium, suitable for analysing psychological games. 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. A psychological Nash equilibrium, on the other hand, corresponds to a full belief hierarchy. In line with the notion of a traditional Nash equilibrium, a psychological Nash equilibrium too requires every 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*, in line with Perea (2012). Such a simple belief hierarchy is generated by a combination of probabilistic beliefs $\sigma = (\sigma_i)_{i \in I}$ that are *independent* of each other, where $\sigma_i \in \Delta(C_i)$ for all $i \in I$. For every player i, σ_i

[4] Sanna (2016) and Jagau and Perea (2017) provide algorithms that characterize the choices that can be made under common belief in rationality in static psychological games.

thus is a probability measure over player i's choice set. 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_{-i} about his opponents' choices, where $\sigma_{-i} = \prod_{j \neq i} \sigma_j$. In addition, it states that *(ii)* player i believes that every opponent opponent j has belief σ_{-j} about his opponents' choices, *(iii)* that player i believes that every opponent j believes that every other player player $k \neq j$ holds belief σ_{-k} about his opponents' choices, *(iv)* et cetera.

We are now in a position to define a psychological Nash equilibrium.

Definition 2.7 (Psychological Nash equilibrium)**.**
The combination of first-order beliefs $(\sigma_i)_{i \in I}$ constitutes a ***psychological Nash equilibrium*** *if*

$$\forall i \in I : \sigma_i(c_i) > 0 \Rightarrow \forall c_i' \in C_i : u_i(c_i, \beta_i(\sigma)) \geq u_i(c_i', \beta_i(\sigma)).$$

The manner in which we formulate a psychological Nash equilibrium here diverges somewhat from the one in Geanakoplos et al. (1989). Usually, one would denote by the set σ the vector of mixed profiles, where σ_i represents the (randomized) choice for player i. We are however interested in the individual reasoning processes of players and thus their beliefs. Finally, it should be pointed out that a psychological Nash equilibrium has a natural link to the concept of common belief in rationality, analogously to how a standard Nash equilibrium relates to common belief in rationality. Namely, 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.

3 Surprise exam: static situation

With these tools at hand, let us turn to the central game in this paper: the surprise exam. As reviewed earlier, the Surprise Exam Paradox has been considered many times in the past, in many different shapes and forms. We too shall consider two different forms 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.

Let us first consider the static scenario of the paradox. A teacher announces on a Friday to his student that *next week* on either Thursday or Friday he intends to give the student an exam. However, he will not announce the exact day to the student. Namely, the goal of the teacher is to surprise the student. The student himself takes a passive role in the game, yet his beliefs matter for the utility of the teacher.

3.1 Surprise by giving or not giving exam

A first form of the game could consider 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. The corresponding game is portrayed in matrix form in Table 1 where $0 < \eta \leq 1$. In this table (and in the tables to come), the rows correspond to the teacher's possible choices, whereas the columns capture the teacher's *extreme second-order expectations*. The extreme second-order expectations are enough to represent the teacher's utility in matrix form. Namely, the teacher does not care for all information conveyed in his second-order beliefs. To surprise the student, only the expectation about what the student believes the teacher to choose is relevant for the teacher's utility. We can furthermore assume that utility depends linearly on said second-order expectations. As such, the extremes of the distribution of the teacher's expectations are sufficient to represent the teacher's utility in Table

Table 1: *Surprise by giving or not giving exam*

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

1. This class of games is what Jagau and Perea (2017) refer to as a belief-linear expectation-based games.

If the teacher chooses to give the exam on Thursday and the student expects him to do so, the teacher receives 0 utility. However, if the student would believe the teacher will give the exam on Friday, the teacher receives utility of 1. If the teacher chooses Friday and the student believes the teacher will give the exam on Thursday, the teacher gets η. This η corresponds to a small surprise: the teacher still receives the highest amount of utility if he surprises the student by giving the exam on Thursday. If the student believes, on the other hand, that the teacher will give the exam on Friday, the teacher gets 0 utility. This is akin to the type of surprise game Geanakoplos (1996) considers.

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 (partially) surprise the student. The epistemic model in Table 2, with its corresponding beliefs diagram in Figure 1, provides an answer to this. In this epistemic model, we see that the teacher has a type t_1 and a type t_1', each deeming one type of the student possible. Type t_2 of the student holds the belief that the teacher is of type t_1' and chooses Friday and type t_2' thinks the teacher is of type t_1 and chooses Thursday. To show that a type of the teacher expresses common belief in rationality it is sufficient to show that every type in the model expresses 1-fold belief in rationality.

Let us start at type t_2 of the student. Type t_2 of the student holds the belief that the teacher is of type t_1' and will give the exam on Friday. This is a reasonable belief to hold for type t_2, in the sense that it expresses 1-fold belief in the opponent's rationality: type t_1' namely believes the student believes the teacher will choose to give the exam on Thursday. If that is the case, then it is indeed optimal for the teacher, given he has beliefs induced by type t_1', to give the exam on Friday, as $\eta > 0$. Type t_2' too believes in the teacher's rationality: t_2' believes the teacher is of type t_1 and gives the exam on Thursday. Type t_1 of the teacher the student believes that the teacher will give the exam on Friday. Indeed, then it is optimal for the teacher to give the exam on Thursday ($1 > 0$) and hence type t_2' also believes in the opponent's rationality.

Types t_1 and t_1' by definition believe in the opponent's rationality, as the student does not have

Table 2: Epistemic model for "Surprise by giving or not giving exam"

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)

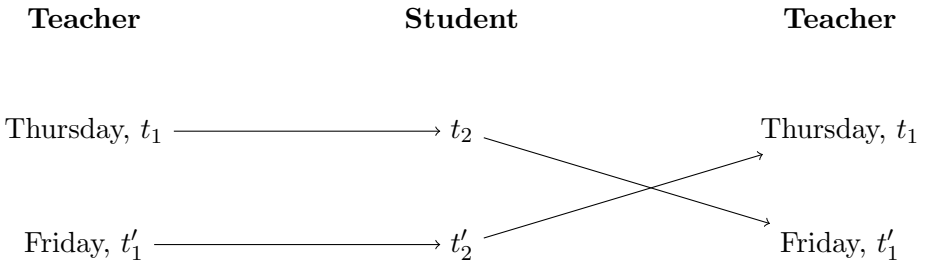

Figure 1: Beliefs diagram for "Surprise by giving or not giving exam"

any choices to make. Since every type in the model expresses 1-fold belief in rationality, it follows that every type in fact expresses *common belief in rationality*. Since type t_1 believes the student believes the teacher will give the exam on Friday, the teacher, given he is of type t_1, can rationally choose to give the exam on Thursday yet still *fully* surprise the student (i.e. getting a utility 1). Moreover, this belief is part of a belief hierarchy that is reasonable in the sense that it expresses common belief in rationality. 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. So giving the exam on Thursday or on Friday can both reasonably come as a (full) surprise to the student.

This result differs significantly from the findings in Geanakoplos (1996), where the concept of a psychological Nash equilibrium is applied to the game in Table 1. In fact, this game has, for a given η, a unique psychological Nash equilibrium given by the belief σ_1 where $\sigma_1(Th) = \frac{1}{\eta+1}$. The proof is elementary, and is left to the reader. In this equilibrium, the teacher will get $u_1(Th, \beta_1(\sigma_1)) = \frac{\eta}{\eta+1}$ from choosing Thursday. From choosing Friday, the teacher will also get $u_1(Fr, \beta_1(\sigma_1)) = \frac{\eta}{\eta+1}$. Believing to surprise the student with probability $\frac{1}{\eta+1}$ by choosing Friday or with probability $\frac{\eta}{\eta+1}$ by choosing Thursday is the best the teacher can hope for. Since neither of these probabilities will ever be equal to 1, a psychological Nash equilibrium will never allow for a full surprise. This is in stark contrast to what we found under the concept of common belief in rationality, where we found a belief hierarchy that supports a choice leading to full surprise.

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, and so forth. 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. In this version of the paradox there is only one psychological Nash equilibrium to reason from for the teacher. There is however no particular argument why the teacher should hold the sort of beliefs as prescribed by the psychological Nash equilibrium. Even more so, depending on the value of η, the distribution of probabilities in the teacher's belief might be rather arbitrary. It would be rather unnatural to think that the student would be correct about such arbitrary beliefs. The belief hierarchies under common belief in rationality do not suffer from the same problem. The epistemic model we constructed is just one example of a set of belief

Table 3: *Only surprise possible by giving exam*

		Beliefs Student	
		$Thursday$	$Friday$
Teacher	$Thursday$	0	1
	$Friday$	0	0

hierarchies that express common belief in rationality under which the teacher can either fully or partially surprise the student. More in particular, the psychological Nash equilibrium corresponds to only one of those possible belief hierarchies.

3.2 Only surprise possible from giving the exam

A second version of the game we can consider is portrayed in Table 3. This situation is perhaps closer to the actual crux of the paradox. Here the teacher can only surprise the student the moment he gives the exam. As a result, once Friday has come about and the exam is still not given, the student *knows* that the exam happens on that day, giving the teacher a utility of 0. It can be shown that the only possible psychological Nash equilibrium here is when the teacher believes the student thinks with probability one that the teacher will give the exam on Thursday. Hence, no surprise would be possible at all. Namely, consider a scenario in which $\sigma_1(Th) \neq 1$. This implies that the teacher would think the student believes the teacher 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 and consequently fully believe the teacher will choose Thursday. Hence, we must have $\sigma_1(Th) = 1$, a contradiction. The correct beliefs assumption here implies that the student *knows* what the teacher is thinking and can thus predict the rational choices that the teacher may consider. This defeats any purpose of surprise, which fully depends on being able to do something that an opponent will not be able to predict.

It is of course paradoxical to announce to give a surprise exam, but not being able to surprise the student. So let us resume with what epistemic game theory tells us about this problem: can we find a type for the teacher that expresses common belief in rationality such that he can still surprise the student? Indeed, there is a belief hierarchy that expresses common belief in rationality and such that the teacher can give the exam on Thursday and *fully* surprise the student. Table 4 shows an epistemic model that includes a type that fits this requirement. In fact, it is identical to the epistemic model in Table 2 (See also Figure 1). Like before, let us start at type t_2 of the student in the model, who believes that the teacher is of type t_1' and will give the exam on Friday. Friday can only be optimal to choose as long as the teacher believes that the student expects the

Table 4: Epistemic model for "Only surprise possible by giving exam"

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)

teacher to give the exam on Thursday. Type t_1' indeed believes that the student believes the teacher will give the exam on Thursday. Hence type t_2 believes in the opponent's rationality. Type t_2' of the student believes the teacher is of type t_1 and will give the exam on Thursday. Whatever the teacher believes, Thursday is always a rational choice as its minimum expected utility is equal to the maximum utility of Friday, being 0. So type t_2' also believes in the opponent's rationality. Types t_1 and t_1' of the teacher always believe in the opponent's rationality by construction, since they do not have to assign probabilities to choices, but only to types. Hence, every type in the epistemic model believes in the opponent's rationality. Consequently, every type also expresses common belief in rationality. Since type t_1 of the teacher believes the student believes that the teacher will give the exam on Friday, the teacher's type t_1 can *fully* surprise the student by choosing Thursday and still express common belief in rationality.

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. 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 therefore must 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. The idea that the student can reasonably doubt the validity of the teacher's announcement is what allows the teacher to believe to be able to vindicate his announcement in the first place. The belief hierarchies described here in a game-theoretic setting manage to capture this idea.

It appears to be the case, however, that the teacher can only believe he can fully surprise the student if he believes the student believes with certainty that the teacher will choose Friday and thus forgo a possible surprise. To formally show this, we will introduce in Section 4 the notion of caution in psychological games.

4 Cautious reasoning in the Surprise Exam Paradox

Much like in traditional game theory, it might be too much of a stretch to assume that players in the game may completely disregard a choice c_j from a choice-set C_j of the opponent in his beliefs. Even though a player might be fairly certain about what his opponent is going to do, some doubt may always remain about the other choices available to an opponent. In other words, the player may consider a first choice *infinitely more likely* to be chosen than a second, but nevertheless consider the second choice as well. Such reasoning is captured by the notion of caution.

In traditional games, several methods have been utilised to capture cautious beliefs. The first is Selten's (1975) *trembling hand* argument where a belief for player i about player j's choice does not consist of a single probability distribution, but a whole sequence $(b_i^n)_{n\in\mathbb{N}}$. In this sequence, every element of the sequence assigns positive probability to every possible choice for player j. As a result, every belief in the sequence in cautious. It is however not an effective method to use when trying to make exact statements about players' preferences, as we can only use arguments that rely on the long-run behaviour of such a sequence. In epistemic game theory the usage of lexicographic belief systems is prominent. First introduced by Blume et al. (1991a, 1991b), this concept also entails sequences of beliefs, though organized into finitely many different levels. However, as is argued in Mourmans (2017), lexicographic beliefs contain insufficient information to consistently capture preferences of cautious reasoners in *psychological* games if one relies on epistemic models..

We are thus in need of a method to capture cautious beliefs without having to rely on infinite sequences, yet that does allow us to derive unambiguous preferences over choices. Fortunately, there does exist a method that is able to combine both. The idea of *non-standard analysis* goes

back to at least Robinson (1973) but was first introduced to game theory by Hammond (1994). Its use is similar to the trembling-hand argument in the sense that they also assign very small numbers to events that are highly unlikely to happen. However, instead of relying on infinite sequences, non-standard analysis entails assigning an *infinitesimal* ϵ, which is a non-real, positive number. Insights in the usefulness of non-standard analysis to game theory have been provided by Hammond (1994), who showed equivalence results between probability systems from non-standard analysis and lexicographic probability systems, and more recently by Halpern (2010), who amongst other things showed that this equivalence only holds as long as the state-space is finite.

As both Hammond (1994) and Halpern (2010) define it, infinitesimals can be found on an extended field of numbers $\mathbb{R}^*$, also called the non-Archimedean field. This field *contains* the real line $\mathbb{R}$, but also hyperreal numbers that do not satisfy the so-called Archimedean property. The Archimedean property entails that for each positive real number $r \in \mathbb{R}$, there exists another real, positive number s such that $r > s$.[5] Consequently, a real number can never truly become infinitely small. An infinitesimal, that is on the extended field $\mathbb{R}^*$, however can. That is, we have an infinitesimal $\epsilon \in \mathbb{R}^*$ if $\epsilon > 0$ and $\epsilon < r$ for all $r \in \mathbb{R}$ with $r > 0$.

Though the field $\mathbb{R}^*$ features numerous complexities (Halpern, 2010), there is a property of the field that is important to highlight here: if we have $r, s \in \mathbb{R}^*$, $r, s > 0$ such that $\frac{s}{r}$ is an infinitesimal, then we say s is *infinitely smaller* than r. The closest real number to $\frac{s}{r}$ is thus 0. This closest real number always exists, and is referred to as the standard part of $\frac{s}{r}$. In other words, it should be case that if we have $st(s/r) = 0$, where $st(s/r)$ denotes the standard part of $\frac{s}{r}$, then for all $a \in \mathbb{R}$ and $a > 0$ we have $a \cdot s < r$. This property is especially important when trying to capture cautious beliefs in psychological games by non-standard probability distributions. It allows us to quantitatively establish when one event is deemed infinitely more likely to occur than another. More specifically to the setting of game-theory, we can now say that if a choice-type combination is assigned an infinitely smaller probability compared to another choice-type combination, it is deemed *infinitely less likely* to occur.

We can use these characterisations to adapt the notions of caution and primary belief in rationality (Perea, 2012), which is akin to Brandenburger (1992)'s concept of permissibility, to fit with psychological games. In order to do so, we first need to formally define what a non-standard probability distribution and an epistemic model based on such distributions entail.

Definition 4.1 (Non-standard probability distribution)**.**
Consider a finite set X. A ***non-standard probability distribution*** *p on X assigns probabilities $p(x) \in \mathbb{R}^*$, where $p(x) \geq 0$, such that $\sum_{x \in X} p(x) = 1$. Then $\Delta^*(X)$ denotes the set of all non-standard probability distributions over X.*

The leap to an epistemic model with *non-standard beliefs* is now easy to make.

Definition 4.2 (Epistemic model with non-standard beliefs)**.**
Consider a psychological game G. 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 of opponents' choice-type combinations $C_{-i} \times T_{-i}$.*

A cautious player does not rule out any choice for an opponent. One subtlety in defining caution is however that a player may consider multiple types for the opponent is his belief. As

[5] More specifically, the argument goes that every ordered field F contains the set of natural numbers $n \in \mathbb{N}$. The Archimedean property entails that we can find for every real, positive number $r \in \mathbb{R}$ a natural number $n \in \mathbb{N}$ such that $r > \frac{1}{n}$.

such, caution should be defined for each type that is *deemed possible*. Given an epistemic model $M = (T_i, b_i)_{i \in I}$ with non-standard beliefs, a type t_j of an opponent j is deemed possible by player i if $b_i(t_i)(c_j, t_j) > 0$ for some $c_j \in C_j$. In the same epistemic model, we say player i deems a choice-type combination (c_j, t_j) possible for player j if $b_i(t_i)(c_j, t_j) > 0$. Note here that $b_i(t_i)(c_j, t_j) > 0$ is also possible when its standard part is zero. In this case, (c_j, t_j) receives infinitesimal probability.

Caution is then defined as follows.

Definition 4.3 (Cautious type)**.**
Consider an epistemic model $M = (T_i, b_i)_{i \in I}$ with non-standard beliefs and a type t_i for player i within the model. ***Type t_i is cautious*** *if, whenever it deems possible an opponent's type t_j for some player j, then for every $c_j \in C_j$ it deems the choice-type pair (c_j, t_j) possible.*

The notion of optimality remains the same. The only difference is that the belief hierarchy now contains non-standard beliefs instead of standard beliefs. However, we cannot maintain the same concept of believing in the opponent's rationality here. Requiring a type to be cautious and to believe in an opponent's rationality may be incompatible. Namely, believing in an opponent's rationality implies assigning positive probability only to opponent's choices that are optimal for the opponent, yet caution requires one to consider all the opponent's choices, including the suboptimal ones. We can however adapt a weaker form of believing in an opponent's rationality. Very akin to the concept of *permissibility* as developed by Brandenburger (1992) and Börgers (1994), we consider the notion of *primary belief in rationality* (similar to Perea (2012)).

Definition 4.4 (Primary belief in an opponent's rationality)**.**
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, $b_i(t_i)(c_j, t_j) \in \mathbb{R}_+$ only if c_j is optimal for t_j.*

By $\mathbb{R}_+$ we denote the set of all positive, real numbers and by $\mathbb{R}_+^*$ the set of all positive numbers on the extended field of real number $\mathbb{R}^*$. Note here that $\epsilon \in \mathbb{R}_+^*$, but $\epsilon \notin \mathbb{R}_+$. Just like with common belief in rationality, we can now iterate belief in caution.

Definition 4.5 (Common full belief in caution)**.**
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.*

If we do a similar iteration process for primary belief in rationality, we get common full belief in primary belief in rationality.

Definition 4.6 (Common full belief in primary belief in rationality)**.**
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, similarly to a rational choice under common belief in rationality, 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 is cautious and expresses common full belief in caution and primary belief in rationality.

Let us return to the game depicted in Table 3, where only surprise from giving the exam is possible. It turns out that if the student is a cautious reasoner, there is little the teacher can do to surprise the student if there is common full belief in caution and primary belief in rationality. To formally see why, consider some epistemic model $M = (T_i, b_i)_{i \in I}$, with a cautious type t_2^* for the student. Suppose the student with type t_2^*, with some *non-standard* (real or non-real), positive probability, believes that the teacher will choose Friday and with some non-standard, positive (real or non-real) probability believes that the teacher will choose Thursday (an example can be found in Table 5). Hence, $b_2(t_2^*)(Fr) > 0$ and $b_2(t_2^*)(Th) > 0$. Under such beliefs, the only optimal choice for the teacher is Thursday. If the student primarily believes in the teacher's rationality and has a belief hierarchy that expresses common full belief in caution, then he must believe that it is infinitely more likely that the exam is going to happen on Thursday than on Friday. Under common full belief in caution and primary belief in rationality, it follows that the teacher can only rationally give the exam on Thursday, while believing the student primarily believes that the teacher will in fact give the exam on Thursday. Hence, (almost) no surprise is possible under cautious reasoning. Table 5 illustrates this, where it may be verified that both t_1 and t_2 express common full belief in caution and primary belief in rationality by the previous discussion.

In Section 3.2 we observed that the teacher thinks he can fully surprise the student if he believes the student thinks the teacher will wait with giving the exam until Friday and thus forgoing the possibility to surprise him. In other words, the believed reasonable doubt the student has in the teacher being able to simultaneously giving the exam and surprising the student, is what allows the teacher to do exactly that. However, this doubt that the teacher believes the student has about him being able to surprise the student, only manages to resolve the paradox to the extent that the student is *not* believed to be cautious. We have shown that the student must believe with certainty that the teacher will give the exam on Friday if the teacher wants to fully surprise the student. In essence this result is very much in the spirit of the findings by Quine (1953). However, the game-theoretic setting allows us to formally capture how the student can justifiably believe the teacher can rationally choose to forgo the possible utility from surprising the student by choosing to give the exam on Friday. The notion of common belief in rationality formalizes how the teacher can eventually vindicate his announcement of simultaneously giving an exam and surprising the student. Moreover, the notion of common full belief in caution and primary belief in rationality provides a condition under which the paradox cannot be resolved. This discussion does not argue against the possibility of the teacher expecting the student to believe the teacher may consider both Thursday and Friday. That is, it is well possible for the teacher to be able to rationally choose either Thursday or Friday, if he is indifferent between the two. This however only occurs

Table 5: Epistemic model for "Only surprise possible by giving exam" with cautious beliefs

	$T_1 = \{t_1\}$		
Types	$T_2 = \{t_2\}$		
Beliefs for Teacher	$b_1(t_1')$	$=$	t_2'
Beliefs for Student	$b_2(t_2')$	$=$	$(1-\epsilon)(Th, t_1) + \epsilon(Fr, t_1)$

if the teacher fully expects the student to believe the teacher will give the exam on Thursday. As a consequence, common full belief in caution cannot be satisfied, as the student must be able to fully believe in the teacher choosing Thursday. Hence, a possible plan of the teacher of making it *common knowledge* that he is considering a distribution over both Thursday and Friday such that the student must guess this distribution will not help him at all. Namely, if the student knows the teacher is considering a distribution over his options, the student must primarily believe the teacher will consider Thursday with almost probability one, giving the teacher no room for surprise.

How likely it is for a player in a (psychological) game to have cautious beliefs is a matter beyond the scope of this paper, though an interesting avenue for future (experimental) research.

5 Dynamic Surprise Exam Paradox

In this section we will try to answer whether our findings from the static versions also translate to earlier days once we consider dynamic psychological games in which the teacher announces to give a surprise exam on a day possibly before Thursday. More particularly, we will consider the scenario in which the teacher can give the exam on Wednesday, Thursday or Friday. If the teacher gives the exam on Wednesday, the game ends. If he does not, the game continues to the second stage of the game where the teacher has another chance to surprise the student on Thursday. In that regard, the teacher's beliefs about the student's beliefs at each stage of the game may be relevant. Before we analyse this setting, we will first need to establish what such a dynamic psychological game formally entails and how rationality works in dynamic settings. We will consider these notions specifically for the dynamic surprise exam game, to avoid a straying discussion on the many complexities that extending a static psychological game can bring about.

5.1 Description of dynamic surprise exam paradox

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 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 correspond to Wednesday (denoted by $\varnothing$) and Thursday (denoted by h_1). Note that in the three-day scenario there are only three possible 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 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 2 illustrate every combination of a choice and an extreme second-order expectation that are 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, may the exam not be 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

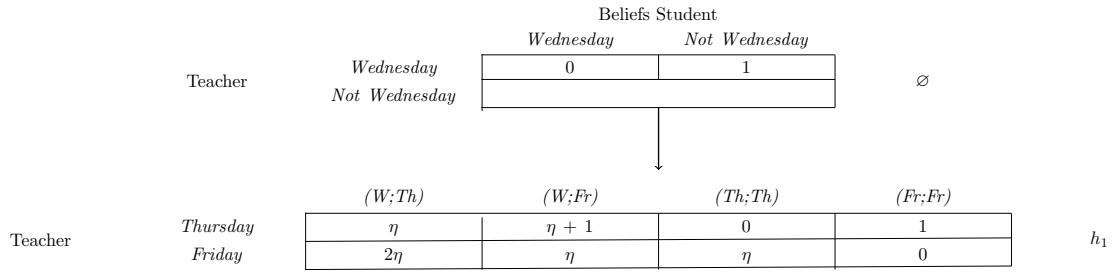

Figure 2: Dynamic situation with surprise possible by giving or not giving exam

he believes the student believes at h_1, but also on what he believes the student believed at $\varnothing$. The depicted utilities in Figure 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 game, 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 1 in Section 3.1.

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 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.[6]. As a result, the utilities for the teacher under the second-order beliefs $(Th; Fr)$ are identical to those under $(Fr; Fr)$ and those under $(Fr; Th)$ are identical to those under $(Th; Th)$

Similarly, we can extend the psychological game in which only surprise is possible from giving the exam from Section 3.2 to include Wednesday as well. The resulting game is depicted in Figure 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 will regardless of the conditional belief hierarchy result in a utility of 0.

Note that we have only defined a dynamic psychological game for the central game in this paper. 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 Surprise Exam Paradox. 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 5.1 (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*

[6]In other words, we assume Bayesian updating here.

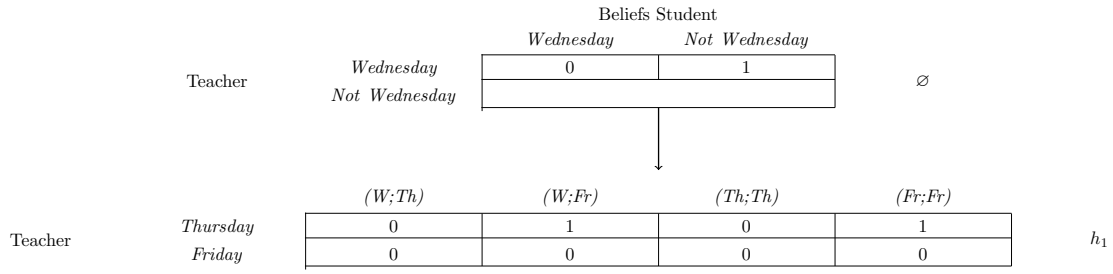

Figure 3: Dynamic situation with surprise only possible by giving exam

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 $S_1(h_1) = \{Th, Fr\}$.

5.2 Common belief in future rationality

The Surprise Exam Paradox highlights the idea of, and potential problems of dealing with, backwards induction. When using backward induction, a player tries to reason at every stage of a dynamic game about what an opponent can reasonably think and choose at the present and the future and takes decisions and beliefs in the past for granted. Dekel et al. (1999) and Asheim and Perea (2005) formally model backwards induction reasoning by the notion of sequential rationalizability. Baltag et al. (2009) as well Penta (2015) propose different concepts as to capture backwards 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, introduced by Perea (2014), that also manages to capture reasoning processes represented by backwards induction: *common belief in future rationality.*

Much like in traditional settings, common belief in future rationality in psychological games entails that at any information set $h \in H$ a player believes that his opponents are rational now and in the future, believes that his opponents believe their opponents are rational now and in the future, et cetera. Just like in the static scenario, also here one of the main divergences from traditional game theory can be found in how we define an optimal strategy. Namely, the expected utility from a strategy at a certain information set $h \in H$ now no longer just depends on the conditional first-order beliefs, but possibly on the entire conditional belief hierarchy. In the Surprise Exam Paradox in particular, we denote the expected utility for the teacher at information set $h \in \{\varnothing, h_1\}$ from a certain strategy $s_1 \in S_1(h)$ by $u_1(s_1, \beta_1(t_1, h))$, where $\beta_1(t_1, h)$ is the conditional belief hierarchy induced by type t_1 at h. In the paradox the only element that is relevant for the teacher's utility from the conditional belief hierarchy $\beta_1(t_1, h)$ is the second-order belief of the teacher. Using the expected utilities, we can define an optimal strategy $s_1 \in \{W, Th, Fr\}$ at information set $\varnothing$ and an optimal strategy $s_1 \in \{Th, Fr\}$ at information set h_1 for the teacher.

Definition 5.2 (Optimal strategy at $\varnothing$ and h_1)**.**
Consider a dynamic epistemic model $M = (T_i, b_i)_{i \in I}$ in the Surprise Exam Paradox and a type $t_1 \in T_1$ in such a model. A ***strategy $s_1 \in \{W, Th, Fr\}$ is optimal for type t_1 at information set*** *$\varnothing$ if $\forall s_1' \in \{W, Th, Fr\} : u_1(s_1, \beta_1(t_1, \varnothing)) \geq u_1(s_1', \beta_1(t_1, \varnothing))$.*

A ***strategy*** $s_1 \in \{Th, Fr\}$ ***is optimal for type*** t_1 ***at information set*** h_1 *if* $\forall s_1' \in \{Th, Fr\}$: $u_1(s_1, \beta_1(t_1, h_1)) \geq u_1(s_1', \beta_1(t_1, h_1))$.

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.

We are now in the position to adapt the idea of common belief in future rationality to fit with the psychological game of the surprise exam. That is, we can define what it means for the student to believe that the teacher chooses optimally at the current stage and at future stages and what it means for the teacher to believe that the student believes he plays optimally at the current and future stage. Subsequently, like in common belief in rationality, we can iterate these arguments to arrive at a state where there is common belief in future rationality.

Definition 5.3 (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.*

In other words, for the student to believe at $\varnothing$ in the teacher's future rationality, he 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 he 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 he 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 Surprise Exam Paradox.

Definition 5.4 (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*.*

Then, we can finally consider what it means for the teacher to have a rational strategy under the reasoning concept considered.

Definition 5.5 (Rational strategy under common belief in future rationality)**.**
We say that strategy s_1 *can be rationally chosen by the teacher under common belief in future*

rationality if there is a dynamic epistemic model $M = (T_i, b_i)_{i \in I}$ and a type $t_1 \in T_1$ for the teacher within that model such that t_1 expresses common belief in future rationality and s_1 is, at every $h \in \{\varnothing, h_1\}$ that it leads to, optimal for t_1.

5.3 Psychological subgame perfect equilibrium

In the literature for dynamic psychological games there is a lot of reliance on equilibrium concepts. Perhaps the most notable of such concepts is the psychological sequential equilibrium by Battigalli and Dufwenberg (2009). As a refinement of the sequential psychological equilibrium by Geanakoplos et al. (1989), it revolves around sequential rationality while allowing for utilities to be determined by updated beliefs about the beliefs of others as well. In the surprise exam game we are considering here there is however no apparent reason for the teacher to believe that the student would change his belief between Wednesday and Thursday if the teacher already believed the student believed he would not give the exam on Wednesday.

The equilibrium concept that is applicable to the class of dynamic psychological games we are considering here is thus 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 also in a dynamic Surprise Exam Paradox 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 Surprise Exam Paradox 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 scenario, by construction σ_1 implies correctness of beliefs.

With the definition of a conditional belief hierarchy that is generated by σ_1 clarified, we can now turn to what it entails for there to be a psychological subgame perfect equilibrium in the Surprise Exam Paradox.

Definition 5.6 (Psychological subgame perfect equilibrium)**.**
Consider a dynamic surprise exam game portrayed in either Figure 2 or in Figure 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 2, but not in the game of Figure 3.

5.4 Dynamic surprise exam: surprise possible 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 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 6. 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 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', he 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 he 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 types t_1 and t_1' only assign positive probabilities to types of the student, we automatically have that both 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

Table 6: Epistemic model for "Dynamic situation with surprise possible by giving or not giving exam"

	$T_1 = \{t_1, t_1'\}$		
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)

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. Then, by not giving the exam on Thursday, some additional utility is gained. As such, 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 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 student 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 would be 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 $\sigma_1(\varnothing)(W) < \frac{1}{(\eta+1)^2}$. Then we can implicitly infer from the relations highlighted above that $u_1((W, \sigma_1), \varnothing) > u_1((NW, \sigma_1), \varnothing)$. Again, however, the student

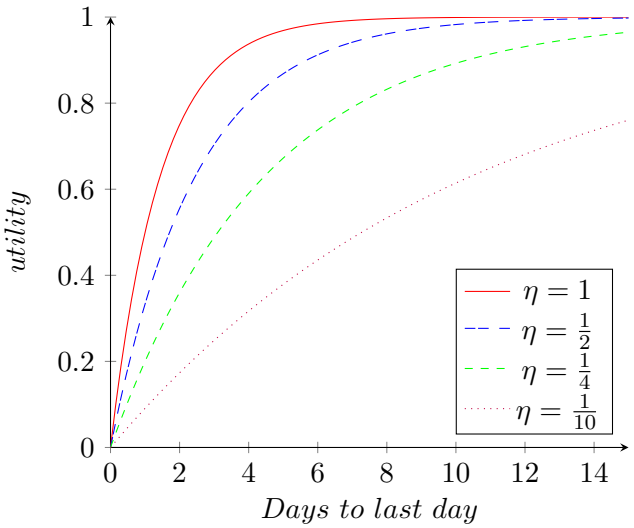

Figure 4: Expected utilities in equilibrium for game in Figure 2

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 $u_1((Th, \sigma_1), h_1) = u_1((Fr, \sigma_1), h_1) = 1 - \frac{1}{(\eta+1)^2}$. Namely, from σ_1 we can also infer what the teacher believes at h_1 what the student believes at h_1 *and* what the student believed at $\varnothing$. 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 to potentially give an exam on, the expected utility will increase even further (Mourmans, 2017). 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. For formal proofs we refer the reader to Geanakoplos (1996) and Mourmans (2017).[7]

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. By slightly modifying the epistemic model in Table 6 this can be accommodated for. For instance, in case of a four day example we could simply extend the epistemic model in Table 6 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, depending on his own type.

[7]This also is the intuition behind the results of Sober (1998), though now in the setting of a psychological game.

However, as Geanakoplos (1996) and Mourmans (2017) point out, the subgame perfect equilibrium remains unique.

Thus, much like in the static scenario 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 scenario the teacher has gained extra options to fully surprise the student. At the other end we have the belief hierarchy that is generated by the psychological subgame perfect equilibrium. 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. Even though the teacher's expected utility in the psychological subgame perfect equilibrium increases with the time horizon and thus his options to surprise the student increase, actual full surprise is still not possible under this concept.

5.5 Dynamic surprise exam: surprise possible 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 3.2, in which the teacher only believes he can possibly surprise 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 3. For the purpose of analysing this game, we can utilise the epistemic model presented in Section 5.4, now repeated in Table 7. 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

Table 7: Epistemic model for "Dynamic situation with surprise possible only by giving the exam"

	$T_1 = \{t_1, t_1'\}$		
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)

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 conditional 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. Moreover, by expanding the game, full surprise on *some* day no longer requires the condition that at Thursday the teacher must expect the student to fully believe the teacher will choose Friday, even though Friday can never lead to a surprise. In other words, the student no longer needs to fully doubt the teacher's announcement of simultaneously giving and exam and surprising him for the teacher to believe that he is able to fully surprise the student.

It is also important to point out here that the issue of believing the teacher may give the exam on the last day may become less of a realistic obstruct for the teacher if the game gets extended to include an arbitrary large number of days above two. Let us for instance consider the extreme situation in which the teacher believes he can fully surprise the student on *any* day before the last day. This is possible if the teacher believes the student believes at any given day that the teacher will give the exam the next day. This would inadvertently lead to a conditional belief hierarchy in which the teacher believes at the second-to-last day (Thursday) that the student believes the teacher will give the exam on the last day (Friday). Again, like in the static scenario of this version, the student should fully doubt the teacher's statement of simultaneously giving the exam and surprising the student, if the teacher is to believe he can fully surprise the student on Thursday. However, one may wonder whether this condition of a non-cautious belief held about the last day is much of a realistic obstacle when it pertains to the teacher trying to fully surprise the student at one of the first days of a sequence of days, especially when the time horizon is sufficiently long.[8]

Referring back to the discussion on cautious reasoning, for Friday to be considered an optimal strategy at $\varnothing$ by the student, he must not only believe that the teacher believes he believes that the teacher will choose Thursday at h_1 with certainty, but also Wednesday at $\varnothing$ with certainty. Only then it is optimal for the teacher to not choose Wednesday at $\varnothing$ and subsequently also Friday at h_1. So the required certainty about the teacher's choice now applies to two days. If we assume the student to be a cautious reasoner, then we know from Section 4 that the student must primarily believe at h_1 that the teacher will give the exam on Thursday. Consequently, the teacher would only be able to surprise the student up to an infinitesimal probability, say ϵ. Then at $\varnothing$, if we further assume that the student's non-standard beliefs at $\varnothing$ are similarly described as at h_1, the student must primarily believe that the teacher will choose Wednesday. Namely, choosing Thursday under such beliefs would give the teacher a utility of *at most* $\epsilon(1-\epsilon)$ whereas choosing Wednesday would give *at least* ϵ. Hence, also in the dynamic surprise exam game, common full belief in caution causes surprise to be virtually impossible.

Also in a dynamic setting, psychological subgame perfect equilibrium again faces the same pitfall as traditional game theory does. In section 3.2 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

[8] Kim and Vadusevan (2017) derive a similar effect of the time horizon when considering the coherency of the teacher's announcement in a Bayesian analysis.

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 conceivable for the teacher that believes the student believes the teacher will give the exam on the present day. The reasonable doubt of the student in the announcement of the teacher is 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. One can question how realistic it is for the teacher to have a simple belief hierarchy that is characterised by the equilibrium in a one-shot scenario like the surprise exam game, even more so when dynamics make the situation more complex.

6 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 him. 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 him 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 is 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. These doubts are however essential to the whole game of the surprise exam, as they allow the teacher to eventually follow up on his announcement.

More generally, equilibria in psychological games may be less realistic than in traditional games. Whereas repeated play in traditional games may move players towards an equilibrium as they learn about their opponents' beliefs because of observed actions, players in psychological games derive utility from their opponents' beliefs, which may not be observable at all. On the other hand, in psychological games decision-makers are forced to consider more explicitly their own belief hierarchy, as their preferences are shaped by it. Consequently, reasoning concepts that explicitly rest on such belief hierarchies, such as common belief in (future) rationality, appear to be even more natural in the setting of psychological games.

Paradoxes make us aware of flaws in our reasoning and bounds in our understanding of certain problems. In that regard, the Surprise Exam Paradox teaches us a special lesson, namely that we may be limiting our understanding of the problem by using a too restrictive manner of reasoning.

References

Asheim, G. and Perea, A. (2005). Sequential and quasi-perfect rationalizability in extensive games. *Games and Economic Behaviour*, 53, 15–42.

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. and Dufwenberg, M. (2009). Dynamic psychological games. *Journal of Economic Theory*, 144, 1–35.

Battigalli, P., Dufwenberg, M., and Smith, A. (2015). Frustration and anger in games. *Working Paper*.

Bernheim, B. D. (1984). Rationalizable strategic behavior. *Econometrica*, 52(4), 1007–1028.

Bjorndahl, A., Halpern, J., and Pass, R. (2016). Language-based games. *Working Paper*.

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.

Brandenburger, A. (1992). Lexicographic probabilities and iterated admissibility. In Dasgupta, P., editor, *Economic Analysis of Markets and Games*, pages 282–290. Cambridge, MA: MIT Press.

Brandenburger, A. and Dekel, E. (1987). Rationalizability and correlated equilibria. *Econometrica*, 55, 1391–1402.

Brandenburger, A. and Dekel, E. (1993). Hierarchies of beliefs and common knowledge. *Journal of Economic Theory*.

Börgers, T. (1994). Weak dominance and approximate common knowledge. *Journal of Economic Theory*, 64, 265–276.

Chow, T. Y. (2011). The Surprise examination or Unexpected hanging paradox. *Working paper*.

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 Behaviour*, 47(2), 269–298.

Dufwenberg, M. J. and Dufwenberg, M. (2016). Lies in disguise a - theoretical analysis of cheating. *Working Paper*.

Ferreira, J. L. and Bonilla, J. Z. (2008). The surprise exam paradox, rationality, and pragmatics: a simple game theoretic analysis. *Journal of Economic Methology*, 15(3).

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 Behaviour*, 1(1), 60–79.

Halpern, J. Y. (2010). Lexicographic probability, conditional probability, and nonstandard probability. *Games and Economic Behaviour*, 68, 155–179.

Hammond, P. J. (1994). *Scientific Philosopher*, volume 1, chapter Elementary non-Archimedean representations of probability for decision theory and games, pages 25–49. Dordrecht: Kluwer.

Harsanyi, J. (1967-1968). Games with imcomplete information played by "Bayesian players". *Management Science*, 14, 159–182 320–334 486–502.

Jagau, S. and Perea, A. (2017). Common belief in rationality in psychological games. *Epicenter working paper No.10*.

Kim, B. and Vadusevan, A. (2017). How to expect a surprising exam. *Synthese*, 194(8), 3101–3133.

Mourmans, N. (2017). Reasoning about the surprise exam paradox: Common belief in future rationality in psychological games. Master's thesis, Maastricht University.

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*. Camebridge: Camebridge University Press.

Perea, A. (2014). Belief in the opponents' future rationality. *Games and Economic Behaviour*, 83, 231–254.

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.

Robinson, A. (1973). Function theory on some nonarchimedean fields. *American Mathematical Monthly:Papers in the Foundations of Mathematics*, 80(6), S87–S109.

Sanna, F. A. (2016). Universal spaces of hierarchies of beliefs: an application to k-th order psychological games. Master's thesis, Universita Commerciale Luigi Bocconi.

Selten, R. (1975). Reexamination of the perfectness concept for equilibrium points in extensive games. *International Journal of Game Theory*, 4, 25–55.

Shaw, R. (1958). The paradox of the unexpected examination. *Mind*, 67, 382–384.

Smullyan, R. M. (1987). *Forever Undecided: A Puzzle Guide to Gödel*. New York: Knopf.

Sober, E. (1998). To give a Surprise Exam, use Game Theory. *Synthese*, 115(3), 355–373.

Tan, T. and Werlang, S. R. C. (1988). The Bayesian foundations of solution concepts of games. *Journal of Economic Theory*, 45, 370–391.

Tan, T. and Werlang, S. R. C. (1992). On Aumann's notion of common knowledge: An alternative approach. *Revista Brasileira de Economia*, 64, 151–166.