When Do Introspection Axioms Matter for Multi-Agent Epistemic Reasoning?

Yifeng Ding
University of California, Berkeley
yf.ding@berkeley.edu

Wesley H. Holliday
University of California, Berkeley
weshholliday@berkeley.edu

Cedegao Zhang
University of California, Berkeley
cedzhang@berkeley.edu

The early literature on epistemic logic in philosophy focused on reasoning about the knowledge or belief of a single agent, especially on controversies about "introspection axioms" such as the 4 and 5 axioms. By contrast, the later literature on epistemic logic in computer science and game theory has focused on multi-agent epistemic reasoning, with the single-agent 4 and 5 axioms largely taken for granted. In the relevant multi-agent scenarios, it is often important to reason about *what agent A believes about what agent B believes about what agent A believes*; but it is rarely important to reason just about *what agent A believes about what agent A believes*. This raises the question of the extent to which single-agent introspection axioms actually matter for multi-agent epistemic reasoning. In this paper, we formalize and answer this question. To formalize the question, we first define a set of multi-agent formulas that we call *agent-alternating formulas*, including formulas like $\Box_a\Box_b\Box_a p$ but not formulas like $\Box_a\Box_a p$. We then prove, for the case of belief, that if one starts with multi-agent K or KD, then adding both the 4 and 5 axioms (or adding the B axiom) does not allow the derivation of any new agent-alternating formulas—in this sense, introspection axioms do not matter. By contrast, we show that such conservativity results fail for knowledge and multi-agent KT, though they hold with respect to a smaller class of *agent-nonrepeating formulas*.

1 Introduction

The classic early works on epistemic logic in philosophy by Hintikka [13] and Lenzen [19] focused on the logic of knowledge and belief for a single agent,[1] especially on controversies about "introspection axioms": for example, if an agent knows p, does she know that she knows p (formalized by the 4 axiom of modal logic, $K_a p \to K_a K_a p$)? If an agent does not know p, does she know that she does not know p (formalized by the 5 axiom of modal logic, $\neg K_a p \to K_a \neg K_a p$)? By contrast, the later literature on epistemic logic in computer science (e.g., [21, 5]) and game theory (e.g., [2]) focused on *multi-agent* epistemic reasoning, especially as required for coordination between agents or strategic reasoning against opponents. In this literature, the single-agent introspection principles formalized by the 4 and 5 axioms are largely taken for granted (for exceptions, see, e.g., [25, 17, 15]). In the relevant multi-agent scenarios, it is often important to reason about *what agent A believes about what agent B believes about what agent A believes* ($B_a B_b B_a p$); but it is rarely important to reason just about *what agent A believes about what agent A believes* ($B_a B_a p$). Consider the following famous examples of multi-agent epistemic reasoning.

Muddy children We assume familiarity with the 3-agent Muddy Children puzzle where two children have mud on their foreheads (see, e.g., § 1.1 of [5]). The following is a derivation in the bimodal version of the minimal normal modal logic K showing how one of the muddy children comes to realize that

[1] Only §§ 4.1-4.6 and § 4.13 of [13] and pp. 59, 66, and 70 of [19] contain discussion of multi-agent formulas.

L.S. Moss (Ed.): TARK 2019
EPTCS 297, 2019, pp. 121–139, doi:10.4204/EPTCS.297.9

she is muddy.[2] Note that (i) no introspection axioms are used, and in fact (ii) modalities occur only "alternatingly," in the sense that no occurrence of a modality for an agent i has scope over another occurrence of a modality for i without an intervening occurrence of some modality for an agent $j \neq i$.

(a) $\Box_1\Box_2((\neg m_1 \wedge \neg m_3) \rightarrow m_2)$ (assumption: 1 knows that 2 knows that at least one child is muddy)

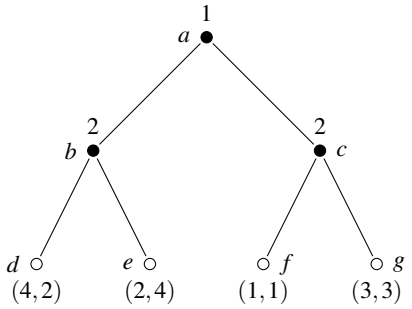
(b) $\Box_1\Box_2\neg m_3$ (assumption: 1 knows that 2 can see 3, who is not muddy)

(c) $\Box_1(\neg m_1 \rightarrow \Box_2\neg m_1)$ (assumption: 1 knows that 2 can see 1)

(d) $\Box_1\neg\Box_2 m_2$ (assumption: 1 knows that 2 did not step forward after the parent's first question)

1. $\Box_2((\neg m_1 \wedge \neg m_3) \rightarrow m_2) \rightarrow (\Box_2(\neg m_1 \wedge \neg m_3) \rightarrow \Box_2 m_2)$ (K axiom)
2. $\Box_1\Box_2((\neg m_1 \wedge \neg m_3) \rightarrow m_2) \rightarrow \Box_1(\Box_2(\neg m_1 \wedge \neg m_3) \rightarrow \Box_2 m_2)$ (from (1) by RM)
3. $\Box_1(\Box_2(\neg m_1 \wedge \neg m_3) \rightarrow \Box_2 m_2)$ (from (a) and (2) by PL)
4. $\Box_1(\neg\Box_2 m_2 \rightarrow \neg\Box_2(\neg m_1 \wedge \neg m_3))$ (from (3) using PL and RM)
5. $\Box_1\neg\Box_2(\neg m_1 \wedge \neg m_3)$ (from (d) and (4) by K and PL)
6. $\neg\Box_2(\neg m_1 \wedge \neg m_3) \rightarrow \neg(\Box_2\neg m_1 \wedge \Box_2\neg m_3)$ (theorem of K)
7. $\Box_1\neg\Box_2(\neg m_1 \wedge \neg m_3) \rightarrow \Box_1\neg(\Box_2\neg m_1 \wedge \Box_2\neg m_3)$ (from (6) by RM)
8. $\Box_1\neg(\Box_2\neg m_1 \wedge \Box_2\neg m_3)$ (from (5) and (7) by PL)
9. $\Box_1\neg\Box_2\neg m_1$ (from (b) and (8) using PL, Nec, and K)
10. $\Box_1 m_1$ (from (c) and (9) using PL, Nec, and K)

Backward induction We assume familiarity with the classic backward induction reasoning in extensive form games (see, e.g., [22, § 6.2]). In [26], Vilks provides a syntactical derivation of backwards induction in the bimodal version of the modal logic KT, which we reproduce below. Again note that (i) no introspection axioms are used, and in fact (ii) modalities occur only "alternatingly" as above.



- $p_1 := ab \wedge \neg ac \wedge bd \wedge \neg be \wedge \neg cf \wedge \neg cg$ (both play left)
- $p_2 := ab \wedge \neg ac \wedge \neg bd \wedge be \wedge \neg cf \wedge \neg cg$ (1 play left, 2 play right)
- $p_3 := \neg ab \wedge ac \wedge \neg bd \wedge \neg be \wedge cf \wedge \neg cg$ (1 play right, 2 play left)
- $p_4 := \neg ab \wedge ac \wedge \neg bd \wedge \neg be \wedge \neg cf \wedge cg$ (both play right)

[2] 'PL' stands for propositional logic, 'Nec' stands for the necessitation rule, and 'RM' stands for the monotonicity rule that if $\varphi \rightarrow \psi$ is a theorem, then so is $\Box_i\varphi \rightarrow \Box_i\psi$. Note that in the derivation, RM is only applied to theorems of the logic. For example, to obtain (4), RM is applied to the theorem $(\varphi \rightarrow \psi) \rightarrow (\neg\psi \rightarrow \neg\varphi)$ where $\varphi := \Box_2(\neg m_1 \wedge \neg m_3)$ and $\psi := \Box_2 m_2$.

- $q \coloneqq d >_1 e \wedge d >_1 f \wedge d >_1 g \wedge e >_1 f \wedge g >_1 e \wedge g >_1 f \wedge e >_2 d \wedge d >_2 f \wedge g >_2 d \wedge e >_2 f \wedge e >_2 g \wedge g >_2 f$ (players' preferences)
- $G \coloneqq (p_1 \vee p_2 \vee p_3 \vee p_4) \wedge q$ (description of the game)

(a) $\Box_1 G$ (assumption: 1 knows the game)

(b) $\Box_1((bd \vee be) \to \Diamond_2 be)$ (assumption: 1 knows that if 2 is at b then 2 considers the move be possible)

(c) $\Box_1((cf \vee cg) \to \Diamond_2 cg)$ (assumption: similar to (b))

(d) $\Box_2((ab \vee ac) \to \Diamond_1 ac)$ (assumption: similar to (b))

(e) $\Box_1((e >_2 d \wedge \Diamond_2 be) \to \neg bd)$ (assumption: follows from assuming 1 knows that 2 is rational)

(f) $\Box_1((g >_2 f \wedge \Diamond_2 cg) \to \neg cf)$ (assumption: similar to (e))

(g) $\big(\Box_1(ab \leftrightarrow be) \wedge \Box_1(ac \leftrightarrow cg) \wedge g >_1 e \wedge \Diamond_1 ac\big) \to \neg ab$ (assumption: follows from 1 being rational)

1. $\Box_1(ab \leftrightarrow be)$ (from (a), (b), and (e) using PL, Nec, and K)
2. $\Box_1(ac \leftrightarrow cg)$ (from (a), (c), and (f) using PL, Nec, and K)
3. $(ab \vee ac) \to \Diamond_1 ac$ (from (d) by T)
4. G (from (a) by T)
5. $ab \vee ac$ (from (4) by PL)
6. $\Diamond_1 ac$ (from (3) and (5) by PL)
7. $\neg ab$ (from (g), (1), (2), (4), and (6) by PL)
8. ac (from (5) and (7) by PL)
9. $ac \leftrightarrow cg$ (from (2) by T)
10. $ac \wedge cg$ (from (8) and (9) by PL)

In general, in typical strategic form games a player needs to reason about the beliefs of her opponents, as which action is best for her depends on her opponents' actions, which in turn depend on their beliefs. On the other hand, reasoning about one's own beliefs seems unnecessary, as the dependencies just mentioned seem to be tight: which action is the best for a player depends on what her opponents' actions are alone, which in turn depend on their beliefs over what their opponents' actions are alone. We can then iterate this reasoning, and it seems there is no place for reasoning about one's own beliefs. In Appendix A we provide a formalization of this idea using Kripke models of games in the style of [24] and [4], where only formulas with no modality scoping immediately over a modality of the same agent are used to ensure that rationalizable strategies are played.[3]

These considerations raise the question of the extent to which single-agent introspection axioms actually matter for multi-agent epistemic reasoning. In particular, as motivated by the above examples, we can ask: in situations where the agents and also the analyst only need to reason about formulas where modalities occur only alternatingly, would the commonly debated introspection axioms still matter, in the sense that assuming them allows us to derive more conclusions?

This question has indeed been partially investigated previously, though motivated not by the question of whether introspection axioms may in practice be "irrelevant" but rather by the goal of devising efficient

[3]We are not arguing that introspection assumptions never matter in multi-agent epistemic reasoning. For example, it is shown in [8, 18] that Aumann's [1] theorem on agreeing to disagree fails without the assumption of positive introspection.

reasoning algorithms for the system K45. In [16], it is explicitly stated (Lemma 5) that when restricted to the fragment of the multi-agent language in which modalities occur only in the agent-alternating way, K and K45 derive the same set of theorems.[4] This facilitates reasoning in K45 since it is also known that every formula is provably equivalent in K45 to an agent-alternating formula,[5] which is then derivable in K45 iff it is derivable K, making the efficient methods of deciding theoremhood in K applicable to K45. Subsequently, the idea of agent-alternating formulas was also used in the axiomatization of refinement quantification logics [10, 9] and in epistemic planning [14, 20, 6].

In this paper, we study the question more systematically. In § 2, we provide multiple ways to define the *agent-alternating formulas*, which include formulas like $\Box_a(\Box_b p \wedge \Box_b \Box_a q)$ but not $\Box_a(\Box_b p \wedge \Box_a q)$. In § 3, we first provide a bisimulation notion for the fragment of agent-alternating formulas and then use it to completely chart the relationships of the modal logics in the well-known "Modal Logic Cube" when restricted to the fragment of agent-alternating formulas. We prove that if one starts with multi-agent K or KD, then adding both the 4 and 5 axioms (or adding the B axiom) does not allow the derivation of any new agent-alternating formula—in this sense, introspection axioms do not matter. By contrast, we show that such conservativity results fail for knowledge and multi-agent KT, though they hold with respect to a smaller class of *agent-nonrepeating formulas* introduced in § 4. In § 5, we report on preliminary investigations of how these results are affected in the presence of a *common belief* operator in the language. Finally, we conclude in § 6 with some directions for future research.

2 Agent-Alternating Formulas

Fix a set A of agents with $|A| \geq 2$ and a countably infinite set Prop of proposition letters.

Definition 2.1. *The language of multi-agent epistemic logic is defined inductively by*

$$\mathscr{L} \ni \varphi ::= p \mid \neg\varphi \mid (\varphi \wedge \varphi) \mid \Box_a \varphi$$

where $p \in \mathsf{Prop}$ *and* $a \in A$. *Connectives* $\rightarrow$, $\vee$, *and* $\Diamond_a$ *are abbreviations as usual.*

We adopt the standard definition of when one formula is a *subformula* of another.

Notation 2.2. For $\varphi, \psi \in \mathscr{L}$, let $\varphi \preccurlyeq \psi$ indicate that φ is a subformula of ψ and $\varphi \prec \psi$ that φ is a proper subformula of ψ.

Intuitively, agent-alternating formulas are those formulas in which an operator $\Box_a$ does not immediately scope over another operator $\Box_a$ of the same agent a. We now offer two ways to precisely capture this intuition, one using immediate subformulas and occurrences, and one using simultaneous induction.

Definition 2.3. For $\alpha, \beta \in \mathscr{L}$, we say α is an *immediate subformula* of β, and write $\alpha \prec \beta$, if β is either $\neg\alpha$, or $(\alpha \wedge \gamma)$ for some $\gamma \in \mathscr{L}$, or $(\gamma \wedge \alpha)$ for some $\gamma \in \mathscr{L}$, or $\Box_a\alpha$ for some $a \in A$. Note that the reflexive and transitive closure of $\prec$ is precisely $\preccurlyeq$.

For any $\varphi \in \mathscr{L}$, an *occurrence type* O of φ is a finite sequence $\langle O_1, O_2, \cdots, O_{\mathrm{len}(O)}\rangle$ of formulas in $\mathscr{L}$ such that $O_{\mathrm{len}(O)} = \varphi$ and for each i between 1 and $\mathrm{len}(O) - 1$, $O_i \prec O_{i+1}$. Let $OC(\varphi)$ be the set of occurrence types of φ and $\leq$ the prefix-extension relation: $O \leq O'$ iff O' is a suffix of O. It is then easy to see that $\langle OC(\varphi), \leq\rangle$ is a (downward-growing) tree.

We call an occurrence type O of φ with $O_1 = \alpha$ an *α-occurrence* of φ. If this α is $\Box_a\beta$ for some $\beta \in \mathscr{L}$ and $a \in A$, then we also call O a *$\Box_a$-occurrence*. We typically denote an α-occurrence by $O[\alpha]$.

[4] The authors refer to [11] for the proof of this lemma, though we are unable to locate an explicit proof there.

[5] In Appendix B, we show the semantic counterpart of this proposition and further show that 4 and 5 are in a sense necessary. See also Theorem 1 of [23] for an early precursor of this result.

Definition 2.4. A formula $\varphi \in \mathscr{L}$ is an *agent-alternating formula* iff for any $a \in A$ and any two different $\Box_a$ occurrences $O[\Box_a\alpha]$ and $O[\Box_a\beta]$ such that $O[\Box_a\alpha] \leq O[\Box_a\beta]$, there is a $b \in A \setminus \{a\}$ and a $\Box_b$-occurrence $O[\Box_b\gamma]$ of φ such that $O[\Box_a\alpha] \leq O[\Box_b\gamma] \leq O[\Box_a\beta]$. In other words, φ is agent alternating iff in the tree $\langle OC(\varphi), \leq \rangle$, between any two $\Box_a$-occurrences, there is a $\Box_b$-occurrence for some $b \in A \setminus \{a\}$.

Example 2.5. Assuming a, b, c are different elements in A, examples of agent-alternating formulas include:

$$\Box_a p,\ \Box_a\Box_b p,\ \Box_a\Box_b\Box_a p,\ \Box_a\Box_b\Box_c p,\ \Box_a(p \wedge \Box_b q).$$

Non-examples include:

$$\Box_a\Box_a p,\ \Box_a\Box_b\Box_a\Box_a p,\ \Box_a(\Box_b\Box_a p \wedge \Box_a q).$$

We now give an equivalent inductive definition of the set of agent-alternating formulas.

Definition 2.6. Define a family $\{\mathscr{L}_{-a}\}_{a\in A}$ of languages through the following simultaneous induction:

$$\mathscr{L}_{-a} \ni \varphi ::= p \mid \Box_x\psi \mid \neg\varphi \mid (\varphi \wedge \varphi)$$

where $p \in \mathsf{Prop}$ and $x \in A \setminus \{a\}$ while $\psi \in \mathscr{L}_{-x}$. Then the language $\mathscr{L}_{alt}$ is defined inductively by

$$\mathscr{L}_{alt} \ni \varphi ::= p \mid \chi \mid \neg\varphi \mid (\varphi \wedge \varphi)$$

where $p \in \mathsf{Prop}$ and $\chi \in \bigcup_{a\in A} \mathscr{L}_{-a}$.

Note that $\bigcup_{a\in A} \mathscr{L}_{-a}$ does not cover all of $\mathscr{L}_{alt}$. For example, when $A = \{a, b\}$ with $a \neq b$, $\Box_a p \wedge \Box_b p$ is in $\mathscr{L}_{alt}$ but not in $\bigcup_{x\in A} \mathscr{L}_{-x}$.

It is not hard to verify that the two definitions above are equivalent, suggesting that our formal definitions captures the intended intuition. Due to limited space, we omit the proof of this equivalence, but the idea is simply to examine the parsing trees of formulas.

Proposition 2.7. *For any $\varphi \in \mathscr{L}$, φ is agent alternating iff $\varphi \in \mathscr{L}_{alt}$.*

3 Collapsing logics by $\mathscr{L}_{alt}$

We now investigate which logics are indistinguishable by formulas in $\mathscr{L}_{alt}$. For any normal modal logic L (defined as a set of formulas in $\mathscr{L}$ satisfying the usual closure properties), let $\mathsf{L}|_{alt} := \mathsf{L} \cap \mathscr{L}_{alt}$. Then the general question is: for which modal logics L and L' are $\mathsf{L}|_{alt}$ and $\mathsf{L}'|_{alt}$ the same?

More specifically, since we are mainly interested in the introspection axioms 4 and 5, we focus on the logics appearing in the classic modal logic cube shown in Figure 1 below.[6] Our main result is that the two shaded areas in Figure 1 are collapsed in $\mathscr{L}_{alt}$ but no other logics are. To establish this result, we need to first develop bisimulation and unraveling concepts for agent-alternating formulas.

Notation 3.1. For convenience, we consider *alt* as an object not in A. Also for any set $\mathscr{L}'$ of formulas, $\mathscr{M}, u \equiv_{\mathscr{L}'} \mathscr{N}, v$ means that for all $\varphi \in \mathscr{L}'$, $\mathscr{M}, u \models \varphi$ iff $\mathscr{N}, v \models \varphi$.

Definition 3.2 (Agent-alternating bisimulation relation)**.** An *agent-alternating bisimulation family* between two models $\mathscr{M}$ and $\mathscr{N}$ is a family of binary relations $\{\leftrightarrows_a\}_{a\in A\cup\{alt\}}$ between $\mathscr{M}$ and $\mathscr{N}$ such that for every $a \in A \cup \{alt\}$ and every $u \in \mathscr{M}$ and $v \in \mathscr{N}$ such that $u \leftrightarrows_a v$:

- (Atom) for all $p \in \mathsf{Prop}$, $u \in V^{\mathscr{M}}(p)$ iff $v \in V^{\mathscr{N}}(p)$;

[6] Figure 1 is reproduced from [7].

- (Zig) for all $b \in A \setminus \{a\}$ and $u' \in R_x^{\mathscr{M}}(u)$, there is $v' \in R_x^{\mathscr{N}}(v)$ such that $u' \leftrightarrows_b v'$;
- (Zag) for all $b \in A \setminus \{a\}$ and $v' \in R_x^{\mathscr{N}}(v)$, there is $u' \in R_x^{\mathscr{M}}(u)$ such that $u' \leftrightarrows_b v'$.

Then we say $\mathscr{M}, u$ is *agent-alternating bisimilar to* $\mathscr{N}, v$ if there is an agent-alternating bisimulation family $\{\leftrightarrows_a\}_{a \in A \cup \{alt\}}$ between $\mathscr{M}$ and $\mathscr{N}$ such that $u \leftrightarrows_{alt} v$.

Lemma 3.3. *For any models $\mathscr{M}$ and $\mathscr{N}$, agent-alternating bisimulation family $\{\leftrightarrows_a\}_{a \in A \cup \{alt\}}$ between $\mathscr{M}$ and $\mathscr{N}$, and $a \in A$, if $u \leftrightarrows_a v$, then $\mathscr{M}, u \equiv_{\mathscr{L}_{-a}} \mathscr{N}, v$, and if $u \leftrightarrows_{alt} v$, then $\mathscr{M}, u \equiv_{\mathscr{L}_{alt}} \mathscr{N}, v$.*

Proof. A simple induction on modal depth. □

Definition 3.4 (Agent-alternating unraveling)**.** Given a model $\mathscr{M} = \langle W^{\mathscr{M}}, \{R_a^{\mathscr{M}}\}_{a \in A}, V^{\mathscr{M}} \rangle$, its *agent-alternating unravelings* are all models of the form $\langle S, \{R_a\}_{a \in A}, V \rangle$ satisfying the following conditions:

- S is the set of all nonempty finite sequences s of pairs in $(A \cup \{alt\}) \times W^{\mathscr{M}}$ such that
  (1) $s_1 \in \{alt\} \times W^{\mathscr{M}}$,
  (2) $s_i \in A \times W^{\mathscr{M}}$ for all $i = 2 \ldots \mathrm{len}(t)$, and
  (3) letting $\langle a_i, w_i \rangle = s_i$ for all $i = 1 \ldots \mathrm{len}(s)$, $w_i R_{a_{i+1}}^{\mathscr{M}} w_{i+1}$ and $a_i \neq a_{i+1}$ for all $i = 1 \ldots \mathrm{len}(s) - 1$;
- for all $a \in A \cup \{alt\}$ and $s \in S$ such that $s_{\mathrm{len}(s)} = \langle a, w \rangle$, for all $b \in A \setminus \{a\}$, $R_b(s) = \{s + \langle b, w' \rangle \mid w \in R_b^{\mathscr{M}}(w))\}$ (note that this is precisely $\{t \in S \mid s = t_{1 \ldots \mathrm{len}(t)-1}, t_{\mathrm{len}(t)} \in \{b\} \times W^{\mathscr{M}}\}$;
- for every $s \in S$ and $p \in \mathsf{Prop}$, $s \in V(p)$ iff $s_{\mathrm{len}(s)} \in (A \cup \{alt\}) \times V^{\mathscr{M}}(p)$.

Let $Alt(\mathscr{M})$ denote the set of all agent alternating unraveling of $\mathscr{M}$. Then for every $\mathscr{N} \in Alt(\mathscr{M})$, we define a family of binary relations between $\mathscr{M}$ and $\mathscr{N}$, which we denote as $\{P_a^{\mathscr{N}}\}_{a \in A \cup \{alt\}}$, by

$$u P_a^{\mathscr{N}} s \iff s_{\mathrm{len}(s)} = \langle a, u \rangle.$$

Lemma 3.5. *For any model $\mathscr{M}$ and $\mathscr{N} \in Alt(\mathscr{M})$, $\{P_a^{\mathscr{N}}\}_{a \in A \cup \{alt\}}$ is an agent-alternating bisimulation family between $\mathscr{M}$ and $\mathscr{N}$. Consequently, by Lemma 3.3, for every $w \in \mathscr{M}$, $\mathscr{M}, w \equiv_{\mathscr{L}_{alt}} \mathscr{N}, \langle \langle alt, w \rangle \rangle$.*

Proof. Immediate from Definition 3.4 and the recursive structure of $\mathscr{L}_{alt}$ as defined in Definition 2.6. □

Now we can formally state our main result.

Theorem 3.6. *Among the systems displayed in Figure 1:*

1. $\mathsf{K}|_{alt} = \mathsf{K4}|_{alt} = \mathsf{K5}|_{alt} = \mathsf{K45}|_{alt} = \mathsf{KB}|_{alt}$;
2. $\mathsf{KD}|_{alt} = \mathsf{KD4}|_{alt} = \mathsf{KD5}|_{alt} = \mathsf{KD45}|_{alt} = \mathsf{KDB}|_{alt}$;
3. *no other collapse happens when restricting to $\mathscr{L}_{alt}$.*

The results are summarized in Figure 2, where systems in the same shaded region in Figure 1 collapse.

Proof. Combining Proposition 3.7, 3.8, and 3.9 below, we have all the collapsing and non-collapsing results in the three layers of Figure 1. To see that the three layers do not collapse, it is enough to observe that the axioms D and T are in $\mathscr{L}_{alt}$. □

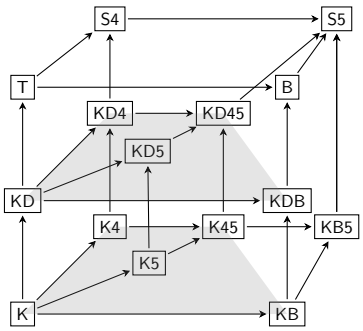


Figure 1: Systems of interest

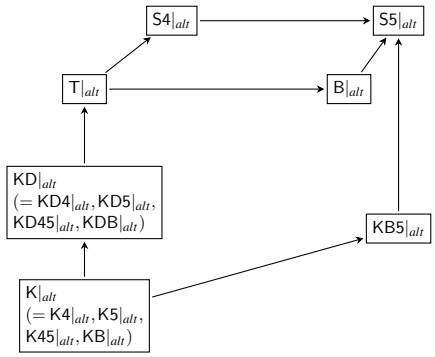


Figure 2: Systems when restricted to $\mathscr{L}_{alt}$

Proposition 3.7 (Collapsing 4 and 5)**.** $\mathsf{K}|_{alt} = \mathsf{K45}|_{alt}$ *and* $\mathsf{KD}|_{alt} = \mathsf{KD45}|_{alt}$

Proof. The right-to-left direction of both equations is trivial. For the left-to-right direction, by completeness, we need only show that for every $\varphi \in \mathscr{L}_{alt}$, if φ is satisfied by a pointed model, then it is also satisfied by a pointed model based on a transitive and Euclidean frame. Further, if the first model is based on a serial frame, then the frame of the second model is also serial. So it is enough to show the following: for every pointed model $\mathscr{M},u$, there exists a pointed model $\mathscr{N},v$ such that:

1. if for every $a \in A$, $R_a^{\mathscr{M}}$ is serial, then for every $a \in A$, $R_a^{\mathscr{N}}$ is also serial;
2. for every $a \in A$, $R_a^{\mathscr{N}}$ is transitive and Euclidean;
3. $\mathscr{M},u \equiv_{\mathscr{L}_{alt}} \mathscr{N},v$.

Now let $\mathscr{N} = \langle S, \{R_a\}_{a\in A}, V\rangle$ be constructed by adding to the definition of being in $Alt(\mathscr{M})$ as in Definition 3.4 the following:

- for all $a \in A$ and $s \in S$ such that $s_{\text{len}(s)} \in \{a\} \times W^{\mathscr{M}}$, $R_a(s) = R_a(s_{1\ldots\text{len}(s)-1})$.

This construction is possible because crucially the definition of being an agent-alternating unraveling of $\mathscr{M}$ is silent on what $R_a(s)$ should be when s ends in $\{a\} \times W^{\mathscr{M}}$ for $a \in A$. Also, when s ends in $\{a\} \times W^{\mathscr{M}}$ for some $a \in A$, $\text{len}(s) > 1$ and $s_{1\ldots\text{len}(s)-1}$ does not end in $\{a\} \times W^{\mathscr{M}}$, which means that $R_a(s_{1\ldots\text{len}(s)-1})$ is defined in Definition 3.4.

Now we can show that $\mathscr{N},\langle\langle alt,u\rangle\rangle$ satisfies all the requirements. It is not hard to see that if $\mathscr{M}$ is serial, then so is $\mathscr{N}$. The key observation is that for any $s \in S$, letting $\langle a,w\rangle = s_{len(s)}$, $R_a(s)$ must include s, and $R_b(s)$ for any $b \in A \setminus \{s\}$ must be nonempty since $R_b^{\mathscr{M}}(w)$ is nonempty. Hence we are done with (1). To see that for every $a \in A$, R_a is transitive and Euclidean, note that for any $s \in S$, letting $\langle x,w\rangle = s_{\text{len}(s)}$, we have the following:

- If $x \neq a$, then for every $t \in R_a(s)$, t ends in $\{a\} \times W^{\mathscr{M}}$, and $t_{1\ldots\text{len}(t)-1} = s$. This means that our construction above applies to t and $R_a(t) = R_a(s)$.
- If $x = a$, then our construction above applies to s: letting $s^0 = s_{1\ldots\text{len}(s)-1}$, s^0 does not end in $\{a\} \times W^{\mathscr{M}}$, and $R_a(s) = R_a(s^0)$ by our definition. Then it is easy to see that for every $t \in R_a(s)$, $t \in R_a(s^0)$, and $t_{1\ldots\text{len}(t)-1}$ is also s^0. This means that t ends in $\{a\} \times W^{\mathscr{M}}$, and our construction above also applies to t. Hence $R_a(t) = R_a(s^0) = R_a(s)$.

Adding the above two points together, we have shown that for every $s \in S$ and $t \in R_a(s)$, $R_a(t) = R_a(s)$. This is precisely transitivity plus Euclideanness.

By Lemma 3.5, $\mathscr{M}, u \equiv_{\mathscr{L}_{alt}} \mathscr{N}, \langle\langle alt, u\rangle\rangle$ since $\{P_a^{\mathscr{N}}\}_{a \in A \cup \{alt\}}$ is an agent-alternating bisimulation family and $u P_{alt}^{\mathscr{N}} \langle\langle alt, u\rangle\rangle$. Thus, all three requirements are satisfied, so we are done. □

Proposition 3.8 (Collapsing B)**.** $\mathsf{K}|_{alt} = \mathsf{KB}|_{alt}$ *and* $\mathsf{KD}|_{alt} = \mathsf{KDB}|_{alt}$

Proof. Following the strategy of the proof of Proposition 3.7, we only need to show that for every pointed model $\mathscr{M}, u$, there exists an agent-alternating unraveling $\mathscr{N} = \langle S, \{R_a\}_{a \in A}, V\rangle$ of $\mathscr{M}$ such that for every $a \in A$, R_a is symmetric.

Indeed, let $\mathscr{N}$ be the agent-alternating unraveling of $\mathscr{M}$ such that for every $a \in A$ and $s \in S$ such that $s_{\mathrm{len}(s)} \in \{a\} \times W^{\mathscr{M}}$, $R_a(s) = \{s_{1 \ldots \mathrm{len}(s)-1}\}$. Then it is easy to see that for every $a \in A$, R_a is symmetric: for every $s, t \in S$, if $s R_a t$, then we have the following.

- If $s_{\mathrm{len}(s)} \in \{a\} \times W^{\mathscr{M}}$, then t must be $s_{1 \ldots \mathrm{len}(s)-1}$ by our construction. By the definition of unraveling, $t R_a s$.
- If $s_{\mathrm{len}(s)} \notin \{a\} \times W^{\mathscr{M}}$, then t must be $s + \langle a, w\rangle$ for some w such that letting $\langle b, w_0\rangle = s_{\mathrm{len}(s)}$, $w_0 R_a^{\mathscr{M}} w$. Then our construction applies to t and $t R_a s$.

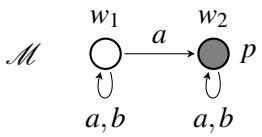
Putting the above two points together, R_a is symmetric, so we are done. □

Proposition 3.9 (Non-collapsing results)**.** $\mathsf{B}|_{alt} \setminus \mathsf{S4}|_{alt}$, $\mathsf{S4}|_{alt} \setminus \mathsf{B}|_{alt}$, $\mathsf{KB5}|_{alt} \setminus \mathsf{S4}|_{alt}$, $\mathsf{KB5}|_{alt} \setminus \mathsf{B}|_{alt}$ *are all nonempty.*

Proof. Let a, b be two different elements in A. In B $(= \mathsf{KTB})$, we have the following theorems.

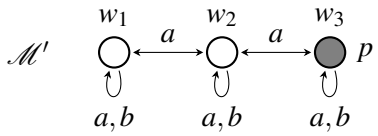
$$\begin{aligned}
&\vdash_{\mathsf{B}} \Box_b \Box_a p \to \Box_a p && (1)[\mathsf{T}]\\
&\vdash_{\mathsf{B}} \Diamond_a \Box_b \Box_a p \to \Diamond_a \Box_a p && (2)[\mathsf{RM}, 1]\\
&\vdash_{\mathsf{B}} \Diamond_a \Box_a p \to p && (3)[\mathsf{B}]\\
&\vdash_{\mathsf{B}} \Diamond_a \Box_b \Box_a p \to p && (4)[\mathsf{MP}, 2, 3].
\end{aligned}$$

Now the last formula, formula (4), is agent-alternating. However, $\nvdash_{\mathsf{S4}} (4)$. Using soundness, it is enough to find an $\mathsf{S4}$ model refuting (4). Consider the following model:



By focusing on the restriction of $\mathscr{M}$ to a and b, respectively, it is easy to see that $\mathscr{M}$ is based on an $\mathsf{S4}$ frame. Indeed, the accessibility relation for b is even an equivalence relation. Now, $\mathscr{M}, w_1 \models \Diamond_a \Box_b \Box_a p$ since $\mathscr{M}, w_2 \models \Box_b \Box_a p$. Also we have $\mathscr{M}, w_1 \not\models p$. Hence $\mathscr{M}, w_1 \not\models (4)$, and thus $\nvdash_{\mathsf{S4}} (4)$. This shows that $\mathsf{B}|_{alt} \setminus \mathsf{S4}|_{alt}$ is nonempty.

In the same spirit, $\Diamond_a \Box_b \Diamond_a p \to \Diamond_a p \in \mathsf{S4}|_{alt} \setminus \mathsf{B}|_{alt}$. The derivation of $\Diamond_a \Box_b \Diamond_a p \to \Diamond_a p$ in $\mathsf{S4}$ is essentially the same as above: using T we can eliminate the $\Box_b$ in between the two $\Diamond_a$'s. A symmetric countermodel of this formula is as follows.



In KB5 we do not have the T axiom. So $\Diamond_a\Box_b\Box_a p \to p$ and $\Diamond_a\Box_b\Diamond_a p \to \Diamond_a p$ are not in KB5. However, we only need to add $\Box_a\Diamond_b(p \vee \neg p)$ to the antecedents. Specifically, note that the formula $(\Diamond_b(p \vee \neg p) \wedge \Box_b q) \to q$ is in KB5. Hence:

- $\Diamond_a(\Diamond_b(p \vee \neg p) \wedge \Box_b\Box_a p) \to p \in \mathsf{KB5}|_{alt} \setminus \mathsf{S4}|_{alt}$;
- $\Diamond_a(\Diamond_b(p \vee \neg p) \wedge \Box_b\Diamond_a p) \to \Diamond_a p \in \mathsf{KB5}|_{alt} \setminus \mathsf{B}|_{alt}$.

Their derivations in KB5 are in the same spirit as above, and $\mathscr{M}$ and $\mathscr{M}'$ can be reused. □

4 Agent-nonrepeating formulas

The above non-collapsing results raise a natural question: is there a smaller fragment defined in the same spirit that also collapses S5 to T? Recall that the non-collapsing results are witnessed by formulas like $\Diamond_a\Box_b\Diamond_a p \to \Diamond_a p$. When $\Box_b$ is factive, agent a is *ipso facto* introspecting since we can eliminate $\Box_b$ by T. In this section we identify a fragment of *agent-nonrepeating formulas* in which this cannot happen and S5 does collapse to T. The key idea is that we need to forbid $\Box_a$ to appear at all in the scope of $\Box_a$. Again, to formalize this idea, we provide an occurrence-based definition and an inductive definition.

Definition 4.1. A formula $\varphi \in \mathscr{L}$ is an *agent-nonrepeating formula* iff for any $a \in A$ and $\Box_a$ occurrence $O[\Box_a\alpha]$, there is no other $\Box_a$ occurrence $O[\Box_a\beta]$ such that $O[\Box_a\beta] \leq O[\Box_a\alpha]$

Definition 4.2. Define a family $\{\mathscr{L}_X\}_{X\subseteq A}$ of fragments of $\mathscr{L}$ through the following simultaneous induction:

$$\mathscr{L}_X \ni \varphi ::= p \mid \Box_x\psi \mid \neg\varphi \mid (\varphi \wedge \varphi)$$

where $p \in \mathsf{Prop}$ and $x \in X$ while $\psi \in \mathscr{L}_{X\setminus\{x\}}$.

The following equivalence is easily verified.

Proposition 4.3. *For any $\varphi \in \mathscr{L}$, $\varphi \in \mathscr{L}_A$ iff φ is agent nonrepeating.*

As before, we need a notion of bisimulation appropriate for the fragment.

Definition 4.4. An *agent-nonrepeating bisimulation family* between two models $\mathscr{M}$ and $\mathscr{N}$ is a family of binary relations $\{\leftrightarrows_X\}_{X\subseteq A}$ between $\mathscr{M}$ and $\mathscr{N}$ such that for every $X \subseteq A$ and every $u \in \mathscr{M}$ and $v \in \mathscr{N}$ such that $u \leftrightarrows_X v$:

- (Atom) for all $p \in \mathsf{Prop}$, $u \in V^{\mathscr{M}}(p)$ iff $v \in V^{\mathscr{N}}(p)$;
- (Zig) for all $x \in X$ and $u' \in R_x^{\mathscr{M}}(u)$, there is $v' \in R_x^{\mathscr{N}}(v)$ such that $u' \leftrightarrows_{X\setminus\{x\}} v'$;
- (Zag) for all $x \in X$ and $v' \in R_x^{\mathscr{N}}(v)$, there is $u' \in R_x^{\mathscr{M}}(u)$ such that $u' \leftrightarrows_{X\setminus\{x\}} v'$.

Then we say $\mathscr{M},u$ is *agent-nonrepeating bisimilar to* $\mathscr{N},v$ if there is an agent-nonrepeating bisimulation family $\{\leftrightarrows_X\}_{X\subseteq A}$ between $\mathscr{M}$ and $\mathscr{N}$ such that $u \leftrightarrows_A v$.

Lemma 4.5. *For any models $\mathscr{M}$ and $\mathscr{N}$, agent-nonrepeating bisimulation family $\{\leftrightarrows_X\}_{X\subseteq A}$ between $\mathscr{M}$ and $\mathscr{N}$, and $X \subseteq A$, if $u \leftrightarrows_X v$, then $\mathscr{M},u \equiv_{\mathscr{L}_X} \mathscr{N},v$. Hence whenever $\mathscr{M},u$ is agent-nonrepeating bisimilar to $\mathscr{N},v$, we have $\mathscr{M},u \equiv_{L_A} \mathscr{N},v$.*

For any logic L, we write $\mathsf{L}|_{nr}$ for $\mathsf{L} \cap \mathscr{L}_A$. We can now prove the desired collapse result.

Theorem 4.6. *For every reflexive pointed model $\mathscr{M}, w$, there is a partition model $\mathscr{N}, w'$ such that $\mathscr{M}, w$ is agent-nonrepeating bisimilar to $\mathscr{N}, w'$. Consequently, $\mathsf{T}|_{nr} = \mathsf{S5}|_{nr}$.*

Proof. Let $\mathscr{M}$ be a reflexive model. We construct $\mathscr{N} = \langle S, \{R_a\}_{a \in A}, V \rangle$. Let S be the set of all nonempty finite sequences s of pairs in $\wp(A) \times W^{\mathscr{M}}$ such that, letting $s = \langle \langle X_i, w_i \rangle \rangle_{i=1 \ldots \mathrm{len}(s)}$, (1) $X_1 = A$, and (2) for all $i = 1 \ldots \mathrm{len}(s) - 1$ and $X_{i+1} \subsetneq X_i$, there is $a \in A$ such that $X_i = X_{i+1} \cup \{a\}$ and $w_i R_a^{\mathscr{M}} w_{i+1}$.

To make the rest of the construction easier, we make a few auxiliary definitions. For each $s \in S$, define $\mathit{LastA}(s)$ to be $* \notin A$ when $\mathrm{len}(s) = 1$ and otherwise the $a \in X \setminus X_0$ with $\langle X, u \rangle = s_{\mathrm{len}(s)}$ and $\langle X_0, u_0 \rangle$ when $s_{\mathrm{len}(s)-1}$. Intuitively, $\mathit{LastA}(s)$ denotes the last accessibility relation used in the sequence s. It is easy to observe from the definition above that for any $s \in S$ with $\langle X, u \rangle = s_{\mathrm{len}(s)}$, $\mathit{LastA}(s) \notin X$, and moreover when $\mathrm{len}(s) > 1$, $s_{\mathrm{len}(s)-1} = \langle X \cup \mathit{LastA}(s), u_0 \rangle$ for a $u_0 \in \mathscr{M}$ such that $u_0 R^{\mathscr{M}}_{\mathit{LastA}(s)} u$.

Then for any $a \in A$ and $s \in S$, define $s_{<a}$ to be s if $a \neq \mathit{LastA}(s)$ and otherwise $s_{1 \ldots \mathrm{len}(s)-1}$. Intuitively this is the a-predecessor of s in S. Now R_a is defined for each $a \in A$ by the condition that $s R_a t$ iff $s_{<a} = t_{<a}$ for all $s, t \in S$. With this definition, it is not hard to compute $R_a(s)$ specifically. For all $s \in S$ such that $\mathrm{len}(s) > 1$, letting $s_{\mathrm{len}(s)} = \langle X, u \rangle$, $s_{\mathrm{len}(s)-1} = \langle X_0, u_0 \rangle$ and $s_0 = s_{1 \ldots \mathrm{len}(s)-1}$, we have the following.

- For all $a \in X$, $R_a(s) = \{s + \langle X \setminus \{a\}, v \rangle \mid v \in R_a^{\mathscr{M}}(u)\} \cup \{s\}$.
- For the $a \in X_0 \setminus X$ (namely $\mathit{LastA}(s)$), $R_a(s) = \{s_0\} \cup \{s_0 + \langle X, u' \rangle \mid u_0 R_a^{\mathscr{M}} u'\}$.
- For $a \in A \setminus X_0$, $R_a(s) = \{s\}$.

For all $s \in S$ such that $\mathrm{len}(s) = 1$, in which case $s = \langle A, u \rangle$ for some $u \in W^{\mathscr{M}}$, we have that $R_a(s) = \{s + \langle A \setminus \{a\}, v \rangle \mid v \in R_a^{\mathscr{M}}(u)\} \cup \{s\}$.

The valuation V is defined as usual. For every $s \in S$ and $p \in \mathsf{Prop}$, $s \in V(p)$ iff $s_{\mathrm{len}(s)} \in \wp(A) \times V^{\mathscr{M}}(p)$. That is, $s \in V(p)$ iff the second coordinate of $s_{\mathrm{len}(s)}$ is in $V^{\mathscr{M}}(p)$.

Then there is a natural family $\{\leftrightarrows_X\}_{X \in \wp(A)}$ of relations between $\mathscr{M}$ and $\mathscr{N}$ defined by

$$u \leftrightarrows_X s \iff s_{\mathrm{len}(s)} = \langle Y, u \rangle \text{ with } X \subseteq Y.$$

Now we are left with two tasks: to show that $\mathscr{N}$ is a partition model and to show that $\{\leftrightarrows_X\}_{X \in \wp(A)}$ is an agent-nonrepeating bisimulation family between $\mathscr{M}$ and $\mathscr{N}$. That $\mathscr{N}$ is a partition model is clear: for any $a \in A$, we defined R_a by an equality condition. Now we show that $\{\leftrightarrows_X\}_{X \in \wp(A)}$ is an agent-nonrepeating bisimulation family. Pick arbitrary $u \in \mathscr{M}, s \in S$, and $X \in \wp(A)$ such that $u \leftrightarrows_X s$. By definition, $s_{\mathrm{len}(s)} = \langle Y, u \rangle$ for some $Y \supseteq X$. The (Atom) clause is trivial. For the (Zig) clause, pick an arbitrary $a \in X$ and $v \in R_a^{\mathscr{M}}(u)$. Then we see that $s + \langle Y \setminus \{a\}, v \rangle$ witnesses the requirement, as $s + \langle Y \setminus \{a\}, v \rangle \in R_a(s)$ and $v \leftrightarrows_{X \setminus \{a\}} s + \langle Y \setminus \{a\}, v \rangle$ because from $X \subseteq Y$ we have $X \setminus \{a\} \subseteq Y \setminus \{a\}$. For the (Zag) clause, we need to use the fact that $\mathscr{M}$ is reflexive. Picking an arbitrary $a \in X$ and $t \in R_a(s)$, we know that $a \in Y$ and hence there are two cases for t:

- $t = s$. Then u itself witnesses the requirement, as $u \leftrightarrows_{X \setminus \{a\}} s$ and $u R_a^{\mathscr{M}} u$.
- $t = s + \langle Y \setminus \{a\}, v \rangle$ for some $v \in \mathscr{M}$ such that $u R_a^{\mathscr{M}} v$. Then clearly v is witnesses the requirement.

In summary, $\mathscr{N}, \langle \langle A, w \rangle \rangle$ is a pointed partition model, and $\mathscr{M}, w$ is agent-nonrepeating bisimilar to it. Hence we are done. □

With the help of the above theorem, we obtain the poset of logics in Figure 3 when restricted to $\mathscr{L}_A$.

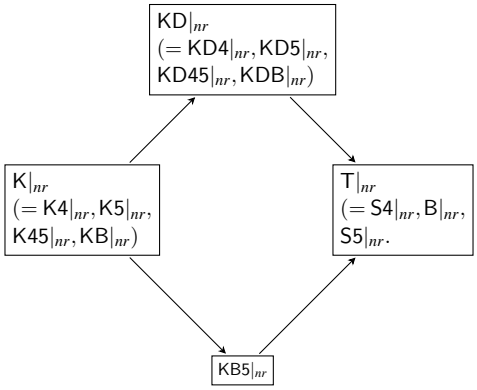


Figure 3: Systems when restricted to $\mathscr{L}_A$

Theorem 4.7. *Among the systems displayed in Figure 1:*

1. $\mathsf{K}|_{nr} = \mathsf{K4}|_{nr} = \mathsf{K5}|_{nr} = \mathsf{K45}|_{nr} = \mathsf{KB}|_{nr}$;
2. $\mathsf{KD}|_{nr} = \mathsf{KD4}|_{nr} = \mathsf{KD5}|_{nr} = \mathsf{KD45}|_{nr} = \mathsf{KDB}|_{nr}$;
3. $\mathsf{T}|_{nr} = \mathsf{S4}|_{nr} = \mathsf{B}|_{nr} = \mathsf{S5}|_{nr}$;
4. *no other collapse happens when restricting to* $\mathscr{L}_A$.

The results are summarized in Figure 3.

Proof. Since $\mathscr{L}_A \subseteq \mathscr{L}_{alt}$, all collapsing results in Theorem 3.6 obtain. This covers (1) and (2). Due to Theorem 4.6, we have (3). Clearly $\mathsf{KB5}|_{nr} \subseteq \mathsf{S5}|_{nr}$ since the T axiom is in $\mathscr{L}_A$. Hence we are left to show that $\mathsf{KB5}|_{nr}$ is not in $\mathsf{KD}|_{nr}$. The witness is simply $(\Diamond_a(p \vee \neg p) \wedge \Box_a p) \to p$. □

5 Allowing the standard common belief operator?

Given its importance in many applications, it is natural to consider adding the standard common belief operator to $\mathscr{L}_{alt}$ and investigate the resulting collapse of logics. In this section, we provide three non-collapsing results for the axioms 4 and 5, and leave a full investigation with possible collapsing results for future work. Given that Cp expresses a potentially infinitary conjunction of formulas where modalities are compounded in arbitrary order, implicitly Cp is not agent alternating: formulas like $\Box_a\Box_a p$ are part of the definition of Cp. Hence it is not surprising that we get many non-collapsing results. Moreover, we face the problem of whether to allow C to be in the scope of or scope over any $\Box_a$ or itself. Again the reason is that if we expand $C\Box_a p$ or CCp syntactically as infinitary formulas, $\Box_a$ will scope over an occurrence of $\Box_a$ immediately. Hence it is not obvious what is the most appropriate definition of an agent-alternating fragment in a language with a common belief operator, and a full investigation would require a hierarchy of fragments, each allowing more interactions between C and other modalities or C itself. Our non-collapsing results about 4 also crucially rely on A being finite. We conjecture that the collapsing situation would change radically when A is infinite.

Now let us fix the language and semantics for the common belief operator.

Definition 5.1. Let $\mathscr{C}$ be defined by adding new clauses $C\varphi$ and $E\varphi$ to $\mathscr{L}$'s context-free grammars. Semantically, $\mathscr{M}, u \models E\varphi$ iff for all $v \in \mathscr{M}$ such that $u(\bigcup_{a\in A} R_a^{\mathscr{M}})v$, $\mathscr{M}, v \models \varphi$, and $\mathscr{M}, u \models C\varphi$ iff for all $v \in \mathscr{M}$ such that $u(\bigcup_{a\in A} R_a^{\mathscr{M}})^+ v$, $\mathscr{M}, v \models \varphi$, where $(\bigcup_{a\in A} R_a^{\mathscr{M}})^+$ means the transitive closure of the

union of relations in $\{R_a^{\mathscr{M}}\}_{a\in A}$. Hence $E\varphi$ formalizes "everyone believes φ," and $C\varphi$ formalizes "it is commonly believed that φ."

Our logics must expand as well, as we need to add the axioms and rules for the C and E operators. To avoid choosing particular axiomatizations, we define logics directly as validities. For any $\mathsf{L} \subseteq \mathscr{L}$, let CL denote the set of formulas in $\mathscr{C}$ that are valid on all frames that validates L. For particular axiomatizations, see [12]. For our purposes, it is enough to note that for any L, the followings formulas are in CL.

$$(Cp \wedge C(p \to q)) \to Cq \qquad (Ep \wedge E(p \to q)) \to Eq$$
$$(C(p \to Ep) \wedge Ep) \to Cp \qquad Ep \to \Box_a p$$
$$Cp \to E(p \wedge Cp) \qquad \bigwedge_{a\in A} \Box_a p \to Ep \quad (\text{when } A \text{ is finite}).$$

Then we can identify at least two E-free fragments: one in which C is not allowed to interact with $\Box_a$ but allowed to interact with C, and one in which C can appear arbitrarily.

Definition 5.2. Let $\mathscr{C}^p$ be the fragment of formulas in $\mathscr{C}$ with C the only appearing modality. Then let $\mathscr{L}_{alt}\mathscr{C}^p$ be the fragment consisting of Boolean combinations of formulas in $\mathscr{L}_{alt}$ and $\mathscr{C}^p$.

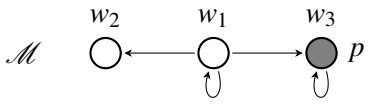
Definition 5.3. Let $\mathscr{C}_{alt}$ and $\mathscr{C}_{-a}$ for any $a \in A$ be defined by adding a new clause $C\varphi$ to $\mathscr{L}_{alt}$ and $\mathscr{L}_{-a}$'s context-free grammars.

For example, $\Box_a C\varphi$ is in $\mathscr{C}_{alt}$ but not in $\mathscr{L}_{alt}\mathscr{C}^p$. Note that $\mathscr{L}_{alt}\mathscr{C}^p \subseteq \mathscr{C}_{alt}$. Hence for non-collapsing results, using $\mathscr{L}_{alt}\mathscr{C}^p$ would be stronger.

Now we present the non-collapsing results. The situation with the 5 axiom is relatively simple. Even in the smaller fragment $\mathscr{L}_{alt}\mathscr{C}^p$ and even with the D axiom, 5 is still important.

Proposition 5.4. $\mathsf{CK5} \cap \mathscr{L}_{alt}\mathscr{C}^p$ *is not contained in* $\mathsf{CKD} \cap \mathscr{L}_{alt}\mathscr{C}^p$. *The formula* $\chi_{\mathsf{5}} = (\Diamond_a p \wedge \Diamond_a \neg p) \to \widehat{C}(p \wedge \widehat{C}\neg p)$ *is the witness.*

Proof. Clearly the following model proves the claim. All accessibility relations are the same, so we are not labeling the arrows.



□

For the 4 axiom we provide two non-collapsing results. First, in $\mathscr{L}_{alt}\mathscr{C}^p$, $\mathsf{CK4}$ does not collapse to CK when A is finite.

Proposition 5.5. $\mathsf{CK4} \cap \mathscr{L}_{alt}\mathscr{C}^p$ *is not contained in* $\mathsf{CK} \cap \mathscr{L}_{alt}\mathscr{C}^p$. *The witness is the formula* $\chi_{\mathsf{4}} = (\bigwedge_{x\in A\setminus\{a\}}(\Box_x\bot \wedge \Box_a\Box_x\bot) \wedge \Box_a p) \to Cp$.

Proof. The idea is essentially the same as the proof of the next proposition, Proposition 5.6. □

The formula in the previous proposition does not separate $\mathsf{CKD4}$ from CKD, as it is trivially valid in CKD for the reason that $\Box_b\bot$ is inconsistent. Here we provide a formula not in $\mathscr{L}_{alt}\mathscr{C}^p$ but in $\mathscr{C}_{alt}$ that separates $\mathsf{CKD4}$ from CKD, again assuming that A is finite.

Proposition 5.6. $\mathsf{CKD4} \cap \mathscr{C}_{alt}$ *is not contained in* $\mathsf{CKD} \cap \mathscr{C}_{alt}$. *The witness is the following formula* χ_{D4}*:*

$$\left(\bigwedge_{b\in A\setminus\{a\}} (\Box_b p \wedge C\Box_b p \wedge \Box_b\Box_a p \wedge C\Box_b\Box_a p) \wedge \Box_a p\right) \to Cp.$$

Proof. Clearly χ_{D4} is in $\mathscr{C}_{alt}$. To see that it is in $\mathsf{CK4}$, recall that the introduction axiom for common belief is

$$(C(\varphi \to E\varphi) \wedge E\Box_a\varphi) \to C\varphi.$$

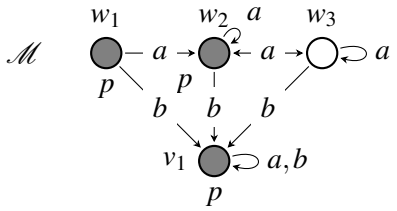
Note that $E\Box_a p$ is derivable from the antecedent of χ_{D4}, as $\Box_b\Box_a p$ for any $b \in A \setminus \{a\}$ is already in the antecedent, and $\Box_a\Box_a p$ follows from $\Box_a p$ by $\mathsf{4}$. Hence we only need $C(p \to Ep)$. It is enough to show CEp and in fact $C\Box_a p$, as for every $b \in A \setminus \{a\}$, $C\Box_b p$ is already in the antecedent of χ_{D4}. By the C-intro axiom again, we only need to derive $E\Box_a p$ and $C(\Box_a p \to E\Box_a p)$. We already dealt with $E\Box_a p$, so we are left with $C(\Box_a p \to E\Box_a p)$. For any $b \in A \setminus \{a\}$, $C(\Box_a p \to \Box_b\Box_a p)$ follows from $C\Box_b\Box_a p$, which is already in the antecedent of χ_{D4}. For the case of $C(\Box_a p \to \Box_a\Box_a p)$, we only need to necessitate $\mathsf{4}$.

Semantically, consider an arbitrary transitive model and a world u in the model. For any v that is reachable from u, there are only the following cases.

- Only R_a is used. Then $\Box_a p$ being true at u is enough to make p true at v, by transitivity.
- Only R_b is used for $b \in A \setminus \{a\}$. Since $\Box_b p$ is true at u, similarly p is true at v.
- The last step is in R_a, and the last non-a step is R_b. Then depending on if there is a step before the last R_b step, $\Box_b\Box_a p$ or $C\Box_b\Box_a p$ being true at u guarantees p's being true at v.
- The last step is in R_b for some $b \in A \setminus \{a\}$, and R_b is not the only relation used. Then $C\Box_b p$ being true at u guarantees that p is true at v.

Hence p is true at v no matter how v is reached from u. Thus, Cp is true at u.

Now to see that χ_{D4} is not in CKD, consider the following model.



This model $\mathscr{M}$ has all relations serial. Note that at any world, a b step moves you to v_1, which makes p and Cp true. Note that $\Box_a p$ is also true at w_1. Hence the antecedent is true at w_1. But clearly Cp is false at w_1, as w_3 is reachable but p is false at w_3. Thus, χ_{D4} is not in CKD. □

6 Discussion

In the introduction, we suggested that $\mathscr{L}_{alt}$ is sufficient to formalize agents' multi-agent epistemic reasoning in many cases, especially in games. As shown in Appendix B, this claim is substantial if we do not assume both introspection axioms $\mathsf{4}$ and $\mathsf{5}$, for then there is a loss of expressivity in moving from $\mathscr{L}$ to $\mathscr{L}_{alt}$. There remains the question of how widely it is true that $\mathscr{L}_{alt}$ is sufficient to formalize multi-agent epistemic reasoning. In concrete games, it may well be that there is a brute fact φ that is not expressible in $\mathscr{L}_{alt}$, yet for agents to do well in this game, they must reason about φ. For example, when twins are playing games, there seems to be motivation for them to introspect and reason about themselves. A formal study of this question would complement our work and contribute to answering the question of to what extent introspection axioms matter for multi-agent epistemic reasoning.

In § 4, we identified one fragment, the fragment of agent-nonrepeating formulas, with respect to which $\mathsf{S5}$ collapses to T. It is not too hard to see that the expressivity of this fragment is extremely

poor. For example, there is a bound on the modal depth of the formulas in this fragment when A is finite. It remains an open question whether there is an expressively more satisfying fragment with a natural syntactic definition that can collapse S5 to T.

In § 5, we noted that a full investigation of which fragments of $\mathscr{C}$ collapse which logics is left for future research. In particular, there are two obvious questions. First, when we are separating CKD4 from CKD, the formula we used is in $\mathscr{C}_{alt}$ but not in $\mathscr{L}_{alt}\mathscr{C}^p$. The question here is whether $\mathscr{L}_{alt}\mathscr{C}^p$ in fact collapses CLD4 to CKD. Second, we did not consider the case where A is infinite. We conjecture that when A is infinite, $\mathscr{L}_{alt}\mathscr{C}^p$ and perhaps even $\mathscr{C}_{alt}$ has the same collapsing power as $\mathscr{L}_{alt}$ does.

The main reason for the complexity of the problem with $\mathscr{C}$ is that C is implicitly not agent alternating and hence our unraveling technique does not apply directly. This motivates the formulation of an agent-alternating common belief operator. Indeed, we need many versions of agent-alternating common belief. For any subsets X and Y of A, we can define an operator ${}^XC^Y$ such that ${}^XC^Y p$ means that for any nonempty agent alternating finite sequence l of elements in A such that $l_1 \in X$ and $l_{len(l)} \in Y$, l_1 believes that l_2 believes that $\cdots$ $l_{len(l)}$ believes that p. The X and Y are here to make sure that ${}^XC^Y$ immediately scopes over and is immediately in the scope of the right modalities. For example, $\Box_a{}^XC^Y\Box_b p$ would be agent alternating iff $a \notin X$ and $b \notin Y$. It is not hard to see that the techniques in § 3 are enough to deal with these operators, through a translation to infinitary languages allowing infinite conjunctions, as our agent-alternating bisimulation families preserves truth values of even infinitary formulas.

Finally, our project can be naturally extended to any extension of the multi-agent doxastic/epistemic language. Natural candidates include languages with dynamic operators, probability operators, or non-standard knowledge operators. The central question to ask in each case is this: what would be a natural agent-alternating fragment or a fragment sufficient for the intended application of those languages, and how does restricting to this fragment affect the landscape of logics? We believe that this type of question will generate interesting results and deepen our understanding of the realm of epistemic logics.

References

[1] Robert J. Aumann (1976): *Agreeing to disagree*. The Annals of Statistics 4(6), pp. 1236–1239, doi:10.1214/aos/1176343654.

[2] Robert J. Aumann (1999): *Interactive epistemology I: Knowledge*. International Journal of Game Theory 28(3), pp. 263–300, doi:10.1007/s001820050111.

[3] B. Douglas Bernheim (1984): *Rationalizable strategic behavior*. Econometrica 52(4), pp. 1007–1028, doi:10.2307/1911196.

[4] Giacomo Bonanno (2002): *Modal logic and game theory: two alternative approaches*. Risk, Decision and Policy 7(3), pp. 309–324, doi:10.1017/s1357530902000704.

[5] Ronald Fagin, Joseph Y. Halpern, Yoram Moses & Moshe Y. Vardi (2003): *Reasoning About Knowledge*. MIT Press.

[6] Liangda Fang, Kewen Wang, Zhe Wang & Ximing Wen (2018): *Knowledge compilation in multi-agent epistemic logics*. arXiv: 1806.10561v2.

[7] James Garson (2018): *Modal Logic*. In Edward N. Zalta, editor: The Stanford Encyclopedia of Philosophy, fall 2018 edition, Metaphysics Research Lab, Stanford University.

[8] John Geanakoplos (1989): *Game theory without partitions, and applications to speculation and consensus*. Cowles Foundation Discussion Papers 914, Cowles Foundation for Research in Economics, Yale University.

[9] James Hales (2016): *Quantifying over epistemic updates*. Ph.D. thesis, The University of Western Australia.

[10] James Hales, Tim French & Rowan Davies (2012): *Refinement quantified logics of knowledge and belief for multiple agents*. In: *Advances in Modal Logic, Volume 9*, pp. 317–338.

[11] Joseph Y Halpern & Gerhard Lakemeyer (2001): *Multi-agent only knowing*. *Journal of Logic and Computation* 11(1), pp. 41–70, doi:10.1093/logcom/11.1.41.

[12] Joseph Y. Halpern & Richard A. Shore (2004): *Reasoning about common knowledge with infinitely many agents*. *Information and Computation* 191(1), pp. 1–40, doi:10.1016/j.ic.2004.01.003.

[13] Jaakko Hintikka (1962): *Knowledge and Belief: An Introduction to the Logic of the Two Notions*. Cornell University Press.

[14] Xiao Huang, Biqing Fang, Hai Wan & Yongmei Liu (2018): *A general multi-agent epistemic planner based on higher-order belief change*. arXiv: 1806.11298v2.

[15] Mamoru Kaneko (2002): *Epistemic logics and their game theoretic applications: Introduction*. *Economic Theory* 19(1), pp. 7–62, doi:10.1007/s001990100202.

[16] Gerhard Lakemeyer & Yves Lespérance (2012): *Efficient reasoning in multiagent epistemic logics*. In: *Proceedings of the 20th European Conference on Artificial Intelligence*, ECAI'12, pp. 498–503, doi:10.3233/978-1-61499-098-7-498.

[17] Philippe Lamarre & Yoav Shoham (1994): *Knowledge, certainty, belief, and conditionalisation (abbreviated version)*. In Jon Doyle, Erik Sandewall & Pietro Torasso, editors: *Principles of Knowledge Representation and Reasoning*, The Morgan Kaufmann Series in Representation and Reasoning, pp. 415–424, doi:10.1016/b978-1-4832-1452-8.50134-2.

[18] Harvey Lederman (2015): *People with common priors can agree to disagree*. *Review of Symbolic Logic* 8(1), pp. 11–45, doi:10.1017/s1755020314000380.

[19] Wolfgang Lenzen (1978): *Recent work in epistemic logic*. *Acta Philosophica Fennica* 30(2), pp. 1–219.

[20] Qiang Liu & Yongmei Liu (2018): *Multi-agent epistemic planning with common knowledge*. In: *Proceedings of the 27th International Joint Conference on Artificial Intelligence, IJCAI-18*, pp. 1912–1920, doi:10.24963/ijcai.2018/264.

[21] John-Jules Ch Meyer & Wiebe van der Hoek (1995): *Epistemic Logic for AI and Computer Science*. Cambridge University Press, doi:10.1017/cbo9780511569852.

[22] Martin J. Osborne & Ariel Rubinstein (1994): *A Course in Game Theory*. MIT Press.

[23] Rohit Parikh & Paul Krasucki (1992): *Levels of knowledge in distributed systems*. *Sadhana* 17(1), pp. 167–191, doi:10.1007/bf02811342.

[24] Robert Stalnaker (1994): *On the evaluation of solution concepts*. *Theory and Decision* 37(1), pp. 49–73, doi:10.1007/bf01079205.

[25] Moshe Y. Vardi (1985): *A model-theoretic analysis of monotonic knowledge*. In: *Proceedings of the 9th International Joint Conference on Artificial Intelligence, IJCAI-85*, pp. 509–512.

[26] Arnis Vilks (1999): *Knowledge of the game, relative rationality, and backwards induction without counterfactuals*. Working Paper no.25, Leipzig Graduate School of Management.

A Rationalizability as agent-alternating common belief of rationality

In this appendix, we sketch a proof that rationality plus agent-alternating common belief of rationality, in which one need not believe that oneself is rational, is enough for all agents to play their rationalizable strategies. This is already implicit in one of the very first definitions of rationalizability by Bernheim [3], where he carefully stipulated that agents cannot formulate "conjectures" about themselves in systems of beliefs that rationalize their actions. We make this more explicit by using modal logic and Kripke models of games in the style of [24] and [4].

We utilize typical notation in game theory in this appendix. Let $G = \langle \{S_a\}_{a\in A}, \{U_a\}_{a\in A}\rangle$ be a strategic form game: for all $a \in A$, S_a is a finite set and U_a is a function from $\Pi_{a\in A} S_a$ to $\mathbb{R}$, which then naturally extends to $\Delta\Pi_{a\in A} S_a$.

Following [4], we first pick a set of distinct proposition letters $\{r_a \mid a \in A\} \subseteq \mathsf{Prop}$. Then a model $\mathscr{M}$ of G is a tuple $\langle W, \{R_a\}_{a\in A}, \{P_a\}_{a\in A}, \{\sigma_a\}_{a\in A}, V\rangle$ satisfying the following properties.

- W is a finite set.
- For any $a \in A$, R_a is serial binary relation on W.
- For any $a \in A$, P_a is a function from W to probability distributions on W such that for any $w \in W$, $P_{a,w}(R_a(w)) = 1$.
- For any $a \in A$, σ_a is a function from W to S_a.
- V is a function from Prop to $\wp(W)$.
- For any $a \in A$, $w \in V(r_a)$ iff $\sigma_a(w)$ is a best response to what a believes her opponents play. Formally, this condition is

$$\forall s_a \in S_a, \sum_{w'\in R_a(w)} P_{a,w}(w') U_a(\sigma_a(w), \sigma_{-a}(w')) \geq \sum_{w'\in R_a(w)} P_{a,w}(w') U_a(s_a, \sigma_{-a}(w')).$$

To formulate agent-alternating common belief of rationality, for each nonempty finite sequence l of elements in A, let ρ_s be the formula $\Box_{l_1}\Box_{l_2}\cdots\Box_{l_{len(l)-1}} r_{l_{len(l)}}$. For example, $\rho_{\langle a\rangle} = r_a$ and $\rho_{\langle a,b,a\rangle} = \Box_a\Box_b r_a$. Then let Γ be the set of ρ_l such that l is agent-alternating: for all $i = 1 \ldots len(l) - 1$, $l_i \neq l_{i+1}$. This Γ then encodes agent-alternating common belief of rationality.

With these definitions, it easily follows that if $\mathscr{M}, w \models \Gamma$, then for each $a \in A$, $\sigma_a(w)$ is a strategy that survives the iterated elimination of strictly dominated strategies. To show this, we can simply adapt the proof in [24]. For each $a \in A$, collect $\sigma_a(w')$ in Q_a for each $w' \in W$ such that there is an agent-alternating path from w to w' with the last move not using R_a. Then it is not hard to see that for any $a \in A$ and $q_a \in Q_a$, q_a is not strictly dominated by any strategy in ΔQ_a assuming any opponent $b \in A \setminus \{a\}$ only plays strategies in Q_b. Indeed, given that q_a is in Q_a, by definition there is $w' \in W$ and $w_1, w_2, \ldots, w_n$ and $a_1, a_2, \cdots, a_n \neq a$ such that $w_1 = w$, $w_n = w'$, $a_n \neq a$, and for all $i = 1 \ldots n-1$, $w_i R_i w_{i+1}$ and $a_i \neq a_{i+1}$. Then for each $w'' \in R_a(w')$, $\sigma_{-a}(w'') \in Q_{-a}$ as w'' is reachable from w using $a_1, a_2, \cdots, a_n, a$ which is still an alternating sequence. Note also that $\mathscr{M}, w \models \Gamma$ and in particular $\mathscr{M}, w \models \Box_{a_1}\Box_{a_2}\cdots\Box_{a_{n-1}} r_{a_n}$. Hence $\mathscr{M}, w' \models r_a$, and q_a is the best response to the mixture m_{-a} of $\langle \sigma_{-a}(w'') \mid w'' \in R_a(w')\rangle$ using $P_{a,w'}$, which is a mixture of Q_{-a} since each $\sigma_{-a}(w'') \in Q_{-a}$. Thus q_a is not strictly dominated by any mixture m_a of Q_a: if m_a strictly dominates q_a on any $q_{-a} \in Q_{-a}$, then m_a strictly dominates q_a on m_{-a}, and then q_a is not the best response to m_{-a}, a contradiction. Thus each Q_a survives each stage of elimination. Now for each $a \in A$, $\sigma_a(w) \in Q_a$, since we can use the trivially agent-alternating path $w_1 = w$ and $a_1 = b \in A \setminus \{a\}$ (recall that $|A| > 1$). Hence the strategy $\sigma_q(w)$ played at w is a strategy that survives iterated elimination of strictly dominated strategies.

We may also express agents' belief that their opponents' actions are independent by proposition letters. Then agent-alternating common belief that agents believe that their opponents' actions are independent can be expressed using modal formulas. To this end, first fix another set $\{r'_a \mid a \in A\} \subseteq \mathsf{Prop}$ of distinct proposition letters such that $\{r'_a \mid a \in A\} \cap \{r_a \mid a \in A\} = \varnothing$. The intended interpretation of r'_a is that a believes that her opponents' actions are independent. Hence we require that for any $a \in A$, $w \in V(r'_a)$ iff $P_{a,w}(\{w' \in R_a(w) \mid \sigma_{-a}(w') = s_{-a}\}) = \Pi_{b\in A\setminus\{a\}} P_{a,w}(\{w' \in R_a(w) \mid \sigma_b(w') = s_{-a}(b)\})$ for any $s_{-a} \in S_{-a}$. Then similar to the definition of ρ_l, for any alternating sequence l of elements in A, we

define ρ'_l with the trailing proposition letter being $r'_{l_{len(l)}}$, and we let Γ' be the set of all such ρ'_l's. With this setup, it is not hard to see, using the same strategy as above, that if $\mathscr{M},w \models \Gamma \cup \Gamma'$, then for any $a \in A$, $\sigma_a(w)$ is a rationalizable strategy for a.

B Expressivity of $\mathscr{L}_{alt}$

In this appendix we study the influence of the introspection axioms on the expressivity of $\mathscr{L}_{alt}$. We first define finite agent-alternating bisimulations so we can give a more quantitative analysis of expressivity.

Definition B.1. By induction on n, let binary relations $\{\leftrightarrows^n_{-a}\}_{a\in A}$ be defined on all pointed models as follows:

- $\mathscr{M},u \leftrightarrows^0_{-a} \mathscr{N},v$ iff $V^{\mathscr{M}}(u) = V^{\mathscr{N}}(v)$.
- $\mathscr{M},u \leftrightarrows^{n+1}_{-a} \mathscr{N},v$ iff
  - (Atom) $V^{\mathscr{M}}(u) = V^{\mathscr{M}}(v)$ and
  - $\forall b \in A \setminus \{a\}$
    - (b-zig) $\forall x \in R_b^{\mathscr{M}}(u)\ \exists y \in R_b^{\mathscr{N}}(v)\ \mathscr{M},x \leftrightarrows^n_{-b} \mathscr{N},y$ and
    - (b-zag) $\forall y \in R_b^{\mathscr{N}}(v)\ \exists x \in R_b^{\mathscr{M}}(u)\ \mathscr{M},x \leftrightarrows^n_{-b} \mathscr{N},y$.

Then let $\leftrightarrows^0_{alt}$ be defined in the same way as $\leftrightarrows^0_{-a}$ for any $a \in A$, and define $\leftrightarrows^{n+1}_{alt}$ by

- (Atom) $V^{\mathscr{M}}(u) = V^{\mathscr{M}}(v)$ and
- $\forall b \in A$
  - (b-zig) $\forall x \in R_b^{\mathscr{M}}(u)\ \exists y \in R_b^{\mathscr{N}}(v)\ \mathscr{M},x \leftrightarrows^n_{-b} \mathscr{N},y$ and
  - (b-zag) $\forall y \in R_b^{\mathscr{N}}(v)\ \exists x \in R_b^{\mathscr{M}}(u)\ \mathscr{M},x \leftrightarrows^n_{-b} \mathscr{N},y$.

Then for $x \in A \cup \{alt\}$, let $\leftrightarrows^\omega_x$ be the intersection of $\{\leftrightarrows^n_x\}_{n\in\mathbb{N}}$.

Proposition B.2. *For all $n \in \mathbb{N}$ and $a \in A$, $\leftrightarrows^n_{-a}$ is an equivalence relation. Consequently, $\leftrightarrows^n_{alt}$ is an equivalence relation for all $n \in \mathbb{N}$.*

Theorem B.3. *For any $n \in \mathbb{N}$ and pointed models $\mathscr{M},u$ and $\mathscr{N},v$:*

- $\mathscr{M},u \equiv_{\mathscr{L}^n_{-a}} \mathscr{N},v$ *if* $\mathscr{M},u \leftrightarrows^n_{-a} \mathscr{N},v$;
- $\mathscr{M},u \equiv_{\mathscr{L}^n_{alt}} \mathscr{N},v$ *if* $\mathscr{M},u \leftrightarrows^n_{alt} \mathscr{N},v$;
- $\mathscr{M},u \equiv_{\mathscr{L}_{-a}} \mathscr{N},v$ *if* $\mathscr{M},u \leftrightarrows^\omega_{-a} \mathscr{N},v$;
- $\mathscr{M},u \equiv_{\mathscr{L}_{alt}} \mathscr{N},v$ *if* $\mathscr{M},u \leftrightarrows^\omega_{alt} \mathscr{N},v$.

Let $\leftrightarrows^n$ denote the usual n-bisimulation relation and $\leftrightarrows^\omega$ the intersection of $\{\leftrightarrows^n\}_{n\in\mathbb{N}}$.

Now we can present our main results in this appendix. First, with both introspection axioms, there is no loss of expressivity at each modal depth in moving from $\mathscr{L}$ to $\mathscr{L}_{alt}$.

Proposition B.4. *Letting* **K45** *be the class of pointed models where each R_a is transitive and Euclidean, $\leftrightarrows^n \cap \mathbf{K45}^2 = \leftrightarrows^n_{alt} \cap \mathbf{K45}^2$ for all $n \in \mathbb{N}$.*

Proof. If $|A| = 1$, then $\leftrightarrows^1 \cap \mathbf{K45}^2 = \leftrightarrows^2 \cap \mathbf{K45}^2 = \leftrightarrows^n \cap \mathbf{K45}^2$ for all $n \geq 1$. It is also easy to see that $\leftrightarrows^1_{alt} = \leftrightarrows^1$. Hence $\leftrightarrows^1_{alt} \cap \mathbf{K45}^2 = \leftrightarrows^n \cap \mathbf{K45}^2$ for all n. The required proposition follows immediately.

Now we assume $|A| > 1$ and prove the claim by induction on n. The case for $n = 0$ is trivial. Now suppose $\leftrightarrows^n \cap \mathbf{K45}^2 = \leftrightarrows^n_{alt} \cap \mathbf{K45}^2$, and let us show that the claim is true for $n+1$. The left-to-right subset relation is trivial. Hence let us pick two arbitrary pointed models $\mathscr{M},u$ and $\mathscr{N},v$ in **K45** such

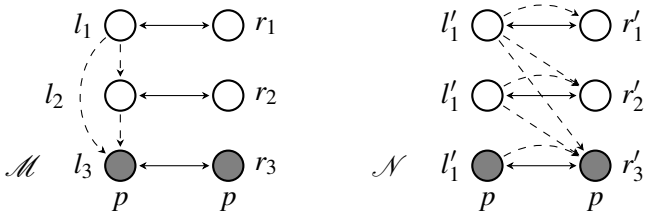
that $\mathscr{M}, u \leftrightarrows_{alt}^{n+1} \mathscr{N}, u$. Now we need to show that $\mathscr{M}, u \leftrightarrows^{n+1} \mathscr{N}, v$. That they have the same atomic valuation is trivial. Now pick an arbitrary $b \in A$ and u' such that $uR_b^{\mathscr{M}} u'$ in $\mathscr{M}$. Our goal is then to find a v' such that $vR_b^{\mathscr{N}} v'$ and $\mathscr{M}, u' \leftrightarrows^n \mathscr{N}, v'$. Pick some $a \in A \setminus \{b\}$ (note that we assumed that $|A| > 1$) so that $b \in A \setminus \{a\}$. By the definition of $\leftrightarrows_{alt}^{n+1}$, $\mathscr{M}, u \leftrightarrows_{-a}^{n+1} \mathscr{N}, v$. Then we obtain a v' such that $vR_b^{\mathscr{N}} v'$ and $\mathscr{M}, u' \leftrightarrows_{-b}^{n} \mathscr{N}, v'$. Now we show that this v' is what we need: $\mathscr{M}, u' \leftrightarrows^n \mathscr{N}, v'$. By the induction hypothesis, it is enough to show that $\mathscr{M}, u' \leftrightarrows_{alt}^{n} \mathscr{N}, v'$.

Thus, pick an arbitrary $x \in A$. We need to show that $\mathscr{M}, u' \leftrightarrows_{-x}^{n} \mathscr{N}, v'$. The case for the atomic valuation is again trivial. Now we need to show y-zig and y-zag for all $y \in A \setminus \{x\}$. When $y \neq b$, they are part of the definition of $\leftrightarrows_{-b}^{n}$, which holds between $\mathscr{M}, u'$ and $\mathscr{N}, v'$. Hence we are left with the case where $y = b$. For b-zig, pick an arbitrary u'' such that $u'R_b^{\mathscr{M}} u''$. Recall that $uR_b^{\mathscr{M}} u'$ and R_b is transitive. Hence $uR_b^{\mathscr{M}} u''$. Applying $\mathscr{M}, u \leftrightarrows_{-a}^{n+1} \mathscr{N}, v$, we obtain v'' such that $vR_b^{\mathscr{N}} v''$ and $\mathscr{M}, u'' \leftrightarrows_{-b} \mathscr{N}, v''$. But $R_b^{\mathscr{N}}$ is Euclidean and $vR_b^{\mathscr{N}} v'$ too. Hence $vR^{\mathscr{N}} v''$. Thus this v'' witnesses the b-zig clause for $\mathscr{M}, u' \leftrightarrows_{-x}^{n} \mathscr{N}, v'$. b-zag is shown symmetrically, where the transitivity of $R_b^{\mathscr{N}}$ and the Euclideanness of $R_b^{\mathscr{M}}$ are used. The zag clause for $\mathscr{M}, u \leftrightarrows^{n+1} \mathscr{N}, v$ is also shown symmetrically. □

However, if we consider frame classes corresponding to the modal logic cube as in Figure 1, having both introspection properties is necessary.

Proposition B.5. *Letting* **S4** *(resp.* **KD5**, **B***) be the class of pointed models where each R_a is reflexive and transitive (resp. serial and Euclidean, reflexive and symmetrical), we have $\leftrightarrows_{alt}^{\omega} \cap \mathbf{S4}^2 \not\subseteq \leftrightarrows^2 \cap \mathbf{S4}^2$, $\leftrightarrows_{alt}^{\omega} \cap \mathbf{KD5}^2 \not\subseteq \leftrightarrows^2 \cap \mathbf{KD5}^2$, and $\leftrightarrows_{alt}^{\omega} \cap \mathbf{B}^2 \not\subseteq \leftrightarrows^2 \cap \mathbf{B}^2$.*

Proof. The following two models deal with the **S4** case. Reflexive loops are omitted. The dashed arrows represent relations for a, and the solid ones represent relations for all agents in A beside a.

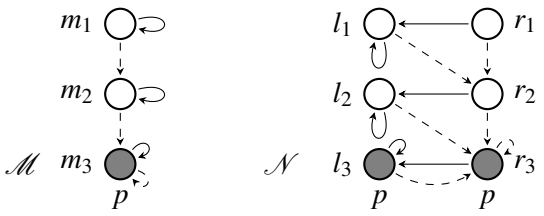


Then we have an agent-alternating family $\{\leftrightarrows_a\}_{a \in A \cup \{alt\}}$ of bisimulations. For any $i \in \{1,2,3\}$ and $b \in A \setminus \{a\}$:

- $l_i \leftrightarrows_a l'_i, r'_i$ and $r_i \leftrightarrows_a l'_i, r'_i$;
- $l_i \leftrightarrows_b l'_i$ and $r_i \leftrightarrows_b r'_i$.

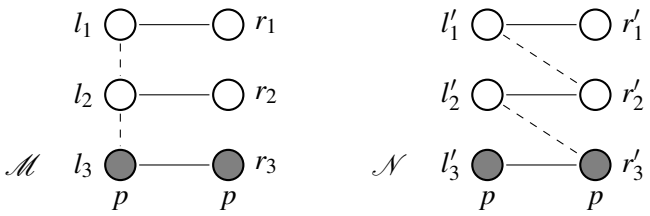
Essentially the nodes on the same level are connected by $\leftrightarrows_a$, and the left column in $\mathscr{M}$ is connected to the left column of $\mathscr{N}$ by $\leftrightarrows_b$, and similarly the right column in $\mathscr{M}$ is connected to the right column of $\mathscr{N}$ by $\leftrightarrows_b$. Finally it enough to just connect l_1 with l'_1 by $\leftrightarrows_{alt}$. Then it is not hard to check that $\{\leftrightarrows_a\}_{a \in A \cup \{alt\}}$ is indeed an agent-alternating bisimulation family. By a simple induction, this clearly implies that $\mathscr{M}, l_1 \leftrightarrows_{alt}^{\omega} \mathscr{N}, l'_1$. But of course $\mathscr{M}, l_1 \not\leftrightarrows^2 \mathscr{N}, l'_1$ since $\mathscr{M}, l_1 \models \Diamond_a \Diamond_a p$ but $\mathscr{N}, l'_1 \not\models \Diamond_a \Diamond_a p$.

The following two models deal with the **KD5** case.



This case is easier. For each $i \in \{1,2,3\}$, $m_i \leftrightarrows_a l_i, r_i$ and $m_i \leftrightarrows_b r_i$. Then connecting m_1 with r_1 by $\leftrightarrows_{alt}$, we have an agent-alternating bisimulation family. Hence $\mathscr{M}, m_1 \leftrightarrows_{alt} \mathscr{N}, r_1$. However, we have $\mathscr{M}, m_1 \models \Diamond_a \Diamond_a p$ and $\mathscr{N}, r_1 \not\models \Diamond_a \Diamond_a p$.

Finally, the following two models deal with the **B** case. Again, the reflexive loops are omitted from the diagram. The agent-alternating bisimulation family and the formula to refute $\leftrightarrows^2$ we need to use are the same as we used in the **S4** case.



□