Moorean Phenomena in Epistemic Logic*

Wesley H. Holliday and Thomas F. Icard, III

Department of Philosophy
Stanford University
Stanford, California, USA

Abstract

A well-known open problem in epistemic logic is to give a syntactic characterization of the *successful* formulas. Semantically, a formula is successful if and only if for any pointed model where it is true, it remains true after deleting all points where the formula was false. The classic example of a formula that is not successful in this sense is the "Moore sentence" $p \wedge \neg\Box p$, read as "p is true but you do not know p." Not only is the Moore sentence unsuccessful, it is *self-refuting*, for it never remains true as described. We show that in logics of knowledge and belief for a single agent (extended by **S5**), Moorean phenomena are the source of all self-refutation; moreover, in logics for an *introspective* agent (extending **KD45**), Moorean phenomena are the source of all unsuccessfulness as well. This is a distinctive feature of such logics, for with a non-introspective agent or multiple agents, non-Moorean unsuccessful formulas appear. We also consider how successful and self-refuting formulas relate to the *Cartesian* and *learnable* formulas, which have been discussed in connection with Fitch's "paradox of knowability." We show that the Cartesian formulas are exactly the formulas that are not *eventually* self-refuting and that not all learnable formulas are successful. In an appendix, we give syntactic characterizations of the successful and the self-refuting formulas.

Keywords: epistemic logic, successful formulas, Moore sentence

1 Introduction

According to the epistemic interpretation of modal logic, the points in a modal model represent ways the world might be, consistent with an agent's information. In this context, "learning" a formula amounts to eliminating those points in the model where the formula is false. The resulting submodel represents the agent's information state after learning has occurred. Some formulas—though not all—remain true whenever they are learned. A well-known open problem [12,4,5,6,7,2] in epistemic logic is to give a syntactic characterization of these *successful* formulas. Partial results have been obtained (Section 2), but a full solution has proven elusive.

The classic example of an *unsuccessful* formula is the Moore sentence $p \wedge \neg\Box p$, read as "p is true but you do not know p." This example is a second-person variation of G.E. Moore's famous puzzle [16] involving the paradoxical first-person assertion, "p is true but I do not believe p." Hintikka devoted a chapter of his seminal 1962 monograph *Knowledge and Belief* [13] to an analysis of such sentences, including the second-person Moore sentence. Hintikka observed its unsuccessfulness as follows: "You may come to know that what I said *was* true, but saying it in so many words

* In L. Beklemishev, V. Goranko and V. Shehtman, eds., *Advances in Modal Logic*, Volume 8, 178-199, College Publications, 2010. Minor changes to the published version have been made in this version.

has the effect of making what is being said false" [13, p. 69]. Yet the formal question of unsuccessfulness did not arise for Hintikka. Only with the advent of Dynamic Epistemic Logic (see, e.g., [7]) and the idea of learning as model reduction have the Moorean phenomena and unsuccessfulness been formally related.

The reason the Moore sentence is unsuccessful is that it can only be true in a model if p is true at some point and false at some other point accessible from the first. When the agent learns the sentence, all points where p is false are eliminated from the model, including all witnesses for $\neg\Box p$, so the sentence becomes false. This shows that the Moore sentence is not only unsuccessful, it is *self-refuting*, for it always becomes false when learned. Related to the self-refuting property of the Moore sentence is the fact that the sentence cannot be known. Indeed, the Moore sentence is at the root of Fitch's famous "paradox of knowability" [11,3,9]: if there is an unknown truth, then there is an unknowable truth. For if p is true but unknown, then the Moore sentence $p \wedge \neg\Box p$ is true, but the Moore sentence cannot be known, because $\Box(p \wedge \neg\Box p)$ is inconsistent with standard assumptions about knowledge.

While the Moore sentence is conspicuously unsuccessful, other unsuccessful formulas are less conspicuous. The formula $\neg\,(p \vee q) \vee (p \wedge (\Box p \vee \Diamond q))$ is also unsuccessful, as we show in Example 5.12 below, but is the reason Moorean? While on the surface this formula looks unlike a Moore sentence, it is in fact possible to transform the formula to reveal its Moorean character. Indeed, we will prove that for a wide range of logics, such a transformation is possible for every unsuccessful formula. However well-disguised, their nature is always Moorean.

In Section 2 we establish notation, give the definitions of *successful*, *self-refuting*, etc., and review what is already known in the literature on successful formulas. In Section 3 we show that in logics of knowledge and belief for a single agent (extended by **S5**), Moorean phenomena are the source of all self-refutation; moreover, in logics for an *introspective* agent (extending **KD45**), Moorean phenomena are the source of all unsuccessfulness as well. This is a distinctive feature of such logics, for as we show in Section 4, in logics for a non-introspective agent or multiple agents, non-Moorean unsuccessful formulas appear. Finally, in Section 5 we relate successful and self-refuting formulas to the *Cartesian* and *learnable* formulas, which have been discussed in connection with Fitch's paradox, and to the *informative*, *eventually* self-refuting, *super-successful* formulas, which we introduce here. In Appendix A we give syntactic characterizations of the successful and the self-refuting formulas.

2 Preliminaries and Previous Results

Throughout we work with a fixed, unimodal language with $\Box$ and its dual $\Diamond$, and an infinite set Prop of propositional variables. The expression $\Diamond^+\varphi$ abbreviates $\Diamond\varphi \wedge \varphi$. We use the standard semantics of modal logic, where a model is a triple $\langle W, R, V\rangle$ with W any set of points, $R \subseteq W \times W$ any relation, and $V : \mathsf{Prop} \to \wp(W)$ a valuation function. A pointed model is a pair $\mathcal{M}, w$ with $\mathcal{M}$ a model and $w \in W$. The satisfaction relation $\models$ between pointed models and formulas is defined as usual.

It is typical to take **KD45** as a logic of belief and **S5** as a logic of knowledge. Apart from the assumptions that one does not believe or know anything inconsistent, and that what is known is true, these logics assume *positive* (axiom **4**) and

negative (axiom **5**) *introspection*: if one believes something, then one believes that one believes it; if one does not believe something, then one believes that one does not believe it; and *mutatis mutandis* for knowledge. Most work on successful formulas has assumed **S5**. Since our results apply to a wider range of logics, we assume through Section 3 that we are working with at least **KD45**, so all of our models will be serial, transitive, and Euclidean. We call such models *quasi-partitions*.

Where $\mathcal{M}'$ is a submodel (not necessarily proper) of $\mathcal{M}$, we write $\mathcal{M}' \subseteq \mathcal{M}$. Rather than studying formulas preserved under arbitrary submodels, we will study formulas preserved under a special way of taking submodels, as in the following.

Definition 2.1 Given a model $\mathcal{M} = \langle W, R, V\rangle$, the *relativization of* $\mathcal{M}$ *to* φ is the (possibly empty) submodel $\mathcal{M}_{|\varphi} = \langle W_{|\varphi}, R_{|\varphi}, V_{|\varphi}\rangle$ of $\mathcal{M}$, where $W_{|\varphi} = \{w \in W : \mathcal{M}, w \vDash \varphi\}$, $R_{|\varphi}$ is R restricted to $W_{|\varphi}$, and $V_{|\varphi}(p) = V(p) \cap W_{|\varphi}$.

Definition 2.2 A formula φ is *successful* (in logic **L**) iff for every pointed model (of **L**), $\mathcal{M}, w \vDash \Diamond^+\varphi$ implies $\mathcal{M}_{|\varphi}, w \vDash \varphi$. A formula is *unsuccessful* (in **L**) iff it is not successful. A formula is *self-refuting* (in **L**) iff for every pointed model (of **L**), $\mathcal{M}, w \vDash \Diamond^+\varphi$ implies $\mathcal{M}_{|\varphi}, w \nvDash \varphi$. [1]

In the standard definitions of successful and self-refuting formulas, where **L** is assumed to be **S5**, the precondition only requires that φ be true at w. Since we are also working with **KD45**, we additionally require that φ be true at an accessible point, so that $\mathcal{M}_{|\varphi}$ is a quasi-partition provided $\mathcal{M}$ is. Our definition reduces to the standard one in the case of **S5**. In either case, unsatisfiable formulas are self-refuting and successful. In the case of **KD45**, a satisfiable formula such as $p \wedge \Box\neg p$, read as "p is true but you believe $\neg p$," is self-refuting and successful, since $\Diamond^+ (p \wedge \Box\neg p)$ is unsatisfiable. While it may be more intuitive to require the satisfiability of $\Diamond^+\varphi$ for a successful φ, we will follow the standard definition in not requiring satisfiability.

The following lemma relates success and self-refutation across different logics.

Lemma 2.3 *Let* **L** *be a sublogic of* **L**′*. If* φ *is unsuccessful in* **L**′*, then* φ *is unsuccessful in* **L***, and if* φ *is self-refuting in* **L***, then* φ *is self-refuting in* **L**′*.*

Proof. Immediate from Definition 2.2, given that models of **L**′ are models of **L**.□

An obstacle to giving a simple syntactic characterization of the set of successful formulas is its lack of closure properties. Successful formulas are not closed under negation (take $\neg p \vee \Box p$), conjunction (take p and $\neg\Box p$), or implication (use $\varphi \to \bot$) [6]. We show in Proposition 5.11 that they are also not closed under disjunction. Conversely, if a negated formula is successful, the unnegated formula may be unsuccessful (take $\neg (p \wedge \Diamond\neg p)$), if a conjunction is successful, some of the conjuncts may be unsuccessful (take $(p \wedge \Diamond\neg p) \wedge \neg p$), and if a disjunction is successful, some or even all of the disjuncts may be unsuccessful (Proposition 5.9 and Example 5.2).

By contrast, the formulas preserved under arbitrary submodels are well-behaved. The following result was proved independently by van Benthem and Visser [20,8]. A formula is *universal* iff it can be constructed using only literals, $\wedge$, $\vee$, and $\Box$.

[1] The term 'successful' is used by Gerbrandy [12], by analogy with the success postulate of belief revision, while the term 'self-refuting' is used by van Benthem [3]. Self-refuting formulas have also been called *strongly unsuccessful* [2].

Theorem 2.4 *A formula is preserved under submodels (of all relational models) iff it is equivalent (in* **K***) to a universal formula.*

Similarly, a formula is preserved under model extensions iff it is equivalent to an *existential* formula, constructed using only literals, $\wedge$, $\vee$, and $\Diamond$ [8].

Lemma 2.3 and the right-to-left direction of Theorem 2.4 give the following.

Corollary 2.5 *Universal formulas are successful in any normal modal logic.*

As noted by van Benthem [4], for any model $\mathcal{M}$ and formula φ, there is a universal formula φ' such that $\mathcal{M}_{|\varphi} = \mathcal{M}_{|\varphi'}$.[2] However, this result assumes the relation for $\mathcal{M}$ is at least a quasi-partition. For example, given the model in Figure 1, where the relation is not Euclidean, there is no universal formula ψ such that $\mathcal{M}_{|\Diamond p} = \mathcal{M}_{|\psi}$. This is symptomatic of the fact, established in Section 4, that in logics without both axioms **4** and **5**, there are non-Moorean sources of unsuccessfulness.

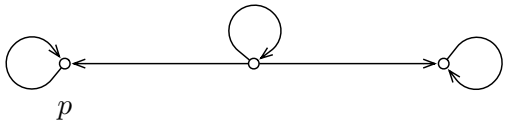

Fig. 1.

Gerbrandy [12] proved a proposition similar to Corollary 2.5, but in a slightly different formal context and with $\Diamond$ only as a defined operator. He showed that if all instances of $\Box$ are within the scope of an even number of negations, then the formula is successful. However, the converse of Corollary 2.5 does not hold, even in the case of formulas containing no nested modal operators. For example, the formula $\Diamond p \vee p$ is successful in **K**, but it is not equivalent to any universal formula. Yet other connections with universal formulas do hold. Qian [18] showed that a formula containing no nested modal operators is successful in **K** iff it can be transformed using a certain algorithm into a universal formula; moreover, a "homogeneous" formula, containing no nested modal operators and no propositional variables outside the scope of modal operators, is successful in **K** iff it is equivalent to a universal formula. The restriction to **K** is essential, for the formula $\Diamond p$ is successful in extensions of **KD45** (not in **K**), but it is not equivalent to any universal formula.

As far as we know, there are no other published results on the syntax of successful formulas in the basic modal language. However, successful formulas are often studied in the more general context of Dynamic Epistemic Logic (DEL) [1,12,7], where model changing operations are formalized in the logic itself by adding dynamic operators to the language. In the simplest fragment of DEL extending **S5**, known as Public Announcement Logic (PAL), there are sentences of the form $[\varphi]\psi$, read as "after the *announcement* of φ, ψ is true," for which satisfaction is defined:

$$\mathcal{M}, w \vDash [\varphi]\psi \text{ iff } \mathcal{M}, w \vDash \varphi \text{ implies } \mathcal{M}_{|\varphi}, w \vDash \psi.$$

An advantage of the PAL setting is that it allows for simple definitions of success and self-refutation [6]. Successful formulas are those for which $[\varphi]\varphi$ is valid (in **S5**),

[2] The result in [4] is given for the case of multi-agent epistemic logic with a common knowledge operator, assuming $\mathcal{M}$ is finite. The proof for the single-agent case is trivial and requires no assumption of finiteness.

and self-refuting formulas are those for which $[\varphi]\neg\varphi$ is valid (in **S5**). While we will not deal directly with this language, our work will apply indirectly to PAL given the following result, due to Plaza [17], which relates PAL to the basic modal language.

Theorem 2.6 *Every formula in the language of* PAL *is equivalent* (*over partitions*) *to a formula in the basic modal language.*

Direct results on successful formulas in PAL with multiple knowledge modalities and the *common knowledge* operator have been obtained by van Ditmarsch and Kooi in [6], which also presents an analysis of the role of (un)successful formulas in a variety of scenarios involving information change.

Another benefit of the PAL definitions of successful and self-refuting is that they give an upper bound on the complexity of checking a formula for these properties. The following result including a lower bound for the success problem is due to Johan van Benthem in correspondence.

Theorem 2.7 *The success problem for* **S5** *is* coNP*-complete.*

Proof. For an upper bound, φ is successful iff $[\varphi]\,\varphi$ is valid, and the validity problem for single-agent PAL is coNP-complete [14]. For a lower bound, we use the following reduction of the validity problem for **S5**, which is coNP-complete, to the success problem for **S5**: φ is valid iff for a new variable p, $\psi \equiv p \wedge (\Diamond\neg p \vee \varphi)$ is successful. From left to right, if $\mathcal{M}, w \vDash \psi$ then $\mathcal{M}_{|\psi}, w \vDash p$, and since φ is valid, $\mathcal{M}_{|\psi}, w \vDash \varphi$, so $\mathcal{M}_{|\psi}, w \vDash \psi$. From right to left, take any pointed model $\mathcal{M}, w$. Extend the language with a new variable p and extend $\mathcal{M}$ to $\mathcal{M}'$ with one new point v, related to all other points, with the same valuation as w for all variables of the old language. Make p true everywhere except at v. Then since $\mathcal{M}', w \vDash p \wedge \Diamond\neg p$, we have $\mathcal{M}', w \vDash \psi$, and since ψ is successful, we have $\mathcal{M}'_{|\psi}, w \vDash \psi$. But v is eliminated in $\mathcal{M}'_{|\psi}$, so $\mathcal{M}'_{|\psi}, w \nvDash \Diamond\neg p$. Hence $\mathcal{M}'_{|\psi}, w \vDash \varphi$. We conclude that $\mathcal{M}, w \vDash \varphi$, for $\mathcal{M}'_{|\psi}$ and $\mathcal{M}$ differ only with respect to the valuation of p, which φ does not contain. □

A similar argument shows that φ is valid iff for a new variable p, $p \wedge (\Diamond\neg p \vee \neg\varphi)$ is self-refuting, so the self-refutation problem is coNP-complete given the upper bound from PAL. From left to right the reduction is obvious, while from right to left the same extension of $\mathcal{M}$ to $\mathcal{M}'$ works. Similar arguments also show that the success and self-refutation problems are PSPACE-complete for multimodal **S5**.

3 The Moorean Source of Unsuccessfulness

To show that Moorean phenomena are the source of all unsuccessfulness in logics for an introspective agent, we proceed via normal forms for such logics. The following normal form derives from Carnap [10], and Proposition 3.2 is standard.

Definition 3.1 A formula is in *normal form* iff it is a disjunction of conjunctions of the form $\delta \equiv \alpha \wedge \Box\beta_1 \wedge ... \wedge \Box\beta_n \wedge \Diamond\gamma_1 \wedge ... \wedge \Diamond\gamma_m$ where α and each γ_i are conjunctions of literals and each β_i is a disjunction of literals.

Proposition 3.2 *For every formula φ, there is a formula φ' in normal form such that φ and φ' are equivalent in* **K45**.

Theorem 1.7.6.4 of [15] gives the analogue of Proposition 3.2 for **S5**. Inspection of the proof shows that the necessary equivalences hold in **K45**.

We use the following notation and terminology in this section and Appendix A.

Definition 3.3 Given $\delta \equiv \alpha \wedge \Box\beta_1 \wedge ... \wedge \Box\beta_n \wedge \Diamond\gamma_1 \wedge ... \wedge \Diamond\gamma_m$ in normal form, we define $\delta^{\alpha} \equiv \alpha$, $\delta^{\alpha\Box} \equiv \alpha \wedge \Box\beta_1 \wedge ... \wedge \Box\beta_n$, and similarly for $\delta^{\alpha\Diamond}$, $\delta^{\Box}$, $\delta^{\Box\Diamond}$, and $\delta^{\Diamond}$.

Definition 3.4 Where χ is a conjunction or disjunction of literals, let $L(\chi)$ be the set of literals in χ. A set of literals is *open* iff no literal in the set is the negation of any of the others.

Definition 3.5 A conjunction $\delta \equiv \alpha \wedge \Box\beta_1 \wedge ... \wedge \Box\beta_n \wedge \Diamond\gamma_1 \wedge ... \wedge \Diamond\gamma_m$ in normal form is **KD45**-*clear* iff: (i) $L(\alpha)$ is open; (ii) there is an open set of literals $\{l_1, ..., l_n\}$ with $l_i \in L(\beta_i)$; and (iii) for every γ_k there is a set of literals $\{l_1, ..., l_n\}$ with $l_i \in L(\beta_i)$ such that $\{l_1, ..., l_n\} \cup L(\gamma_k)$ is open. A disjunction in normal form is **KD45**-clear iff at least one of its disjuncts is **KD45**-clear.

In the following, by "clear" and "satisfiable" we mean **KD45**-clear and satisfiable in a quasi-partition, respectively. In the case of **S5**-clear, we must require in clause (ii) of Definition 3.5 that $\{l_1, ..., l_n\} \cup L(\alpha)$ is open, in which case (i) is unnecessary. In both cases, the following lemma holds with the appropriate definition of clarity.

Lemma 3.6 *A formula in normal form is satisfiable iff it is clear.*

Proof. [Sketch] Suppose φ in normal form is satisfiable. Then there is some disjunct δ of φ satisfied at a pointed model. Read off the appropriate open sets of literals from the current point for clarity condition (i), from some accessible point for (ii), and from witnesses for each $\Diamond\gamma_k$ in δ for (iii). In the other direction, suppose there are open sets of literals as described. Construct a model where δ^{α} is true at the root point w, which is possible by clarity condition (i). For each $\Diamond\gamma_k$, add a point v accessible from w and extend the valuation such that all conjuncts of γ_k and some disjunct of each β_i are true at v, which is possible by (iii). If there are no $\Diamond\gamma_k$ formulas in δ, add an accessible point v and extend the valuation such that some disjunct of each β_i is true at v, which is possible by (ii). Then φ is true at w. □

The following definition fixes the basic class of Moore conjunctions, as well as a wider class of Moorean conjunctions. The intuition is that a Moore conjunction simultaneously asserts a lack of information about another fact being asserted, and a Moorean conjunction is one that behaves like a Moore conjunction in some context.

We write $\sim \varphi$ for the negation of φ in negation normal form, with $\neg$ applying only to literals.

Definition 3.7 Let δ be a conjunction in normal form.

(i) δ is a *Moore conjunction* iff $\delta \wedge \Diamond\delta^{\alpha}$ is not clear or there is a $\Diamond\gamma_k$ conjunct in δ such that $\delta \wedge \Diamond(\delta^{\alpha} \wedge \gamma_k)$ is not clear.

(ii) δ is a *Moorean conjunction* iff there is a $\Diamond\gamma_k$ conjunct in δ such that $\delta \wedge \Diamond\delta^{\alpha} \wedge \Box(\sim\delta^{\alpha} \vee \sim\gamma_k)$ is clear.

Assuming **S5**, $\delta \wedge \Diamond\delta^{\alpha}$ may be replaced by δ in both (i) and (ii).

The definition of a Moore conjunction generalizes from the paradigmatic case of $p \wedge \Diamond \neg p$ to include formulas such as $p \wedge \Diamond q \wedge \Box (q \rightarrow \neg p)$. The formulas $p \wedge \neg p$ and $p \wedge \Box \neg p$ are also Moore conjunctions, since for these δ the formula $\delta \wedge \Diamond \delta^{\alpha}$ is not clear, but for the same reason they are not Moorean conjunctions. An example of a Moorean conjunction that is not a Moore conjunction is $p \wedge \Diamond q$. We consider this formula Moorean because in a context where the agent knows that q implies $\neg p$, so $\Box (q \rightarrow \neg p)$ holds, learning $p \wedge \Diamond q$ has the same effect as learning $p \wedge \Diamond \neg p$. By contrast, the formula $p \wedge \Diamond q \wedge \Diamond (p \wedge q)$ rules out the Moorean context with its last conjunct, and it is not Moorean according to Definition 3.7(ii).

In the proofs of Lemma 3.9 and Theorem 3.13 below, we will assume without loss of generality that all quasi-partitions considered are *chained*, i.e., $\forall w, v \in W : w \neq v \Rightarrow wRv \vee vRw$,[3] in which case the following basic facts hold.

Lemma 3.8 *Where $\mathcal{M}$ is a chained quasi-partition and δ is a conjunction in normal form:*

(i) $\mathcal{M}, w \vDash \delta \Rightarrow \mathcal{M} \vDash \delta^{\Box \Diamond}$;

(ii) $\mathcal{M}, w \vDash \delta \Rightarrow \mathcal{M}_{|\delta} = \mathcal{M}_{|\delta^{\alpha}}$ *and* $\mathcal{M}_{|\delta} \vDash \delta^{\alpha}$;

(iii) $\mathcal{M}_{|\delta}, w \vDash \delta \Rightarrow \mathcal{M}_{|\delta} \vDash \delta$.

We now prove the main lemma used in the proof of Theorem 3.13.

Lemma 3.9 *Let δ be a conjunction in normal form. The following hold for both* **KD45** *and* **S5**.

(i) *δ is self-refuting if and only if it is a Moore conjunction.*

(ii) *δ is unsuccessful if and only if it is a Moorean conjunction.*

Proof. ($\Leftarrow (i)$) Suppose δ is a Moore conjunction. Case 1: $\delta \wedge \Diamond \delta^{\alpha}$ is not clear. By Lemmas 3.6 and 3.8(i), $\delta \wedge \Diamond \delta^{\alpha}$ is clear iff $\Diamond^{+} \delta$ is satisfiable, so in this case $\Diamond^{+} \delta$ is unsatisfiable and hence δ is self-refuting. Case 2: $\delta \wedge \Diamond \delta^{\alpha}$ is clear, so suppose $\mathcal{M}, w \vDash \Diamond^{+} \delta$. For the $\Diamond \gamma_k$ in δ such that $\delta \wedge \Diamond (\delta^{\alpha} \wedge \gamma_k)$ is not clear, $\neg \delta \vee \Box (\neg \delta^{\alpha} \vee \neg \gamma_k)$ is valid by Lemma 3.6, so $\mathcal{M}, w \vDash \Box (\neg \delta^{\alpha} \vee \neg \gamma_k)$ given $\mathcal{M}, w \vDash \delta$. Then since $\Box (\neg \delta^{\alpha} \vee \neg \gamma_k)$ is universal, it is preserved under submodels by Theorem 2.4, so $\mathcal{M}_{|\delta}, w \vDash \Box (\neg \delta^{\alpha} \vee \neg \gamma_k)$. From Lemma 3.8(ii), $\mathcal{M}_{|\delta}, w \vDash \Box \delta^{\alpha}$, so $\mathcal{M}_{|\delta}, w \vDash \Box \neg \gamma_k$. Since $\Diamond \gamma_k$ is a conjunct in δ, $\mathcal{M}_{|\delta}, w \nvDash \delta$. Since $\mathcal{M}$ was arbitrary, δ is self-refuting.

($(i) \Rightarrow$) We prove the contrapositive.[4] Suppose δ is not a Moore conjunction. Then $\delta \wedge \Diamond \delta^{\alpha}$ is clear, and for every $\Diamond \gamma_k$ in δ, $\delta \wedge \Diamond (\delta^{\alpha} \wedge \gamma_k)$ is clear $(*)$. If there are no $\Diamond \gamma_k$ conjuncts in δ, then δ is a universal formula with $\Diamond^{+} \delta$ satisfiable, so it is not self-refuting by Theorem 2.4. Suppose there are $\Diamond \gamma_k$ conjuncts in δ. We claim that $\delta' \equiv \delta \wedge \Box \beta_{n+1} \wedge ... \wedge \Box \beta_{n+j}$ is clear where $\{\beta_{n+1}, ..., \beta_{n+j}\} = L(\delta^{\alpha})$. Given assumption $(*)$ and clarity condition (iii) for each $\delta \wedge \Diamond (\delta^{\alpha} \wedge \gamma_k)$, we have that for all $\Diamond \gamma_k$ in δ there is a set $\{l_1, ..., l_n\}$ with $l_i \in L(\beta_i)$ such that $\{l_1, ..., l_n\} \cup L(\delta^{\alpha} \wedge \gamma_k)$ is open. Taking $\{l_{n+1}, ..., l_{n+j}\} = L(\delta^{\alpha})$, $\{l_1, ..., l_{n+j}\} \cup L(\gamma_k) = \{l_1, ..., l_n\} \cup L(\delta^{\alpha} \wedge \gamma_k)$ is open, which gives clarity conditions (i), (ii) and (iii) for δ'. Since δ' is clear, suppose $\mathcal{N}, w \vDash \delta'$. From the fact that $\vDash \delta' \leftrightarrow (\delta \wedge \Box \delta^{\alpha})$ we have $\mathcal{N}, w \vDash \delta \wedge \Box \delta^{\alpha}$. Given

[3] Thanks to Dr. Yanjing Wang for catching the omission of $w \neq v$ in the published version.

[4] The following argument establishes something stronger than we need for Lemma 3.9, but we use it to establish Corollary 5.3 below. Compare the ($\Rightarrow$) direction of the proof of Theorem A.3 in Appendix A.

$\mathcal{N}, w \vDash \Box\delta^{\alpha}$ and the assumption that $\mathcal{N}$ is a chained quasi-partition, $\mathcal{N} \vDash \delta^{\alpha}$; given $\mathcal{N}, w \vDash \delta$ and Lemma 3.8(i), $\mathcal{N} \vDash \delta^{\Box\Diamond}$. Hence $\mathcal{N} \vDash \delta$, in which case $\mathcal{N}, w \vDash \Diamond^{+}\delta$ and $\mathcal{N}_{|\delta} = \mathcal{N}$. It follows that $\mathcal{N}_{|\delta}, w \vDash \delta$, so δ is not self-refuting.

$((ii) \Leftarrow)$ Suppose δ is a Moorean conjunction. Since for some $\Diamond\gamma_k$ in δ, $\chi \equiv \delta \wedge \Diamond\delta^{\alpha} \wedge \Box(\sim\delta^{\alpha} \vee \sim\gamma_k)$ is clear, there is a model with $\mathcal{M}, w \vDash \chi$. Given $\mathcal{M}, w \vDash \delta \wedge \Diamond\delta^{\alpha}$, we have $\mathcal{M}, w \vDash \Diamond^{+}\delta$ by Lemma 3.8(i). Given $\mathcal{M}, w \vDash \Box(\sim\delta^{\alpha} \vee \sim\gamma_k)$, by the same reasoning as in Case 2 of $((i) \Leftarrow)$, $\mathcal{M}_{|\delta}, w \nvDash \delta$. Therefore δ is unsuccessful.

$((ii) \Rightarrow)$ We prove the contrapositive. Suppose $\mathcal{M}, w \vDash \Diamond^{+}\delta$ and δ is not a Moorean conjunction. Then for all $\Diamond\gamma_k$ in δ, $\chi_k \equiv \delta \wedge \Diamond\delta^{\alpha} \wedge \Box(\sim\delta^{\alpha} \vee \sim\gamma_k)$ is not clear. To show $\mathcal{M}_{|\delta}, w \vDash \delta$, it suffices to show $\mathcal{M}_{|\delta}, w \vDash \delta^{\Diamond}$, since $\delta^{\alpha\Box}$ is universal and therefore preserved under submodels. Consider some $\Diamond\gamma_k$ conjunct in δ. It follows from our assumption that χ_k is unsatisfiable, whence $(\delta \wedge \Diamond\delta^{\alpha}) \rightarrow \Diamond(\delta^{\alpha} \wedge \gamma_k)$ is valid. Then from $\mathcal{M}, w \vDash \Diamond^{+}\delta$ we obtain $\mathcal{M}, w \vDash \Diamond(\delta^{\alpha} \wedge \gamma_k)$, so there is a v with wRv and $\mathcal{M}, v \vDash (\delta^{\alpha} \wedge \gamma_k)$. By Lemma 3.8(ii), v is retained in $\mathcal{M}_{|\delta}$. Since γ_k is propositional, $\mathcal{M}_{|\delta}, v \vDash \gamma_k$ and hence $\mathcal{M}_{|\delta}, w \vDash \Diamond\gamma_k$. Since $\Diamond\gamma_k$ was arbitrary, $\mathcal{M}_{|\delta}, w \vDash \delta^{\Diamond}$ and hence $\mathcal{M}_{|\delta}, w \vDash \delta$. Since $\mathcal{M}$ was arbitrary, δ is successful. □

Lemma 3.9 gives necessary and sufficient conditions for the successfulness of a conjunction in normal form. We now introduce an apparently stronger notion.

Definition 3.10 A formula φ is *super-successful* (in **L**) iff for every pointed model (of **L**), $\mathcal{M}, w \vDash \Diamond^{+}\varphi$ implies $\mathcal{M}', w \vDash \varphi$ for every $\mathcal{M}'$ such that $\mathcal{M}_{|\varphi} \subseteq \mathcal{M}' \subseteq \mathcal{M}$.

If φ is super-successful and $\mathcal{M}, w \vDash \varphi$, then as points that are not in $\mathcal{M}_{|\varphi}$ are eliminated from $\mathcal{M}$, φ remains true at w. Since we take the elimination of points as an agent's acquisition of new information, this means that φ remains true as the agent approaches, by way of the incremental acquisition of new information, the epistemic state of $\mathcal{M}_{|\varphi}$ wherein the agent knows φ. Intuitively, we can say that a super-successful formula remains true while an agent is "on the way" to learning it.

We will use the next lemma in the proof of Theorem 3.13.

Lemma 3.11 *If δ is a successful conjunction in normal form, δ is super-successful.*

Proof. Suppose δ is not super-successful, so there is a pointed model such that $\mathcal{M}, w \vDash \Diamond^{+}\delta$ and an $\mathcal{M}'$ such that $\mathcal{M}_{|\varphi} \subseteq \mathcal{M}' \subseteq \mathcal{M}$ and $\mathcal{M}', w \nvDash \delta$. Since $\mathcal{M}' \subseteq \mathcal{M}$ and $\delta^{\alpha\Box}$ is preserved under submodels, we must have $\mathcal{M}', w \nvDash \delta^{\Diamond}$. But then $\mathcal{M}_{|\delta}, w \nvDash \delta^{\Diamond}$ given that $\mathcal{M}_{|\delta} \subseteq \mathcal{M}'$ and $\delta^{\Diamond}$ is preserved under extensions. Hence $\mathcal{M}_{|\delta}, w \nvDash \delta$, so δ is unsuccessful. □

We now lift the definition of Moore and Moorean to arbitrary formulas.

Definition 3.12 Let φ be an arbitrary formula.

(i) φ is a *Moore sentence* iff any normal form of φ is a disjunction of Moore conjunctions.

(ii) φ is a *Moorean sentence* iff any normal form of φ contains a Moorean conjunction as a disjunct.

The following theorem gives necessary conditions for self-refuting and unsuccessful formulas. In Appendix A we strengthen this result with conditions that are sufficient as well as necessary.

Theorem 3.13 *Let φ be an arbitrary formula.*

(i) *If φ is self-refuting in any sublogic of* **S5**, *then φ is a Moore sentence.*

(ii) *If φ is unsuccessful in any extension of* **KD45**, *then φ is a Moorean sentence.*

Proof. By Lemma 2.3 it suffices to show the consequent of (i) for φ that is self-refuting in **S5** and the consequent of (ii) for φ that is unsuccessful in **KD45**. Since φ is self-refuting (resp. unsuccessful) iff any equivalent normal form of φ is self-refuting (resp. unsuccessful), let us assume that φ is already in normal form.

(i) Suppose φ is not a Moore sentence, so by Definition 3.12 there is a disjunct δ of φ that is not a Moore conjunction. Then by Lemma 3.9, δ is not self-refuting, so there is a pointed model with $\mathcal{M}, w \vDash \Diamond^+\delta$ and $\mathcal{M}_{|\delta}, w \vDash \delta$. By Lemma 3.8(iii), $\mathcal{M}_{|\delta} \vDash \delta$. It follows that $\left(\mathcal{M}_{|\delta}\right)_{|\varphi} = \mathcal{M}_{|\delta}$, since all points in $\mathcal{M}_{|\delta}$ satisfy one of the disjuncts of φ, namely δ. Then given $\mathcal{M}_{|\delta}, w \vDash \Diamond^+\delta$, we have $\mathcal{M}_{|\delta}, w \vDash \Diamond^+\varphi$ and hence $\left(\mathcal{M}_{|\delta}\right)_{|\varphi}, w \vDash \varphi$. Therefore φ is not self-refuting.

(ii) We prove something stronger. Suppose φ is not a Moorean sentence, so by Definition 3.12 no disjunct of φ is a Moorean conjunction. Then each disjunct of φ is successful by Lemma 3.9. Consider a pointed model such that $\mathcal{M}, w \vDash \Diamond^+\varphi$, so for some disjunct δ of φ, we have $\mathcal{M}, w \vDash \delta$. Since φ is a disjunction, $\mathcal{M}_{|\delta} \subseteq \mathcal{M}_{|\varphi}$. By Lemma 3.11, δ is super-successful, so for any $\mathcal{M}'$ with $\mathcal{M}_{|\varphi} \subseteq \mathcal{M}' \subseteq \mathcal{M}$, we have $\mathcal{M}', w \vDash \delta$ and hence $\mathcal{M}', w \vDash \varphi$. Since $\mathcal{M}$ was arbitrary, φ is super-successful. □

4 Unsuccessfulness in Other Logics

We now consider the sources of unsuccessfulness in logics for an agent without introspection (logics without axioms **4** and **5**) and in logics for multiple agents.

From an epistemic perspective, the most interesting (normal) proper sublogics of **S5** are obtained by dropping axiom **5** and adding something weaker in its place. Indeed, logics such as **S4**, **S4.x** for $\mathbf{x} = \mathbf{2}, \mathbf{3}, \mathbf{4}$, etc., have been proposed as logics of knowledge. Call logics **L** and **L′** *comparable* if **L** is a sublogic of **L′** or *vice versa*.

Proposition 4.1 *For any normal, proper sublogic* **L** *of* **S5**, *comparable to* **S4.4**, *there is a formula* (*consistent with* **S5**) *that is unsuccessful in* **L** *but is not Moorean.* [5]

Proof. First, we claim that $\varphi \equiv \Diamond p \wedge \Diamond \neg p$ is unsuccessful in **S4.4** and hence in any sublogic of **S4.4** by Lemma 2.3. In the **S4.4** model $\mathcal{M}$ in Figure 2, φ is true at the left point, but in $\mathcal{M}_{|\varphi}$, the right point is eliminated, so φ becomes false at the left point. Note that the formula is already in normal form and is not Moorean.

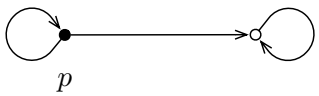

Fig. 2.

Next, we claim that $\mathcal{M}$ is a model of any logic **L** that is a proper extension of **S4.4** and a proper sublogic of **S5**. Suppose not, so there is a theorem φ of **L** with $\mathcal{M} \nvDash \varphi$. Change φ to φ' by substituting $(p \wedge \neg p)$ for any propositional variable q

[5] **S4.4** is **S4** plus $\varphi \rightarrow (\Diamond\Box\varphi \rightarrow \Box\varphi)$. It is a proper extension of **S4.3** and therefore also of **S4.2** [21].

other than p. Since $\mathbf{L}$ is normal and therefore closed under substitution, φ' is also a theorem of $\mathbf{L}$ and therefore of **S5**. Moreover, since for all variables q other than p, $\mathcal{M} \models \neg q$, the substitution of $(p \wedge \neg p)$ for q preserves (un)satisfiability in $\mathcal{M}$, so $\mathcal{M} \not\models \varphi'$. Hence φ' is not a theorem of **S4.4**. But by a result of Zeman [21], for any formula ψ, containing exactly one variable, that is a theorem of **S5** but not a theorem of **S4.4**, adding ψ to **S4.4** gives **S5**. Hence $\mathbf{L}$ is **S5**, a contradiction. Since $\mathcal{M}$ models any logic between **S4.4** and **S5**, φ is unsuccessful in these logics. □

Proposition 4.1 shows that **S5** is unique among the typical logics of knowledge insofar as all of its unsuccessful formulas are Moorean. The counterexample for the weaker logics shows that without negative introspection, one can come to know p by being truly told, "You do not know whether or not p," a surprising case of unsuccessfulness. The following proposition, although weaker than Proposition 4.1, shows that non-Moorean unsuccessful formulas appear if we weaken logics of knowledge and belief in other ways as well.

Proposition 4.2 *For any sublogic* $\mathbf{L}$ *of* **KTB** *or* **KD5**, *there is a formula* (*consistent with* **S5**) *that is unsuccessful in* $\mathbf{L}$ *but is not Moorean.*

Proof. For **KTB** consider $\varphi \equiv \Diamond p \wedge \Diamond q$ and the model $\mathcal{M}$ in Figure 3. In $\mathcal{M}_{|\varphi}$, the left and right points are eliminated, so φ becomes false at the center point.

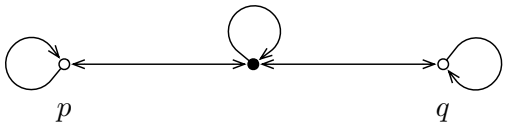

Fig. 3.

For **KD5** consider $\psi \equiv \neg q \vee (\Box p \wedge \Diamond\Diamond q)$ and the model $\mathcal{M}$ in Figure 4. Only the right point is eliminated in $\mathcal{M}_{|\psi}$, so ψ becomes false at the left point.

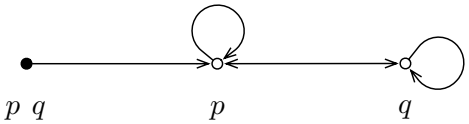

Fig. 4.

The formula ψ is equivalent in **KD45** to the normal form $\neg q \vee (\Box p \wedge \Diamond q)$, which is not Moorean, since neither disjunct is a Moorean conjunction. [6] □

There are other formulas that one may wish to categorize as Moorean, perhaps even as Moore sentences, but which are inconsistent with **KD45**.

Example 4.3 The formula $\Box p \wedge \Diamond\neg\Box p$ is a kind of "higher order" Moore sentence, which is satisfiable on intransitive frames but is also self-refuting over such frames. Similarly, the formula $\Diamond p \wedge \Diamond\neg\Diamond p$ is satisfiable yet self-refuting on non-Euclidean frames. The first formula says that the agent is not aware of what he believes, while the second says that he is not aware of what he does not believe. These formulas are the very witnesses of a failure to validate axioms **4** and **5**, respectively.

[6] Note that while the right disjunct is not a Moorean conjunction, in **S5** (but not in **KD45**) it implies the Moorean conjunction $p \wedge \Diamond q$. This is an instance of the more general fact that in **S5** some successful formulas imply unsuccessful formulas.

A natural question is whether there are non-Moorean sources of unsuccessfulness in languages more expressive than the basic modal language. Consider, for example, a language with multiple modalities, as in multi-agent epistemic logic. Without giving a formal definition of Moorean in the multi-agent case, it is nonetheless clear that there are more ways to be Moorean in the multi-agent case than in the single-agent case, even assuming quasi-partitions for each relation. For example, there are self-refuting, *indirect* Moore sentences such as $\Box_a p \wedge \Diamond_b \neg p$, which imply single-agent Moore sentences (in this case, $p \wedge \Diamond_b \neg p$) in multi-agent **S5**. There are also self-refuting, *higher order* Moore sentences such as $\Box_a p \wedge \Diamond_b \neg \Box_a p$ and $\Diamond_a p \wedge \Diamond_b \neg \Diamond_a p$, which resemble the higher order, single-agent Moore sentences consistent with logics weaker than **KD45**, noted in Example 4.3. However, not all unsuccessful formulas in the multi-agent context have a Moorean character.

Example 4.4 In the model $\mathcal{M}$ in Figure 5, the formula $\varphi \equiv \Diamond_a p \wedge \Diamond_a \Diamond_b \neg p$ is true at the left point but false at the right point, since $\Diamond_a p$ is false at the right point.

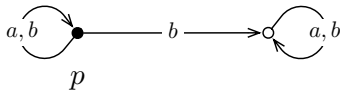

Fig. 5.

As a result, in the model $\mathcal{M}_{|\varphi}$ the right point is eliminated, in which case φ becomes false at the left point because $\Diamond_a \Diamond_b \neg p$ becomes false there.

The formula $\Diamond_a p \wedge \Diamond_a \Diamond_b \neg p$ does not resemble any single-agent Moorean formula, yet it is nonetheless unsuccessful. Given the connection that we have observed between introspection and Moorean phenomena in the single-agent case, we can see why there should be non-Moorean unsuccessful formulas in the multi-agent case: agents do not have introspective access to each other's knowledge or beliefs.

5 Related Classes of Formulas

Theorem 3.13 shows that in a wide range of logics, every self-refuting formula is a Moore sentence (i) and every unsuccessful formula is a Moorean sentence (ii). However, neither the converse of (i) nor the converse of (ii) holds in general. In this section, we relate the failures of the converses of (i) and (ii) to other classes of formulas. In Theorems A.3 and A.6 in Appendix A, we overcome these failures and give syntactic characterizations of the self-refuting and unsuccessful formulas as the *strong* Moore sentences and *strong* Moorean sentences, respectively.

For simplicity we assume in this section that we are working with **S5** only.

5.1 Informative, Cartesian, and eventually self-refuting formulas

Definition 5.1 A formula φ is (*potentially*) *informative* iff there is a pointed model such that $\mathcal{M}, w \vDash \varphi$ and $\mathcal{M}_{|\varphi} \neq \mathcal{M}$. Otherwise φ is *uninformative*. A formula φ is *always informative* iff for all pointed models such that $\mathcal{M}, w \vDash \varphi$, $\mathcal{M}_{|\varphi} \neq \mathcal{M}$.

Note that if a formula is not always informative, then it is not self-refuting, for there is a model such that $\mathcal{M}, w \vDash \varphi$ but $\mathcal{M}_{|\varphi} = \mathcal{M}$, so $\mathcal{M}_{|\varphi}, w \vDash \varphi$. This observation

explains some of the counterexamples to the converse of Theorem 3.13(i), Moore sentences that are not self-refuting.

Example 5.2 The formula $(p \wedge \Diamond\neg p) \vee (q \wedge \Diamond\neg q)$ is a Moore sentence, but it is not always informative (take a model with only two connected points, one of which satisfies p but not q and the other of which satisfies q but not p) and hence not self-refuting. The formula $(p \wedge \Diamond\neg p) \vee (\neg p \wedge \Diamond p)$ is a Moore sentence, but it is uninformative and hence successful, but not self-refuting. These examples show that neither the self-refuting nor the unsuccessful formulas are closed under disjunction.

From our previous results, we have the following corollary.

Corollary 5.3 *A conjunction in normal form is always informative iff it is self-refuting.*

Proof. By inspection of the proof of Lemma 3.9(i). □

However, a Moore sentence that is always informative may not be self-refuting.

Example 5.4 The formula $\varphi \equiv (p \wedge \Diamond\neg p) \vee (p \wedge q \wedge \Diamond\neg q)$ is always informative, for if $\mathcal{M}, w \vDash (p \wedge \Diamond\neg p)$, then there is a witness to $\Diamond\neg p$ that is eliminated in $\mathcal{M}_{|\varphi}$, and if $\mathcal{M}, w \vDash (p \wedge q \wedge \Diamond\neg q)$, then either the witness to $\Diamond\neg q$ does not satisfy $(p \wedge \Diamond\neg p)$, in which case it does not satisfy φ and is eliminated in $\mathcal{M}_{|\varphi}$, or it does satisfy $(p \wedge \Diamond\neg p)$, in which case there is a witness to $\Diamond\neg p$ that is eliminated in $\mathcal{M}_{|\varphi}$. In either case, $\mathcal{M}_{|\varphi} \neq \mathcal{M}$, so φ is always informative. However, φ is not self-refuting, as shown by the partition model $\mathcal{M}$ with $W = \{w, v, u\}$, $V(p) = \{w, v\}$, and $V(q) = \{w, u\}$. We have $\mathcal{M}, w \vDash p \wedge q \wedge \Diamond\neg q$, and given $W_{|\varphi} = \{w, v\}$, also $\mathcal{M}_{|\varphi}, w \vDash p \wedge q \wedge \Diamond\neg q$.

Interestingly, the formula φ is *self-refuting within two steps*; given $\mathcal{M}, w \vDash \varphi$ and $\mathcal{M}_{|\varphi}, w \vDash \varphi$, we have $\left(\mathcal{M}_{|\varphi}\right)_{|\varphi}, w \nvDash \varphi$. Let $\delta_1 \equiv (p \wedge \Diamond\neg p)$ and $\delta_2 \equiv (p \wedge q \wedge \Diamond\neg q)$ be the disjuncts of φ. All $\neg p$-points are eliminated in $\mathcal{M}_{|\varphi}$, so $\mathcal{M}_{|\varphi} \vDash \neg\delta_1$. Hence $\left(\mathcal{M}_{|\varphi}\right)_{|\varphi} = \left(\mathcal{M}_{|\varphi}\right)_{|\delta_2}$. But since $\mathcal{M}_{|\varphi}, w \vDash \delta_2$ and δ_2 is self-refuting, $\left(\mathcal{M}_{|\varphi}\right)_{|\delta_2}, w \nvDash \delta_2$. Since δ_1 is existential, $\left(\mathcal{M}_{|\varphi}\right)_{|\delta_2}, w \nvDash \delta_1$, so we conclude that $\left(\mathcal{M}_{|\varphi}\right)_{|\varphi}, w \nvDash \varphi$.

Example 5.4 points to the interest of self-refutation "in the long run."

Definition 5.5 Given a model $\mathcal{M}$, we define $\mathcal{M}_{|^n\varphi}$ recursively by $\mathcal{M}_{|^0\varphi} = \mathcal{M}$, $\mathcal{M}_{|^{n+1}\varphi} = \left(\mathcal{M}_{|^n\varphi}\right)_{|\varphi}$. A formula φ is *self-refuting within n steps* iff for all pointed models, if $\mathcal{M}, w \vDash \varphi$, then $\mathcal{M}_{|^m\varphi}, w \nvDash \varphi$ for some $m \leq n$; φ is *eventually self-refuting* iff for all pointed models, if $\mathcal{M}, w \vDash \varphi$, then there is an n such that $\mathcal{M}_{|^n\varphi}, w \nvDash \varphi$.

Using Definition 5.5, we can generalize Corollary 5.3. First, recall Fitch's paradox, mentioned in Section 1, which we can restate as: *if all truths are knowable, then all truths are (already) known.* One proposal for avoiding the paradox [19] is to restrict the claim that all truths are knowable to the claim that all *Cartesian* truths are knowable, in which case it does not follow that all truths are known.

Definition 5.6 φ is *Cartesian* iff $\Box\varphi$ is satisfiable. [7]

[7] The term 'Cartesian' is due to Tennant [19], though his definition is in terms of consistency rather than satisfiability. The term 'knowable' seems more natural, but 'knowable' is used in a different sense in the literature [9], for what we call 'learnable' in Definition 5.13.

The class of formulas φ for which $\Box\varphi$ is satisfiable seems a natural object of study in its own right. The following proposition establishes a connection between these formulas, defined "statically," and the other formulas classes that we have defined "dynamically" in terms of model transformations. It also generalizes Corollary 5.3.

Proposition 5.7 *The following are equivalent:*

(i) φ *is always informative.*

(ii) φ *is not Cartesian.*

(iii) φ *is eventually self-refuting.*

Proof. Suppose φ is always informative and $\mathcal{M}, w \vDash \varphi$. Without loss of generality, assume $\mathcal{M}$ is chained (as above Lemma 3.8). Since φ is always informative, $\mathcal{M}_{|\varphi} \neq \mathcal{M}$. Hence there is a point v such that wRv and $\mathcal{M}, v \nvDash \varphi$, in which case $\mathcal{M}, w \nvDash \Box\varphi$. Since $\mathcal{M}$ was arbitrary, $\Box\varphi$ is unsatisfiable, so φ is not Cartesian.

Suppose φ is not Cartesian, and without loss of generality, assume φ is in normal form. We prove by induction that for all $n \geq 0$, if $\mathcal{M}_{|^n\varphi}, w \vDash \varphi$, then there are $\Diamond\gamma_1, ..., \Diamond\gamma_{n+1}$ distinct subformulas of φ with $\mathcal{M}_{|^{n+1}\varphi} \vDash \neg\Diamond\gamma_1 \wedge ... \wedge \neg\Diamond\gamma_{n+1}$ $(*)$. It follows that since φ has some finite number n of distinct $\Diamond\gamma_k$ subformulas, we must have $\mathcal{M}_{|^m\varphi}, w \nvDash \varphi$ for some $m \leq n$. For the base case, assume $\mathcal{M}, w \vDash \varphi$. Since φ is not Cartesian, $\mathcal{M}_{|\varphi}, w \nvDash \Box\varphi$. Let v be such that wRv and $\mathcal{M}_{|\varphi}, v \nvDash \varphi$, and note that $\mathcal{M}, v \vDash \varphi$, for otherwise v would not have been retained in $\mathcal{M}_{|\varphi}$. From the fact that universal formulas are preserved under submodels, it follows that for some $\Diamond\gamma_1$ in φ, we have $\mathcal{M}, v \vDash \Diamond\gamma_1$ but $\mathcal{M}_{|\varphi}, v \nvDash \Diamond\gamma_1$. Hence $\mathcal{M}_{|\varphi} \vDash \neg\Diamond\gamma_1$ by Lemma 3.8(i). For the inductive step, assume that $\mathcal{M}_{|^{n+1}\varphi}, w \vDash \varphi$. Since w is retained in $\mathcal{M}_{|^{n+1}\varphi}$, we have $\mathcal{M}_{|^n\varphi}, w \vDash \varphi$, so by the induction hypothesis there are $\Diamond\gamma_1, ..., \Diamond\gamma_{n+1}$ distinct subformulas of φ for which $(*)$ holds. By the same reasoning as before, since φ is Cartesian, $\mathcal{M}_{|^{n+2}\varphi}, w \nvDash \Box\varphi$, so there is some z with wRz and some $\Diamond\gamma_{n+2}$ in φ such that $\mathcal{M}_{|^{n+1}\varphi}, z \vDash \Diamond\gamma_{n+2}$ but $\mathcal{M}_{|^{n+2}\varphi} \vDash \neg\Diamond\gamma_{n+2}$. Moreover, since $\Diamond\gamma_k$ formulas are preserved under extensions, $\mathcal{M}_{|^{n+2}\varphi} \vDash \neg\Diamond\gamma_1 \wedge ... \wedge \neg\Diamond\gamma_{n+1}$ as well. Finally, given $\mathcal{M}_{|^{n+1}\varphi}, z \vDash \Diamond\gamma_{n+2}$ and $(*)$, $\Diamond\gamma_{n+2}$ is distinct from $\Diamond\gamma_1, ..., \Diamond\gamma_{n+1}$.

Suppose φ is not always informative. Then there is a model with $\mathcal{M}, w \vDash \varphi$ and $\mathcal{M}_{|\varphi} = \mathcal{M}$, in which case $\mathcal{M}_{|^n\varphi} = \mathcal{M}$ for all n, so φ is not eventually self-refuting.□

Corollary 5.8 φ *is eventually self-refuting iff it is self-refuting within* n *steps, with* n *bounded by the number of distinct diamond formulas in a normal form of* φ*.*

Proof. By inspection of the proof of the previous proposition. □

5.2 Successful, super-successful, and learnable formulas

The converse of Theorem 3.13(ii) fails because a formula φ in normal form that contains a Moorean conjunction as a disjunct may be successful. For example, it is easy to see that if the disjunction of the non-Moorean conjunctions in φ is a consequence of the disjunction of the Moorean conjunctions in φ, then φ is not only successful but super-successful, given Lemma 3.11. Examples of such φ include $(p \wedge \Diamond\neg p) \vee p$ and $(p \wedge \Diamond\neg p) \vee \Diamond\neg p$. There are also successful formulas ψ that contain Moorean conjunctions as disjuncts, but do not meet the condition of φ. These formulas nevertheless manage to be successful by a kind of compensation: when one disjunct

of ψ goes from true at $\mathcal{M}, w$ to false at $\mathcal{M}_{|\psi}, w$, another disjunct compensates by going from false at $\mathcal{M}, w$ to true at $\mathcal{M}_{|\psi}, w$. The formula $(p \wedge \Diamond\neg p) \vee \Box p$ exhibits this kind of compensation against a Moore conjunction, while the formula in the proof of the following proposition does so against a Moorean conjunction.

Proposition 5.9 *Not all successful formulas are super-successful.*

Proof. The formula $\varphi \equiv (p \wedge \Diamond q) \vee \Box p$ is successful but not super-successful. Suppose $\mathcal{M}, w \vDash \varphi$, and without loss of generality assume that $\mathcal{M}$ is chained. Case 1: $\mathcal{M}, w \vDash \Box p$. Then $\mathcal{M}_{|\varphi} = \mathcal{M}$, so $\mathcal{M}_{|\varphi}, w \vDash \varphi$. Case 2: $\mathcal{M}, w \nvDash \Box p$. Then $\mathcal{M}_{|\varphi} = \mathcal{M}_{|p \wedge \Diamond q}$, and given $\mathcal{M}_{|p\wedge\Diamond q} \vDash \Box p$, we again have $\mathcal{M}_{|\varphi}, w \vDash \varphi$. Hence φ is successful. But φ is not super-successful, as shown by the partition model $\mathcal{N}$ with $W = \{w, v, u\}$, $V(p) = \{w\}$, and $V(q) = \{u\}$. We begin with $\mathcal{N}, w \vDash \varphi$, for while $\mathcal{N}, w \nvDash \Box p$, it holds that $\mathcal{N}, w \vDash p \wedge \Diamond q$. Moreover, given $W_{|\varphi} = \{w\}$, we still have $\mathcal{N}_{|\varphi}, w \vDash \varphi$, for while $\mathcal{N}_{|\varphi}, w \nvDash p \wedge \Diamond q$, it holds that $\mathcal{N}_{|\varphi}, w \vDash \Box p$. However, in the extension $\mathcal{N}'$ of $\mathcal{N}_{|\varphi}$ with $W' = \{w, v\}$, we now have $\mathcal{N}', w \nvDash \varphi$. The fact that $u \notin W'$ gives $\mathcal{N}', w \nvDash p \wedge \Diamond q$, and the fact that $v \in W'$ gives $\mathcal{N}', w \nvDash \Box p$. □

Proposition 5.10 *A formula φ is super-successful iff for a propositional variable p that does not occur in φ, $\varphi \vee p$ is successful.*

Proof. ($\Rightarrow$) Suppose φ is super-successful and $\mathcal{M}, w \vDash \varphi \vee p$, where p does not occur in φ. If $\mathcal{M}, w \vDash p$, then $\mathcal{M}_{|\varphi\vee p}, w \vDash p$, and if $\mathcal{M}, w \vDash \varphi$, then given $\mathcal{M}_{|\varphi} \subseteq \mathcal{M}_{|\varphi\vee p}$ and the assumption that φ is super-successful, $\mathcal{M}_{|\varphi\vee p}, w \vDash \varphi$. In either case, $\mathcal{M}_{|\varphi\vee p}, w \vDash \varphi \vee p$, so $\varphi \vee p$ is successful.

($\Leftarrow$) Suppose φ is not super-successful, so there is an $\mathcal{M} = \langle W, R, V\rangle$ with $w \in W$ such that $\mathcal{M}, w \vDash \varphi$, and an $\mathcal{M}' = \langle W', R', V'\rangle$ such that $\mathcal{M}_{|\varphi} \subseteq \mathcal{M}' \subseteq \mathcal{M}$ and $\mathcal{M}', w \nvDash \varphi$. Let $\mathcal{M}^* = \langle W, R, V^*\rangle$ be the same as $\mathcal{M}$ except that $V^*(p) = W' \setminus W_{|\varphi}$. Then since p does not occur in φ, we have $\mathcal{M}^*, w \vDash \varphi \vee p$ given $\mathcal{M}, w \vDash \varphi$. Moreover, since $W^*_{|\varphi\vee p} = W'$, we have $\mathcal{M}^*_{|\varphi\vee p}, w \nvDash \varphi$ given $\mathcal{M}', w \nvDash \varphi$. From the assumption that $\mathcal{M}, w \vDash \varphi$, it follows that $w \in W_{|\varphi}$, in which case $w \notin V^*(p)$ and hence $\mathcal{M}^*_{|\varphi\vee p}, w \nvDash p$. But then $\mathcal{M}^*_{|\varphi\vee p}, w \nvDash \varphi \vee p$, so $\varphi \vee p$ is unsuccessful. □

By Proposition 5.10, complexity and syntactic characterization results for successful formulas carry over immediately to super-successful formulas.

From the previous propositions we obtain a surprising failure of closure.

Corollary 5.11 *The set of successful formulas is not closed under disjunction.*

Proof. Immediate from Propositions 5.9 and 5.10. □

Example 5.12 The formula $\neg p \wedge \neg q$ is successful, and by the proof of Proposition 5.9, $(p \wedge \Diamond q) \vee \Box p$ is also successful. However, the disjunction of these formulas, $\chi \equiv (\neg p \wedge \neg q) \vee ((p \wedge \Diamond q) \vee \Box p)$ is unsuccessful, as shown by the model $\mathcal{N}$ in the proof of Proposition 5.9. We begin with $\mathcal{N}, w \vDash \chi$, since while $\mathcal{N}, w \nvDash \neg p \wedge \neg q$ (and hence $\mathcal{N}_{|\chi}, w \nvDash \neg p \wedge \neg q$), we have already seen that $\mathcal{N}, w \vDash (p \wedge \Diamond q) \vee \Box p$. However, $\mathcal{N}_{|\chi} = \mathcal{N}'$, and we have already seen that $\mathcal{N}', w \nvDash (p \wedge \Diamond q) \vee \Box p$.

By contrast, the set of super-successful formulas is closed under disjunction, since $\mathcal{M}_{|\varphi\vee\psi}$ is an extension of $\mathcal{M}_{|\varphi}$ and $\mathcal{M}_{|\psi}$.

Another consequence of the previous results concerns the relation between successful and *learnable* formulas. Like the Cartesian formulas, the learnable formulas have been discussed in connection with Fitch's paradox [3,9].

Definition 5.13 A formula φ is (*always*) *learnable* iff for all pointed models, if $\mathcal{M}, w \vDash \varphi$, then there is some ψ such that $\mathcal{M}_{|\psi}, w \vDash \Box\varphi$.[8]

Formulas that are learnable (and satisfiable) are Cartesian according to Definition 5.6, but the converse does not hold [3]. For example, the formula $p \wedge \Diamond q$ is Cartesian, since $\Box(p \wedge \Diamond q)$ is satisfiable, but it is not (always) learnable, for if $\mathcal{M}, w \vDash \Box(p \to \neg q)$ and $\mathcal{M}' \subseteq \mathcal{M}$, then $\mathcal{M}', w \nvDash \Box(p \wedge \Diamond q)$.

All successful formulas are learnable [9], since if $\mathcal{M}, w \vDash \varphi$ and φ is successful, then not only $\mathcal{M}_{|\varphi}, w \vDash \varphi$ but also $\mathcal{M}_{|\varphi}, w \vDash \Box\varphi$. For if v is retained in $\mathcal{M}_{|\varphi}$, then $\mathcal{M}, v \vDash \varphi$, in which case $\mathcal{M}_{|\varphi}, v \vDash \varphi$ by the successfulness of φ. Hence for a successful φ, we can always take ψ in Definition 5.13 to be φ itself. On the other hand, it is natural to ask whether some unsuccessful formulas are learnable as well.

Corollary 5.14 *Not all learnable formulas are successful.*

Proof. Let $\delta_1 \vee \delta_2$ be an unsuccessful disjunction with δ_1 and δ_2 successful, as given by Corollary 5.11. If $\mathcal{M}, w \vDash \delta_1 \vee \delta_2$, then $\mathcal{M}, w \vDash \delta_i$ for $i = 1$ or $i = 2$. Since δ_i is successful, we have $\mathcal{M}_{|\delta_i}, w \vDash \Box\delta_i$, in which case $\mathcal{M}_{|\delta_i}, w \vDash \Box(\delta_1 \vee \delta_2)$. Since $\mathcal{M}$ was arbitrary, $\delta_1 \vee \delta_2$ is (always) learnable. □

6 Conclusion

From a technical point of view, we have studied the question of when a modal formula is preserved under relativizing models to the formula itself. In the epistemic interpretation of modal logic, this preservation question takes on a new significance: it concerns whether an agent retains knowledge of what is learned, which requires that what is learned remain true. We have shown (Theorem 3.13) that for an introspective agent, the only true sentences that may become false when learned are variants of the Moore sentence. For an agent without introspection or multiple agents without introspective access to each other's knowledge or beliefs, there are non-Moorean sources of unsuccessfulness (Propositions 4.1 and 4.2, Example 4.4).

In connection with our study of Moorean phenomena, we have observed a number of related results. We saw that the sentences that always provide information to an agent, no matter the agent's prior epistemic state, are exactly those sentences that cannot be known—and will eventually become false if repeated enough (Proposition 5.7); we saw that there are sentences that always remain true when they are learned, but whose truth value may oscillate while an agent is on the way to learning them (Proposition 5.9); and we saw that there are sentences that sometimes become false when learned directly, but which an agent can always come to know indirectly by learning something else (Corollary 5.14).

In Appendix A, we return to the problem with which we began. We give syntactic characterizations of the self-refuting and unsuccessful formulas as the *strong* Moore sentences and *strong* Moorean sentences, respectively.

[8] The term 'learnable' is used by van Benthem [3]. Balbiani et al. [9] use the term 'knowable'.

Acknowledgement

We wish to thank Hans van Ditmarsch and the anonymous referees for their helpful comments on an earlier draft of this paper, and Johan van Benthem for stimulating our interest in the topic of successful formulas.

References

[1] Baltag, A., L. Moss and S. Solecki, *The logic of public announcements, common knowledge and private suspicions*, in: I. Gilboa, editor, *Proceedings of the 7th Conference on Theoretical Aspects of Rationality and Knowledge (TARK 98)*, Morgan Kaufmann, 1998 pp. 43–56.

[2] Baltag, A., H. van Ditmarsch and L. Moss, *Epistemic logic and information update*, in: P. Adriaans and J. van Benthem, editors, *Philosophy of Information*, North-Holland, 2008 pp. 361–456.

[3] van Benthem, J., *What one may come to know*, Analysis **64** (2004), pp. 95–105.

[4] van Benthem, J., *One is a lonely number: on the logic of communication*, in: Z. Chatzidakis, P. Koepke and W. Pohlers, editors, *Logic Colloquium '02*, ASL & A.K. Peters, 2006 pp. 96–129.

[5] van Benthem, J., *Open problems in logical dynamics*, in: D. Gabbay, S. Goncharov and M. Zakharyashev, editors, *Mathematical Problems from Applied Logic I*, Springer, 2006 pp. 137–192.

[6] van Ditmarsch, H. and B. Kooi, *The secret of my success*, Synthese **151** (2006), pp. 201–232.

[7] van Ditmarsch, H., W. van der Hoek and B. Kooi, "Dynamic Epistemic Logic," Springer, 2008.

[8] Andréka, H., I. Németi and J. van Benthem, *Modal languages and bounded fragments of predicate logic*, Journal of Philosophical Logic **27** (1998), pp. 217–274.

[9] Balbiani, P., A. Baltag, H. van Ditmarsch, A. Herzig, T. Hoshi and T. de Lima, *'Knowable' as 'known after an announcement'*, The Review of Symbolic Logic **1** (2008), pp. 305–334.

[10] Carnap, R., *Modalities and quantification*, The Journal of Symbolic Logic **11** (1946), pp. 33–64.

[11] Fitch, F. B., *A logical analysis of some value concepts*, The Journal of Symbolic Logic **28** (1963), pp. 135–142.

[12] Gerbrandy, J., "Bisimulations on Planet Kripke," Ph.D. thesis, University of Amsterdam (1999), ILLC Dissertation Series DS-1999-01.

[13] Hintikka, J., "Knowledge and Belief: An Introduction to the Logic of the Two Notions," College Publications, 2005.

[14] Lutz, C., *Complexity and succinctness of public announcement logic*, in: P. Stone and G. Weiss, editors, *Proceedings of the Fifth International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS 06)* (2006), pp. 137–143.

[15] Meyer, J.-J. Ch. and W. van der Hoek, "Epistemic Logic for AI and Computer Science," Cambridge Tracts in Theoretical Computer Science **41**, Cambridge University Press, 1995.

[16] Moore, G. E., *A reply to my critics*, in: P. Schilpp, editor, *The Philosophy of G.E. Moore*, The Library of Living Philosophers **4**, Northwestern University, 1942 pp. 535–677.

[17] Plaza, J., *Logics of public communications*, in: M. Emrich, M. Pfeifer, M. Hadzikadic and Z. Ras, editors, *Proceedings of the 4th International Symposium on Methodologies for Intelligent Systems: Poster Session Program*, Oak Ridge National Laboratory, ORNL/DSRD-24, 1989 pp. 201–216.

[18] Qian, L., *Sentences true after being announced*, in: *Proceedings of the Student Session of the 1st North American Summer School in Logic, Language, and Information (NASSLLI)*, Stanford University, 2002.

[19] Tennant, N., "The Taming of The True," Oxford: Clarendon Press, 1997.

[20] Visser, A., J. van Benthem, D. de Jongh and G. R. R. de Lavalette, *NNIL, A study in intuitionistic propositional logic*, Logic Group Preprint Series 111, Philosophical Institute, Utrecht University (1994).

[21] Zeman, J. J., *A study of some systems in the neighborhood of S4.4*, Notre Dame Journal of Formal Logic **12** (1971), pp. 341–357.

A Appendix

In this appendix, we address the problem of giving syntactic characterizations of the successful and self-refuting formulas. Given Theorem 2.6, Lemma 3.6, and the PAL definitions of successful and self-refuting, there is a trivial characterization of both classes: φ is successful iff the result of reducing the PAL formula $\neg[\varphi]\varphi$ to normal form in the basic modal language is not clear; φ is self-refuting iff the result of reducing the PAL formula $\neg[\varphi]\neg\varphi$ to normal form in the basic modal language is not clear. Although by Proposition 2.7, the PAL definitions of successful and self-refuting lead to optimal methods for checking for these properties, they do not provide much insight beyond the semantic definitions of the formula classes. In Theorems A.3 and A.6 below, we give syntactic characterizations of the self-refuting and unsuccessful formulas that reveal more than the trivial characterization.

We state the following results for **S5** with the standard definitions of successful ($\forall \mathcal{M}, w : \mathcal{M}, w \vDash \varphi \Rightarrow \mathcal{M}_{|\varphi}, w \vDash \varphi$) and self-refuting ($\forall \mathcal{M}, w : \mathcal{M}, w \vDash \varphi \Rightarrow \mathcal{M}_{|\varphi}, w \nvDash \varphi$). With minor changes the results also hold for **KD45**, given the modified definitions using precondition $\Diamond^+\varphi$ instead of φ.

We will continue to use the abbreviations δ^α, $\delta^{\Box\Diamond}$, etc., from Definition 3.3. We will also use $\hat{\delta}^{\Box\Diamond}$ to denote a conjunct of $\delta^{\Box\Diamond}$, i.e., a $\Box\beta_i$ or $\Diamond\gamma_k$ in δ. As before, $\sim\varphi$ is the negation of φ in negation normal form, with $\neg$ applying only to literals.

Definition A.1 φ is a *strong Moore sentence* iff for any normal form φ^* of φ, no disjunct of φ^* is *compensated in* φ^*. A disjunct $\delta \equiv \alpha \wedge \Box\beta_1 \wedge ... \wedge \Box\beta_n \wedge \Diamond\gamma_1 \wedge ... \wedge \Diamond\gamma_m$ of φ^* is compensated in φ^* iff there is a disjunct δ' of φ^*, a subset S of the disjuncts of φ^*, a sequence of disjuncts $\sigma_{\gamma_1}, ..., \sigma_{\gamma_m}$ (not necessarily distinct) from S, and for every $\sigma \notin S$ a conjunct $\hat{\sigma}^{\Box\Diamond}$ of $\sigma^{\Box\Diamond}$, such that $\chi \equiv \delta' \wedge \delta^\alpha \wedge \chi_\Box \wedge \chi_\Diamond$ is clear, where

$$\chi_\Box \equiv \bigwedge_{\sigma\in S} \sigma^{\Box\Diamond} \wedge \bigwedge_{\sigma\notin S} \sim \hat{\sigma}^{\Box\Diamond} \wedge \bigwedge_{\sigma\in S, j\leq n} \Box\left(\sim \sigma^\alpha \vee \beta_j\right);$$

$$\chi_\Diamond \equiv \bigwedge_{k\leq m} \Diamond\left(\sigma^\alpha_{\gamma_k} \wedge \gamma_k\right).$$

Note that χ is in normal form, so the definition of clarity properly applies. The reason for the use of $\sim \hat{\sigma}^{\Box\Diamond}$ is that $\sim \sigma^{\Box\Diamond}$ may be a disjunction, in which case χ would not be in normal form. Hence we pull the disjunction out of the formula and add an existential quantifier to the condition, since $\sigma^{\Box\Diamond}$ is false iff there is some conjunct $\hat{\sigma}^{\Box\Diamond}$ of $\sigma^{\Box\Diamond}$ that is false.

As we will see in the proof of Theorem A.3, the existence of a compensated disjunct in φ is equivalent to there being a model in which φ has a "successful update," in the sense that $\mathcal{M}, w \vDash \varphi$ and $\mathcal{M}_{|\varphi}, w \vDash \varphi$. For φ to have a successful update in $\mathcal{M}$, there must be a disjunct δ' of φ true at a point w in $\mathcal{M}$ and a disjunct δ of φ (possibly distinct from δ') true at w in $\mathcal{M}_{|\varphi}$. What is required for δ to be true at w in $\mathcal{M}_{|\varphi}$ is that its propositional part δ^α is true at w in $\mathcal{M}$, that its universal part $\delta^\Box$ is already true at w in $\mathcal{M}$ or *becomes* true at w in $\mathcal{M}_{|\varphi}$, and that its existential part $\delta^\Diamond$ remains true at w in $\mathcal{M}_{|\varphi}$. This is exactly what χ captures.

Proposition A.2 *Every strong Moore sentence is a Moore sentence.*

Proof. By Definitions A.1 and 3.12, it suffices to show that for φ in normal form, if a disjunct δ of φ is not compensated in φ, then δ is a Moore conjunction. To prove the contrapositive, suppose δ is not a Moore conjunction. Then by Definition 3.7, $\delta \wedge \Diamond \delta^{\alpha}$ is clear and $\delta \wedge \Diamond (\delta^{\alpha} \wedge \gamma_k)$ is clear for every $\Diamond \gamma_k$ conjunct in δ. It follows that $\delta \wedge \bigwedge_{k \leq m} \Diamond (\delta^{\alpha} \wedge \gamma_k)$ is clear and hence satisfiable. Given $\mathcal{M}, w \vDash \delta \wedge \bigwedge_{k \leq m} \Diamond (\delta^{\alpha} \wedge \gamma_k)$, let S be the set of disjuncts in φ such that $\mathcal{M}, w \vDash \bigwedge_{\sigma \in S} \sigma^{\Box \Diamond} \wedge \bigwedge_{\sigma \notin S} \neg \sigma^{\Box \Diamond}$. For every $\sigma \notin S$, pick a conjunct $\hat{\sigma}^{\Box \Diamond}$ of $\sigma^{\Box \Diamond}$ such that $\mathcal{M}, w \vDash \bigwedge_{\sigma \in S} \sigma^{\Box \Diamond} \wedge \bigwedge_{\sigma \notin S} \sim \hat{\sigma}^{\Box \Diamond}$. Since $\delta \in S$, let the sequence $\sigma_{\gamma_1}, ..., \sigma_{\gamma_m}$ of disjuncts from S be such that $\sigma_{\gamma_i} \equiv \delta$ for $1 \leq i \leq m$. We claim that for the χ as in Definition A.1 based on these choices, $\mathcal{M}, w \vDash \chi$. From the fact that $\mathcal{M}, w \vDash \bigwedge_{k \leq m} \Diamond (\delta^{\alpha} \wedge \gamma_k)$ and our choice of $\sigma_{\gamma_1}, ..., \sigma_{\gamma_m}$, we have $\mathcal{M}, w \vDash \chi_{\Diamond}$. Since $\mathcal{M}, w \vDash \delta^{\Box}$, it is immediate that $\mathcal{M}, w \vDash \bigwedge_{\sigma \in S, j \leq n} \Box (\sim \sigma \vee \beta_j)$. Together with $\mathcal{M}, w \vDash \bigwedge_{\sigma \in S} \sigma^{\Box \Diamond} \wedge \bigwedge_{\sigma \notin S} \sim \hat{\sigma}^{\Box \Diamond}$, this gives $\mathcal{M}, w \vDash \chi_{\Box}$. Finally, setting $\delta' \equiv \delta$, we have $\mathcal{M}, w \vDash \chi$. Hence χ is clear, so δ is compensated in φ. □

We can now generalize Theorem 3.13(i).

Theorem A.3 *φ is self-refuting if and only if φ is a strong Moore sentence.*

Proof. φ is self-refuting iff any equivalent normal form of φ is self-refuting, so let us assume φ is already in normal form.

($\Leftarrow$) We prove the contrapositive. Suppose φ is not self-refuting, so there is a pointed model such that $\mathcal{M}, w \vDash \varphi$ and $\mathcal{M}_{|\varphi}, w \vDash \varphi$. Given $\mathcal{M}_{|\varphi}, w \vDash \varphi$, there is a disjunct δ of φ such that $\mathcal{M}_{|\varphi}, w \vDash \delta$. We claim that δ is compensated in φ, so φ is not a strong Moore sentence. It suffices to show the satisfiability of an appropriate χ as in Definition A.1.

Since $\mathcal{M}, w \vDash \varphi$, there is a disjunct δ' of φ such that $\mathcal{M}, w \vDash \delta'$. This gives the first conjunct of χ. Since $\mathcal{M}_{|\varphi}, w \vDash \delta$, we have $\mathcal{M}, w \vDash \delta^{\alpha}$ because δ^{α} is propositional and hence preserved under extensions. This gives the second conjunct of χ.

Next we claim $\mathcal{M}, w \vDash \chi_{\Box}$. Let S be the set of the disjuncts of φ such that $\mathcal{M}, w \vDash \bigwedge_{\sigma \in S} \sigma^{\Box \Diamond} \wedge \bigwedge_{\sigma \notin S} \neg \sigma^{\Box \Diamond}$, and let $\hat{\sigma}^{\Box \Diamond}$ be a false conjunct of each $\sigma^{\Box \Diamond}$ for $\sigma \notin S$. For *reductio*, suppose $\mathcal{M}, w \nvDash \bigwedge_{\sigma \in S, j \leq n} \Box (\sim \sigma^{\alpha} \vee \beta_j)$, where $\Box \beta_1 \wedge ... \wedge \Box \beta_n \equiv \delta^{\Box}$. Then there is some v with wRv and $\mathcal{M}, v \vDash \sigma^{\alpha} \wedge \neg \beta_j$ for some $\sigma \in S$ and $j \leq n$. Since $\sigma \in S$, we have $\mathcal{M}, w \vDash \sigma^{\Box \Diamond}$. It follows by Lemma 3.8(i) that $\mathcal{M}, v \vDash \sigma^{\Box \Diamond}$, in which case $\mathcal{M}, v \vDash \sigma$ given $\mathcal{M}, v \vDash \sigma^{\alpha}$. Hence v is retained in $\mathcal{M}_{|\varphi}$. But then given $\mathcal{M}_{|\varphi}, v \nvDash \beta_j$, we have $\mathcal{M}_{|\varphi}, w \nvDash \Box \beta_j$, which contradicts the assumption that $\mathcal{M}_{|\varphi}, w \vDash \delta$. We conclude that $\mathcal{M}, w \vDash \chi_{\Box}$, which gives the third conjunct of χ.

Finally, we claim $\mathcal{M}, w \vDash \chi_{\Diamond}$. Given $\mathcal{M}_{|\varphi}, w \vDash \delta^{\Diamond}$, take an arbitrary $\Diamond \gamma_k$ in δ, and let v be such that wRv and $\mathcal{M}_{|\varphi}, v \vDash \gamma_k$. It follows that $\mathcal{M}, v \vDash \sigma$ for some disjunct σ of φ, which we label as σ_{γ_k}, for otherwise v would not be retained in $\mathcal{M}_{|\varphi}$. Since γ_k is propositional, $\mathcal{M}, v \vDash \gamma_k$ given $\mathcal{M}_{|\varphi}, v \vDash \gamma_k$. Therefore $\mathcal{M}, w \vDash \Diamond \left(\sigma_{\gamma_k}^{\alpha} \wedge \gamma_k\right)$. Then from the fact that $\mathcal{M}, v \vDash \sigma_{\gamma_k}^{\Box \Diamond}$, we have $\mathcal{M}, w \vDash \sigma_{\gamma_k}^{\Box \Diamond}$ by Lemma 3.8(i), so $\sigma_{\gamma_k} \in S$. Since $\Diamond \gamma_k$ was arbitrary, $\mathcal{M}, w \vDash \chi_{\Diamond}$, which gives the final conjunct of χ.

($\Rightarrow$) Again we prove the contrapositive. Suppose δ is not a strong Moore sen-

tence, so there is some δ that is compensated in φ, for which an appropriate χ as in Definition A.1 is clear. Where $\mathcal{M}, w \vDash \chi$, we claim that $\mathcal{M}_{|\varphi}, w \vDash \delta$. We will show $\mathcal{M}_{|\varphi}, w \vDash \delta^{\alpha}$, $\mathcal{M}_{|\varphi}, w \vDash \delta^{\Box}$, and $\mathcal{M}_{|\varphi}, w \vDash \delta^{\Diamond}$ separately.

Given $\mathcal{M}, w \vDash \delta$, we have $\mathcal{M}_{|\varphi}, w \vDash \delta^{\alpha}$ since δ^{α} is propositional.

Next, for any v retained in $\mathcal{M}_{|\varphi}$, we have $\mathcal{M}, v \vDash \sigma^{\alpha}$ for some disjunct σ of φ. It must be that $\sigma \in S$, for otherwise $\mathcal{M} \vDash \neg\sigma$ given $\mathcal{M}, w \vDash \chi_{\Box}$ and Lemma 3.8(i). Then from the fact that $\mathcal{M}, w \vDash \bigwedge_{\sigma \in S, j \leq n} \Box(\sim \sigma^{\alpha} \vee \beta_j)$, we have $\mathcal{M}_{|\varphi}, v \vDash \beta_j$ for all $j \leq n$. Since v was arbitrary, $\mathcal{M}_{|\varphi}, w \vDash \delta^{\Box}$.

Finally, given $\mathcal{M}, w \vDash \chi_{\Diamond}$, for any $\Diamond\gamma_k$ in δ we have $\mathcal{M}, w \vDash \Diamond(\sigma_{\gamma_k} \wedge \gamma_k)$ with $\sigma_{\gamma_k} \in S$. Let v be such that wRv and $\mathcal{M}, v \vDash \sigma_{\gamma_k}^{\alpha} \wedge \gamma_k$. Since $\sigma_{\gamma_k} \in S$, $\mathcal{M}, w \vDash \sigma_{\gamma_k}^{\Box\Diamond}$. It follows by Lemma 3.8(i) that $\mathcal{M}, v \vDash \sigma_{\gamma_k}^{\Box\Diamond}$, in which case $\mathcal{M}, v \vDash \sigma_{\gamma_k}$ given $\mathcal{M}, v \vDash \sigma_{\gamma_k}^{\alpha}$. Hence v is retained in $\mathcal{M}_{|\varphi}$. Since γ_k is propositional, we have $\mathcal{M}_{|\varphi}, v \vDash \gamma_k$ given $\mathcal{M}, v \vDash \gamma_k$, whence $\mathcal{M}_{|\varphi}, w \vDash \Diamond\gamma_k$. Since γ_k was arbitrary, $\mathcal{M}_{|\varphi}, w \vDash \delta^{\Diamond}$.

We conclude that $\mathcal{M}_{|\varphi}, w \vDash \delta$, in which case $\mathcal{M}_{|\varphi}, w \vDash \varphi$. Given $\mathcal{M}, w \vDash \chi$, we have $\mathcal{M}, w \vDash \delta$ and hence $\mathcal{M}, w \vDash \varphi$, so φ is not self-refuting. □

Finally, we will prove an analogous generalization of Theorem 3.13(ii).

Definition A.4 φ is a *strong Moorean sentence* iff for any normal form φ^* of φ, there is a disjunct δ and non-empty sets S and T of disjuncts of φ^*, with for every $\theta \in T$, a $\Diamond\gamma_\theta$ in θ, such that $\chi \equiv \delta \wedge \chi_1 \wedge \chi_2 \wedge \chi_3$ is clear, where

$$t(\theta) \equiv \theta^{\alpha\Diamond} \wedge \bigwedge_{\sigma \in S, \Box\beta \text{ in } \theta} \Box(\sim \sigma^{\alpha} \vee \beta);$$

$$\chi_1 \equiv \bigwedge_{\theta \in T} t(\theta) \wedge \bigwedge_{\theta \notin T} \neg t(\theta); \; \chi_2 \equiv \bigwedge_{\sigma \in S} \sigma^{\Box\Diamond} \wedge \bigwedge_{\sigma \notin S} \neg\sigma^{\Box\Diamond};$$

$$\chi_3 \equiv \bigwedge_{\sigma \in S, \theta \in T} \Box(\sim \sigma^{\alpha} \vee \sim \gamma_\theta).$$

The formula χ is not yet in normal form, due to the $\neg t(\theta)$ conjuncts in χ_1 and the $\neg\sigma^{\Box\Diamond}$ conjuncts in χ_2, so strictly the definition of clarity does not apply. However, it is straightforward to put χ into normal form using the same method involving $\hat{\sigma}^{\Box\Diamond}$ as in Definition A.1, together with some distribution of $\wedge$ and $\vee$. Since in this case the necessary modifications add four existential quantifiers to the definition, for simplicity we do not write them out. When we say that χ is clear, strictly we mean that the modified formula is clear.

As we will see in the proof of Theorem A.6, the clarity of χ is equivalent to there being a model in which φ has an unsuccessful update. For φ to have an unsuccessful update in $\mathcal{M}$, there must be some disjunct δ in φ such that $\mathcal{M}, w \vDash \delta$, but no disjunct δ' such that $\mathcal{M}_{|\varphi}, w \vDash \delta'$. To ensure that there are no such δ', we need only keep track of those disjuncts θ of φ whose propositional part θ^{α} and existential part $\theta^{\Diamond}$ are true in $\mathcal{M}$ and whose universal part $\theta^{\Box}$ was already true in $\mathcal{M}$ or *becomes* true in $\mathcal{M}_{|\varphi}$, since all other disjuncts will be false in $\mathcal{M}_{|\varphi}$. This is the purpose of χ_1. For each such θ, we must ensure that there is a diamond formula $\Diamond\gamma_\theta$ in θ that becomes false in $\mathcal{M}_{|\varphi}$, because none of its witnesses satisfy any of the disjuncts that are satisfied somewhere in $\mathcal{M}$. This is the purpose of χ_3 and χ_2.

Proposition A.5 *Every strong Moorean sentence is a Moorean sentence.*

Proof. Assume φ is a strong Moorean sentence, so an appropriate χ as in Definition A.4 is clear. Note that the distinguished disjunct δ of φ^* must be a member of both S and T. Then given that $\delta \wedge \chi_3$ is clear, we have that $\delta \wedge \Box(\sim \delta^{\alpha} \vee \sim \gamma_{\delta})$ is clear, where $\Diamond\gamma_{\delta}$ is a conjunct of δ. Hence δ is a Moorean conjunction by Definition 3.7, in which case φ is a Moorean sentence by Definition 3.12. □

Theorem A.6 *φ is unsuccessful if and only if φ is a strong Moorean sentence.*

Proof. As before, let us assume that φ is already in normal form.

($\Leftarrow$) Suppose φ is a strong Moorean sentence, so an appropriate χ as in Definition A.4 is clear. Consider a pointed model such that $\mathcal{M}, w \vDash \chi$. For *reductio*, assume $\mathcal{M}_{|\varphi}, w \vDash \varphi$. Then there exists a disjunct θ in φ such that $\mathcal{M}_{|\varphi}, w \vDash \theta$.

We claim that $\theta \in T$. For if $\theta \notin T$, then given $\mathcal{M}, w \vDash \chi_1$ there are two cases. Case 1: $\mathcal{M}, w \nvDash \theta^{\alpha\Diamond}$. Then $\mathcal{M}_{|\varphi}, w \nvDash \theta^{\alpha\Diamond}$ since $\theta^{\alpha\Diamond}$ is existential. Case 2: $\mathcal{M}, w \vDash \Diamond(\sigma^{\alpha} \wedge \neg\beta)$ for some $\sigma \in S$ and $\Box\beta$ in θ. Let v be such that wRv and $\mathcal{M}, v \vDash \sigma^{\alpha} \wedge \neg\beta$. Since $\sigma \in S$, we have $\mathcal{M}, w \vDash \sigma^{\Box\Diamond}$, in which case $\mathcal{M}, v \vDash \sigma$ given Lemma 3.8(i) and $\mathcal{M}, v \vDash \sigma^{\alpha}$. Hence v is retained in $\mathcal{M}_{|\varphi}$, and since β is propositional, $\mathcal{M}_{|\varphi}, v \nvDash \beta$, so $\mathcal{M}_{|\varphi}, w \nvDash \Box\beta$ and $\mathcal{M}_{|\varphi}, w \nvDash \theta^{\Box}$. In both cases, $\mathcal{M}_{|\varphi}, w \nvDash \theta$, a contradiction. Therefore $\theta \in T$.

Given $\mathcal{M}_{|\varphi}, w \vDash \theta$, for every $\Diamond\gamma$ in θ there is a v with wRv and $\mathcal{M}_{|\varphi}, v \vDash \gamma$. Since v was retained in $\mathcal{M}_{|\varphi}$, we have $\mathcal{M}, v \vDash \sigma$ for some $\sigma \in S$. Then given $\mathcal{M}, w \vDash \chi_3$ and $\theta \in T$, we have $\mathcal{M}, v \vDash \neg\gamma$. Since γ is propositional, $\mathcal{M}_{|\varphi}, v \vDash \neg\gamma$, a contradiction. We conclude that $\mathcal{M}_{|\varphi}, w \nvDash \varphi$, so φ is unsuccessful.

($\Rightarrow$) Suppose φ is unsuccessful, so there is a pointed model with $\mathcal{M}, w \vDash \varphi$ but $\mathcal{M}_{|\varphi}, w \nvDash \varphi$. To show that φ is a strong Moorean sentence, it suffices to show that an appropriate χ as in Definition A.4 is satisfiable. Given $\mathcal{M}, w \vDash \varphi$, we can read off from w the disjunct δ and sets S and T such that $\mathcal{M}, w \vDash \delta \wedge \chi_1 \wedge \chi_2$. It only remains to show $\mathcal{M}, w \vDash \chi_3$. For *reductio*, suppose there is a $\theta \in T$ such that for all $\Diamond\gamma$ in θ, $\mathcal{M}, w \nvDash \bigwedge_{\sigma \in S} \Box(\sim \sigma^{\alpha} \vee \sim \gamma)$, i.e., $\mathcal{M}, w \vDash \Diamond(\sigma^{\alpha} \wedge \gamma)$ for some $\sigma \in S$. Then we claim that $\mathcal{M}_{|\varphi}, w \vDash \theta$. As before, we take the three parts of θ separately.

For $\theta^{\Diamond}$, consider any $\Diamond\gamma$ in θ and let v be such that wRv and $\mathcal{M}, v \vDash \sigma^{\alpha} \wedge \gamma$. Given $\mathcal{M}, w \vDash \chi_2$ and the fact that $\sigma \in S$, we have $\mathcal{M}, w \vDash \sigma^{\Box\Diamond}$, so $\mathcal{M}, v \vDash \sigma$ by Lemma 3.8(i). Hence v is retained in $\mathcal{M}_{|\varphi}$. Since $\mathcal{M}, v \vDash \gamma$ and γ is propositional, $\mathcal{M}_{|\varphi}, v \vDash \gamma$ and therefore $\mathcal{M}_{|\varphi}, w \vDash \Diamond\gamma$. Since $\Diamond\gamma$ was arbitrary, $\mathcal{M}_{|\varphi}, w \vDash \theta^{\Diamond}$.

For $\theta^{\Box}$, take any v retained in $\mathcal{M}_{|\varphi}$ such that wRv, and we have $\mathcal{M}, v \vDash \sigma^{\alpha}$ for some $\sigma \in S$. It follows that for any $\Box\beta$ in θ, we have $\mathcal{M}, v \vDash \beta$ given $\mathcal{M}, w \vDash \chi_1$. Since β is propositional, $\mathcal{M}_{|\varphi}, v \vDash \beta$. Since v and β were arbitrary, $\mathcal{M}_{|\varphi}, w \vDash \delta^{\Box}$.

Finally, for θ^{α}, since by assumption $\theta \in T$ and $\mathcal{M}, w \vDash \chi_1$, we have $\mathcal{M}, w \vDash \theta^{\alpha}$ and hence $\mathcal{M}_{|\varphi}, w \vDash \theta^{\alpha}$.

We have shown $\mathcal{M}_{|\varphi}, w \vDash \theta$ and hence $\mathcal{M}_{|\varphi}, w \vDash \varphi$, which contradicts our initial assumption. It follows that for every $\theta \in T$ there is a $\Diamond\gamma_{\theta}$ in θ such that $\mathcal{M}, w \vDash \chi_3$.□