On the epistemic foundation for iterated weak dominance: an analysis in a logic of individual and collective attitudes

Emiliano Lorini

On the Epistemic Foundation for Iterated Weak Dominance: An Analysis in a Logic of Individual and Collective attitudes

Emiliano Lorini

Abstract This paper proposes a logical framework for representing static and dynamic properties of different kinds of individual and collective attitudes. A complete axiomatization as well as a decidability result for the logic are given. The logic is applied to game theory by providing a formal analysis of the epistemic conditions of iterated deletion of weakly dominated strategies (IDWDS), or iterated weak dominance for short. The main difference between the analysis of the epistemic conditions of iterated weak dominance given in this paper and other analysis is that we use a *semi-qualitative* approach to uncertainty based on the notion of plausibility first introduced by Spohn, whereas other analysis are based on a *quantitative* representation of uncertainty in terms of probabilities.

Keywords Epistemic logic · Epistemic game theory · Belief revision · Iterated weak dominance

1 Introduction

The fundamental concept of game theory is the concept of solution which is, at the same time, a prescriptive notion, in the sense that it prescribes how rational agents in a given interaction *should* play, and a predictive one, in the sense that it allows us to predict how the agents *will* play. There exist many different solution concepts both for games in normal form and for games in extensive form (e.g., Nash Equilibrium, iterated deletion of strongly dominated strategies, iterated deletion of weakly dominated strategies, correlated equilibrium, backward induction, forward induction, etc.) and new ones have been proposed in the recent years (see, e.g., [30]). A major issue

E. Lorini (✉)
IRIT-CNRS, Toulouse, France
e-mail: emiliano.lorini@irit.fr

we face when we want to use some solution concept in order either to predict human behavior or to build some practical applications (e.g., for computer security or for multi-agent systems) is to evaluate its significance. Some of the questions that arise in these situations are, for instance: given certain assumptions about the agents such as the assumption that they are rational (e.g., utility maximizers), under which conditions will the agents converge to equilibrium? Are these conditions realistic? Are they too strong for the domain of application under consideration? There is a branch of game theory, called epistemic game theory, which can help to answer these questions (see [42] for a general introduction to the research in this area). Indeed, the aim of epistemic game theory is to provide an analysis of the necessary and/or sufficient epistemic conditions of the different solution concepts, that is, the assumptions about the epistemic states of the players that are necessary and/or sufficient to ensure that they will play according to the prescription of the solution concept. Typical epistemic conditions which have been considered are, for example, the assumption that players have common belief (or common knowledge) about the rationality of every player,[1] the assumption that every player knows the choices of the others,[2] or the assumption that players are logically omniscient.[3]

The aim of this paper is to propose a new logic, called PDL-A (*Propositional Dynamic Logic of individual and collective Attitudes*), in which the epistemic conditions of different solution concepts for normal form games can be formally specified. Our logic PDL-A is a combination of van Benthem et al.'s variant of *Propositional Dynamic Logic* PDL which gives an epistemic interpretation to programs [53] with Spohn's theories of uncertainty and belief change [45]. The interesting aspect of this logic is that it allows us to describe both the static and the dynamic properties of different kinds of individual and collective epistemic attitudes such as knowledge, belief, graded belief, robust belief and common belief which provide the epistemic foundations of different solution concepts in game theory.

In this work we mainly concentrate on the logical characterization in PDL-A of the epistemic conditions of iterated deletion of weakly dominated strategies (IDWDS) (also called 'iterated weak dominance' or 'iterated admissibility').

Iterated weak dominance is an important solution concept in game theory which is distinguished from iterated strong dominance. Although there have been some works in economics investigating the epistemic conditions of iterated weak dominance, they are far less studied and understood than the epistemic conditions of iterated strong dominance. The fundamental difference between the analysis of iterated weak dominance given in this paper and other analysis is that we use a *semi-qualitative* approach to uncertainty based on the notion of plausibility introduced by Spohn [45], whereas existing analysis of the epistemic conditions of iterated weak dominance are based on a *quantitative* representation of uncertainty in terms of probabilities (see, e.g.,

[1] This is the typical condition of iterated deletion of strongly dominated strategies (also called iterated strong dominance).

[2] This condition is required in order to ensure that the agents will converge to a Nash equilibrium.

[3] See [56] for an interesting analysis of iterated strong dominance after relaxing the assumption of logical omniscience.

[20, 23, 32, 48]). Spohn's theory of uncertainty and belief change, generally referred to as 'κ calculus', has been largely used in Artificial Intelligence (AI) (Goldszmidt & Pearl [28] refer to it as 'rank-based system' and 'qualitative probabilities'). It provides an elegant and relatively simple approach designed to reason about both the static aspects and the dynamic aspects of epistemic attitudes. Our approach is semi-qualititative in the sense that we assume that beliefs of agents are ranked by a finite number of non-negative integers providing a qualitative scale for degrees of belief. Specifically, each integer corresponds to a linguistic quantifier such as I *weakly* believe that φ, I *mildly* believe that φ, I *strongly* believe that φ, etc. However, our approach is not purely qualitative because it allows us to say *how much* a given agent believes that a certain proposition φ is true. The distinction between purely quantitative, semi-qualitative and purely qualitative approaches to uncertainty has been widely discussed in the AI literature (see, e.g., [41, 55]). While in purely quantitative approaches belief states are characterized by classical probabilistic measures or by alternative numerical accounts, such as lexicographic probabilities [14, 20, 23] or conditional probabilities [12], in semi-qualitative approaches, such as Spohn's theory, belief states are described by rough qualitative measures assigning orders of magnitude. Finally, purely qualitative approaches do not use any numerical representation of uncertainty but simply a plausibility ordering on possible worlds structures inducing an epistemic-entrenchment-like ordering on propositions. Purely qualitative approaches have been used both in the area of belief revision and in the area of logics of belief change (see, e.g., [9, 27, 51]).

The rest of the paper is organized as follows. Section 2 provides an informal introduction to the concept of iterated weak dominance and to the epistemic conditions of this solution concept. Section 3 presents the syntax and the semantics of the logic PDL-A. Section 4 is devoted to the formalization in PDL-A of the previous different kinds of individual and collective attitudes, which are fundamental building blocks for the analysis of the epistemic conditions of iterated weak dominance. In Section 5 a complete axiomatization as well as a decidability result for PDL-A are given. In Section 6 the logic PDL-A is extended with constructions to describe information about agents' choices and is used to provide an analysis of the epistemic conditions of iterated weak dominance. Related works on the analysis of the epistemic conditions of iterated weak dominance are discussed in Section 7.

2 Epistemic Conditions of Iterated Weak Dominance: Some Intuitions

Given a game in normal form Γ and a player i in this game, a strategy a of player i is a weakly dominated strategy if and only if there is another strategy b of player i such that: (1) no matter what strategies the other players will choose, playing b is for i at least as good as playing a and (2) there exists at least one strategy profile of the other players such that, if the others play this strategy then playing b is for player i better than playing a. The so-called iterated weak dominance is a procedure that starts with a given game in normal form and, at each step, for every player in the game removes all his weakly dominated strategies, thereby generating a subgame of the original game, and that repeats this process again and again. The strategy profiles that survive

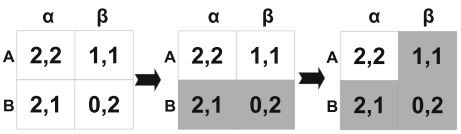

Fig. 1 Example of iterated deletion of weakly dominated strategies

after this iteration of removal of weakly dominated strategies are the equilibria of the game.

Consider the game in Fig. 1 with two players Row and Column (the payoff on the left-hand side being the payoff of Row and the payoff on the right-hand side being the payoff of Column). In the initial game the strategy B of Row is weakly dominated by the strategy A and is deleted from the game. In the resulting game the strategy β of Column is weakly dominated by the strategy α and is also deleted from the game. Therefore, the solution of the game is the strategy profile (A, α).

There are at least two requirements that should be met in order to be able to predict that the players will act according to the prediction of iterated weak dominance. Let us illustrate them with the aid of the example in Fig. 1. Assume that both Row and Column are rational players (i.e., they are utility maximizers) and that they have a common belief about this.

If we assume that Row is rational then his only reason for discarding strategy B and for deciding not to play it is that he envisages the possibility that Column will play β (even though playing β clashes with the hypothesis that Column is rational and that she believes that Row is rational, as we assumed common belief in rationality, and hence that Row will not play B). Thus, *the first requirement is that* the beliefs of Row and, more generally, *the beliefs of every player in the game must be cautious, in the sense that every player must envisage all possible choices of the other players* (see, e.g., [38, Chapter 8] and also [2, 4, 18, 19, 22, 44] for further discussion about this requirement).

Moreover, if we assume that Column is rational then her reason for discarding strategy β is that she considers the situation in which Row plays the admissible (i.e., non-weakly dominated) strategy A strictly more plausible than the situation in which Row plays the inadmissible strategy B. More generally, in order to guarantee that the players in a game will act according to the prescription of the solution concept, each of them should believe that the situations in which the other players play an admissible strategy are strictly more plausible than the situations in which they play an inadmissible one. If we assume that a certain strategy of a player is incompatible with the player's rationality if and only if it is inadmissible, the previous observation leads to the following *second requirement: every player in a game must have a robust belief about the rationality of the other players*, in the sense that he will continue to believe that the others will play rationally as long as he does not learn something which is incompatible with this fact. This concept of 'robust belief' is one of the key elements of Stalnaker's analysis of the epistemic conditions of iterated weak dominance [47, 48].[4]

[4]Related concepts are the concept of 'assumption' [23], 'strong belief' [12] and 'full belief' [2]. See [3] for a comparative analysis of these four concepts.

The previous two requirements will be formally specified in Section 6, in which a logical analysis of the epistemic conditions of iterated weak dominance will be provided. But before moving to game theory we present in the next section the logic PDL-A.

3 Logical Framework

This section presents the syntax and the semantics of the logic PDL-A. Technically, this logic is an extension of the logic E-PDL (*Epistemic Propositional Dynamic Logic*) proposed by van Benthem et al. [53] with special constructions for representing plausibility orderings over possible worlds and with dynamic operators for representing the effects of an operation of belief conditioning in the sense of Spohn. Generally speaking, PDL-A can be seen as a combination of E-PDL with Spohn's rank-based system.

3.1 Syntax

Assume a countable set of atomic propositions describing facts $Prop = \{p, q, \ldots\}$, a finite set of agents $Agt = \{i, j, \ldots\}$ and a set of natural numbers $Num = \{0, \ldots, \mathsf{max}\}$, with $\mathsf{max} \in \mathbb{N} \setminus \{0\}$. For example, suppose $Num = \{0, 1, 2, 3, 4, 5\}$. *Num* can be interpreted as a *qualitative* scale where 0 means 'null', 1 means 'very low', 2 means 'low', 3 means 'medium', 4 means 'high' and 5 means 'very high'.[5] For the sake of simplicity we assume that *Num* is finite. This assumption is crucial to be able to provide a complete axiomatization of the logic PDL-A (see Section 5). A generalization of our results to the case where *Num* is infinite is postponed to future work.

$2^{Agt*} = 2^{Agt} \setminus \emptyset$ is the set of all non-empty sets of agents (*alias* coalitions). Elements of 2^{Agt*} are denoted by symbols $J, H, \ldots$ For notational convenience, the coalition $Agt \setminus \{i\}$ is denoted by $-i$.

The language of PDL-A is defined by the following grammar in Backus-Naur Form (BNF):

$$\pi ::= i \mid \pi_1; \pi_2 \mid \pi_1 \cup \pi_2 \mid \pi^* \mid ?\varphi$$
$$\varphi, \psi ::= p \mid \mathsf{exc}_{i,\mathsf{h}} \mid \neg\varphi \mid \varphi \wedge \psi \mid [\pi]\varphi \mid [*_i^{\alpha}\varphi]\psi$$

where p ranges over *Prop*, h ranges over *Num*, i ranges over *Agt* and α ranges over $Num \setminus \{0\}$. The other Boolean constructions $\top$, $\bot$, $\vee$, $\rightarrow$ and $\leftrightarrow$ are defined from p, $\neg$ and $\wedge$ in the standard way. We define *Obj* to be the set of all Boolean combinations of atomic propositions in *Prop* and we call the elements of *Obj* ontic (or objective) facts in order to distinguish them from 'epistemic facts' about agents' mental states.

[5] It has to be noted that Spohn's notion of plausibility is measured on the set of ordinals. Here, for simplicity, it is assumed that plausibility is measured on the integer scale *Num*.

Knowledge constructs (or programs) π correspond to the basic constructions of Propositional Dynamic Logic (PDL) [33]: sequential composition (;), non-deterministic choice ($\cup$), iteration (*) and test (?). A given knowledge program π corresponds to a specific configuration of the agents' epistemic states.

The formula $[\pi]\varphi$ has to be read "φ is true, according to the knowledge program π". For the atomic case, the operator $[i]$ represents the standard S5-notion, partition-based and fully introspective notion of knowledge that is commonly used both in computer science [26] and economics [6]. $[i]\varphi$ has to be read "φ is true according to what agent i knows" or more simply "agent i knows that φ is true", which just means that "φ is true in all worlds that agent i envisages". Sequential composition ; allows to represent an agent's knowledge over his knowledge and over other agents' knowledge. For instance, $[i;j]\varphi$ means that "φ is true according to what agent i knows about agent j's knowledge". or more simply "agent i knows that agent j knows that φ is true". Non-deterministic choice $\cup$ allows to represent the notion of shared knowledge. For instance, $[i \cup j]\varphi$ means that "both agent i and agent j know that φ is true". By means of iteration * one can represent higher order knowledge of arbitrary depth. For instance, $[(i;j)^*]\varphi$ means that "agent i knows that agent j knows φ is true, agent i knows that agent j knows that agent i knows that agent j knows φ is true, and so on, ad infinitum". The test operation ? has the usual meaning as in PDL: $[?\varphi]\psi$ means that "if φ is true then ψ is true".

As we will show in Section 4.1, the operators $[i]$ captures a form of 'absolutely unrevisable belief', that is, a form of belief which is stable under belief revision with any new *evidence*. A similar property for the notion of knowledge has been advanced by the so-called *defeasibility* (or *stability*) *theory of knowledge* [34, 43, 49]. According to this theory, a given piece of information φ is part of an agent's knowledge only if the agent's justification to believe that φ is true is sufficiently strong that it is not capable of being defeated by evidence that the agent does not possess. As pointed out by [9], two different interpretations of the term 'evidence' have been given in the context of this theory, each giving a different interpretation of what *knowledge* is. The first one [49] defines knowledge as a form of belief which is stable under belief revision with 'any piece of *true* information', while the second one [43] gives a stronger definition of knowledge as a form of belief which is stable under belief revision with 'any piece of information'. The concept formalized by the operators $[i]$ captures the latter notion of knowledge in a stronger sense. In Section 4.1, we will introduce the notion of safe belief which corresponds to the former notion of knowledge.

The formula $\left[*_i^\alpha \varphi\right]\psi$ has to be read "after agent i has learnt that φ is true with a degree of firmness α, ψ will be true" (or "after agent i has revised his beliefs with φ and with a degree of firmness α, ψ will be true"). As we will show in Section 3.2, technically an epistemic event $*_i^\alpha \varphi$ amounts to an operation of beliefs' conditionalization in Spohn's sense [45], where the parameter α measures the extent to which agent i will believe that φ is true after learning that φ is true. The epistemic event $*_i^\alpha \varphi$ is supposed to be public, i.e., if agent i learns that φ is true then all other agents know this. This assumption could be easily relaxed by using action models as introduced in [7, 8], which would allow us to model private and semi-private epistemic events.

The language of PDL-A contains special atoms of the form $\mathsf{exc}_{i,\mathsf{h}}$ which are used to rank the worlds that agent i considers possible at a given world according to their *plausibility* degree for the agent. Starting from [29], ranking among possible worlds have been extensively used in belief revision theory. We here use the notion of plausibility first introduced by Spohn [45]. Following Spohn's theory, the worlds that are assigned the smallest numbers are the most plausible, according to the beliefs of the individual. That is, the number h assigned to a given world rather captures the degree of *exceptionality* of this world, where the exceptionality degree of a world is nothing but the opposite of its plausibility degree (i.e., the exceptionality degree of a world decreases when its plausibility degree increases). Therefore, formula $\mathsf{exc}_{i,\mathsf{h}}$ can be read alternatively as "the current world has for agent i a degree of exceptionality equal to h" or "the current world has for agent i a degree of plausibility equal to $\mathsf{max} - \mathsf{h}$".

Before turning into semantics, we provide some abbreviations that will be used in the rest of the paper. We define $\langle i \rangle$ to be the dual of $[i]$, that is:

$$\langle i \rangle \varphi \stackrel{\mathtt{def}}{=} \neg [i] \neg \varphi$$

Formula $\langle i \rangle \varphi$ means that "φ is compatible with agent i's knowledge".

3.2 Semantics

The semantics of the logic PDL-A is a possible world semantics with a special function for exceptionality.

Definition 1 (**Model**) PDL-A models are tuples $M = \langle W, \{\mathcal{E}_i : i \in Agt\}, \kappa, \mathcal{V} \rangle$ where:

- W is a nonempty set of possible worlds or states;
- every $\mathcal{E}_i$ is an equivalence relation between worlds in W;
- $\kappa : W \times Agt \longrightarrow Num$ is a total function mapping worlds and agents to natural numbers in Num such that:

  (**Constr1**) for every $w \in W$ and for every $i \in Agt$, there is v such that $w \mathcal{E}_i v$ and $\kappa(v, i) = 0$;

- $\mathcal{V} : W \longrightarrow 2^{Prop}$ is a valuation function.

As usual, $p \in \mathcal{V}(w)$ means that proposition p is true at world w. The equivalence relations $\boldsymbol{\mathcal{E}}_i$, which are used to interpret the epistemic operators $[i]$, can be viewed as functions from W to 2^W. Therefore, we can write $\mathcal{E}_i(w) = \{v \in W : (w, v) \in \mathcal{E}_i\}$. The set $\mathcal{E}_i(w)$ is agent i's *information set* at world w: the set of worlds that agent i envisages at world w. As $\mathcal{E}_i$ is an equivalence relation, if $(w, v) \in \mathcal{E}_i$ then agent i has the same information set at w and v (i.e., agent i has the same knowledge at w and v).

The function κ provides a plausibility grading of the possible worlds for each agent i and is used to interpret the atomic formulas $\mathsf{exc}_{i,\mathsf{h}}$. $\kappa(w, i) = \mathsf{h}$ means that, according to agent i the world w has a degree of exceptionality h or, alternatively, according to agent i the world w has a degree of plausibility $\mathsf{max} - \mathsf{h}$. (Remember that the degree of plausibility of a world for an agent is the opposite of its exceptionality degree for

the agent). The function κ allows to rank an agent's envisaged worlds according to their plausibility degree: among the worlds agent i envisages at world w (i.e., agent i's information set at w), there are worlds that i considers more plausible than others. For example, suppose that $\mathcal{E}_i(w) = \{w, v, u\}$, $\kappa(w, i) = 2$, $\kappa(u, i) = 1$ and $\kappa(v, i) = 0$. This means that at world w agent i envisages the three worlds w, v and u. Moreover, according to agent i, the world v is strictly more plausible than the world u and the world u is strictly more plausible than the world w (as $\mathsf{max}-0 > \mathsf{max}-1 > \mathsf{max}-2$).

(**Constr1**) is a *normality* constraint for the plausibility grading which ensures that an agent can always envisage a world with a minimal degree of exceptionality 0. This constraint is important because it ensures that an agent's beliefs are consistent e.g., an agent cannot believe φ and $\neg\varphi$ at the same time (see Section 4.1 for more details).

As in PDL, the accessibility relation for atomic knowledge programs is generalized to all kinds of knowledge programs. Given a PDL-A-model $M = \langle W, \{\mathcal{E}_i : i \in Agt\}, \kappa, \mathcal{V}\rangle$ we define:

$$\begin{aligned} \mathcal{E}_{\pi_1;\pi_2} &= \mathcal{E}_{\pi_1} \circ \mathcal{E}_{\pi_2} \\ \mathcal{E}_{\pi_1\cup\pi_2} &= \mathcal{E}_{\pi_1} \cup \mathcal{E}_{\pi_2} \\ \mathcal{E}_{\pi^*} &= (\mathcal{E}_{\pi})^* \\ \mathcal{E}_{?\varphi} &= \{(w, w) : w \in ||\varphi||\} \end{aligned}$$

where $||\varphi|| = \{w \in W : M, w \models \varphi\}$.

Definition 2 (**Truth conditions**) Given a PDL-A-model M, a world w and a formula φ, $M, w \models \varphi$ means that φ is true at world w in M. The rules defining the truth conditions of formulas are:

- $M, w \models p$ *iff* $p \in \mathcal{V}(w)$
- $M, w \models \mathsf{exc}_{i,\mathsf{h}}$ *iff* $\kappa(w, i) = \mathsf{h}$
- $M, w \models \neg\varphi$ *iff not* $M, w \models \varphi$
- $M, w \models \varphi \wedge \psi$ *iff* $M, w \models \varphi$ and $M, w \models \psi$
- $M, w \models [\pi]\varphi$ *iff* $M, v \models \varphi$ for all v such that $(w, v) \in \mathcal{E}_\pi$
- $M, w \models [*_i^\alpha\varphi]\psi$ *iff* $M^{*_i^\alpha\varphi}, w \models \psi$

where the updated model $M^{*_i^\alpha\varphi}$ is defined according to the Definition 4 below.

The epistemic event $*_i^\alpha\varphi$ (i.e., agent i learns that φ is true) updates agent i's information set by modifying the exceptionality degree that i ascribes to his envisaged worlds. Before defining this model update, we follow [45] and lift the exceptionality of a possible world to the exceptionality of a formula viewed as a set of worlds.

Definition 3 (**Exceptionality degree of a formula**) Given a PDL-A model $M = \langle W, \{\mathcal{E}_i : i \in Agt\}, \kappa, \mathcal{V}\rangle$, let $||\varphi||_{w,i} = \{v \in W : v \in ||\varphi|| \text{ and } (w, v) \in \mathcal{E}_i\}$ be the set of worlds that agent i envisages at w and in which φ is true. The exceptionality degree of formula φ for agent i at world w, denoted by $\kappa_{w,i}(\varphi)$, is defined as follows:

$$\kappa_{w,i}(\varphi) = \begin{cases} \min\limits_{v\in||\varphi||_{w,i}} \kappa(v, i) & \text{if } ||\varphi||_{w,i} \neq \emptyset \\ \mathsf{max} & \text{if } ||\varphi||_{w,i} = \emptyset \end{cases}$$

The exceptionality degree of a formula φ captures the extent to which φ is considered to be exceptional by the agent. As expected, the *plausibility* degree of a formula φ is defined as $\mathsf{max} - \kappa_{w,i}(\varphi)$. The plausibility degree of a formula φ captures the extent to which φ is considered to be plausible by the agent.

Definition 4 (**Update**) Given a PDL-A-model $M = \langle W, \mathcal{E}, \kappa, \mathcal{V}\rangle$, $M^{*_i^{\alpha}\varphi}$ is the model such that for all $w \in W$ and for all $j \in Agt$:

$$
\begin{aligned}
W^{*_i^{\alpha}\varphi} &= W \\
\mathcal{E}_j^{*_i^{\alpha}\varphi} &= \mathcal{E}_j \\
\kappa^{*_i^{\alpha}\varphi}(w,i) &= \begin{cases} \kappa(w,i) - \kappa_{w,i}(\varphi) & \text{if } M, w \models \varphi \\ Cut(\alpha + \kappa(w,i) - \kappa_{w,i}(\neg\varphi)) & \text{if } M, w \models \neg\varphi \wedge \langle i\rangle\varphi \\ \kappa(w,i) & \text{if } M, w \models [i]\neg\varphi \end{cases} \\
\kappa^{*_i^{\alpha}\varphi}(w,j) &= \kappa(w,j) \quad \text{if } i \neq j \\
\mathcal{V}^{*_i^{\alpha}\varphi} &= \mathcal{V}
\end{aligned}
$$

where

$$
Cut(x) = \begin{cases} x & \text{if } 0 \leq x \leq \mathsf{max} \\ \mathsf{max} & \text{if } x > \mathsf{max} \end{cases}
$$

The epistemic event $*_i^{\alpha}\varphi$ does not affect the objective world. This is the reason why the valuation function $\mathcal{V}$ is not altered by it (see the definition of $\mathcal{V}^{*_i^{\alpha}\varphi}$). Moreover, it modifies agent i's plausibility ordering but does not modify the plausibility orderings of the agents different from i (see the definition of $\kappa^{*_i^{\alpha}\varphi}$). In particular, it induces a kind of belief conditioning in Spohn's sense [45]. Agent i's plausibility ranking over his envisaged worlds is updated as follows.

1. For every world w in which φ is true, i.e., $M, w \models \varphi$, the degree of exceptionality of w for i decreases from $\kappa(w,i)$ to $\kappa(w,i) - \kappa_{w,i}(\varphi)$, which is the same thing as saying that, degree of plausibility of w for i increases from $\mathsf{max} - \kappa(w,i)$ to $\mathsf{max} - (\kappa(w,i) - \kappa_{w,i}(\varphi))$. (Note that, by Definition 3, we have $\kappa(w,i) - \kappa_{w,i}(\varphi) \leq \kappa(w,i)$).
2. For every world w in which φ is false:
   (a) if at w agent i envisages a world in which φ is true, i.e., $M, w \models \neg\varphi \wedge \langle i\rangle\varphi$, then the degree of exceptionality of w for i changes from $\kappa(w,i)$ to $Cut(\alpha + \kappa(w,i) - \kappa_{w,i}(\neg\varphi))$;
   (b) if at w agent i does not envisage a world in which φ is true, i.e. $M, w \models [i]\neg\varphi$, then the degree of exceptionality of w for i does not change.

The preceding condition 1 ensures the intuitive requirement that belief revision with φ leaves the plausibility ranking in the φ-part of agent i's information set unchanged. In other words, if v and u are worlds in which φ is true and agent i considers v more plausible (or less exceptional) than u then, after revising his beliefs with φ, agent i will still consider v more plausible than u. More formally, for all $v, u \in ||\varphi||$ we

have that if $\kappa(u,i) > \kappa(v,i)$ and $(u,v) \in \mathcal{E}_i$ then $\kappa^{*_i^\alpha \varphi}(u,i) > \kappa^{*_i^\alpha \varphi}(v,i)$.[6] The degree of firmness α in the preceding condition 2(a) measures the extent to which agent i will believe that φ is true after revising his beliefs with φ. Indeed, as we will show in Section 4.1, in Spohn's theory of uncertainty the strength of the belief that φ is true is defined by the exceptionality degree of the negation of φ (i.e., $\kappa_{w,i}(\neg\varphi)$). Consequently, condition 2(a) guarantees that, if agent i envisages a world in which an objective formula φ is true then, after revising his beliefs with formula φ and with a degree of firmness α, he will believe φ with strength α. This property will become clearer in Section 4.1 in which the concept of graded belief (i.e., believing something with a certain strength) will be formally defined and its logical properties will be studied (see, in particular, Proposition 7).

Note that the reason why in Section 3.1 we assumed that α must be different from 0 is to guarantee that if an agent envisages a world in which a given objective formula φ is true then, after learning that φ is true, he will believe φ. Again, this property of the belief revision operation will become clearer in Section 4.1 where the concept of belief will be clearly defined (see, again, Proposition 7).

The function *Cut* is a minor technical device, taken from [5], which ensures that the new plausibility assignment fits into the finite set of natural numbers *Num*. Finally, the preceding condition 2(b) guarantees that the agent's i plausibility ordering over worlds does not change, if i learns something that he does not envisage.[7]

Remark 1 It is straightforward to verify that the operation $*_i^\alpha \varphi$ preserves the constraint (**Constr1**) over PDL-A-models. Therefore, if M is a PDL-A-model then $M^{*_i^\alpha \varphi}$ is a PDL-A-model too.

In what follows we write $\models \varphi$ to mean that φ is *valid* (φ is true in all PDL-A-models).

4 Varieties of Individual and Collective Attitudes

In the following two sections a variety of individual and collective attitudes will be defined, and their logical properties and logical relationships will be studied. We consider three kinds of individual attitudes, in addition to the concept of *knowledge* formalized by the operator $[i]$, namely *belief, graded belief* and *robust belief*. Furthermore, we consider three kinds of collective attitudes, namely *common knowledge, common belief* and *common robust belief*. Of course, we do not claim that this analysis is exhaustive. For instance, we do not consider the collective counterpart of

[6] As for the $\neg\varphi$-part, due to the fact that *Num* is finite, we can only say that: if v and u are worlds in which φ is false, agent i considers v more plausible than u and, according to agent i, u has a degree of exceptionality equal or lower than $\mathsf{max} - \alpha$ then, after revising his beliefs with φ, agent i will still consider v more plausible than u. More formally, for all $v, u \in ||\neg\varphi||$ we have that if $\kappa(u,i) > \kappa(v,i)$ and $(u,v) \in \mathcal{E}_i$ and $\kappa(u,i) \leq \mathsf{max} - \alpha$ then $\kappa^{*_i^\alpha \varphi}(u,i) > \kappa^{*_i^\alpha \varphi}(v,i)$.

[7] Note that the three conditions 1, 2(a) and 2(b) cover all cases. Indeed, the third condition $[i]\neg\varphi$ is equivalent to $\neg\varphi \wedge [i]\neg\varphi$, because $[i]\neg\varphi \rightarrow \neg\varphi$ is valid.

graded belief, namely the concept *common graded belief*, as a logical analysis of this concept goes beyond the objectives of the present work.

The concepts of belief, graded belief, robust belief and common belief will be fundamental building blocks for the analysis of the epistemic conditions of iterated weak dominance that we will carry out in Section 6. The concept of graded belief is also essential to understand the meaning of the parameter α in the belief revision operator $\left[*_i^{\alpha}\varphi\right]$). Finally, the concepts of common knowledge and common robust belief are defined here because we are interested: (1) in comparing them with the concept of common belief and (2) in understanding the similarities between the dynamic properties of robust belief and the dynamic properties of common robust belief, and between the dynamic properties of knowledge and the dynamic properties of common knowledge.

4.1 Individual Attitudes

Following [45], we say that agent i believes that φ is true, denoted by $\mathsf{Bel}_i\varphi$, if and only if φ is true in all worlds that i considers minimally exceptional (or maximally plausible). Let us define the belief operator Bel_i as follows:

$$\mathsf{Bel}_i\varphi \stackrel{\texttt{def}}{=} [i](\mathsf{exc}_{i,0} \rightarrow \varphi)$$

As the following Proposition 1 highlights, the previous abbreviation correctly characterizes this notion of belief. Given a PDL-A model $M = \langle W, \{\mathcal{E}_i : i \in Agt\}, \kappa, \mathcal{V}\rangle$ and a world w in M, let $\mathcal{B}_i = \{(w, v) : (w, v) \in \mathcal{E}_i \text{ and } \kappa(v, i) = 0\}$ be the accessibility relation for agent i's belief and $\mathcal{B}_i(w) = \{v \in W : (w, v) \in \mathcal{B}_i\}$ be the corresponding i's belief state at world w.

Proposition 1 *For every PDL-A model M and for every world w in M, $M, w \models \mathsf{Bel}_i\varphi$ if and only if $M, v \models \varphi$ for all $v \in \mathcal{B}_i(w)$.*

The following concept of 'graded belief' is taken from Spohn.[8] We say that at world w agent i believes that φ with strength equal to h, denoted by $\mathsf{Bel}_i^{\mathsf{h}}\varphi$, if and only if the degree of exceptionality of $\neg\varphi$ for agent i at w (i.e., $\kappa_{w,i}(\neg\varphi)$) is equal to h. In formal terms we define:

$$\mathsf{Bel}_i^{\mathsf{h}}\varphi \stackrel{\texttt{def}}{=} \begin{cases} \langle i\rangle(\mathsf{exc}_{i,\mathsf{h}} \wedge \neg\varphi) \wedge [i](\mathsf{exc}_{i,<\mathsf{h}} \rightarrow \varphi) & \text{if } \mathsf{h} < \mathsf{max} \\ [i](\mathsf{exc}_{i,<\mathsf{h}} \rightarrow \varphi) & \text{if } \mathsf{h} = \mathsf{max} \end{cases}$$

[8] A modal logic analysis of this concept has been given by Aucher [5] (see also [35, 54]). A relevant difference between Aucher's approach and our approach is that he introduces graded belief operators in the syntax right away, whereas we build them from the special atomic formulae $\mathsf{exc}_{i,\mathsf{h}}$. However, the added value of working with the $\mathsf{exc}_{i,\mathsf{h}}$-constructs is that they provide a simple extension of van Benthem et al.'s logic presented in [53]. This simple extension allows us to formalize a variety of individual and collective epistemic attitudes that have been studied in the literature (see Section 4) and that are not expressible in Aucher's logic. (For instance, Aucher's logic does not incorporate the concepts of common belief and common knowledge).

where $\mathsf{exc}_{i,<\mathsf{h}} \stackrel{\mathtt{def}}{=} \bigvee_{\mathsf{k} \in Num: 0 \leq \mathsf{k} < \mathsf{h}} \mathsf{exc}_{i,\mathsf{k}}$ for all $\mathsf{h} \in Num$ such that $\mathsf{h} \geq 1$, and $\mathsf{exc}_{i,<0} \stackrel{\mathtt{def}}{=} \bot$. The following proposition highlights that the preceding definition of the graded belief operator is indeed correct.

Proposition 2 *For every PDL-A model M, for every world w in M and for every $\mathsf{h} \in Num$, $M, w \models \mathsf{Bel}_i^{\mathsf{h}} \varphi$ if and only if $\kappa_{w,i}(\neg\varphi) = \mathsf{h}$.*

As we have emphasized above, graded belief is a fundamental concept of Spohn's theory of uncertainty and belief change, as it justifies the definition of belief revision we have given in Section 3.2 (Definition 4). It is worth noting that the graded belief operator $\mathsf{Bel}_i^{\mathsf{h}}$ is an operator of strong necessity (or actual necessity) in the sense of possibility theory [25].

For every $\mathsf{h} \in Num$, we moreover provide the following definition:

$$\mathsf{Bel}_i^{\geq \mathsf{h}} \varphi \stackrel{\mathtt{def}}{=} \bigvee_{\mathsf{k} \in Num: \mathsf{k} \geq \mathsf{h}} \mathsf{Bel}_i^{\mathsf{k}} \varphi$$

$\mathsf{Bel}_i^{\geq \mathsf{h}} \varphi$ has to be read "agent believes that φ is true with strength *at least* h". It is worth noting that, when $\mathsf{h} \geq 1$, the operator $\mathsf{Bel}_i^{\geq \mathsf{h}}$ is a normal operator, as it can be interpreted by means of the following accessibility relation:

$$\mathcal{B}_i^{<\mathsf{h}} = \{(w, v) : (w, v) \in \mathcal{E}_i \text{ and } \kappa(v, i) < \mathsf{h}\}.$$

In particular, given a PDL-A model $M = \langle W, \{\mathcal{E}_i : i \in Agt\}, \kappa, \mathcal{V}\rangle$, we have that $M, w \models \mathsf{Bel}_i^{\geq \mathsf{h}} \varphi$ if and only if $M, v \models \varphi$ for all v such that $(w, v) \in \mathcal{B}_i^{<\mathsf{h}}$. The operator $\mathsf{Bel}_i^{\geq \mathsf{h}}$ will be a key element in the analysis of the epistemic conditions of iterated weak dominance that we will conduct in Section 6.4. Specifically, it will be necessary in order to define the notion of *perfect rationality* on which the concept of iterated weak dominance is based. For notational convenience, we define $\widehat{\mathsf{Bel}}_i^{\geq \mathsf{h}}$ to be the dual operator of $\mathsf{Bel}_i^{\geq \mathsf{h}}$, that is, $\widehat{\mathsf{Bel}}_i^{\geq \mathsf{h}} \varphi \stackrel{\mathtt{def}}{=} \neg \mathsf{Bel}_i^{\geq \mathsf{h}} \neg \varphi$.

The following Propositions 3–7 capture some interesting properties of the preceding types of individual attitudes. For instance, the following Proposition 3 highlights that modal operators for belief and graded belief with strength *at least* h are normal.

Proposition 3 *For every $\Box \in \{\mathsf{Bel}_i : i \in Agt\} \cup \left\{\mathsf{Bel}_i^{\geq \mathsf{h}} : i \in Agt, \mathsf{h} \in Num \setminus \{0\}\right\}$ we have:*

$$\models (\Box\varphi \wedge \Box\psi) \rightarrow \Box(\varphi \wedge \psi) \tag{1a}$$

$$\textit{If } \models \varphi \text{ then } \models \Box\varphi \tag{1b}$$

According to the following Proposition 4, belief is characterized by the normal modal logic system KD45.

Proposition 4 *For every $i \in Agt$ we have:*

$$\models \neg(\mathsf{Bel}_i\varphi \wedge \mathsf{Bel}_i\neg\varphi) \tag{2a}$$
$$\models \mathsf{Bel}_i\varphi \rightarrow \mathsf{Bel}_i\mathsf{Bel}_i\varphi \tag{2b}$$
$$\models \neg\mathsf{Bel}_i\varphi \rightarrow \mathsf{Bel}_i\neg\mathsf{Bel}_i\varphi \tag{2c}$$

Note that the item (2a) follows from the normality constraint (**Constr1**) over PDL-A models given in Section 3.2.

The following Proposition 5 highlights some basic relationships between knowledge, belief and graded beliefs with different strengths.

Proposition 5 *For every $i \in Agt$ we have:*

$$\models \mathsf{Bel}_i^{\mathsf{h}}\varphi \rightarrow \neg\mathsf{Bel}_i^{\mathsf{k}}\varphi \text{ if } \mathsf{h} \neq \mathsf{k} \tag{3a}$$
$$\models [i]\varphi \rightarrow \mathsf{Bel}_i^{\mathsf{max}}\varphi \tag{3b}$$
$$\models [i]\varphi \rightarrow \mathsf{Bel}_i\varphi \tag{3c}$$
$$\models \mathsf{Bel}_i\varphi \leftrightarrow \mathsf{Bel}_i^{\geq 1}\varphi \tag{3d}$$

According to the item (3a), an agent cannot believe the same thing with different strengths. Moreover, knowing that φ implies believing that φ with maximal strength max (3b); knowing that φ implies believing that φ (3c); believing that φ coincides believing that φ with strength at least 1 (3d).

The following Proposition 6 captures the basic decomposability properties of the operators of graded belief.

Proposition 6 *For every $i \in Agt$ we have:*

$$\models \left(\mathsf{Bel}_i^{\mathsf{h}}\varphi \wedge \mathsf{Bel}_i^{\mathsf{k}}\psi\right) \rightarrow \mathsf{Bel}_i^{\geq \max\{\mathsf{h},\mathsf{k}\}}(\varphi \vee \psi) \tag{4a}$$
$$\models \left(\mathsf{Bel}_i^{\mathsf{h}}\varphi \wedge \mathsf{Bel}_i^{\mathsf{k}}\psi\right) \rightarrow \mathsf{Bel}_i^{\min\{\mathsf{h},\mathsf{k}\}}(\varphi \wedge \psi) \tag{4b}$$

According to the validity (4a), the degree of belief of $\varphi \vee \psi$ is at least equal to the maximum of the degree of belief of φ and ψ. According to the validity (4b), the degree of belief of $\varphi \wedge \psi$ is equal to the minimum of the degree of belief of φ and ψ. Similar properties for graded belief are given in possibility theory [25]. Note that second validity uses the "definite" value $\min\{\mathsf{h}, \mathsf{k}\}$ while the first validity uses the "at least" construction $\geq \max\{\mathsf{h}, \mathsf{k}\}$ because the minimun of the union of two sets is equal to the minimum of the minima of the two sets, while the minimun of the intersection of two sets is at least equal to the maximum of the minima of the two sets but not necessarily equal.

Finally, the following Proposition 7 is about the dynamic properties of belief and graded belief.

Proposition 7 *For every $i \in Agt$ and for every $\alpha \in Num \setminus \{0\}$ we have:*

$$\models \langle i \rangle \varphi \rightarrow \left[*_i^{\alpha} \varphi\right] \mathsf{Bel}_i \varphi \text{ if } \varphi \in Obj \tag{5a}$$

$$\models \langle i \rangle \varphi \rightarrow \left[*_i^{\alpha} \varphi\right] \mathsf{Bel}_i^{\alpha} \varphi \text{ if } \varphi \in Obj \tag{5b}$$

$$\models [i]\psi \rightarrow \left[*_i^{\alpha} \varphi\right] [i]\psi \text{ if } \psi \in Obj \tag{5c}$$

$$\models \left[*_i^{\alpha} \varphi\right] [i]\psi \rightarrow [i]\psi \text{ if } \psi \in Obj \tag{5d}$$

Item (5a) highlights a basic property of belief revision in the sense of AGM theory [1], namely the so-called success postulate: if φ is an objective fact and agent i envisages a world in which φ is true then, after learning that φ is true, agent i believes that φ is true.[9] Note that this property does not hold in general but only for formulas in Obj. Indeed, if we drop the restriction to Boolean formulas, the match to AGM success postulate does not work anymore. For instance, if φ is a Moore-like sentence of the form $p \wedge \neg \mathsf{Bel}_i p$, the formula $\langle i \rangle \varphi \rightarrow \left[*_i^{\alpha} \varphi\right] \mathsf{Bel}_i \varphi$ is clearly not valid.

Item (5b) clarifies the role of the degree of firmness α in the operation of belief revision: if φ is an objective fact and agent i envisages a world in which φ is true then, after revising his beliefs with formula φ and with a degree of firmness α, agent i believes that φ is true with strength equal to α. Finally, item (5c) captures the fundamental property of knowledge that we have discussed in Section 3.1: if ψ is an objective fact and agent i knows that ψ then, after learning a new fact φ, he will continue to know that ψ is true. In this sense, knowledge is stable under belief revision with any piece of information. The last item (5d) is a no-learning principle for knowledge: if ψ is an objective fact and agent i will know that ψ after revising his knowledge with φ then, it means that i already knows ψ.

Before moving to the analysis of collective attitudes, we follow Stalnaker [48] in defining a notion of robust belief *relative to* a specific formula φ, in the sense of belief which is stable under belief revision with φ:

$$\mathsf{RBel}_i(\varphi, \psi) \stackrel{\mathtt{def}}{=} \mathsf{Bel}_i \psi \wedge \left[*_i^{\alpha} \varphi\right] \mathsf{Bel}_i \psi$$

where α is any arbitrary value in $Num \setminus \{0\}$ (e.g., $\alpha = 1$). The construction $\mathsf{RBel}_i(\varphi, \psi)$ has to be read "agent i has a robust belief that ψ relative to φ". One reason why the value of the parameter α can be taken arbitrarily is that we have the following validity, for all $\alpha, \alpha' \in Num \setminus \{0\}$ and for all $\psi \in Obj$:

$$\models \left[*_i^{\alpha} \varphi\right] \mathsf{Bel}_i \psi \leftrightarrow \left[*_i^{\alpha'} \varphi\right] \mathsf{Bel}_i \psi$$

Therefore, the definition of relative robust belief $\mathsf{RBel}_i(\varphi, \psi)$ is independent from the value of α in case of objective formulas (i.e., $\mathsf{Bel}_i \psi \wedge \left[*_i^{\alpha} \varphi\right] \mathsf{Bel}_i \psi$ is logically equivalent to $\mathsf{Bel}_i \psi \wedge \left[*_i^{\alpha'} \varphi\right] \mathsf{Bel}_i \psi$ for all $\alpha, \alpha' \in Num \setminus \{0\}$ and for all $\psi \in Obj$). This concept of relative robust belief, as well as the concept of graded belief defined above, will be fundamental in the logical

[9] The only difference with AGM theory is the condition $\langle i \rangle \varphi$. AGM assumes that the new information φ must incorporated in the belief base, whereas we here assume that φ must incorporated in the belief base *only if* agent i envisages a world in which φ is true.

analysis of the epistemic conditions of iterated weak dominance we will carry out in Section 6.

Remark 2 It is worth noting that our concept of relative robust belief, represented by the formula $\mathsf{RBel}_i(\varphi, \psi)$, is related to the concept of robust (or strong) belief defined by [10]. According to Baltag & Smets, φ is a strong belief if and only if φ is epistemically possible and moreover all epistemically possible φ-states are strictly more plausible than all epistemically possible $\neg\varphi$-states. More formally, we can say that at world w agent i has the robust belief that ψ (in Baltag & Smets's sense), denoted by $\mathsf{RBel}_i^{B\&S}\psi$, if and only if (i) $||\psi||_{w,i} \neq \emptyset$ and (ii) for all $v \in ||\psi||_{w,i}$ and for all $u \in ||\neg\psi||_{w,i}$, $\kappa(v, i) < \kappa(u, i)$. The operator $\mathsf{RBel}_i^{B\&S}\psi$ can be syntactically expressed in the logic PDL-A. In particular, for every PDL-A model M and for all $w \in W$ we have:

$$M, w \models \mathsf{RBel}_i^{B\&S}\psi \text{ if and only if}$$

$$M, w \models \mathsf{Bel}_i\psi \wedge \bigwedge\nolimits_{\mathsf{h}\in Num\setminus\{0\}} (\langle i\rangle(\neg\psi \wedge \mathsf{exc}_{i,\mathsf{h}}) \rightarrow [i](\psi \rightarrow \mathsf{exc}_{i,<\mathsf{h}})).$$

Note that for all objective facts $\psi \in Obj$, we have the following validity:

$$\models \Big(\mathsf{RBel}_i^{B\&S}\psi \wedge \langle i\rangle(\varphi \wedge \psi)\Big) \rightarrow \mathsf{RBel}_i(\varphi, \psi).$$

In other words, if ψ is robust belief in Baltag & Smets's sense and, according to the agent's knowledge, φ and ψ are consistent, then the agent has a robust belief that ψ relative to φ, that is, the belief that ψ is stable under belief revision with φ. In other words, a robust belief in Baltag & Smets's sense can only be defeated by evidence (truthful or not) that is known to contradict it.

Following Baltag & Smets [9] we moreover define the following concept of 'safe belief'. In [48] Stalnaker calls it 'absolutely robust belief' in order to distinguish it from the preceding concept of 'relative robust belief', while in [49] he uses it to formally characterize Lehrer's defeasibility analysis of knowledge [36]. We say that agent i has the safe belief that φ, denoted by $\mathsf{SBel}_i\varphi$, if and only if φ is true in all worlds that i envisages and that are at least as plausible as the current world. In formal terms, we define:

$$\mathsf{SBel}_i\varphi \stackrel{\mathtt{def}}{=} \bigwedge_{\mathsf{h}\in Num} (\mathsf{exc}_{i,\mathsf{h}} \rightarrow [i](\mathsf{exc}_{i,\leq\mathsf{h}} \rightarrow \varphi))$$

with $\mathsf{exc}_{i,\leq\mathsf{h}} \stackrel{\mathtt{def}}{=} \bigvee_{\mathsf{k}\in Num:0\leq\mathsf{k}\leq\mathsf{h}} \mathsf{exc}_{i,\mathsf{k}}$. As the following Proposition 8 highlights, the previous abbreviation correctly characterizes this notion of safe belief. Given a PDL-A model $M = \langle W, \{\mathcal{E}_i : i \in Agt\}, \kappa, \mathcal{V}\rangle$ and a world w in M, let $\mathcal{SB}_i = \{(w, v) : (w, v) \in \mathcal{E}_i \text{ and } \kappa(v, i) \leq \kappa(w, i)\}$ be the accessibility relation for agent i's safe belief and let $\mathcal{SB}_i(w) = \{v \in W : (w, v) \in \mathcal{SB}_i\}$ be the corresponding set of i's $\mathcal{SB}_i$-accessible worlds at world w.[10]

[10]Note that the relation $\mathcal{SB}_i$ matches exactly to the plausibility relation in the pure qualitative accounts on belief revision in Dynamic Epistemic Logic (DEL) [9, 51].

Proposition 8 *For every PDL-A model M and for every world w in M, $M, w \models \mathsf{SBel}_i\varphi$ if and only if $M, v \models \varphi$ for all $v \in \mathcal{SB}_i(w)$.*

For notational convenience, we define $\widehat{\mathsf{SBel}}_i$ to be the dual operator of SBel_i, that is, $\widehat{\mathsf{SBel}}_i\varphi \stackrel{\text{def}}{=} \neg\mathsf{SBel}_i\neg\varphi$.

As the items (6a)–(6c) in the following Proposition 9 highlight, safe belief is characterized by the normal modal logic system S4.3 which exactly corresponds to Stalnaker's logic S4.3 in his defeasibility analysis of knowledge [49]. The item (6d) in Proposition 9 captures the characteristic property of safe belief: if ψ is an objective fact, agent i safely believes that ψ and φ is *true* then, after learning that φ, he will continue to safely believe that ψ is true. In this sense, a safe belief is stable under belief revision with any piece of *true* information.

Proposition 9 *For every $i \in Agt$ and for every $\alpha \in Num \setminus \{0\}$ we have:*

$$\models \mathsf{SBel}_i\varphi \rightarrow \varphi \tag{6a}$$

$$\models \mathsf{SBel}_i\varphi \rightarrow \mathsf{SBel}_i\mathsf{SBel}_i\varphi \tag{6b}$$

$$\models \left(\widehat{\mathsf{SBel}}_i\varphi \wedge \widehat{\mathsf{SBel}}_i\psi\right) \rightarrow \left(\widehat{\mathsf{SBel}}_i\left(\varphi \wedge \widehat{\mathsf{SBel}}_i\psi\right) \vee \widehat{\mathsf{SBel}}_i\left(\psi \wedge \widehat{\mathsf{SBel}}_i\varphi\right)\right) \tag{6c}$$

$$\models (\varphi \wedge \mathsf{SBel}_i\psi) \rightarrow \left[*_i^\alpha\varphi\right]\mathsf{SBel}_i\psi \text{ if } \psi \in Obj \tag{6d}$$

Note that, differently from knowledge, safe belief is not necessarily stable under belief revision with false information. Indeed, $\mathsf{SBel}_i\psi \rightarrow \left[*_i^\alpha\varphi\right]\mathsf{SBel}_i\psi$ is invalid even for $\psi \in Obj$.

4.2 Collective Attitudes

Given a PDL-A model $M = \langle W, \{\mathcal{E}_i : i \in Agt\}, \kappa, \mathcal{V}\rangle$, let $\mathcal{E}_J = \bigcup_{i\in J}\mathcal{E}_i$, $\mathcal{B}_J = \bigcup_{i\in J}\mathcal{B}_i$ and $\mathcal{SB}_J = \bigcup_{i\in J}\mathcal{SB}_i$. We define a world v to be *$\mathcal{E}_J$-reachable from world w*, denoted by $(w, v) \in \mathcal{E}_J^+$, if and only if there exist worlds $w_0, \ldots, w_n$ such that $w_0 = w$, $w_n = v$ and for all $0 \leq k \leq n-1$, there exists $i \in J$ such that $(w_k, w_{k+1}) \in \mathcal{E}_i$. In other words, for every $J \in 2^{Agt*}$, $\mathcal{E}_J^+$ is the defined to be the *transitive closure* of $\mathcal{E}_J$. Similarly, we define $\mathcal{B}_J^+$ and $\mathcal{SB}_J^+$ to be the transitive closures of $\mathcal{B}_J$ and $\mathcal{SB}_J$.

We define three types of collective attitudes that are interpreted by means of the relations $\mathcal{E}_J^+$, $\mathcal{B}_J^+$ and $\mathcal{SB}_J^+$. The first two correspond to the well-known concepts of common knowledge and common belief and are represented, respectively, by the operators CK_J and CBel_J. The third one, represented by the operator CSBel_J, corresponds to the concept of common safe belief that has been rather neglected in the logical literature up to now.[11] We define it here because we are interested in comparing it with the concepts of common knowledge and common belief, in the same

[11] The only exception is Baltag et al. [11] who study a notion of "common stable true belief" which similar to our notion of "common safe belief".

way as in Section 4.1 we compared safe belief with knowledge and belief. For all $J \in 2^{Agt*}$ we define:

$$\mathsf{CK}_J\varphi \stackrel{\mathtt{def}}{=} \left[\left(\bigcup_{i\in J} i\right)^*\right]\varphi$$

$$\mathsf{CBel}_J\varphi \stackrel{\mathtt{def}}{=} \left[\left(\bigcup_{i\in J}(i;\, ?\mathsf{exc}_{i,0})\right)^*\right]\varphi$$

$$\mathsf{CSBel}_J\varphi \stackrel{\mathtt{def}}{=} \left[\left(\bigcup_{\mathsf{h}\in Num, i\in J}(?\mathsf{exc}_{i,\mathsf{h}};\, i;\, ?\mathsf{exc}_{i,\leq \mathsf{h}})\right)^*\right]\varphi$$

As the following proposition highlights, the preceding three abbreviations correctly characterize the concepts of common knowledge, common belief and common safe belief.

Proposition 10 *For every PDL-A model M and for every world w in M:*

- $M, w \models \mathsf{CK}_J\varphi$ *if and only if* $M, v \models \varphi$ *for all* v *such that* $(w, v) \in \mathcal{E}_J^+$,
- $M, w \models \mathsf{CBel}_J\varphi$ *if and only if* $M, v \models \varphi$ *for all* v *such that* $(w, v) \in \mathcal{B}_J^+$,
- $M, w \models \mathsf{CSBel}_J\varphi$ *if and only if* $M, v \models \varphi$ *for all* v *such that* $(w, v) \in \mathcal{SB}_J^+$.

The following Proposition 11 highlights the basic logical relationships between common knowledge, common belief and common safe belief.

Proposition 11 *For every $J \in 2^{Agt*}$ we have:*

$$\models \mathsf{CK}_J\varphi \rightarrow \mathsf{CSBel}_J\varphi \tag{7a}$$

$$\models \mathsf{CSBel}_J\varphi \rightarrow \mathsf{CBel}_J\varphi \tag{7b}$$

According to the item (7a), J's common knowledge that φ entails J's common safe belief that φ whereas, according to the item (7b), J's common safe belief that φ entails J's common belief that φ.

Finally, the following Proposition 12 is about the dynamic properties of common knowledge, common belief and common safe belief. Let

$$[*_J\varphi]\,\psi \stackrel{\mathtt{def}}{=} \left[*_{i_1}^{\alpha_1}\varphi\right]\ldots\left[*_{i_{card(J)}}^{\alpha_{card(J)}}\varphi\right]\psi$$

where $(i_1, \ldots, i_{card(J)})$ is any arbitrary ordering of the elements of J, and $\alpha_1, \ldots, \alpha_{card(J)}$ are any arbitrary values in $Num \setminus \{0\}$. The construction $[*_J\varphi]\psi$ has to be read "after every agent in J has learnt that φ is true, ψ will be true".

Proposition 12 *For every $J \in 2^{Agt*}$ we have:*

$$\models \mathsf{CK}_J \bigwedge_{i\in J}\langle i\rangle\varphi \rightarrow [*_J\varphi]\mathsf{CBel}_J\varphi \text{ if } \varphi \in Obj \tag{8a}$$

$$\models \mathsf{CK}_J\psi \rightarrow [*_J\varphi]\mathsf{CK}_J\psi \text{ if } \psi \in Obj \tag{8b}$$

$$\models (\mathsf{CSBel}_J\varphi \wedge \mathsf{CSBel}_J\psi) \rightarrow [*_J\varphi]\mathsf{CSBel}_J\psi \text{ if } \psi \in Obj \tag{8c}$$

The item (8a) is the collective counterpart of the item (5a) of Proposition 7: if φ is an objective fact and the agents in J have common knowledge that each of them envisages a world in which φ is true then, after learning that φ is true, the agents in J acquire the common belief that φ. This principle can be viewed as the collective counterpart of the *success postulate* of AGM theory [1]. The item (8b) is the collective counterpart of the inertial principle for knowledge (Proposition 7, item 5c): if the agents in J have common knowledge that the objective fact ψ is true then, after learning that φ, they will continue to have common knowledge that ψ. The item (8c) is the collective counterpart of the inertial principle for safe belief (Proposition 9, item 6d): if the agents in J have the common safe belief that the objective fact ψ is true and that φ is true then, after learning that φ is true, they will continue to have the common safe belief that ψ. In this sense, common safe belief is stable under belief revision with any piece of information that the agents commonly and safely believe to be *true*.

5 Axiomatization and Decidability

In this section, we provide a complete axiomatization as well as a decidability result for the logic PDL-A. This logic has so-called reduction axioms which allow us to eliminate all the dynamic operators of belief revision from formulas. That elimination provides a decidable procedure for checking whether a given formula is PDL-A valid. Moreover it provides an axiomatics.

Let PDL-A$^-$ be the fragment of the logic PDL-A without the operators $\left[*_i^\alpha \varphi\right]$ and PDL-A^{--} be the fragment of PDL-A$^-$ without the special atoms $\mathsf{exc}_{i,\mathsf{h}}$. That is, let the language of PDL-A$^-$ be the set of formulas defined by the following BNF:

$$\varphi ::= p \mid \mathsf{exc}_{i,\mathsf{h}} \mid \neg\varphi \mid \varphi_1 \wedge \varphi_2 \mid [\pi]\varphi$$

and let the language of PDL-A^{--} be the set of formulas defined by the following BNF:

$$\varphi ::= p \mid \neg\varphi \mid \varphi_1 \wedge \varphi_2 \mid [\pi]\varphi$$

where p ranges over *Prop*, h ranges over *Num*, i ranges over *Agt* and π ranges over the set of knowledge programs as defined in Section 3.1.

Proposition 13 *The following formulas are PDL-A valid for all* $\mathsf{h}, \mathsf{k} \in$ *Num and for all* $i \in$ *Agt.*

$$\langle i\rangle \mathsf{exc}_{i,0} \qquad \textbf{(NormPlaus)}$$

$$\bigvee_{\mathsf{h}\in Num} \mathsf{exc}_{i,\mathsf{h}} \qquad \textbf{(ComplPlaus)}$$

$$\mathsf{exc}_{i,\mathsf{h}} \rightarrow \neg\mathsf{exc}_{i,\mathsf{k}} \text{ if } \mathsf{h} \neq \mathsf{k} \qquad \textbf{(UniquePlaus)}$$

Proposition 14 *The following equivalences are PDL-A valid for all* $p \in Atm$, $\mathsf{h} \in Num$, $\alpha \in Num \setminus \{0\}$, $i, j \in Agt$ *such that* $i \neq j$:

$$\left[*_i^\alpha \varphi\right] p \leftrightarrow p \tag{R1}$$

$$\begin{aligned}\left[*_i^\alpha \varphi\right] \mathsf{exc}_{i,\mathsf{h}} \leftrightarrow \quad & \Bigg(\Bigg(\varphi \wedge \bigvee_{\mathsf{l},\mathsf{m} \in Num\setminus\{0\}:\mathsf{l}-\mathsf{m}=\mathsf{h}} \left(\mathsf{Bel}_i^{\mathsf{m}} \neg\varphi \wedge \mathsf{exc}_{i,\mathsf{l}}\right)\Bigg) \vee \\ & \Bigg(\neg\varphi \wedge \langle i\rangle\varphi \wedge \bigvee_{\mathsf{l},\mathsf{m} \in Num:Cut(\alpha+\mathsf{l}-\mathsf{m})=\mathsf{h}} \left(\mathsf{Bel}_i^{\mathsf{m}} \varphi \wedge \mathsf{exc}_{i,\mathsf{l}}\right)\Bigg) \\ & \vee([i]\neg\varphi \wedge \mathsf{exc}_{i,\mathsf{h}})\Bigg)\end{aligned} \tag{R2}$$

$$\left[*_i^\alpha \varphi\right] \mathsf{exc}_{j,\mathsf{h}} \leftrightarrow \mathsf{exc}_{j,\mathsf{h}} \tag{R3}$$

$$\left[*_i^\alpha \varphi\right] \neg\psi \leftrightarrow \neg \left[*_i^\alpha \varphi\right] \psi \tag{R4}$$

$$\left[*_i^\alpha \varphi\right] (\psi_1 \wedge \psi_2) \leftrightarrow \left(\left[*_i^\alpha \varphi\right] \psi_1 \wedge \left[*_i^\alpha \varphi\right] \psi_2\right) \tag{R5}$$

$$\left[*_i^\alpha \varphi\right] [\pi]\psi \leftrightarrow \left[F^{*_i^\alpha \varphi}(\pi)\right] \left[*_i^\alpha \varphi\right] \psi \tag{R6}$$

where for all $j \in Agt$:

$$\begin{aligned} F^{*_i^\alpha \varphi}(j) &= j \\ F^{*_i^\alpha \varphi}(\pi_1; \pi_2) &= F^{*_i^\alpha \varphi}(\pi_1); F^{*_i^\alpha \varphi}(\pi_2) \\ F^{*_i^\alpha \varphi}(\pi_1 \cup \pi_2) &= F^{*_i^\alpha \varphi}(\pi_1) \cup F^{*_i^\alpha \varphi}(\pi_2) \\ F^{*_i^\alpha \varphi}(?\psi) &= ?\left[*_i^\alpha \varphi\right] \psi \\ F^{*_i^\alpha \varphi}(\pi^*) &= \left(F^{*_i^\alpha \varphi}(\pi)\right)^* \end{aligned}$$

As the rule of replacement of equivalents preserves validity, the equivalences of Proposition 14 together with this allow to reduce every PDL-A formula to an equivalent PDL-A$^-$ formula. Call *red* the mapping which iteratively applies the above equivalences from the left to the right, starting from one of the innermost modal operators. *Red* pushes the dynamic operators inside the formula, and finally eliminates them when facing an atomic formula.

Proposition 15 *Let* φ *be a formula in the language of PDL-A. Then*

1. *red*(φ) *has no dynamic operators* $[\pi]$
2. *red*$(\varphi) \leftrightarrow \varphi$ *is PDL-A valid.*

The first item of Proposition 15 is clear. The second item is proved using Proposition 14 and the rule of replacement of equivalents.

Theorem 1 *Satisfiability in PDL-A is decidable.*

Theorem 2 *The validities of PDL-A are completely axiomatized by*

- *all principles of classical propositional logic*
- *axiomatization of* PDL *for the operators* $[\pi]$

$$([\pi]\varphi \wedge [\pi](\varphi \rightarrow \psi)) \rightarrow [\pi]\psi \qquad (\mathbf{K}_{[\pi]})$$

$$[\pi_1; \pi_2]\varphi \leftrightarrow [\pi_1][\pi_2]\varphi \qquad (\mathbf{Seq})$$

$$[\pi_1 \cup \pi_2]\varphi \leftrightarrow [\pi_1]\varphi \wedge [\pi_2]\varphi \qquad (\mathbf{Choice})$$

$$[?\psi]\varphi \leftrightarrow (\psi \rightarrow \varphi) \qquad (\mathbf{Test})$$

$$[\pi^*]\varphi \leftrightarrow (\varphi \wedge [\pi][\pi^*]\varphi) \qquad (\mathbf{FixPoint})$$

$$(\varphi \wedge [\pi^*](\varphi \rightarrow [\pi])) \rightarrow [\pi^*]\varphi \qquad (\mathbf{Induction})$$

$$\frac{\varphi}{[\pi]\varphi} \qquad (\mathbf{Nec}_{[\pi]})$$

- *axioms T, 4 and B for the epistemic operators* $[i]$

$$[i]\varphi \rightarrow \varphi \qquad (\mathbf{T}_{[i]})$$

$$[i]\varphi \rightarrow [i][i]\varphi \qquad (\mathbf{4}_{[i]})$$

$$\varphi \rightarrow [i]\langle i\rangle\varphi \qquad (\mathbf{B}_{[i]})$$

- *the schemas of Proposition 13*
- *the reduction axioms of Proposition 14*
- *rule of replacement of equivalents*

$$\frac{\psi_1 \leftrightarrow \psi_2}{\varphi \leftrightarrow \varphi[\psi_1/\psi_2]} \qquad (\mathbf{REP})$$

where $\varphi[\psi_1/\psi_2]$ *is the formula that results from* φ *by replacing zero or more occurrences of* ψ_1*, in* φ*, by* ψ_2*.*

6 Application to Game Theory

In this section we provide an application of the logic PDL-A to game theory. We first introduce in Section 6.1 two important solution concepts of game theory: the procedure of Iterated Deletion of Strongly Dominated Strategies (IDSDS procedure), and the procedure of n-rounds of Iterated Deletion of Weakly Dominated Strategies followed by Iterated Deletion of Strongly Dominated Strategies (DWDS^n-IDSDS

procedure). In Section 6.2 we provide an extension of the logic PDL-A with information about agents' behaviors (i.e., which strategy a given agent is currently playing). We call PDL-A$^+$ the resulting logic. In Section 6.4 the logic PDL-A$^+$ is used to formally characterize different forms of rationality which have been discussed in the field of epistemic game theory: weak rationality, strong rationality and perfect rationality. Section 6.4 is the main contribution of this second part of the paper. Several theorems about the epistemic characterization of IDSDS and DWDS^n-IDSDS will be provided. Proofs of these theorems are collected in a technical annex at the end of the paper.

6.1 Iterated Weak and Strong Dominance

Let us first introduce the notion of normal form game.

Definition 5 (**Normal form game**) A normal form game is a tuple $\Gamma = \{\{S_i : i \in Agt\}, \{U_i : i \in Agt\}\}$ where:

- S_i is agent i's finite set of strategies;
- $U_i : \prod_{i \in Agt} S_i \longrightarrow \mathbb{R}$ is agent i's utility function assigning a real number (the utility value for i) to every combination of agents' actions (alias strategy profiles).

For every agent $i \in Agt$, the elements of S_i are denoted by symbols $a_i, b_i, \ldots$ For every coalition $J \in 2^{Agt*}$, we define the set of strategies for the coalition J to be $S_J = \prod_{i \in J} S_i$. For notational convenience we write S instead of S_{Agt}. For every coalition J, elements of S_J are denoted by $s_J, s'_J, \ldots$ For simplicity, elements of S are denoted by $s, s', \ldots$ Given $s_J \in S_J$ and $i \in J$, we note $s_J[i]$ the position of s_J corresponding to agent i. In what follows we write $s \leq_i s'$ instead of $U_i(s) \leq U_i(s')$ and $s <_i s'$ instead of $U_i(s) < U_i(s')$ to mean respectively that "the strategy profile s' is for agent i at least as good as the strategy profile s" and "the strategy profile s' is for agent i better than the strategy profile s".

Definition 6 (**Subgame**) Given two games $\Gamma = \{\{S_i : i \in Agt\}, \{U_i : i \in Agt\}\}$ and $\Gamma' = \{\{S'_i : i \in Agt\}, \{U'_i : i \in Agt\}\}$, Γ' is a subgame of Γ if and only if:

- for every $i \in Agt$, $S'_i \subseteq S_i$;
- for every $i \in Agt$, $U'_i = U_i|_{\prod_{i \in Agt} S'_i}$ where U'_i is the restriction of U_i to the set of strategy profiles $\prod_{i \in Agt} S'_i$.

A strategy a_i of a given player i is a strongly dominated strategy if and only if, there exists another strategy b_i of i such that, for all strategies s_{-i} of the other players, playing b_i while the others play s_{-i} is for i better than playing a_i while the others play s_{-i}. An example of strongly dominated strategy is cooperation in the Prisoner Dilemma (PD) game: whether the opponent chooses to cooperate or defect, defection yields a higher payoff than cooperation. More formally:

Definition 7 (**Strongly dominated strategies**) Given a game $\Gamma = \{\{S_i : i \in Agt\}, \{U_i : i \in Agt\}\}$, the set

$$\mathsf{SD}_i^{\Gamma} = \{a_i \in S_i : \exists b_i \in S_i \text{ s.t. } \forall s_{-i} \in S_{-i}, \langle a_i, s_{-i}\rangle <_i \langle b_i, s_{-i}\rangle\}$$

is the set of strategies of player i that are strongly dominated in Γ.

A strategy a_i of a given player i is a weakly dominated strategy if and only if, there exists another strategy b_i of i such that, for all strategies s_{-i} of the others, playing b_i while the others play s_{-i} is for i at least as good as playing a_i while the others play s_{-i} and there is at least one strategy s'_{-i} of the others such that playing b_i while the others play s'_{-i} is for i better than playing a_i while the others play s'_{-i}. More formally:

Definition 8 (**Weakly dominated strategies**) Given a game $\Gamma = \{\{S_i : i \in Agt\}, \{U_i : i \in Agt\}\}$, the set

$$\begin{aligned}\mathsf{WD}_i^{\Gamma} = \{a_i \in S_i : \exists b_i \in S_i \text{ s.t. } \forall s_{-i} \in S_{-i}, \langle a_i, s_{-i}\rangle \leq_i \langle b_i, s_{-i}\rangle \text{ and}\\ \exists s'_{-i} \in S_{-i} \text{ s.t.} \langle a_i, s'_{-i}\rangle <_i \langle b_i, s'_{-i}\rangle\}\end{aligned}$$

is the set of strategies of player i that are weakly dominated in Γ.

The so-called Iterated Deletion of Strongly Dominated Strategies (IDSDS) (or iterated strong dominance) is a procedure that starts with the original game Γ and, at each step, for every player i removes from the game all i's strongly dominated strategies, thereby generating a subgame of the original game, and that repeats this process again and again. The IDSDS procedure can be inductively defined as follows.

Definition 9 (**IDSDS Procedure**) Given a game $\Gamma = \{\{S_i : i \in Agt\}, \{U_i : i \in Agt\}\}$, Iterated Deletion of Strongly Dominated Strategies (IDSDS) is the procedure defined recursively as follows.

For all $i \in Agt$: let $S_{i,0}^{IDSDS} = S_i$ and $\Gamma^0 = \Gamma$, for $m \geq 1$, let $S_{i,m}^{IDSDS} = S_{i,m-1}^{IDSDS} \setminus \mathsf{SD}_i^{\Gamma^{m-1}}$, where Γ^{m-1} is the subgame of Γ with strategy sets $S_{i,m-1}^{IDSDS}$.

For every $m \geq 0$ and $J \in 2^{Agt*}$, let $S_{J,m}^{IDSDS} = \prod_{i\in J} S_{i,m}^{IDSDS}$ and let $S_m^{IDSDS} = S_{Agt,m}^{IDSDS}$.

Finally, let $S_i^{IDSDS} = \bigcap_{m\in\mathbb{N}} S_{i,m}^{IDSDS}$. For every $J \in 2^{Agt*}$, let $S_J^{IDSDS} = \prod_{i\in J} S_i^{IDSDS}$ and $S^{IDSDS} = S_{Agt}^{IDSDS}$.

The procedure of n-rounds of Iterated Deletion of Weakly Dominated Strategies followed by Iterated Deletion of Strongly Dominated Strategies (DWDS^n-IDSDS) is a procedure that starts with the original game Γ and, at each step, for every player i removes from the game all i's weakly dominated strategies thereby generating a subgame of the original game, and that repeats this process for n rounds. Then, after n rounds it applies Iterated Deletion of Strongly Dominated Strategies starting with the game Γ^n. The DWDS^n-IDSDS procedure can be inductively defined as follows.

Definition 10 (**DWDS^n-IDSDS Procedure**) Given a game $\Gamma = \{\{S_i : i \in Agt\}, \{U_i : i \in Agt\}\}$, n-iteration of Deletion of Weakly Dominated Strategies

(DWDS^n) followed by Iterated Deletion of Strongly Dominated Strategies (IDSDS) is the procedure defined recursively as follows.

For all $i \in Agt$: let $S_{i,0}^{DWDS^n-IDSDS} = S_i$ and $\Gamma^0 = \Gamma$, for $1 \leq m \leq n$, let $S_{i,m}^{DWDS^n-IDSDS} = S_{i,m-1}^{DWDS^n-IDSDS} \setminus \mathsf{WD}_i^{\Gamma^{m-1}}$, for $m > n$, let $S_{i,m}^{DWDS^n-IDSDS} = S_{i,m-1}^{DWDS^n-IDSDS} \setminus \mathsf{SD}_i^{\Gamma^{m-1}}$, where Γ^{m-1} is the subgame of Γ with strategy sets $S_{i,m-1}^{DWDS^n-IDSDS}$.

For every m such that $0 \leq m \leq n$ and $J \in 2^{Agt*}$, let $S_{J,m}^{DWDS^n-IDSDS} = \prod_{i\in J} S_{i,m}^{DWDS^n-IDSDS}$ and $S_m^{DWDS^n-IDSDS} = S_{Agt,m}^{DWDS^n-IDSDS}$.

Finally, let $S_i^{DWDS^n-IDSDS} = \bigcap_{m\in\mathbb{N}} S_{i,m}^{DWDS^n-IDSDS}$. For every $J \in 2^{Agt*}$, let $S_J^{DWDS^n-IDSDS} = \prod_{i\in J} S_i^{DWDS^n-IDSDS}$ and $S^{DWDS^n-IDSDS} = S_{Agt}^{DWDS^n-IDSDS}$.

Note that, if $n = 0$, DWDS^n-IDSDS is nothing but IDSDS. Moreover, if $n = \infty$, DWDS^n-IDSDS corresponds to the procedure of iterated admissibility or iterated deletion of weakly dominated strategies [20, 23].

6.2 PDL-A with Information About Players' Choices

The logic PDL-A is here extended with special constructions of the form $pl_i(a_i)$ whose meaning is "agent i plays (or chooses) the strategy a_i". We call PDL-A$^+$ the resulting logic. For every $J \in 2^{Agt*}$, we define $pl_J(s_J)$ ("the agents in J play the collective strategy s_J") as follows:

$$pl_J(s_J) \stackrel{\mathtt{def}}{=} \bigwedge_{i\in J} pl_i(s_J[i])$$

For simplicity, we write $pl(s)$ instead of $pl_{Agt}(s)$.

For every coalition of agents $J \in 2^{Agt*}$, we define the set Beh_J of information about J's choices:

$$Beh_J = \left\{ \bigvee_{s_J\in \mathbf{s}_J} pl_J(s_J) : \mathbf{s}_J \subseteq S_J \right\}.$$

For example, the formula $pl_J(s_J) \vee pl_J(s'_J)$ in Beh_J means that "the agents in J play either the collective strategy s_J or the collective strategy s'_J".

Definition 11 (PDL-A$^+$ model) PDL-A$^+$-models are tuples $M' = \langle M, \{\mathcal{A}_i : i \in Agt\}\rangle$ where:

- M is a PDL-A model;
- for every agent i, $\mathcal{A}_i : W \longrightarrow S_i$ is a total function mapping each world w to the strategy played by agent i at w.

$\mathcal{A}_i(w) = a_i$ means that at w agent i plays the strategy a_i.

Functions $\mathcal{A}_i$ are easily generalized to functions $\mathcal{A}_J : W \longrightarrow S_J$, by postulating that $\mathcal{A}_J(w) = s_J$ if and only if $\mathcal{A}_i(w) = s_J[i]$ for every $i \in J$.

Given a PDL-A$^+$ model M and a world w, the truth condition of $pl_i(a_i)$ is:

$$M, w \models pl_i(a_i) \text{ iff } \mathcal{A}_i(w) = a_i$$

PDL-A$^+$ models are assumed to satisfy the following two constraints. For every $w, v \in W$, $i \in Agt$, $a_i \in S_i$ and $s_{-i} \in S_{-i}$:

(**Constr2**) if $\mathcal{A}_i(w) = a_i$ and $(w, v) \in \mathcal{E}_i$ then $\mathcal{A}_i(v) = a_i$;
(**Constr3**) there is u such that $(w, u) \in \mathcal{E}_i$ and $\mathcal{A}_{-i}(u) = s_{-i}$.[12]

According to the Constraint (**Constr2**), an agent i chooses the strategy a_i if and only if he knows this. According to the Constraint (**Constr3**), for every strategy s_{-i} of the other players an agent i envisages a world in which this strategy is played. This corresponds to the second requirement we have discussed in Section 2, namely the assumption that the beliefs of the players are cautious.

The notions of validity and satisfiability in PDL-A$^+$ are defined in the usual way. For every PDL-A$^+$ formula φ, we say that φ is *valid*, denoted again by $\models \varphi$, if φ is true in all PDL-A$^+$-models. We say that φ is *satisfiable* if $\neg\varphi$ is not valid. We moreover say that two PDL-A$^+$ formulas φ and ψ are *compatible*, denoted by $Comp(\varphi, \psi)$, if $\varphi \wedge \psi$ is satisfiable.

Note that the logic PDL-A$^+$ is completely axiomatized by the axioms and rules of inference of the logic PDL-A plus the following axiom schemas:

$$\bigvee_{a_i \in S_i} pl_i(a_i) \qquad \textbf{(Active)}$$

$$pl_i(a_i) \rightarrow \neg pl_i(b_i) \text{ if } a_i \neq b_i \qquad \textbf{(UniqueAct)}$$

$$pl_i(a_i) \rightarrow [i]pl_i(a_i) \qquad \textbf{(ActAware)}$$

$$\langle i \rangle pl_{-i}(s_{-i}) \qquad \textbf{(PossStr)}$$

The decidability of PDL-A$^+$ follows straightforwardly from the decidability of PDL-A and the fact that the set of axioms differentiating PDL-A$^+$ from PDL-A is finite (remember that every strategy set S_i is assumed to be finite).

6.3 Variants of Rationality

The following formula characterizes a notion of weak rationality which is commonly supposed in the epistemic analysis of games (see, e.g., [52]):

$$\mathsf{WRat}_i(a_i) \stackrel{\mathtt{def}}{=} \bigwedge_{b_i \neq a_i} \left(\bigvee_{s_{-i} \in S_{-i} : \langle b_i, s_{-i} \rangle \leq_i \langle a_i, s_{-i} \rangle} \widehat{\mathsf{Bel}}_i pl_{-i}(s_{-i}) \right)$$

[12] Note that this constraint ensures that every function $\mathcal{A}_i$ is an onto function.

This means that the strategy a_i is a *weakly* rational choice for the agent i, i.e., $\mathsf{WRat}_i(a_i)$, if and only if, for every strategy b_i different from a_i, there exists a joint strategy s_{-i} of the other agents that he considers maximally plausible such that, playing a_i while the others play s_{-i} is for agent i at least as good as playing b_i while the others play s_{-i}. This means that weak rationality simply consists in not choosing a strategy that is strongly dominated within the set of worlds that the agent considers maximally plausible. The following abbreviations WRat_i and $\mathsf{AllWRat}_J$ have to be read respectively "agent i is weakly rational" and "all agents in the group J are weakly rational":

$$\mathsf{WRat}_i \stackrel{\mathtt{def}}{=} \bigvee_{a_i \in S_i} (pl_i(a_i) \wedge \mathsf{WRat}_i(a_i))$$

$$\mathsf{AllWRat}_J \stackrel{\mathtt{def}}{=} \bigwedge_{i \in J} \mathsf{WRat}_i$$

The preceding notion of weak rationality has been distinguished from a slightly stronger notion of rationality, called strong rationality (see, e.g., [17]). The strategy a_i is a *strongly* rational choice for the agent i, i.e., $\mathsf{SRat}_i(a_i)$, if and only if, for each strategy b_i different from a_i either (1) there is a joint strategy s_{-i} of the other agents that i considers maximally plausible such that playing a_i while the others play s_{-i} is for i better than playing b_i while the others play s_{-i} or (2) there is no joint strategy s_{-i} of the other agents that i considers maximally plausible such that playing b_i while the others play s_{-i} is for i better than playing a_i while the others play s_{-i}. This means that strong rationality simply consists in not choosing a strategy that is weakly dominated within the set of worlds that the agent considers maximally plausible:

$$\mathsf{SRat}_i(a_i) \stackrel{\mathtt{def}}{=} \bigwedge_{b_i \neq a_i} \left(\left(\bigvee_{s_{-i} \in S_{-i} : \langle b_i s_{-i} \rangle <_i \langle a_i, s_{-i} \rangle} \widehat{\mathsf{Bel}}_i pl_{-i}(s_{-i}) \right) \vee \left(\bigwedge_{s_{-i} \in S_{-i} : \langle a_i, s_{-i} \rangle <_i \langle b_i, s_{-i} \rangle} \mathsf{Bel}_i \neg pl_{-i}(s_{-i}) \right) \right)$$

The following abbreviations SRat_i and $\mathsf{AllSRat}_J$ have to be read respectively "agent i is strongly rational" and "all agents in the group J are strongly rational":

$$\mathsf{SRat}_i \stackrel{\mathtt{def}}{=} \bigvee_{a_i \in S_i} (pl_i(a_i) \wedge \mathsf{SRat}_i(a_i))$$

$$\mathsf{AllSRat}_J \stackrel{\mathtt{def}}{=} \bigwedge_{i \in J} \mathsf{SRat}_i$$

Stalnaker has introduced an even stronger notion of rationality, called perfect rationality [47, 48]. Roughly, Stalnaker's notion expresses that in cases two or more actions maximize utility, the agent should consider, in choosing between them, how he should act if he learned that he was in error. And if the two actions are still tied, the agent considers how he should act if he learned that he was making an error of a higher degree (see also [15]). We here consider a slightly different variant of Stalnaker's notion of perfect rationality that can be conceived as a lexicographic

refinement of the preceding concept of strong rationality. Our notion of perfect rationality can be inductively defined by means of the preceding notion of strong rationality and of the notion of graded belief introduced in Section 4.1. The basic idea is the following. We say that the strategy a_i is a *strongly* rational choice for the agent i at *level* 1, denoted by $\mathsf{SRat}_i^1(a_i)$, if and only if a_i is a *strongly* rational choice according to agent i's current beliefs. This notion of *level* 1-strong rationality coincides with the notion of strong rationality defined above. Moreover, for every $\mathsf{h} \in \mathit{Num}$ such that $1 < \mathsf{h} \leq \mathsf{max}$, we say that the strategy a_i is a *strongly* rational choice for the agent i *at level* h, denoted by $\mathsf{SRat}_i^{\mathsf{h}}(a_i)$, if and only if:

- a_i is a strongly rational choice for the agent i at level $\mathsf{h}-1$, and
- strategy a_i is an admissible choice according to agent i's graded beliefs with strenght at least h,[13] after having discarded all strategies b_i different from a_i that are not strongly rational choices at level $\mathsf{h}-1$.

Finally, we say that the strategy a_i is a *perfectly* rational choice for the agent i, denoted by $\mathsf{PRat}_i(a_i)$, if and only if:

- a_i is a strongly rational choice for the agent i at level max, and
- strategy a_i is an admissible choice according to agent i's knowledge, after having discarded all strategies b_i different from a_i that are not strongly rational choices at level max.

In other words, a_i is a perfectly rational choice for the agent i if and only if: a_i is not weakly dominated within the set of epistemic alternatives that i considers exceptional with degree at most 0; and a_i is not weakly dominated within the set of epistemic alternatives that i considers exceptional with degree at most 1, after having discarded all weakly dominated strategies within the set of epistemic alternatives that i considers exceptional with degree at most 0; and so on until level max. Formally speaking, we define:

$$\mathsf{SRat}_i^1(a_i) \stackrel{\mathtt{def}}{=} \mathsf{SRat}_i(a_i)$$

for all $\mathsf{h} \in \mathit{Num}$ such that $1 < \mathsf{h} \leq \mathsf{max}$:

$$\mathsf{SRat}_i^{\mathsf{h}}(a_i) \stackrel{\mathtt{def}}{=} \mathsf{SRat}_i^{\mathsf{h}-1}(a_i) \wedge \bigwedge_{b_i \neq a_i} \Bigg(\mathsf{SRat}_i^{\mathsf{h}-1}(b_i) \rightarrow \Bigg(\Bigg(\bigvee_{s_{-i} \in S_{-i} : \langle b_i, s_{-i} \rangle <_i \langle a_i, s_{-i} \rangle} \widehat{\mathsf{Bel}}_i^{\geq \mathsf{h}} pl_{-i}(s_{-i}) \Bigg) \vee \Bigg(\bigwedge_{s_{-i} \in S_{-i} : \langle a_i, s_{-i} \rangle <_i \langle b_i, s_{-i} \rangle} \mathsf{Bel}_i^{\geq \mathsf{h}} \neg pl_{-i}(s_{-i}) \Bigg) \Bigg) \Bigg)$$

[13] Remember from Section 4.1 that the graded belief operator $\mathsf{Bel}_i^{\geq \mathsf{h}}$ can be interpreted by means of the binary relation $\mathcal{B}_i^{<\mathsf{h}} = \{(w, v) : (w, v) \in \mathcal{E}_i \text{ and } \kappa(v, i) < \mathsf{h}\}$. Thus, strategy a_i is an admissible choice according to agent i's graded beliefs with strenght at least h if and only if, strategy a_i is a strongly rational choice with respect to agent i's set of envisaged worlds with exceptionality *at most* $\mathsf{h}-1$.

Finally:

$$\mathsf{PRat}_i(a_i) \stackrel{\mathtt{def}}{=} \mathsf{SRat}_i^{\max}(a_i) \wedge \bigwedge_{b_i \neq a_i} \Bigg(\mathsf{SRat}_i^{\max}(b_i) \rightarrow \Bigg(\Bigg(\bigvee_{s_{-i} \in S_{-i} : \langle b_i, s_{-i} \rangle <_i \langle a_i, s_{-i} \rangle} \langle i \rangle pl_{-i}(s_{-i}) \Bigg) \vee \Bigg(\bigwedge_{s_{-i} \in S_{-i} : \langle a_i, s_{-i} \rangle <_i \langle b_i, s_{-i} \rangle} [i] \neg pl_{-i}(s_{-i}) \Bigg) \Bigg) \Bigg)$$

The following abbreviations PRat_i and $\mathsf{AllPRat}_J$ have to be read respectively "agent i is perfectly rational" and "all agents in the group J are perfectly rational":

$$\mathsf{PRat}_i \stackrel{\mathtt{def}}{=} \bigvee_{a_i \in S_i} (pl_i(a_i) \wedge \mathsf{PRat}_i(a_i))$$

$$\mathsf{AllPRat}_J \stackrel{\mathtt{def}}{=} \bigwedge_{i \in J} \mathsf{PRat}_i$$

The following Proposition 16 highlights the logical relationships between the three notions of rationality (weak, strong and perfect).

Proposition 16 *For every $i \in Agt$ we have:*

$$\models \mathsf{SRat}_i \rightarrow \mathsf{WRat}_i \tag{9a}$$

$$\models \mathsf{PRat}_i \rightarrow \mathsf{SRat}_i \tag{9b}$$

The following Proposition 17 is about properties of positive and negative introspection for the three forms of rationality. If an agent is/is not weakly/strongly/perfectly rational then, he knows this.

Proposition 17 *For every $i \in Agt$ we have:*

$$\models \mathsf{WRat}_i \rightarrow [i]\mathsf{WRat}_i \tag{10a}$$

$$\models \mathsf{SRat}_i \rightarrow [i]\mathsf{SRat}_i \tag{10b}$$

$$\models \mathsf{PRat}_i \rightarrow [i]\mathsf{PRat}_i \tag{10c}$$

$$\models \neg\mathsf{WRat}_i \rightarrow [i]\neg\mathsf{WRat}_i \tag{10d}$$

$$\models \neg\mathsf{SRat}_i \rightarrow [i]\neg\mathsf{SRat}_i \tag{10e}$$

$$\models \neg\mathsf{PRat}_i \rightarrow [i]\neg\mathsf{PRat}_i \tag{10f}$$

6.4 Epistemic Conditions of Solution Concepts

The following Theorem 3 is the qualitative version of a probabilistic-based result of Stalnaker [46] who has been the first to use probabilistic Kripke structures in order

to characterize the IDSDS procedure in terms of common belief of weak rationality (see [16, 17] for some recent discussion of Stalnaker's result). A similar result has also been proved, with differing degrees of formality, by Bernheim [13], Pearce [40], Brandenburger & Dekel [21], Tan & Werlang [50] and Lorini & Schwarzentruber [37]. According to the Theorem 3, if the players have common belief that every agent is weakly rational then the strategy profile which is played must survive IDSDS.

Theorem 3 *Let $s \notin S^{IDSDS}$. Then:*

$$\models \mathsf{CBel}_{Agt}\mathsf{AllWRat}_{Agt} \rightarrow \neg pl(s)$$

According to the following Theorem 4, if the players have common belief that every agent is perfectly rational, then the strategy profile which is played must survive one iteration of DWDS followed by IDSDS. This is called the Dekel-Fudenberg procedure as it appeared for the first time in [24]. Characterizations of the epistemic conditions of this solution concept have also been given in a probabilistic setting by Stalnaker [47] as well as by Brandenburger [22] and Börgers [18].[14]

Theorem 4 *Let $s \notin S^{DWDS^1-IDSDS}$. Then:*

$$\models \mathsf{CBel}_{Agt}\mathsf{AllPRat}_{Agt} \rightarrow \neg pl(s)$$

According to the following Theorem 5, if the players have common belief that every player (1) is perfectly rational and (2) has a robust belief that all other players are perfectly rational relative to any compatible information about their choices, then the strategy profile which is played must survive two iterations of DWDS followed by IDSDS. A similar theorem has been stated before by Stalnaker [48], even though he did not provide a formal proof for it. The main difference between Theorem 5 and Stalnaker's result is that Stalnaker's analysis is given in quantitative setting based on probabilities whereas the representation of uncertainty used here is semi-qualitative (see Section 7 for further discussion).

Theorem 5 *Let $s \notin S^{DWDS^2-IDSDS}$. Then:*

$$\models \mathsf{CBel}_{Agt}\left(\mathsf{AllPRat}_{Agt} \wedge \mathsf{AllRBelPRat}_{Agt}\right) \rightarrow \neg pl(s)$$

where for every $J \in 2^{Agt*}$:

$$\mathsf{AllRBelPRat}_J \stackrel{\mathtt{def}}{=} \bigwedge_{i \in J} \left(\bigwedge_{\substack{\chi_{-i} \in Beh_{-i}: \\ Comp(\chi_{-i}, \mathsf{AllPRat}_{-i})}} \mathsf{RBel}_i(\chi_{-i}, \mathsf{AllPRat}_{-i}) \right)$$

[14] Börgers' characterization uses the concept of approximate common knowledge by Monderer & Samet [39] instead of common belief. (Roughly speaking φ is approximate common knowledge if and only if, everybody assigns high probability to φ, everybody assigns high probability to the fact that everybody assigns high probability to φ, and so on).

$\mathsf{AllRBelPRat}_J$ has to be read "every player in J has a robust belief that all other players are perfectly rational relative to any compatible information about their choices". It is worth noting that the hypothesis $\mathsf{CBel}_{Agt}(\mathsf{AllPRat}_{Agt} \wedge \mathsf{AllRBelPRat}_{Agt})$ of Theorem 5 requires that, for every player $i \in Agt$ and for every strategy s_{-i} of the other players, if the strategy s_{-i} is admissible (i.e., it is not weakly dominated) then i envisages a world in which the other agents play the strategy s_{-i} and they are all perfectly rational. In particular, we have the following validity:[15]

$$\models \mathsf{CBel}_{Agt}(\mathsf{AllPRat}_{Agt} \wedge \mathsf{AllRBelPRat}_{Agt}) \rightarrow \bigwedge_{i \in Agt} \bigwedge_{s_{-i} \in S_{-i} : s_{-i} \in S_{-i,1}^{DWDS^2 - IDSDS}} \langle i \rangle (pl_{-i}(s_{-i}) \wedge \mathsf{AllPRat}_{-i})$$

Therefore, the model in which the formula $\mathsf{CBel}_{Agt}(\mathsf{AllPRat}_{Agt} \wedge \mathsf{AllRBelPRat}_{Agt})$ is satisfied must be 'sufficiently rich', as for every player i and for every admissible strategy of the other players there must be a world envisaged by i in which this strategy is played and the other players are perfectly rational. This richness condition, which is called by Brandenburger et al. [20, 23] the 'completeness assumption', has been explicitly spelled out in [42, Definition 7.11, page 304].[16]

The preceding Theorem 5 can be generalized to n-iteration of deletion of weakly dominated strategies. But before generalizing Theorem 5, we need to introduce the concept of k-order robust belief about perfect rationality. For the case $k = 1$, we define:

$$\mathsf{AllPRatRBelPRat}_{J,1} \stackrel{\text{def}}{=} \mathsf{AllPRat}_J \wedge \mathsf{AllRBelPRat}_J$$

and for all $k > 1$:

$$\mathsf{AllPRatRBelPRat}_{J,k} \stackrel{\text{def}}{=} \mathsf{AllPRat}_J \wedge \bigwedge_{i \in J} \left(\bigwedge_{\substack{\chi_{-i} \in Beh_{-i}: \\ Comp(\chi_{-i}, \mathsf{AllPRatRBelPRat}_{-i,k-1})}} \mathsf{RBel}_i(\chi_{-i}, \mathsf{AllPRatRBelPRat}_{-i,k-1}) \right)$$

[15] This validity can be proved by using Proposition 2 given in the technical annex at the end of the paper (Section A.3).

[16] It is worth noting that the completeness assumption together with the fact that there is common belief that every player is perfectly rational and has a robust belief à la Baltag & Smets [10] (see Section 4.1) about the perfect rationality of the other players are also sufficient conditions for two iterations of DWDS followed by IDSDS. Indeed, in a way similar to the proof of Theorem 5, one can prove the following validity for all $s \notin S^{DWDS^2 - IDSDS}$:

$$\models \left(\mathsf{ComplAss} \wedge \mathsf{CBel}_{Agt} \left(\mathsf{AllPRat}_{Agt} \wedge \bigwedge_{i \in Agt} \mathsf{RBel}_i^{B\&S} \mathsf{AllPRat}_{-i} \right) \right) \rightarrow \neg pl(s)$$

with

$$\mathsf{ComplAss} \stackrel{\text{def}}{=} \bigwedge_{i \in Agt} \bigwedge_{s_{-i} \in S_{-i} : s_{-i} \in S_{-i,1}^{DWDS^2 - IDSDS}} \langle i \rangle (pl_{-i}(s_{-i}) \wedge \mathsf{AllPRat}_{-i}).$$

$\mathsf{AllPRatRBelPRat}_{J,k}$ has to be read "every player in J is perfectly rational and has a k-order robust belief that all other players are perfectly rational relative to any compatible information about their choices".

According to the following Theorem 6, if the players have common belief that every player is perfectly rational and has a k-order robust belief that all other players are perfectly rational relative to any compatible information about their choices, then the strategy profile which is played must survive $k+1$ iterations of DWDS followed by IDSDS.

Theorem 6 *Let $s \notin S^{DWDS^{k+1}-IDSDS}$ and $k > 0$. Then:*

$$\models \mathsf{CBel}_{Agt}\mathsf{AllPRatRBelPRat}_{Agt,k} \rightarrow \neg pl(s)$$

Note that, when $k = 1$, Theorem 6 and Theorem 5 coincide.

Before concluding, note that, because every $\mathcal{A}_i$ is a total function, the epistemic conditions given in the antecedents of Theorems 3, 4, 5 and 6 are respectively sufficient conditions of the equilibria $IDSDS$, $DWDS^1 - IDSDS$, $DWDS^2 - IDSDS$ and $DWDS^n - IDSDS$. In particular,

$$\models \mathsf{CBel}_{Agt}\mathsf{AllWRat}_{Agt} \rightarrow \bigvee_{s \in S^{IDSDS}} pl(s)$$

$$\models \mathsf{CBel}_{Agt}\mathsf{AllPRat}_{Agt} \rightarrow \bigvee_{s \in S^{DWDS^1-IDSDS}} pl(s)$$

$$\models \mathsf{CBel}_{Agt}(\mathsf{AllPRat}_{Agt} \wedge \mathsf{AllRBelPRat}_{Agt}) \rightarrow \bigvee_{s \in S^{DWDS^2-IDSDS}} pl(s)$$

$$\models \mathsf{CBel}_{Agt}\mathsf{AllPRatRBelPRat}_{Agt,k} \rightarrow \bigvee_{s \in S^{DWDS^{k+1}-IDSDS}} pl(s) \text{ for } k > 0$$

7 Related Work

As pointed out in the introduction, the main difference between the present approach and alternative epistemic characterizations of iterated weak dominance is that we use a *semi-qualitative* approach to uncertainty based on the notion of plausibility introduced by Spohn [45], whereas existing epistemic analysis of iterated weak dominance are based on a *quantitative* representation of uncertainty in terms of probabilities. In this sense, the representation of uncertainty used in this paper is relatively more simple than the representation of uncertainty used in other approaches. For instance, in [47] Stalnaker presents a result similar to the preceding Theorem 4 whereas in [48] he discusses a result similar to the preceding Theorem 5. Differently from the present approach, Stalnaker uses rich semantic structures combining probability measures over possible worlds, representing the uncertainty of players, with plausibility orderings over epistemic alternatives, in order to model belief revision policies.

Brandenburger et al. [23] (see also [20]) provide an epistemic characterization of iterated admissibility where uncertainty is represented using lexicographic probability systems (LPSs). An LPS assigns to every player a finite sequence of probability measures $(p_1, \ldots, p_n)$ with non-overlapping supports. The probability p_1 corresponds to a player's initial hypothesis about the behavior of the others, p_2 corresponds to the player's secondary hypothesis, and so on. The interpretation given to this sequence of probability measure is that for any $1 \leq k < n$ the hypothesis at level k is *infinitely more likely* than the hypothesis at level $k + 1$. In their analysis of the epistemic conditions of iterated weak dominance based on LPSs, Brandeburger et al. define a concept of 'assumption' that is similar to the concept of robust belief used here and in Stalnaker's analysis. The idea is that a given player assumes that an event (or state of affairs) φ is true if and only if, according to the player, φ is infinitely more likeky than $\neg\varphi$.

In their logical characterization of iterated admissibility based on the concept of "all the agents know" [31], Halpern & Pass [32] consider probability structures of the form $\langle \Omega, \mathbf{s}, \mathcal{F}, \mathcal{PR}_1, \ldots, \mathcal{PR}_n \rangle$, where Γ is a set of states, $\mathbf{s}$ is a function associating each state in Ω to a strategy profile of a given game Γ, $\mathcal{F}$ is a Ω-algebra over Ω, and for each player i in the game Γ, $\mathcal{PR}_i$ associates with each state ω in Ω a probability distribution $\mathcal{PR}_i(\omega)$ on $(\Omega, \mathcal{F})$.

Battigalli & Siniscalchi [12] analyze the epistemic conditions of the concept of forward induction, that has been shown to be tightly related to the concept of iterated admissibility. Their analysis is based on conditional probability systems, a generalization of classical Bayesian probabilities that allow them to model the update and/or the revision of the players' beliefs, in the course of an extensive game. The fundamental concept of Battigalli & Siniscalchi's analysis is "strong belief" that is tightly related to our and Stalnaker's concept of "robust belief" and to Brandeburger et al.'s concept of "assumption". (See [3] for a comparison between these three concepts).

A work that is similar in spirit to our approach is Baltag et al.'s analysis of the epistemic conditions of backward induction based on a purely qualitative notion of plausibility [11]. There are two main similarities between our approach and theirs. First of all, we share with them the idea of analyzing the epistemic conditions of solution concepts by using a relatively simpler representation of uncertainty based either on a purely qualitative approach or on a semi-qualitative one. Secondly, although Baltag et al.'s analysis and our analysis are focused on two different solution concepts, they employ a similar conceptual apparatus. For instance, Baltag et al.'s characterization of the epistemic conditions of backward induction employs a concept of "robust belief" that, as shown in Section 4.1, is closely connected to our concept of "relative robust belief".[17]

[17] Baltag et al.'s analysis too is largely inspired by Stalnaker [48].

8 Conclusive Remarks

In this paper we have developed a logical analysis of the epistemic conditions of iterated weak dominance in a semi-qualitative framework based on Spohn's theory of uncertainty and belief change. One might wonder whether the same kind of analysis could be made by using purely qualitative structures $M = \langle W, \{\mathcal{E}_i : i \in Agt\}, \{\preceq_i : i \in Agt\}, \mathcal{V}\rangle$ which result from replacing the plausibility grading function κ with a family of total preorders $\preceq_i$ over possible worlds, where $w \preceq_i v$ means that v is for agent i at least as plausible as w, and $w \equiv_i v$ and $w \prec_i v$ mean respectively that ($w \preceq_i v$ and $v \preceq_i w$) and ($w \preceq_i v$ and $v \not\preceq_i w$). Let us consider this issue in more detail. Our analysis is mainly based on the notion of perfect rationality, as defined in Section 6.4, which is based on the notion of graded belief when $\mathsf{max} > 1$.[18]

In order to define the concept of graded belief, it is necessary to rank the possible worlds that an agent envisages according to their degree of exceptionality (or plausibility) so that we can identify the set of worlds of rank 0, the set of worlds of rank 1, the set of worlds of rank 2, and so on. The total preorder $\preceq_i$ over possible worlds would be sufficient to make such a kind of ranking, as from a total order over a set of elements we can build a corresponding ranking over the elements in that set. Specifically, for all $i \in Agt$ we could define:

$$Rank_i^0 = \{v \in W : \nexists u \in \mathcal{E}_i(v) \text{ such that } v \prec_i u\}$$

and for all $\mathsf{h} \geq 1$:

$$Rank_i^{\mathsf{h}} = \{v \in W \setminus \bigcup\nolimits_{\mathsf{k}<\mathsf{h}} Rank_i^{\mathsf{k}} : \nexists u \in \mathcal{E}_i(v) \setminus \bigcup\nolimits_{\mathsf{k}<\mathsf{h}} Rank_i^{\mathsf{k}} \text{ such that } v \prec_i u\}.$$

Given the preceding ranking over possible worlds, the plausibility grading function κ is definable as follows: for all $i \in Agt$ and for all $w \in W$, $\kappa(w, i) = \mathsf{h}$ if and only if $w \in Rank_i^{\mathsf{h}}$.

Thus, while in our semi-qualitative approach the plausibility ranking is directly given by the function κ in the definition of a PDL-A model, in a purely qualitative approach, such as the one presented in [9, 51], it would be *induced* by the plausibility ordering $\preceq_i$ over the set of possible worlds.

Although from a semantic point of view, it seems clear that the same kind of analysis could be made after replacing the plausibility grading function κ with a family of total preorders $\preceq_i$, it is not clear at all what the resulting logic would look like. More generally, it is not clear how to build a decidable logic with a complete axiomatization which is interpreted by means of purely qualitative structures of the form $M = \langle W, \{\mathcal{E}_i : i \in Agt\}, \{\preceq_i : i \in Agt\}, \mathcal{V}\rangle$ and which allows us to represent in the object language the epistemic conditions of iterated weak dominance, namely

[18] Note that, when $\mathsf{max} = 1$, the notion of perfectly rationality as defined in Section 6.4 and, consequently, our analysis of the sufficient condition for iterated weak dominance only require the operators of knowledge $[i]$ and belief Bel_i. However, this does not mean that the graded belief operator $\mathsf{Bel}_i^{\mathsf{h}}$ is useless in general. It only means that it becomes unnecessary in the binary case, i.e., when it is assumed that agents rank possible worlds according to a two-value scale $Num = \{0, 1\}$ for degrees of belief.

the concept common belief, the concept of robust belief about perfect rationality and the concept of graded belief on which the definition of perfect rationality is based. However, another important reason for choosing a semi-qualitative approach to uncertainty rather than a purely qualitative one is that the graded belief operator $\mathsf{Bel}_i^{\mathsf{h}}\varphi$, on which our analysis of the epistemic foundation of iterated weak dominance is based, is traditionally interpreted by means of the plausibility grading function κ (see the seminal work by Spohn [45] and also [5, 35, 54]). It would be a non-standard and unnatural choice to intrepret it via the total preorder $\preceq_i$.

Another issue we intend to study in future research is a generalization of the approach to belief change presented in Section 3.2. Due to space restrictions, we only considered in this work an operation of belief change based on Spohn's concept of belief conditioning (Definition 4). We believe that our approach is flexible enough to allow us to model or at least to approximate other kinds of belief revision operation such as, e.g., the concepts of lexicographic upgrade and conservative upgrade in the sense of [51].

Appendix: Some Proofs

A.1 Proof of Theorem 1

Satisfiability in PDL-A is decidable.

Proof First of all note that the logic PDL-A^{--} is nothing but the variant of PDL where each atomic knowledge program i is interpreted by an equivalence relation $\mathcal{E}_i$. This logic can be embedded into PDL extended with converse, by simulating every atomic knowledge programs i with a composite program $(a \cup a^{-1})^*$, where a is an arbitrary atomic program interpreted by a binary relation $\mathcal{R}_a$ (not necessarily an equivalence relation!) and a^{-1} is the converse of a. PDL with converse is decidable [33]. It follows that PDL-A^{--} is decidable too.

Moreover, note that the problem of satisfiability in PDL-A^{-} is reducible to the problem of *global* logical consequence in PDL-A^{--}, where the special atoms $\mathsf{exc}_{i,\mathsf{h}}$ are just elements of the set of propositional variables *Prop* and the set of global axioms Γ is the set of all formulas of Proposition 13. That is, we have $\models_{\text{PDL-A}^{-}} \varphi$ if and only if $\Gamma \models_{\text{PDL-A}^{--}} \varphi$. Observe that Γ is finite (because *Num* is finite). It is a routine task to verify that the problem of global logical consequence in PDL-A^{--} with a finite number of global axioms is reducible to the problem of satisfiability in PDL-A^{--}. In particular, if $\Gamma = \{\chi_1, \ldots, \chi_n\}$, we have $\Gamma \models_{\text{PDL-A}^{--}} \varphi$ if and only if $\models_{\text{PDL-A}^{--}} \mathsf{CK}_{Agt}(\chi_1 \wedge \ldots \wedge \chi_n) \rightarrow \varphi$, where CK_{Agt} is the common knowledge operator defined in Section 4.2. As the problem of satisfiability checking in PDL-A^{--} is decidable, it follows that the problem of satisfiability checking in the logic PDL-A^{-} is decidable too.

Proposition 15 ensures that *red* provides an effective procedure for reducing a PDL-A formula φ into an equivalent PDL-A^{-} formula $red(\varphi)$. As PDL-A^{-} is decidable, PDL-A is decidable too. □

A.2 Proof of Theorem 4

Let $s \notin S^{DWDS^1-IDSDS}$. Then:

$$\models \mathsf{CBel}_{Agt}\mathsf{AllPRat}_{Agt} \rightarrow \neg pl(s)$$

Proof The proof is by induction.

Base case For all $s \notin S_1^{DWDS^1-IDSDS}$ we prove that:

(A1) $\models \mathsf{CBel}_{Agt}\mathsf{AllPRat}_{Agt} \rightarrow \neg pl(s)$

To prove (A1), it is sufficient to prove the following validity (B1), as $\mathsf{CBel}_{Agt}\mathsf{AllPRat}_{Agt} \rightarrow \mathsf{AllPRat}_{Agt}$ is valid by the item (10f) in Proposition 17.[19] For all $s \notin S_1^{DWDS^1-IDSDS}$ we have that:

(B1) $\models \mathsf{AllPRat}_{Agt} \rightarrow \neg pl(s)$

And to prove (B1), it is sufficient to prove that if $s[i] \notin S_{i,1}^{DWDS^1-IDSDS}$ then:

(C1) $\models \mathsf{PRat}_i \rightarrow \neg pl_i(s[i])$

Let us prove (C1) by reductio ad absurdum. We assume that $s[i] \notin S_{i,1}^{DWDS^1-IDSDS}$ and $M, w \models \mathsf{PRat}_i$ and $M, w \models pl_i(s[i])$ for some arbitrary model M and world w in M. We are going to show that these three facts are inconsistent. $s[i] \notin S_{i,1}^{DWDS^1-IDSDS}$ implies that:

(D1) there is $b_i \in S_i$ such that: (1) for all $s'_{-i} \in S_{-i}$ we have $\langle s[i], s'_{-i}\rangle \leq_i \langle b_i, s'_{-i}\rangle$ and (2) there is $s''_{-i} \in S_{-i}$ such that $\langle s[i], s''_{-i}\rangle <_i \langle b_i, s''_{-i}\rangle$.

$M, w \models \mathsf{PRat}_i$ and $M, w \models pl_i(s[i])$ together imply:

(E1) $M, w \models \mathsf{PRat}_i(s[i])$.

By the Constraint (**Constr3**), (D1) implies that:

(F1) there is $b_i \in S_i$ such that: (1) for all $s'_{-i} \in S_{-i}$ we have $\langle s[i], s'_{-i}\rangle \leq_i \langle b_i, s'_{-i}\rangle$ and (2) there are $s''_{-i} \in S_{-i}$ and $u \in W$ and $\mathsf{h} \in Num$ such that $(w, u) \in \mathcal{E}_i$ and $\mathcal{A}_{-i}(u) = s''_{-i}$ and $\kappa(u, i) = \mathsf{h}$ and $\langle s[i], s'_{-i}\rangle <_i \langle b_i, s''_{-i}\rangle$.

(F1) implies that:

(G1) $M, w \not\models \mathsf{PRat}_i(s[i])$.

But (G1) and (E1) are in contradiction.

Inductive case For $m > 1$, we assume that if $s \notin S_m^{DWDS^1-IDSDS}$ then:

(Inductive Hypothesis) $\models \mathsf{CBel}_{Agt}\mathsf{AllPRat}_{Agt} \rightarrow \neg pl(s)$

[19] Indeed, $\mathsf{CBel}_{Agt}\mathsf{AllPRat}_{Agt}$ implies $\bigwedge_{i\in Agt} \mathsf{Bel}_i\mathsf{PRat}_i$ which in turn implies $\bigwedge_{i\in Agt}\langle i\rangle\mathsf{PRat}_i$. The latter implies $\bigwedge_{i\in Agt}\mathsf{PRat}_i$ (by the validity (10f) in Proposition 17).

We are going to prove that if $s \notin S_{m+1}^{DWDS^1-IDSDS}$ then:

(A2) $\models \mathsf{CBel}_{Agt}\mathsf{AllPRat}_{Agt} \rightarrow \neg pl(s)$

Let us take an arbitrary model M and world w and assume that $M, w \models \mathsf{CBel}_{Agt}\mathsf{AllPRat}_{Agt}$ and $M, w \models pl(s)$. We are going to show that $s \in S_{m+1}^{DWDS^1-IDSDS}$.

From $M, w \models \mathsf{CBel}_{Agt}\mathsf{AllPRat}_{Agt}$, by the validity (10f) in Proposition 17, it follows that:

(B2) $M, w \models \mathsf{AllPRat}_{Agt}$

By the validity (9b) in Proposition 16, (B2) implies that:

(C2) $M, w \models \mathsf{AllSRat}_{Agt}$

Moreover we have the following validity by the property $\models \mathsf{CBel}_{Agt}\varphi \rightarrow \mathsf{Bel}_i\mathsf{CBel}_{Agt}\varphi$ for every $i \in Agt$:

(D2) $\models \mathsf{CBel}_{Agt}\mathsf{AllPRat}_{Agt} \rightarrow \bigwedge_{i\in Agt} \mathsf{Bel}_i\mathsf{CBel}_{Agt}\mathsf{AllPRat}_{Agt}$

Therefore, from $M, w \models \mathsf{CBel}_{Agt}\mathsf{AllPRat}_{Agt}$ we infer that:

(E2) $M, w \models \bigwedge_{i\in Agt} \mathsf{Bel}_i\mathsf{CBel}_{Agt}\mathsf{AllPRat}_{Agt}$

By the inductive hypothesis, Axiom K and the rule of necessitation for the belief operator Bel_i, from (E2) it follows that if $s' \notin S_m^{DWDS^1-IDSDS}$ then:

(F2) $M, w \models \bigwedge_{i\in Agt} \mathsf{Bel}_i\neg pl(s')$

From (F2), (C2) and $M, w \models pl(s)$ it follows that for every $i \in Agt$ and for all $b_i \in S_i$ either there is $s' \in S_m^{DWDS^1-IDSDS}$ such that $\langle b_i, s'_{-i}\rangle <_i \langle s[i], s'_{-i}\rangle$ or for all $s' \in S_m^{DWDS^1-IDSDS}$ we have $\langle b_i, s'_{-i}\rangle \leq_i \langle s[i], s'_{-i}\rangle$. The latter implies that for every $i \in Agt$ we have $s[i] \in S_{i,m+1}^{DWDS^1-IDSDS}$ which is equivalent to $s \in S_{m+1}^{DWDS^1-IDSDS}$. □

A.3 Proof of Theorem 5

Let $s \notin S^{DWDS^2-IDSDS}$. Then:

$$\models \mathsf{CBel}_{Agt}(\mathsf{AllPRat}_{Agt} \wedge \mathsf{AllRBelPRat}_{Agt}) \rightarrow \neg pl(s)$$

Proof The proof is by induction. The proof of the inductive case goes exactly as the proof of the inductive case in the proof of Theorem 4.

Here we only prove the base case.

Base case For all $s \notin S_2^{DWDS^2-IDSDS}$ we prove that:

(A) $\models \mathsf{CBel}_{Agt}(\mathsf{AllPRat}_{Agt} \wedge \mathsf{AllRBelPRat}_{Agt}) \rightarrow \neg pl(s)$

To prove (A), it is sufficient to prove the following validity (B), as $\mathsf{CBel}_{Agt}(\mathsf{AllPRat}_{Agt} \wedge \mathsf{AllRBelPRat}_{Agt}) \rightarrow (\mathsf{AllPRat}_{Agt} \wedge \mathsf{AllRBelPRat}_{Agt})$ is valid.[20]

For all $s \notin S_2^{DWDS^2-IDSDS}$ we have that:

(B) $\quad \models (\mathsf{AllPRat}_{Agt} \wedge \mathsf{AllRBelPRat}_{Agt}) \rightarrow \neg pl(s)$

And to prove (B), it is sufficient to prove that if $s[i] \notin S_{i,2}^{DWDS^2-IDSDS}$ then:

(C) $\quad \models (\mathsf{PRat}_i \wedge \bigwedge_{\chi_{-i} \in Beh_{-i}:Comp(\chi_{-i},\mathsf{AllPRat}_{-i})} \mathsf{RBel}_i(\chi_{-i}, \mathsf{AllPRat}_{-i})) \rightarrow \neg pl_i(s[i])$

Let us prove (C) by reductio ad absurdum. We assume that $s[i] \notin S_{i,2}^{DWDS^2-IDSDS}$ and $M, w \models \mathsf{PRat}_i$ and $M, w \models \bigwedge_{\chi_{-i} \in Beh_{-i}:Comp(\chi_{-i},\mathsf{AllPRat}_{-i})} \mathsf{RBel}_i(\chi_{-i}, \mathsf{AllPRat}_{-i})$ and $M, w \models pl_i(s[i])$ for some arbitrary PDL-A$^+$ model $M = \langle W, \{\mathcal{E}_i : i \in Agt\}, \kappa, \{\mathcal{A}_i : i \in Agt\}, \mathcal{V}\rangle$ and world w in M. We are going to show that these three facts are inconsistent.

The rest of the proof makes use of the following Lemma 1.

Lemma 1 *Let $M = \langle W, \{\mathcal{E}_i : i \in Agt\}, \kappa, \{\mathcal{A}_i : i \in Agt\}, \mathcal{V}\rangle$ be a* PDL-A$^+$ *model. If $M, w \models \bigwedge_{\chi_{-i} \in Beh_{-i}:Comp(\chi_{-i},\mathsf{AllPRat}_{-i})} \mathsf{RBel}_i(\chi_{-i}, \mathsf{AllPRat}_{-i})$ and $s_{-i} \in S_{-i,1}^{DWDS^2-IDSDS}$ and $s'_{-i} \notin S_{-i,1}^{DWDS^2-IDSDS}$ then $\kappa_{w,i}(pl_{-i}(s_{-i})) < \kappa_{w,i}(pl_{-i}(s'_{-i}))$.* □

Proof In order to prove Lemma 1, we first prove the following Lemma 2.

Lemma 2 *Let $\chi_{-i} = \bigvee_{s_{-i} \in \mathbf{s}_{-i}} pl_{-i}(s_{-i})$ for some $\mathbf{s}_{-i} \subseteq S_{-i}$. Then, $Comp(\chi_{-i}, \mathsf{AllPRat}_{-i})$ if and only if there exists $s_{-i} \in \mathbf{s}_{-i}$ such that $s_{-i} \in S_{-i,1}^{DWDS^2-IDSDS}$.* □

Proof ($\Leftarrow$) We first prove the right-to-left direction of the equivalence, after assuming that the set of strategy profiles is $S = \{s_1, \ldots, s_n\}$ for some $n \in \mathbb{N}$. Suppose that $s_{-i} \in S_{-i,1}^{DWDS^2-IDSDS}$ with $s_{-i} \in \mathbf{s}_{-i}$. We can exhibit the following PDL-A$^+$ model $M^* = \langle W^*, \{\mathcal{E}_i^* : i \in Agt\}, \kappa^*, \{\mathcal{A}_i^* : i \in Agt\}, \mathcal{V}^*\rangle$ where:

- $W^* = \{w_1, \ldots, w_n\}$;
- for all $i \in Agt$, $\mathcal{E}_i^* = \{(w_h, w_{h'}) : w_h, w_{h'} \in W^* \text{ and } s_h[i] = s_{h'}[i]\}$;
- for all $w_h \in W^*$ and for all $i \in Agt$, $\mathcal{A}_i^*(w_h) = s_h[i]$;
- for all $i \in Agt$ and for all $w_h \in W^*$, $\kappa^*(w_h, i) = 0$;
- for all $w_h \in W^*$, $\mathcal{V}^*(w_h) = Prop$.

---

[20] Indeed, $\mathsf{CBel}_{Agt}(\mathsf{AllPRat}_{Agt} \wedge \mathsf{AllRBelPRat}_{Agt})$ implies that $\bigwedge_{i \in Agt} \mathsf{Bel}_i(\mathsf{PRat}_i \wedge \bigwedge_{\chi_{-i} \in Beh_{-i}:Comp(\chi_{-i},\mathsf{AllPRat}_{-i})} \mathsf{RBel}_i(\chi_{-i}, \mathsf{AllPRat}_{-i}))$. The latter implies that $\bigwedge_{i \in Agt} \mathsf{PRat}_i$ (by the validity (10f) in Proposition 17). Moreover, it implies that $\bigwedge_{i \in Agt} \bigwedge_{\chi_{-i} \in Beh_{-i}:Comp(\chi_{-i},\mathsf{AllPRat}_{-i})} \mathsf{RBel}_i(\chi_{-i}, \mathsf{AllPRat}_{-i})$. To see this, just note that for any PDL-A model $M = \langle W, \mathcal{E}, \kappa, \mathcal{V}\rangle$, worlds w, v in M and $\alpha \in Num \setminus \{0\}$ we have that if $v \in \mathcal{B}_i(w)$ then $\mathcal{B}_i^{*_i^\alpha \varphi}(w) = \mathcal{B}_i^{*_i^\alpha \varphi}(v)$. Hence, we have $\models \mathsf{Bel}_i\left[*_i^\alpha \varphi\right]\mathsf{Bel}_i\psi \rightarrow [*_i^\alpha \varphi]\mathsf{Bel}_i\psi$. By the preceding validity and the validity $\models \mathsf{Bel}_i\mathsf{Bel}_i\varphi \rightarrow \mathsf{Bel}_i\varphi$, we have $\models \mathsf{Bel}_i\mathsf{RBel}_i(\varphi, \psi) \rightarrow \mathsf{RBel}_i(\varphi, \psi)$. Finally, $\bigwedge_{i \in Agt}(\mathsf{PRat}_i \wedge \bigwedge_{\chi_{-i} \in Beh_{-i}:Comp(\chi_{-i},\mathsf{AllPRat}_{-i})} \mathsf{RBel}_i(\chi_{-i}, \mathsf{AllPRat}_{-i}))$ is equivalent to $\mathsf{AllPRat}_{Agt} \wedge \mathsf{AllRBelPRat}_{Agt}$.

It is straightforward to verify that $M^*, w^* \models pl_{-i}(s_{-i}) \wedge \mathsf{AllPRat}_{-i}$ where w^* is a world in W^* such that $\mathcal{A}^*_{-i}(w^*) = s_{-i}$. Therefore, model M^* satisfies $pl_{-i}(s_{-i}) \wedge \mathsf{AllPRat}_{-i}$. It follows that M^* satisfies $\chi_{-i} \wedge \mathsf{AllPRat}_{-i}$ too.

($\Rightarrow$) The left-to-right direction of the equivalence can be proved by reductio ad absurdum. We assume that: (1) $s_{-i} \notin S^{DWDS^2-IDSDS}_{-i,1}$ for all $s_{-i} \in \mathbf{s}_{-i}$ and (2) there exists $s'_{-i} \in \mathbf{s}_{-i}$ such that $M, w \models pl_{-i}(s'_{-i}) \wedge \mathsf{AllPRat}_{-i}$ for some $\mathsf{PDL\text{-}A}^+$ model M and world w in M. From the assumption (1), it follows that there is $j \in Agt \setminus \{i\}$ such that $s'_{-i}[j] \notin S^{DWDS^2-IDSDS}_{j,1}$. From the definition of PRat_j, by the Constraint (**Constr3**) over $\mathsf{PDL\text{-}A}^+$ models, we can prove that if $s'_{-i}[j] \notin S^{DWDS^2-IDSDS}_{j,1}$ and $M, w \models pl_j(s'_{-i}[j])$ then $M, w \models \neg\mathsf{PRat}_j$. Hence, from the initial assumptions it follows that $M, w \models \neg\mathsf{PRat}_j$. The latter is in contradiction with the assumption (2). □

Now assume that $s_{-i} \in S^{DWDS^2-IDSDS}_{-i,1}$ and $s'_{-i} \notin S^{DWDS^2-IDSDS}_{-i,1}$. By Lemma 2, it follows that $Comp(pl_{-i}(s_{-i}) \vee pl_{-i}(s'_{-i}), \mathsf{AllPRat}_{-i})$. Moreover, assume that $M, w \models \bigwedge_{\chi_{-i} \in Beh_{-i}: Comp(\chi_{-i}, \mathsf{AllPRat}_{-i})} \mathsf{RBel}_i(\chi_{-i}, \mathsf{AllPRat}_{-i})$. From the latter assumption it follows that $M, w \models \mathsf{RBel}_i(pl_{-i}(s_{-i}) \vee pl_{-i}(s'_{-i})\mathsf{AllPRat}_{-i})$. Hence $M, w \models [*^{\alpha}_i pl_{-i}(s_{-i}) \vee pl_{-i}(s'_{-i})]\mathcal{B}el_i\mathcal{A}ll\mathcal{P}\mathcal{R}at_{-i}$ for any $\alpha \in Num \setminus \{0\}$. By Lemma 2, from the assumption $s'_{-i} \notin S^{DWDS^2-IDSDS}_{-i,1}$ it follows that $pl_{-i}(s'_{-i}) \wedge \neg\mathsf{AllPRat}_{-i}$ is valid. By the Constraint (**Constr3**) over $\mathsf{PDL\text{-}A}^+$ models and the truth condition of the belief revision operator $[*^{\alpha}_i \varphi]$, it follows that $M, w \models [*^{\alpha}_i pl_{-i}(s_{-i}) \vee pl_{-i}(s'_{-i})]\mathcal{B}el\, i(pl_{-i}(s_{-i}) \wedge \neg pl_{-i}(s'_{-i}))$. The latter implies $\kappa_{w,i}(pl_{-i}(s_{-i})) < \kappa_{w,i}(pl_{-i}(s'_{-i}))$. This completes the proof of Lemma 1.

From $M, w \models \bigwedge_{\chi_{-i} \in Beh_{-i}: Comp(\chi_{-i}, \mathsf{AllPRat}_{-i})} \mathsf{RBel}_i(\chi_{-i}, \mathsf{AllPRat}_{-i})$, by Lemma 1, it follows that:

(D) if $s'_{-i} \in S^{DWDS^2-IDSDS}_{-i,1}$ and $s''_{-i} \notin S^{DWDS^2-IDSDS}_{-i,1}$ then $\kappa_{w,i}(pl_{-i}(s'_{-i})) < \kappa_{w,i}(pl_{-i}(s''_{-i}))$.

$s[i] \notin S^{DWDS^2-IDSDS}_{i,2}$ implies that:

(E1) $s[i] \notin S^{DWDS^2-IDSDS}_{i,1}$ or

(E2) $s[i] \in S^{DWDS^2-IDSDS}_{i,1}$ and $s[i] \notin S^{DWDS^2-IDSDS}_{i,2}$

We split the proof in the two subcases: (E1) and (E2).

Proof for the Case (E1) $s[i] \notin S^{DWDS^2-IDSDS}_{i,1}$ implies that:

(F1) there is $b_i \in S^{DWDS^2-IDSDS}_{i}$ such that: (1) $b_i \neq s[i]$ and (2) $\langle s[i], s'_{-i}\rangle <_i \langle b_i, s'_{-i}\rangle$ for some $s'_{-i} \in S^{DWDS^2-IDSDS}_{-i}$ and (3) $\langle s[i], s''_{-i}\rangle \leq_i \langle b_i, s''_{-i}\rangle$ for all $s''_{-i} \in S^{DWDS^2-IDSDS}_{-i}$.

From (F1) by the Constraint (**Constr3**) it follows that:

(G1) there are $b_i \in S_i^{DWDS^2-IDSDS}$ and $s'_{-i} \in S_{-i}^{DWDS^2-IDSDS}$ and $v \in W$ such that: (1) $b_i \neq s[i]$ and (2) $(w, v) \in \mathcal{E}_i$ and (3) $M, v \models pl_{-i}(s'_{-i})$ and (4) $\langle s[i], s'_{-i}\rangle <_i \langle b_i, s'_{-i}\rangle$ and (5) for all $u \in W$ such that $(w, u) \in \mathcal{E}_i$ and for all $s''_{-i} \in S_{-i}^{DWDS^2-IDSDS}$: if $M, u \models pl_{-i}(s''_{-i})$ then $\langle s[i], s''_{-i}\rangle \leq_i \langle b_i, s''_{-i}\rangle$.

But (G1) is in contradiction with $M, w \models \mathsf{PRat}_i$ and $M, w \models pl_i(s[i])$.

Proof for the case (E2) (E2) implies that:

(F2) there is $b_i \in S_{i,1}^{DWDS^2-IDSDS}$ such that: (1) $b_i \neq s[i]$ and (2) $\langle s[i], s'_{-i}\rangle <_i \langle b_i, s'_{-i}\rangle$ for some $s'_{-i} \in S_{-i,1}^{DWDS^2-IDSDS}$ and (3) $\langle s[i], s'_{-i}\rangle \leq_i \langle b_i, s''_{-i}\rangle$ for all $s''_{-i} \in S_{-i,1}^{DWDS^2-IDSDS}$.

By the Constraint (**Constr3**), (F2) together with $M, w \models \mathsf{PRat}_i$ and $M, w \models pl_i(s[i])$ imply that:

(G2) there are $s'_{-i} \in S_{-i,1}^{DWDS^2-IDSDS}$ and $s''_{-i} \notin S_{-i,1}^{DWDS^2-IDSDS}$ such that $\kappa_{w,i}(pl_{-i}(s'_{-i})) \leq \kappa_{w,i}(pl_{-i}(s''_{-i}))$.

But (G2) is in contradiction with (D).

A.4 Proof of Theorem 6

Let $s \notin S^{DWDS^{k+1}-IDSDS}$ and $k > 0$. Then:

$$\models \mathsf{CBel}_{Agt}\mathsf{AllPRatRBelPRat}_{Agt,k} \rightarrow \neg pl(s)$$

Proof The proof is by induction. The proof of the inductive case goes exactly as the proof of the inductive case in the proof of Theorem 4.

Here we only prove the base case.

Base case For all $s \notin S_{k+1}^{DWDS^{k+1}-IDSDS}$ we prove that:

(A1) $\models \mathsf{CBel}_{Agt}\mathsf{AllPRatRBelPRat}_{Agt,k} \rightarrow \neg pl(s)$

To prove (A1), it is sufficient to prove the following validity (B1), as $\mathsf{CBel}_{Agt}\mathsf{AllPRatRBelPRat}_{Agt,k} \rightarrow \mathsf{AllPRatRBelPRat}_{Agt,k}$ is valid (the proof of this validity is similar to the one given in the proof of Theorem 5 for the validity $\mathsf{CBel}_{Agt}\mathsf{AllPRat}_{Agt} \rightarrow \mathsf{AllPRat}_{Agt}$).

For all $s \notin S_{k+1}^{DWDS^{k+1}-IDSDS}$ we have that:

(B1) $\models \mathsf{AllPRatRBelPRat}_{Agt,k} \rightarrow \neg pl(s)$

We prove something more general than (B1), namely we prove that for all $J \in 2^{Agt*}$ and for all $s_J \notin S_{J,k+1}^{DWDS^{k+1}-IDSDS}$:

(C1) $\models \mathsf{AllPRatRBelPRat}_{J,k} \rightarrow \neg pl_J(s_J)$

The proof of (C1) is again by induction.

Base case For all $s_J \notin S_{J,2}^{DWDS^2-IDSDS}$ we have to prove that:

(A2) $\models (\mathsf{AllPRat}_J \wedge \mathsf{AllRBelPRat}_J) \rightarrow \neg pl_J(s_J)$

In the proof of Theorem 5 we have proved something stronger than (A2), namely we have proved that if $s[i] \notin S_{i,2}^{DWDS^2-IDSDS}$ then $\models (\mathsf{PRat}_i \wedge \bigwedge_{\chi_{-i} \in Beh_{-i}: Comp(\chi_{-i}, \mathsf{AllPRat}_{-i})} \mathsf{RBel}_i(\chi_{-i}, \mathsf{AllPRat}_{-i})) \rightarrow \neg pl_i(s[i])$.

Inductive case Let m be an integer such that $m > 1$. Let us assume that for all $J \in 2^{Agt*}$, if $s_J \notin S_{J,m}^{DWDS^m-IDSDS}$ then:

(Inductive Hypothesis) $\models \mathsf{AllPRatRBelPRat}_{J,m-1} \rightarrow \neg pl_J(s_J)$

We are going to prove that if $s_J \notin S_{J,m+1}^{DWDS^{m+1}-IDSDS}$ then:

(A3) $\models \mathsf{AllPRatRBelPRat}_{J,m} \rightarrow \neg pl_J(s_J)$

The proof is by reduction ad absurdum. We take an arbitrary PDL-A$^+$ model M and a world w in M. We assume that $M, w \models \mathsf{AllPRatRBelPRat}_{J,m}$ and $M, w \models pl_J(s_J)$ and $s_J \notin S_{J,m+1}^{DWDS^{m+1}-IDSDS}$. We are going to show that these three facts are inconsistent.

The rest of the proof is based on the following Lemma 3. □

Lemma 3 *Let m be an integer such that $m > 1$* and let $\chi_{-i} = \bigvee_{s_{-i} \in \mathbf{s}_{-i}} pl_{-i}(s_{-i})$ *for some $\mathbf{s}_{-i} \subseteq S_{-i}$. Then, $Comp(\chi_{-i}, \mathsf{AllPRatRBelPRat}_{-i,m-1})$ if and only if there exists $s_{-i} \in \mathbf{s}_{-i}$ such that $s_{-i} \in S_{-i,m}^{DWDS^m-IDSDS}$.*

Proof of Sketch The proof of Lemma 3 is again by induction. We only prove the base case (i.e., when $m = 2$).

($\Leftarrow$) We first prove the right-to-left direction of the equivalence, after assuming that the set of strategy profiles is $S = \{s_1, \ldots, s_n\}$ for some $n \in \mathbb{N}$. Suppose that $s_{-i} \in S_{-i,2}^{DWDS^2-IDSDS}$ with $s_{-i} \in \mathbf{s}_{-i}$. We can exhibit the following PDL-A$^+$ model $M^* = \langle W^*, \{\mathcal{E}_i^* : i \in Agt\}, \kappa^*, \{\mathcal{A}_i^* : i \in Agt\}, \mathcal{V}^* \rangle$ where:

- $W^* = \{w_1, \ldots, w_n\}$;
- for all $i \in Agt$, $\mathcal{E}_i^* = \{(w_h, w_{h'}) : w_h, w_{h'} \in W^* \text{ and } s_h[i] = s_{h'}[i]\}$;
- for all $w_h \in W^*$ and for all $i \in Agt$, $\mathcal{A}_i^*(w_h) = s_h[i]$;
- for all $i \in Agt$ and for all $w_h \in W^*$:
  1. $\kappa^*(w_h, i) = 0$ if and only if, for all $j \in Agt \setminus \{i\}$, $s_h[j] \in S_{j,1}^{DWDS^2-IDSDS}$,
  2. $\kappa^*(w_h, i) = \mathsf{max}$ if and only if there is $j \in Agt \setminus \{i\}$ such that $s_h[j] \notin S_{j,1}^{DWDS^2-IDSDS}$;
- for all $w_h \in W^*$, $\mathcal{V}^*(w_h) = Prop$.

With the help of Lemma 2 in Section A.3, it is straightforward to verify that $M^*, w^* \models pl_{-i}(s_{-i}) \wedge \mathsf{AllPRat}_{-i} \wedge \bigwedge_{j \in Agt \setminus \{i\}} \Big(\bigwedge_{\substack{\chi_{-j} \in Beh_{-j}: \\ Comp(\chi_{-j}, \mathsf{AllPRat}_{-j})}}$

$\mathsf{RBel}_j(\chi_{-j}, \mathsf{AllPRat}_{-j})\big)$ where w^* is a world in W^* such that $\mathcal{A}^*_{-i}(w^*) = s_{-i}$. Therefore, model M^* satisfies $pl_{-i}(s_{-i}) \wedge \mathsf{AllPRatRBelPRat}_{-i,1}$. It follows that M^* satisfies $\chi_{-i} \wedge \mathsf{AllPRatRBelPRat}_{-i,1}$ too.

($\Rightarrow$) The left-to-right direction of the equivalence can be proved by reductio ad absurdum. We assume that: (1) $s_{-i} \notin S^{DWDS^2-IDSDS}_{-i,2}$ for all $s_{-i} \in \mathbf{s}_{-i}$ and (2) there exists $s'_{-i} \in \mathbf{s}_{-i}$ such that $M, w \models pl_{-i}(s'_{-i}) \wedge \mathsf{AllPRat}_{-i} \wedge \bigwedge_{j \in Agt \setminus \{i\}} (\bigwedge_{\substack{\chi_{-j} \in Beh_{-j}: \\ Comp(\chi_{-j}, \mathsf{AllPRat}_{-j})}} \mathsf{RBel}_j(\chi_{-j}, \mathsf{AllPRat}_{-j}))$ for some PDL-A$^+$ model M and world w in M. From the assumption (1), it follows that there is $j \in Agt \setminus \{i\}$ such that $s'_{-i}[j] \notin S^{DWDS^2-IDSDS}_{j,2}$. From the definition of $\mathsf{PRat}_j(s'_{-i}[j])$, by the Constraint (**Constr3**) over PDL-A$^+$ models and with the help of Lemma 1 in Section A.3, we can prove that if $s'_{-i}[j] \notin S^{DWDS^2-IDSDS}_{j,2}$ and $M, w \models \bigwedge_{\substack{\chi_{-j} \in Beh_{-j}: \\ Comp(\chi_{-j}, \mathsf{AllPRat}_{-j})}} \mathsf{RBel}_j(\chi_{-j}, \mathsf{AllPRat}_{-j})$ then $M, w \models \neg \mathsf{PRat}_j(s'_{-i}[j])$. Therefore, from the initial assumption that $M, w \models pl_{-i}(s'_{-i})$ it follows that $M, w \models \neg \mathsf{PRat}_j$. The latter is in contradiction with the assumption (2), namely with $M, w \models \mathsf{AllPRat}_{-i}$.

$M, w \models \mathsf{AllPRatRBelPRat}_{J,m}$ is equivalent to:

(B3) $M, w \models \mathsf{AllPRat}_J \wedge \bigwedge_{i \in J} \bigwedge_{\chi_{-i} \in Beh_{-i}: Comp} (\chi_{-i}, \mathsf{AllPRatRBelPRat}_{-i,m-1}) \times \mathsf{RBel}_i(\chi_{-i}, \mathsf{AllPRatRBelPRat}_{-i,m-1})$

By inductive hypothesis, (B3) implies that for all $s''_{-i} \notin S^{DWDS^m-IDSDS}_{-i,m}$:

(C3) $M, w \models \bigwedge_{i \in J} \bigwedge_{\chi_{-i} \in Beh_{-i}: Comp} \times (\chi_{-i}, \mathsf{AllPRatRBelPRat}_{-i,m-1}) \mathsf{RBel}_i(\chi_{-i} \neg pl_{-i}(s''_{-i}))$

From (C3) and Lemma 3 it follows that for all $i \in J$:

(D3) if $s_{-i}\prime \in S^{DWDS^m-IDSDS}_{-i,m}$ and $s_{-i}\prime\prime \notin S^{DWDS^m-IDSDS}_{-i,m}$ then $\kappa_{w,i}(pl_{-i}(s_{-i}\prime)) < \kappa_{w,i}(pl_{-i}(s_{-i}\prime\prime))$.

(The proof of the preceding item (D3) is similar to the proof of the item (D) in the proof of Theorem 5 in Section A.3).

The rest of the proof proceeds as the proof of Theorem 5 in Section A.3 (starting from item (D)). For this reason, we do not repeat it here. □

References

1. Alchourrón, C.E., Gärdenfors, P., Makinson, D. (1985). On the logic of theory change: partial meet contraction and revision functions. *Journal of Symbolic Logic*, *50*(2), 510–530.
2. Asheim, G., & Dufwenberg, M. (2003). Admissibility and common belief. *Games and Economic Behavior*, *42*, 208–234.
3. Asheim, G., & Søvik, Y. (2005). Preference-based belief operators. *Mathematical Social Sciences*, *50*, 61–82.
4. Asheim, G.B. (2001). Proper rationalizability in lexicographic beliefs. *International Journal of Game Theory*, *30*, 453–478.

5. Aucher, G. (2005). A combined system for upyear logic and belief revision. In *Proceedings of PRIMA 2004 of LNAI* (Vol. 3371, pp. 1–18). Springer-Verlag.
6. Aumann, R. (1999). Interactive epistemology, I. knowledge. *International Journal of Game Theory*, *28*(3), 263–300.
7. Baltag, A., Moss, L., Solecki, S. (1998). The logic of public announcements, common knowledge and private suspicions. In *Proceedings of TARK'98* (pp. 43–56). Morgan Kaufmann.
8. Baltag, A., & Moss, L.S. (2004). Logics for epistemic programs. *Synthese*, *139*(2), 165–224.
9. Baltag, A., & Smets, S. (2008). A qualitative theory of dynamic interactive belief revision. In *Proceedings of LOFT 7, texts in logic and games* (Vol. 3, pp. 13–60). Amsterdam University Press.
10. Baltag, A., & Smets, S. (2009). Talking your way into agreement: belief merge by persuasive communication. In *Proceedings of the second multi-agent logics, languages, and organisations federated workshops (MALLOW), CEUR workshop proceedings* (Vol. 494). CEUR-WS.org.
11. Baltag, A., Smets, S., Zvesper, J.A. (2009). Keep 'hoping' for rationality: a solution to the backward induction paradox. *Synthese*, *169*(2), 301–333.
12. Battigalli, P., & Siniscalchi, M. (2002). Strong belief and forward induction reasoning. *Journal of Economic Theory*, *106*(2), 356–391.
13. Bernheim, D. (1986). Axiomatic characterizations of rational choice in strategic environments. *Scandinavian Journal of Economics*, *88*, 473–488.
14. Blume, L.E., Brandenburger, A., Dekel, E. (1991). Lexicographic probabilities and choice under uncertainty. *Econometrica*, *59*, 61–79.
15. Board, O. (1998). Belief revision and rationalizability. In *Proceedings of TARK'98* (pp. 201–213). Morgan Kaufmann.
16. Board, O. (2002). Knowledge, beliefs, and game-theoretic solution concepts. *Oxford Review of Economic Policy*, *18*, 418–432.
17. Bonanno, G. (2008). A syntactic approach to rationality in games with ordinal payoffs. In *Proceedings of LOFT 2008, texts in logic and games series* (pp. 59–86). Amsterdam University Press.
18. Börgers, T. (1994). Weak dominance and approximate common knowledge. *Journal of Economic Theory*, *64*, 265–276.
19. Börgers, T., & Samuelson, L. (1992). Cautious utility maximization and iterated weak dominance. *International Journal of Game Theory*, *21*, 13–25.
20. Brandeburger, A. (2007). The power of paradox: some recent developments in interactive epistemology. *International Journal of Game Theory*, *35*(4), 465–492.
21. Brandeburger, A., & Dekel, E. (1987). Rationalizability and correlated equilibria. *Econometrica*, *55*, 1391–1402.
22. Brandenburger, A. (1992). Lexicographic probabilities and iterated admissibility. In P. Dasgupta, D. Gale, O. Hart, E. Maskin (Eds.), *Economic analysis of markets and games* (pp. 282–290). MIT Press.
23. Brandenburger, A., Friedenberg, A., Keisler, J. (2008). Admissibility in games. *Econometrica*, *76*, 307–352.
24. Dekel, E., & Fudenberg, D. (1990). Rational behavior with payoff uncertainty. *Journal of Economic Theory*, *52*, 243–267.
25. Dubois, D., & Prade, H. (1998). Possibility theory: qualitative and quantitative aspects. In D. Gabbay & P. Smets (Eds.), *Handbook of defeasible reasoning and uncertainty management systems* (Vol. 1, pp. 169–226). Kluwer.
26. Fagin, R., Halpern, J.Y., Moses, Y., Vardi, M. (1995). *Reasoning about knowledge*. Cambridge: MIT Press.
27. Gärdenfors, P., & Makinson, D. (1988). Revisions of knowledge systems using epistemic entrenchment. In M.Y. Vardi, (Ed.), *Proceedings of TARK 1988* (pp. 83–95). Morgan Kaufmann.
28. Goldszmidt, M., & Pearl, J. (1996). Qualitative probability for default reasoning, belief revision and causal modeling. *Artificial Intelligence*, *84*, 52–112.
29. Grove, A. (1988). Two modellings for theory change. *Journal of Philosophical Logic*, *17*, 157–170.
30. Halpern, J.Y. (2011). Beyond nash equilibrium: solution concepts for the 21st century. In K.R. Apt & E. Gradel (Eds.), *Lectures in game theory for computer scientists* (pp. 264–1278).
31. Halpern, J.Y., & Lakemeyer, G. (2001). Multi-agent only knowing. *Journal of Logic and Computation*, *11*(1), 41–70.

32. Halpern, J.Y., & Pass, R. (2009). A logical characterization of iterated admissibility. In A. Heifetz (Ed.), *Proceedings of TARK 2009* (pp. 146–155).
33. Harel, D., Kozen, D., Tiuryn, J. (2000). *Dynamic logic*: MIT Press.
34. Klein, P.D. (1971). A proposed definition of propositional knowledge. *Journal of Philosophy*, *68*, 471–482.
35. Laverny, N., & Lang, J. (2005). From knowledge-based programs to graded belief-based programs, part II: off-line reasoning. In *Proceedings of IJCAI'05* (pp. 497–502).
36. Lehrer, K., & Paxton, T. (1969). Undefeated justified true belief. *Journal of Philosophy*, *66*, 225–237.
37. Lorini, E., & Schwarzentruber, F. (2010). A modal logic of epistemic games. *Games*, *1*(4), 478–526.
38. Mas-Colell, M., Winston, A., Green J. (1995). *Microeconomic theory*. New York: Oxford University Press.
39. Monderer, D., & Samet, D. (1989). Approximating common knowledge with common beliefs. *Games and Economic Behavior*, *1*, 170–190.
40. Pearce, D. (1984). Rationalizable strategic behavior and the problem of perfection. *Econometrica*, *52*, 1029–1050.
41. Pearl, J. (1993). From conditional oughts to qualitative decision theory. In D. Heckerman & E.H. Mamdani (Eds.), *Proceedings of UAI'93* (pp. 12–22). Morgan Kaufmann.
42. Perea, A. (2012). *Epistemic game theory: reasoning and choice*. Cambridge: Cambridge University Press.
43. Rott, H. (2004). Stability, strength and sensitivity: converting belief into knowledge. *Erkenntnis*, *61*, 469–493.
44. Samuelson, L. (1992). Dominated strategies and common knowledge. *Games and Economic Behavior*, *4*, 284–313.
45. Spohn W. (1988). Ordinal conditional functions: a dynamic theory of epistemic states. In W.L. Harper & B. Skyrms (Eds.), *Causation in decision, belief change and statistics* (pp. 105–134). Kluwer.
46. Stalnaker, R. (1994). On the evaluation of solution concepts. *Theory and Decision*, *37*, 49–73.
47. Stalnaker, R. (1996). Knowledge, belief and counterfactual reasoning in games. *Economics and Philosophy*, *12*, 133–163.
48. Stalnaker, R. (1998). Belief revision in games: forward and backward induction. *Mathematical Social Sciences*, *36*, 31–56.
49. Stalnaker, R. (2006). On logics of knowledge and belief. *Philosophical Studies*, *128*, 169–199.
50. Tan, T., & Werlang, S. (1988). The bayesian foundation of solution concepts of games. *Journal of Economic Theory*, *45*, 370–391.
51. van Benthem, J. (2007). Dynamic logic for belief revision. *Journal of Applied Non-Classical Logics*, *17*, 129–155.
52. van Benthem, J. (2007). Rational dynamics and epistemic logic in games. *International Game Theory Review*, *9*(1), 13–45.
53. van Benthem, J., van Eijck, J., Kooi, B. (2006). Logics of communication and change. *Information and Computation*, *204*(11), 1620–1662.
54. van Ditmarsch, H. (2005). Prolegomena to dynamic logic for belief revision. *Synthese*, *147*(2), 229–275.
55. Weydert, E. (1994). General belief measures. In R.L. de Mántaras & D. Poole (Eds.), *Proceedings of UAI'94* (pp. 575–582). Morgan Kaufmann.
56. Zvesper, J.A. (2010). *Playing with information*. PhD thesis, The Netherlands: University of Amsterdam.