Conditionalization and Total Knowledge

Ian Pratt-Hartmann

Abstract

This paper employs epistemic logic to investigate the philosophical foundations of Bayesian updating in belief revision. By *Bayesian updating*, we understand the tenet that an agent's degrees of belief—assumed to be encoded as a probability distribution—should be revised by conditionalization on the agent's total knowledge up to that time. A familiar argument, based on the construction of a diachronic Dutch book, purports to show that Bayesian updating is the only rational belief-revision policy. We investigate the conditions under which of the premises of this argument might be satisfied. Specifically, we consider the case of an artificial agent whose language (of thought) features a modal operator TK, where $TK\psi$ has the interpretation "My *total knowledge* is ψ". We show that every proposition of the form $TK\psi$ is *epistemically categorical*: it determines, for every proposition φ in the agent's language, whether the agent knows that φ. We argue that, for certain artificial agents employing such a language, the diachronic Dutch book argument for Bayesian updating is on firm ground.

In the days before their connection to fragments of first-order logic was fully understood, modal logics were conceived of primarily as tools for the philosophical analysis of various fundamental concepts: necessity, time, obligation, knowledge. The axioms of these logics were formalizations of philosophical intuition; the ability to determine the precise consequences of those axioms, and the philosophical perspective it provided were the fruits of that formalization. Two developments have since radically altered this situation. The first is the discovery of relational semantics for modal logic, which made it possible to conceive of modal logic primarily as a collection of formalisms for describing relational structures. The second is the increasing influence of Computer Science in logic, which has pushed issues such as decidability, computational complexity and efficient automation, where modal logics have turned out to exhibit striking behaviour, to the fore. Indeed, it is probably correct to say that modal logicians of the younger generation are predominantly motivated by the special model-theoretic and complexity-theoretic characteristics of the formal languages they study. These days, modal logic just isn't—well, *modal*.

The present paper harks back to the older, philosophical tradition in modal logic. Specifically, we employ epistemic logic to investigate the philosophical foundations of Bayesian updating in belief revision. By *Bayesian updating*, we understand the tenet that an agent's degrees of belief—assumed to be encoded as a probability distribution—should be revised by conditionalization on the agent's total knowledge up to that time. We pose two questions regarding this tenet. First: what reasons do we have for adopting it? Second: what, come to think of it, does it mean anyway?

1 Motivation: Bayesian updating

Consider an agent whose beliefs are expressed in some formal language $\mathcal{L}$. We take that agent's *degree of belief in* a proposition φ of $\mathcal{L}$, denoted $p(\varphi)$, to be the price (in, say, £s) he would consider fair for a bet of £1 on φ—i.e. a ticket worth £1 if φ turns out to be true, and nothing otherwise. We assume—idealizing somewhat—that the agent's degrees of belief exist, and satisfy the usual axioms of probability theory. The justification for this assumption is completely standard, and we do not discuss it further. (See, e.g. Paris [7], pp. 17 ff. or Halpern [2], pp. 17 ff. for a textbook account.) If the agent's degree of belief in a proposition ψ is non-zero, we take his *conditional degree of belief in* φ *given* ψ, denoted $p(\varphi|\psi)$, to be the price he would consider fair for a bet of £1 on φ *conditional on* ψ—i.e. a ticket worth £1 if $\varphi \wedge \psi$ turns out to be true, nothing if $\neg\varphi \wedge \psi$ turns out to be true, and his money back if $\neg\psi$ turns out to be true. Again, we assume that conditional degrees of belief exist, and are given by the familiar equation $p(\varphi|\psi) = p(\varphi \wedge \psi)/p(\psi)$. The justification for this assumption, based on synchronic Dutch book arguments, is again standard, and again we do not discuss it further. (See, e.g. Teller [9] or Jeffrey [3].)

How should such an agent *revise* his beliefs in response to some new evidence ψ, where ψ is assumed to be a proposition of $\mathcal{L}$ such that $p(\psi) > 0$? The most familiar updating strategy is *conditionalization*: the agent's new degree of belief in φ after learning ψ, denoted $p_\psi(\varphi)$, is set equal to $p(\varphi|\psi)$. A familiar argument purports to show that an agent who fails to revise his beliefs by conditionalization is vulnerable to a *diachronic Dutch book* as follows. Suppose to the contrary that $p_\psi(\varphi) < p(\varphi|\psi)$. Select any positive a such that $a(1 - p(\psi)) < p(\varphi|\psi) - p_\psi(\varphi)$. We offer the agent a bet of £1 on φ conditional on ψ, which he will buy for £$p(\varphi|\psi)$, together with a bet of £a on ψ, which he will buy for £$ap(\psi)$. If and when ψ turns out true, we offer to buy from the agent a bet of £1 on φ, which he will sell for £$p_\psi(\varphi)$. Now consider the agent's position after all bets are settled. If ψ turns out false, the conditional bet becomes void, and the agent simply loses the £$ap(\psi)$ he paid for the bet on ψ. If ψ turns out true, he makes a nett gain of £$a(1 - p(\psi))$ by betting on ψ; however, he makes a nett loss of £$(p(\varphi|\psi) - p_\psi(\varphi))$ by—in effect—buying and selling back the bet on φ. Since we have chosen a so the latter amount is greater than the former, we win. An analogous argument applies if $p_\psi(\varphi) > p(\varphi|\psi)$.

This argument rests on various concealed assumptions. To tease some of them out, let us examine a case in which conditionalization apparently fails. The well-known *Monty Hall paradox* is as follows. (For a collection of such problems, see Bar-Hillel and Falk [1].) A TV game-show contestant, whom we will take to be our agent, must guess which of three doors a prize is behind.

The game-show host then opens one of the other doors that the prize is *not* behind (there will always be one), and gives the contestant the opportunity to change his guess to the third door. (If the contestant guessed correctly, we assume the game-show host opens one of the other two doors at random.) A rational contestant will accept this offer, since he and his like will on average bag the prize two-thirds of the time. Yet Bayesian updating seems to give the wrong answer here. For let a, b and c stand for the propositions that the prize is behind the respective doors, and let p encode the contestant's initial degrees of belief. Since the contestant initially has no information about where the prize is, we can assume that $p(a) = p(b) = p(c) = 1/3$, whence, by simple calculation, $p(a|\neg b) = p(c|\neg b) = 1/2$. Now suppose, without loss of generality, the contestant guesses the first door (proposition a), and the game-show host opens the second door to reveal that the prize is not behind it. If the contestant conditionalizes on the new information (proposition $\neg b$), he will see no advantage in switching doors, since he will assign an equal probability to a and c.

There are many ways to see what is going wrong here, and we do not propose to rehearse them all. Rather, we draw attention to the fact that all would be well if only the agent could represent facts about what he *knows*. For, letting $K\neg b$ stand for the proposition that the agent comes to know that $\neg b$, it is reasonable to suppose that $p(K\neg b|a) = 1/2$ (if the prize is behind the first door, game-show host opens one of the other two at random); likewise, $p(K\neg b|\neg a) = 1/2$ (if the prize is not behind the first door, game-show host opens whichever of the other two it is also not behind). Hence:

$$p(a|K\neg b) = \frac{p(K\neg b|a)p(a)}{p(K\neg b|a)p(a) + p(K\neg b|\neg a)p(\neg a)} = \frac{1/2 \cdot 1/3}{1/2 \cdot 1/3 + 1/2 \cdot 2/3} = 1/3\ .$$

Therefore, $p(c|K\neg b) = 2/3$; so conditionalizing on $K\neg b$ leads the contestant to accept the game-show host's offer and switch choices to the third door.

Thus, we have a way of reconciling the Monty Hall paradox with Bayesian updating: enrich the agent's language of thought so that he can conditionalize not on the objective information he obtains, but rather, on the subjective fact of his having obtained that information. (I use 'subjective' here in the sense of 'about the agent's state of mind'; subjective facts, in this sense, are just as real as objective ones.) Once this is done, conditionalization appears to give the correct results. But, why was this process of epistemic ascent (as we might call it) necessary in this particular case? And what assurance do we have that it will be sufficient in general? More concretely: suppose our game-show contestant does, as we recommend, conditionalize on $K\neg b$ (rather than on $\neg b$). Why can we not construct a diachronic Dutch book against him by exploiting his failure to conditionalize on $\neg b$?

The answer is that the Dutch book argument works only if we (the bookmakers) know what the agent's degrees of belief are really going to be if and when ψ turns out true; only then can we decide which bets to offer to buy and sell, and at what prices. But taking ψ to be the proposition $\neg b$ in the Monty Hall paradox (as presented above) fails to satisfy this requirement. To see this, suppose—consistently with $\neg b$—that the prize is in fact behind the first door (i.e. the contestant's original guess was correct). Then the game-show host might open either of the other doors, whence the contestant might learn either of $\neg b$ or $\neg c$. That is: we, the bookmakers, cannot infer the contestant's degrees of belief given merely that $\neg b$ is true. Now contrast this with the case where ψ is $K\neg b$. In the context of the described scenario, the proposition $K\neg b$ determines precisely which propositions the contestant knows. Hence, as long as the contestant's degrees of belief are determined by what he has come to know, the proposition $\psi = K\neg b$ allows us to construct the Dutch book as explained in the above argument. That is why the argument sanctions conditionalization on the proposition $K\neg b$, but not on the proposition $\neg b$. And that is why, in this case, we need to enrich the contestant's language thus far, and no further.

This explanation takes the Dutch book argument to rest on the following two assumptions:

> **A1:** The agent's degrees of belief at any time are a function of what propositions in $\mathcal{L}$ he has come to know by that time.
>
> **A2:** At any time, there exists a true proposition ψ in $\mathcal{L}$, such that the truth of ψ determines, for the bookmaker, what the agent will have come to know by that time.

In the context of *human* reasoning, assumption **A1**—the view that our evidence arrives in the form of *Protokollsätze* of whose truth we can be certain—is widely rejected (see, e.g. Jeffrey [4]). However, in the context of *artificial agents*, with digitized, and hence propositionally describable, perceptual inputs, **A1** is almost certain to be satisfied; and an argument restricted to this context is still of interest. Objection: does this not mean that such artificial agents are encumbered with literally millions of 'tiny' *Protokollsätze* about their transducers? Answer: yes; but why *encumbered*? In fact, exactly this sort of Bayesian inference is now routinely and successfully employed in a wide variety of computing problems.

This leaves us with the assumption **A2**, which, as we shall see, presents us with an intriguing logical challenge. Let $\mathcal{L}$ be a formal language containing a modal operator K, where $K\psi$ has the interpretation "The agent knows that ψ at time t". (For simplicity, we fix a time t within the range of times over which **A2** quantifies; reference to this time will then be implicit in the following discussion.) For the scenario of the Monty Hall paradox, this is all we need: it is assumed as part of that scenario that the proposition $K\neg b$, if true, completely determines the agent's state of knowledge at the time in question. But in general, we can expect no such luck: propositions about what the agent does know imply very little about what he does not know. What we seek, then, is a modal operator TK, where $TK\psi$ has the interpretation "The agent's *total knowledge* at time t is ψ", and with the property that $TK\psi$ is *epistemically categorical*: for all propositions $\chi \in \mathcal{L}$, either $TK\psi \rightarrow K\chi$ or $TK\psi \rightarrow \neg K\chi$ is a logical truth. For an artificial agent using such a language, the diachronic Dutch book argument for belief updating by conditionalization on $TK\psi$ would be on firm ground. Can we define such an operator? That is the topic of this paper.

2 Syntax and semantics

We begin by setting out the logical framework within which the ensuing analysis will be carried out.

Definition 1. Assume as given a countable set of *variables*, a countable set of *names* and, for each n $(0 \leq n)$, a countable set of n-ary *predicate letters*. The symbol $=$ is one of the binary predicate letters. We call the 0-ary predicate letters *proposition letters*. A *term* is a variable or a name. An $\mathcal{FOLTK}$*-formula* is a member of the smallest set of expressions satisfying the following rules:

if r is an n-ary predicate letter and $t_1, \ldots, t_n$ are terms, then $r(t_1, \ldots, t_n)$ is a formula of $\mathcal{FOLTK}$;

if φ and ψ are formulas of $\mathcal{FOLTK}$ and x is a variable, then $\varphi \wedge \psi$, $\varphi \vee \psi$, $\neg\varphi$, $\exists x\varphi$, $\forall x\varphi$, $K\varphi$ and $TK\varphi$ are formulas of $\mathcal{FOLTK}$.

A $\mathcal{PCTK}$-formula is an $\mathcal{FOLTK}$-formula involving no occurrences of $\exists$ or $\forall$ and no n-ary relations for $n > 0$. We standardly refer to $\mathcal{FOLTK}$- or $\mathcal{PCTK}$-formulas simply as *formulas*. Formulas involving no occurrences of TK are called *basic*; formulas involving no occurrences of K or TK are called *objective*. Formulas in which every predicate letter appears within the scope of either K or TK are called *subjective*. We use the constants $\top$, $\bot$ and connectives $\rightarrow$, $\leftrightarrow$ as abbreviations with their usual meanings. A formula with no free variables is a *sentence*.

The general semantic framework used here is that of Levesque [5] (see also Levesque and Lakemeyer [6]). Models for $\mathcal{FOLTK}$-formulas are non-empty sets of "interpretations", where an interpretation is just a model of the underlying non-modal language. The most notable features are that names denote rigidly and uniquely, and that the domain of quantification is covered by the names.

Definition 2. An *interpretation* w is a function mapping any n-ary predicate letter r to a set r^w of n-tuples of names, subject to the constraint that $=^w$ is the identity relation on the set of names. (As usual, we assume that there is exactly one 0-tuple of names.)

Let W be a non-empty set of interpretations, let $w \in W$, and let φ be a sentence of $\mathcal{FOLTK}$. We define $W \models_w \varphi$ inductively as follows:

If r is a predicate letter and $a_1, \ldots, a_n$ are names, then $W \models_w r(a_1, \ldots, a_n)$ if and only if $\langle a_1, \ldots, a_n \rangle \in r^w$;

$W \models_w \varphi \wedge \psi$ if and only if $W \models_w \varphi$ and $W \models_w \psi$, and similarly for the other Boolean connectives;

$W \models_w \exists x\varphi$ if and only if $W \models_w \varphi[x/a]$ for some name a, and similarly for the universal quantifier;

$W \models_w K\varphi$ if and only if, for all $w' \in W$, $W \models_{w'} \varphi$;

$\star$ $W \models_w TK\varphi$ if and only if $W \models_w K\varphi$ and $W \models_w \neg K\chi$ for all objective sentences χ such that $\not\models K\varphi \to \chi$.

Here, $\varphi[x/a]$ denotes the result of substituting the name a for every free occurrence of x in φ. (The symbol $\star$ is for future reference.)

Since only *sentences* receive truth-values, we will henceforth notate free variables explicitly. Thus, φ will denote a sentence, and $\psi(\bar{x})$ a formula with no free variables other than $\bar{x}$. Note that if φ is objective, we can write $\models_w \varphi$ for $W \models_w \varphi$, and if φ is subjective, we can write $W \models \varphi$ for $W \models_w \varphi$. More generally, we write $W \models \varphi$ to mean $W \models_w \varphi$ for all $w \in W$, and $\models \varphi$ to mean $W \models \varphi$ for all W. We say φ is *satisfiable* if $W \models_w \varphi$ for some W and some $w \in W$, and we say φ is *valid* if $\models \varphi$. Clearly, the usual S5-axioms for K are valid.

Note that clause $\star$ above, giving the semantics of $TK\varphi$, quantifies only over *objective* sentences χ. Allowing χ to range over *arbitrary* sentences in $\star$ would result in a non-terminating recursive definition of $\models$. (To determine whether $W \models TK\varphi$, we would would need to determine whether, for various sentences χ *involving occurrences of* TK, $\models K\varphi \to \chi$ implies $W \models \neg K\chi$; but to determine either of these things, we would in turn need to determine whether $W' \models_w \chi$ for various W' and $w \in W'$.) Moreover, allowing χ to range over *basic* sentences in $\star$, though it would result in a well-formed definition, would have other undesirable consequences. Consider, for example, any sentence of the form TKp_1. We do not want this sentence to be unsatisfiable, since it seems reasonable that an agent may have simply learned p_1 and nothing else. Yet Kp_1 fails to imply both p_2 and $\neg Kp_2$, so that, without the restriction of χ to objective sentences, clause $\star$ would make TKp_1 entail both $\neg Kp_2$ and $\neg K\neg Kp_2$, which is unsatisfiable on our semantics. Hence the restriction of χ to objective sentences in $\star$.

However, this restriction creates a problem. Consider the following consequence of $\star$.

Lemma 1. *For any sentence φ and any objective sentence χ, $\models TK\varphi \to K\chi$ or $\models TK\varphi \to \neg K\chi$.*

Proof. Let χ be objective. If $\models K\varphi \to \chi$, then $\models K\varphi \to K\chi$; and since $\models TK\varphi \to K\varphi$, we have $\models TK\varphi \to K\chi$. On the other hand, if $\not\models K\varphi \to \chi$, then, from the semantics of the TK-operator, $\models TK\varphi \to \neg K\chi$. □

Lemma 1 states that, as we might say, $TK\varphi$ is *epistemically categorical* for objective sentences χ. Yet we would prefer that $TK\varphi$ were epistemically categorical for *arbitrary* χ, for that is what, according to **A2**, the diachronic Dutch book argument requires. One of our main results (Theorem 4) is that Lemma 1 can be strengthened in just this way.

It is worth pausing to see why this result is surprising. Lemma 1 guarantees that any two agents whose total knowledge is φ know the same objective sentences. But it is easy to construct an example of two agents who know the same objective sentences but who do not know the same basic sentences. Let p be a unary predicate letter, and enumerate the names as $\{c_i\}_{0\leq i}$. Define the interpretation w_0 by setting $\models_{w_0} p(c_j)$ if and only if j is odd; and define the interpretation w_i, for $i \geq 1$ by setting $\models_{w_i} p(c_j)$ if and only if j is odd or $j = 2i$. Assume that all other predicate letters are assigned the empty interpretation. Let $W = \{w_i | i \geq 0\}$ and $W' = \{w_i | i \geq 1\}$. It follows that, for all objective χ, $W \models K\chi$ if and only if $W' \models K\chi$. However, we have $W' \models K\exists x(p(x) \wedge \neg Kp(x))$ but $W \models \neg K\exists x(p(x) \wedge \neg Kp(x))$. The analysis below shows that this sort of situation cannot arise in the presence of total knowledge.

We close this section with a very simple remark.

Lemma 2. *If φ is any formula, then $\models TK\varphi \leftrightarrow TK(K\varphi)$ and $\models TK\varphi \leftrightarrow K(TK\varphi)$.*

Proof. Straightforward from standard S5-identities. □

See Corollary 3 for a similar, but more difficult, result of this kind.

3 The propositional case

We begin with an observation establishing the consistency of certain total-knowledge sentences.

Lemma 3. *If a sentence φ of $\mathcal{FOLTK}$ is objective and satisfiable, then $TK\varphi$ is satisfiable. Indeed, if W is the set of assignments w such that $\models_w \varphi$, then $W \models TK\varphi$.*

Proof. Since φ is satisfiable, $W \neq \emptyset$. By construction, $W \models K\varphi$. Moreover, for any objective χ such that $\not\models K\varphi \to \chi$, we have $\not\models \varphi \to \chi$, so let w be an interpretation such that $\models_w \varphi \wedge \neg\chi$. Since $w \in W$, $W \models \neg K\chi$. Thus, $W \models TK\varphi$. □

The analysis of TK in the propositional case is very easy, and relies on the existence of the following normal-form theorem.

Lemma 4. *Any basic sentence of $\mathcal{PCTK}$ is equivalent to a sentence of the form*

$$\bigvee_{1\leq h\leq l} (K\psi_h \wedge \neg K\chi_{h,1} \wedge \cdots \wedge \neg K\chi_{h,m_h} \wedge \pi_h).$$

in which the ψ_h, $\chi_{h,i}$ and π_h are objective.

Proof. Straightforward from standard S5-identities. □

Thus, K-operators occurring in the scope of other K-operators in basic $\mathcal{PCTK}$ sentences can always be eliminated. Of course, $\mathcal{FOLTK}$ lacks this feature: the embedded K in $K\exists x(p(x) \wedge \neg Kp(x))$ cannot be removed. As a corollary of this normal form lemma, we have the following.

Lemma 5. *Let φ be a satisfiable, basic sentence of $\mathcal{PCTK}$. Then there exists a basic (in fact, objective) sentence ψ of $\mathcal{PCTK}$ such that $\varphi \wedge TK\psi$ is satisfiable.*

Proof. Assume without loss of generality that φ is of the form given in Lemma 4, with the first disjunct, $K\psi_1 \wedge \neg K\chi_{1,1} \wedge \cdots \wedge \neg K\chi_{1,m_1} \wedge \pi_1$, satisfiable. Hence, $\not\models \psi_1 \rightarrow \chi_{1,j}$ for all j $(1 \leq j \leq m_1)$, and $\not\models \psi_1 \rightarrow \neg\pi_1$. Let W be the set of assignments w such that $\models_w \psi_1$. Thus, for all j $(1 \leq j \leq m_1)$, there exists $w_j \in W$ such that $\models_{w_j} \neg\chi_{1,j}$; and there exists $w \in W$ such that $\models_w \pi_1$. By Lemma 3, $W \models TK\psi_1$. Thus,

$$W \models_w TK\psi_1 \wedge K\psi_1 \wedge \neg K\chi_{1,1} \wedge \cdots \wedge \neg K\chi_{1,m_1} \wedge \pi_1,$$

and hence $W \models_w TK\psi_1 \wedge \varphi$. □

Lemma 5 ensures that, in the propositional case, the assumption that there is a sentence which is the agent's total knowledge does not change the finitary logic of knowledge: any (basic) sentence which is satisfiable without this assumption is satisfiable in its presence. We show below that Lemma 5 is false for $\mathcal{FOLTK}$.

4 The restricted first-order case

The analysis of $\mathcal{FOLTK}$ is somewhat involved, and so we approach it in two stages. In the present section, we prove the main theorems of interest, under the assumption that certain formulas occurring in those theorems are *basic*. This assumption is then lifted in Section 5. The material of the present section was presented, in abbreviated form, in Pratt-Hartmann [8].

The following construction is crucial in understanding the behaviour of TK in the first-order case.

Definition 3. A *permutation of individuals* is a function from the set of names to the set of names which is 1–1 and onto. Let f be a permutation of individuals; we extend f to apply to interpretations and formulas as follows. If w is an interpretation, for any n-ary predicate letter r, let $r^{f(w)}$ be the set of tuples $\langle a_1, \ldots, a_n \rangle$ such that $\langle f^{-1}(a_1), \ldots, f^{-1}(a_n) \rangle \in r^w$. If x is a variable, let $f(x) = x$. If $r(t_1, \ldots, t_n)$ is an atomic formula, let $f(r(t_1, \ldots, t_n)) = r(f(t_1), \ldots, f(t_n))$, and let f be defined on non-atomic formulas by $f(\varphi \wedge \psi) = f(\varphi) \wedge f(\psi)$, $f(\varphi \vee \psi) = f(\varphi) \vee f(\psi)$, $f(\neg\varphi) = \neg f(\varphi)$, $f(\exists x\varphi) = \exists x f(\varphi)$, $f(\forall x\varphi) = \forall x f(\varphi)$, $f(K\varphi) = Kf(\varphi)$, $f(TK\varphi) = TKf(\varphi)$.

Thus, when applying f to interpretations and formulas, we switch round the extensions of predicates and the names occurring in formulas in corresponding ways.

Lemma 6. *If f is a permutation of individuals, then f is also 1–1 and onto on the set of interpretations, the set of formulas, the set of basic formulas and the set of objective formulas. Furthermore, for all sentences φ, sets of interpretations W and interpretations $w \in W$, $W \models_w \varphi$ if and only if $f(W) \models_{f(w)} f(\varphi)$.*

Proof. The first part of the lemma is obvious. The second part follows by structural induction on φ. □

Definition 4. Let $\bar{x} = x_1, \ldots, x_n$ be a tuple of distinct variables, and let X be the set of these variables: $X = \{x_1, \ldots, x_n\}$. Let $A = \{a_1, \ldots, a_m\}$ (with the a_i distinct) be a set of names. Let $P_1, \ldots, P_m$ be a set of (possibly empty) disjoint subsets of X and let $P_{m+1}, \ldots, P_{m+l}$ be a partition of $X \setminus \bigcup_{1 \leq i \leq m} P_i$. (Thus, $0 \leq l \leq n$.) A *distribution formula* (for $\bar{x}$ and A) is a satisfiable formula of the form $\delta(\bar{x}) :=$

$$\bigwedge\{x_j = a_i | 1 \leq i \leq m \text{ and } x_j \in P_i\}$$
$$\bigwedge\{x_j = x_k | m+1 \leq i \leq m+l \text{ and } x_j, x_k \in P_i\}$$
$$\bigwedge\{x_j \neq a_i | 1 \leq i \leq m, m+1 \leq i' \leq m+l \text{ and } x_j \in P_{i'}\}$$
$$\bigwedge\{x_j \neq x_k | m+1 \leq i < i' \leq m+l, x_j \in P_i \text{ and } x_k \in P_{i'}\}.$$

For a given $\bar{x}$ and A, denote the set of all such formulas by $\Delta_A(\bar{x})$. If $n = 0$, set $\Delta_A = \{\top\}$.

Intuitively, $\delta(\bar{x})$ assigns every variable in $\bar{x}$ to one of $m + l$ 'boxes' $P_1, \ldots, P_{m+l}$. Variables assigned to the same box are asserted to be identical, and variables assigned to different boxes are asserted to be distinct. Variables assigned to box P_i ($1 \leq i \leq m$) are asserted to be identical to a_i, and variables assigned to any box other than P_i ($1 \leq i \leq m$) are asserted to be distinct from a_i.

Lemma 7. *Let $\bar{x} = x_1, \ldots, x_n$ be a tuple of distinct variables, and A a set of names. Then $\Delta_A(\bar{x})$ is a partition. That is: $\models \forall\bar{x} \bigvee \Delta_A(\bar{x})$, and $\models \forall\bar{x}\neg(\delta(\bar{x}) \wedge \delta'(\bar{x}))$ for distinct $\delta(\bar{x})$, $\delta'(\bar{x}) \in \Delta_A(\bar{x})$.*

Proof. Obvious. □

We note that distribution formulas are *rigid*: they are satisfied by the same tuples regardless of the interpretation. Hence we sometimes write $\models \delta(\bar{a})$ without mentioning W or w.

Lemma 8. *Let φ be a sentence and $\psi(\bar{x})$ a formula. Let C be the set of names occurring in either formula. Then there exists a disjunction $\pi(\bar{x})$ of formulas in $\Delta_C(\bar{x})$ such that, for all tuples $\bar{a}$, $\models \pi(\bar{a})$ if and only if $\models \varphi \to \psi(\bar{a})$.*

Proof. Suppose that $\bar{a}$ and $\bar{a}'$ satisfy the same $\delta(\bar{x})$ in $\Delta_C(\bar{x})$, and that $\models \varphi \to \psi(\bar{a}')$. We claim that $\models \varphi \to \psi(\bar{a})$. For, in that case, the mapping $\bar{a}' \mapsto \bar{a}$ is well-defined, and extends to a permutation of individuals f such that f is the identity on C. Hence $f(\varphi) = \varphi$ and $f(\psi(\bar{a}')) = \psi(\bar{a})$. By Lemma 6, $\models \varphi \to \psi(\bar{a})$.

Now set $\pi(\bar{x}) :=$

$$\bigvee\{\delta(\bar{x}) \in \Delta_C(\bar{x}) : \models \delta(\bar{a}') \text{ for some } \bar{a}' \text{ s.t. } \models \varphi \to \psi(\bar{a}')\}.$$

(As usual, we take $\bigvee\emptyset$ to be $\bot$.) Suppose $\models \varphi \to \psi(\bar{a})$. Since $\Delta_C(\bar{x})$ is a partition, $\models \delta(\bar{a})$ for some $\delta(\bar{x}) \in \Delta_C(\bar{x})$. Therefore, $\models \pi(\bar{a})$. Conversely, suppose $\models \pi(\bar{a})$. Then $\models \delta(\bar{a})$ for some $\delta(\bar{x})$, such that, for some $\bar{a}'$, $\models \delta(\bar{a}')$ and $\models \varphi \to \psi(\bar{a}')$. But $\bar{a}$ and $\bar{a}'$ satisfy the same $\delta(\bar{x})$ in $\Delta_C(\bar{x})$; and we have just shown that, in this case, $\models \varphi \to \psi(\bar{a}')$ implies $\models \varphi \to \psi(\bar{a})$. □

Lemma 9. *Let φ be a sentence and $\psi(\bar{x})$ an objective formula. Let C be the set of names occurring in either formula. Then there exists a disjunction $\pi(\bar{x})$ of formulas in $\Delta_C(\bar{x})$ such that $\models TK\varphi \to \forall\bar{x}(K\psi(\bar{x}) \leftrightarrow \pi(\bar{x}))$.*

Proof. By Lemma 8 (with $K\varphi$ in place of φ), let $\pi(\bar{x})$ be such that, for all tuples $\bar{a}$, $\models \pi(\bar{a})$ if and only if $\models K\varphi \to \psi(\bar{a})$. Let W be any set of interpretations such that $W \models TK\varphi$, and let $\bar{a}$ be a tuple of names with the same arity as $\bar{x}$. Since $\psi(\bar{x})$ is objective, so is $\psi(\bar{a})$, and, by the semantics of TK, $W \models K\psi(\bar{a})$ if and only if $\models K\varphi \to \psi(\bar{a})$. That is: $W \models K\psi(\bar{a})$ if and only if $\models \pi(\bar{a})$. Thus, if $W \models TK\varphi$, then $W \models \forall\bar{x}(K\psi(\bar{x}) \leftrightarrow \pi(\bar{x}))$, as required. □

Lemma 9 is particularly useful in conjunction with the following (almost trivial) observation.

Lemma 10. *Let $\eta(\bar{x})$ be subjective, and let $\pi(\bar{x})$ be rigid. Let $\xi(\bar{y})$ be any formula containing an occurrence of $\eta(\bar{x})$ that does not lie within the scope of a TK-operator. Finally, let $\xi'(\bar{y})$ be the result of substituting, for that occurrence, the formula $\pi(\bar{x})$. If $W \models \forall\bar{x}(\eta(\bar{x}) \leftrightarrow \pi(\bar{x}))$, then $W \models \forall\bar{y}(\xi(\bar{y}) \leftrightarrow \xi'(\bar{y}))$.*

Proof. Routine structural induction on ξ. □

The condition that the occurrence of $\eta(\bar{x})$ being substituted does not occur within the scope of a TK-operator in $\xi(\bar{y})$ is essential in Lemma 10. To see this, let p be a proposition letter. By Lemma 3, TKp is satisfiable, and by Lemma 2, TKp is logically equivalent to $TK(Kp)$. Let $W \models TK(Kp)$, then. Trivially, $W \models Kp \leftrightarrow \top$, and, of course, the sentence $\top$ is rigid. But $W \not\models TK\top$.

We are now in a position to establish our first main result.

Theorem 1. *Let φ be a sentence and $\psi(\bar{x})$ a basic formula. Then there is a disjunction $\pi(\bar{x})$ of elements of $\Delta_C(\bar{x})$ for some C, such that $\models TK\varphi \to \forall\bar{x}(K\psi(\bar{x}) \leftrightarrow \pi(\bar{x}))$.*

Proof. We proceed by induction on the number n of occurrences of K in $\psi(\bar{x})$. The case $n = 0$ is handled by Lemma 9. If $n > 0$, let $K\psi'(\bar{x}')$ be a subformula of $\psi(\bar{x})$, with $\psi'(\bar{x}')$ objective. By Lemma 9, let $\pi'(\bar{x}')$ be such that $\models TK\varphi \to \forall\bar{x}'(K\psi'(\bar{x}') \leftrightarrow \pi'(\bar{x}'))$, and let $\psi''(\bar{x})$ be the result of substituting $\pi'(\bar{x}')$ for $K\psi'(\bar{x}')$ in $\psi(\bar{x})$. Since $\psi(\bar{x})$ is basic, the embedded occurrence of $K\psi'(\bar{x}')$ cannot occur within the scope of a TK-operator, so that, by Lemma 10, $\models TK\varphi \to \forall\bar{x}(K\psi(\bar{x}) \leftrightarrow K\psi''(\bar{x}))$. Since $\psi''(\bar{x})$ has fewer than n occurrences of K, the result follows by inductive hypothesis. □

Theorem 1 restricts $\psi(\bar{x})$ to be basic, so that Lemma 10 can be applied. As explained above, this step is not valid if TK-operators are present. In fact, Theorem 1 does hold in the general case, but the induction is more delicate, and we postpone it until Section 5.

Corollary 1. *For all sentences φ and all basic sentences ψ, $\models TK\varphi \to K\psi$ or $\models TK\varphi \to \neg K\psi$.*

Proof. By Theorem 1, $\models TK\varphi \to \forall\bar{x}(K\psi \leftrightarrow \pi)$, where π is a disjunction of elements of Δ_C for some C (with a 0-tuple of variables). Hence π is $\bot$ or $\top$. □

Corollary 2. *Let φ be a sentence and $\psi(x)$ a basic formula with one free variable. Suppose that $W \models TK\varphi$. Then the set $\{a : W \models K\psi(a)\}$ is finite or cofinite.*

Proof. By Theorem 1, $\models TK\varphi \to \forall x(K\psi(x) \leftrightarrow \pi(x))$, where $\pi(x)$ is a disjunction of elements of $\Delta_C(x)$ for some C (with a single variable x). Clearly, the set of a satisfying $\pi(x)$ is finite or cofinite. □

Recall that, in the propositional case, if φ is satisfiable, then we can find ψ such that $\varphi \wedge TK\psi$ is satisfiable. In the first-order case, this is no longer true.

Theorem 2. *There exists a satisfiable basic sentence φ such that, for all sentences ψ, $\models \varphi \to \neg TK\psi$.*

Proof. Consider the sentence φ^+ given by

$$\begin{aligned}&\exists x(Kp(x) \wedge \forall y \neg r(y,x)) \wedge \\ &\qquad\qquad \forall x(Kp(x) \to \exists y(Kp(y) \wedge r(x,y))) \wedge \\ &\qquad\qquad\qquad \forall x(Kp(x) \to \forall y \forall z(r(y,x) \wedge r(z,x) \to y = z)).\end{aligned}$$

It is obvious that, if $W \models_w \varphi^+$, then, for infinitely many a, $W \models Kp(a)$. Now let φ^- be a similar sentence implying that, for infinitely many a, $\neg Kp(a)$. Let φ be $\varphi^+ \wedge \varphi^-$. It is not difficult to see that φ is satisfiable. On the other hand, by Corollary 2, $\models \varphi \to \neg TK\psi$ for all sentences ψ. □

Thus, in the first-order case, the assumption that there is total knowledge changes the finitary logic of knowledge: basic sentences that are satisfiable in the absence of this assumption may be unsatisfiable in its presence.

Next, we show that total knowledge of any basic sentence is logically equivalent to total knowledge of an objective sentence. We need the following general lemma.

Lemma 11. *Let φ and ψ be sentences such that $\models TK\varphi \to K\psi$, $\models K\psi \to K\varphi$ and $TK\varphi$ is satisfiable. Then $\models TK\varphi \leftrightarrow TK\psi$.*

Proof. Suppose $W \models TK\varphi$. Then $W \models K\psi$. Let χ be objective with $\not\models K\psi \to \chi$. Then $\not\models K\varphi \to \chi$, because $\models K\psi \to K\varphi$. So $W \models \neg K\chi$. Hence $W \models TK\psi$.

Conversely, suppose $W \models TK\psi$. Since $\models K\psi \to K\varphi$, we certainly have $W \models K\varphi$. To establish $W \models TK\varphi$, we suppose χ is objective with $\not\models K\varphi \to \chi$, and show that $W \models \neg K\chi$. Now, $\not\models K\varphi \to \chi$ implies $\models TK\varphi \to \neg K\chi$. This in turn implies $\not\models K\psi \to \chi$, since otherwise, given that $\models TK\varphi \to K\psi$, we would have $\models TK\varphi \to K\chi$, contradicting the hypothesized consistency of $TK\varphi$. Since $W \models TK\psi$ and $\not\models K\psi \to \chi$, we have $W \models \neg K\chi$, as required. □

We mention, as an aside, the following reassuring corollary of Lemma 11.

Corollary 3. *For all sentences φ, $\models TK\varphi \leftrightarrow TK(TK\varphi)$.*

Proof. If $\models TK\varphi$ is unsatisfiable, then $\models TK(TK\varphi)$ is also unsatisfiable, so the result is immediate. On the other hand, if $\models TK\varphi$ is satisfiable, then, since $\models TK\varphi \to K(TK\varphi)$ and $\models K(TK\varphi) \to K\varphi$, the result follows by taking ψ to be $TK\varphi$ in Lemma 11. □

Theorem 3. *Let φ be a basic sentence. Then there exists an objective sentence φ^* such that $\models TK\varphi \leftrightarrow TK\varphi^*$.*

Proof. If φ is already objective or if $TK\varphi$ is unsatisfiable, the result is trivial, so we may assume otherwise. Let $K\psi_1(\bar{x}_1)$ be a subformula of φ, with $\psi_1(\bar{x}_1)$ objective. Then we can find ρ_1 such that $\models TK\varphi \to \rho_1$, where ρ_1 is the sentence $\forall\bar{x}_1(K\psi(\bar{x}_1) \leftrightarrow \pi_1(\bar{x}))$ constructed as in Lemma 9. Let φ_1 be the result of substituting $\pi_1(\bar{x})$ for $K\psi(\bar{x}_1)$ in φ. Since φ is basic, Lemma 10 implies $\models TK\varphi \to (\varphi_1 \leftrightarrow \varphi)$. It follows that $\models TK\varphi \to K(\varphi_1 \wedge \rho_1)$. Similarly, $\models K(\varphi_1 \wedge \rho_1) \to K\varphi$. Since $TK\varphi$ is assumed satisfiable, Lemma 11 implies that $\models TK\varphi \leftrightarrow TK(\varphi_1 \wedge \rho_1)$. If there is a subformula $K\psi_2(\bar{x}_2)$ in φ_1 with $\psi_2(\bar{x}_2)$ objective, we proceed as before, obtaining $\models TK\varphi \leftrightarrow TK(\varphi_2 \wedge \rho_1 \wedge \rho_2)$, and so on, until we eventually obtain $\models TK\varphi \leftrightarrow TK(\varphi_m \wedge \rho_1 \wedge \cdots \wedge \rho_m)$, with φ_m objective and $m \geq 1$.

Now consider in more detail the sentence $\rho_1 \wedge \cdots \wedge \rho_m$. Ignoring the previous numbering, this may be written out as a conjunction of the form

$$\bigwedge_{1 \leq j \leq M} \forall\bar{x}_j(\delta_j(\bar{x}_j) \to K\psi_j(\bar{x}_j)) \wedge \bigwedge_{1 \leq j \leq M'} \forall\bar{x}'_j(\delta'_j(\bar{x}'_j) \to \neg K\psi'_j(\bar{x}'_j)),$$

where the $\delta_j(\bar{x}_j)$, $\delta'_j(\bar{x}'_j)$ are conjunctions of equality and inequality formulas, and the $\psi_j(\bar{x}_j)$, $\psi'_j(\bar{x}'_j)$ are objective. Since the $\delta_j(\bar{x}_j)$ are rigid, we have

$$\models K\forall\bar{x}_j(\delta_j(\bar{x}_j) \to K\psi_j(\bar{x}_j)) \leftrightarrow K\forall\bar{x}_j(\delta_j(\bar{x}_j) \to \psi_j(\bar{x}_j)).$$

Hence we can omit the K from the relevant conjuncts and set φ^* to be

$$\varphi_m \wedge \bigwedge_{1 \leq j \leq M} \forall\bar{x}_j(\delta_j(\bar{x}_j) \to \psi_j(\bar{x}_j)),$$

whence $\models TK\varphi \leftrightarrow TK(\varphi^* \wedge \sigma_1 \wedge \cdots \wedge \sigma_{M'})$ where σ_j is $\forall\bar{x}'_j(\delta'_j(\bar{x}'_j) \to \neg K\psi'_j(\bar{x}'_j))$.

We claim that, if $\bar{a}$ is a tuple of names such that $\models \delta'_j(\bar{a})$, then $\not\models K\varphi^* \to \psi'_j(\bar{a})$. For suppose, to the contrary, that $\models \delta'_j(\bar{a})$, but $\models K\varphi^* \to \psi'_j(\bar{a})$. Since

$\models TK\varphi \to \forall \bar{x}'_j(\delta'_j(\bar{x}'_j) \to \neg K\psi'_j(\bar{x}'_j))$, we have $\models TK\varphi \to \neg K\psi'_j(\bar{a})$. On the other hand, since $\models TK\varphi \to \varphi^*$, we have $\models TK\varphi \to \psi'_j(\bar{a})$, contradicting the satisfiability of $TK\varphi$. Thus, if $\models \delta'_j(\bar{a})$, then $\not\models K\varphi^* \to \psi'_j(\bar{a})$, whence, since $\psi'_j(\bar{a})$ is objective, $\models TK\varphi^* \to \neg K\psi'_j(\bar{a}'_j)$. We have therefore shown

$$\models TK\varphi^* \to \forall \bar{x}'_j(\delta'_j(\bar{x}'_j) \to \neg K\psi'_j(\bar{x}'_j))$$

for all j $(1 \leq j \leq M')$. Hence,

$$\models TK\varphi^* \to K(\varphi^* \wedge \sigma_1 \wedge \cdots \wedge \sigma_{M'}). \tag{1}$$

Trivially,

$$\models K(\varphi^* \wedge \sigma_1 \wedge \cdots \wedge \sigma_{M'}) \to K\varphi^*. \tag{2}$$

And finally, by Lemma 3, $TK\varphi^*$ is satisfiable. Using Lemma 11, (1) and (2) imply $\models TK\varphi^* \leftrightarrow TK(\varphi^* \wedge \sigma_1 \wedge \cdots \wedge \sigma_{M'})$, and so $\models TK\varphi^* \leftrightarrow TK\varphi$. □

5 The unrestricted first-order case

The main task of this section is to strengthen Corollary 1 and Theorem 3 so that the restrictions to basic sentences can be removed.

Definition 5. If $\varphi(\bar{x})$ is a formula, the TK-rank of $\varphi(\bar{x})$ is the maximum depth of nesting of TK-operators in $\varphi(\bar{x})$. A *TK-formula* is any formula of the form $TK\varphi(\bar{x})$.

Thus, the TK-rank of $\varphi(\bar{x})$ is zero if and only if $\varphi(\bar{x})$ is basic.

We may now state and prove the promised result on the epistemic categoricity of the TK-operator.

Theorem 4. *Let φ and ψ be any sentences. Then $\models TK\varphi \to K\psi$ or $\models TK\varphi \to \neg K\psi$.*

Before giving the proof of this theorem, we derive two useful corollaries.

Corollary 4. *Suppose $\models TK\varphi \wedge TK\psi$ is a satisfiable sentence. Then $\models TK\varphi \to TK\psi$.*

Proof. Since $TK\psi \to K\psi$, we certainly cannot have $\models TK\varphi \to \neg K\psi$. By Theorem 4, then, $\models TK\varphi \to K\psi$. We claim that, for any objective formula χ,

$$\models K\varphi \to \chi \text{ implies } \models K\psi \to \chi.$$

For suppose $\models K\varphi \to \chi$, and $\not\models K\psi \to \chi$. Then $\models TK\varphi \to K\chi$, and $\models TK\psi \to \neg K\chi$, contradicting the satisfiability of $TK\varphi \wedge TK\psi$. It then easily follows that $\models TK\varphi \to TK\psi$. The reverse implication follows symmetrically. □

The following result is also a corollary of Theorem 4.

Lemma 12. *Let φ be a sentence and $\psi(\bar{x})$ a formula. Let C be the set of names occurring in either φ or $\psi(\bar{x})$. Define*

$$\pi(\bar{x}) \ := \ \bigvee\{\delta(\bar{x}) \in \Delta_C(\bar{x}) : \models \delta(\bar{a}) \text{ for some } \bar{a} \text{ s.t. } \models TK\varphi \to TK\psi(\bar{a})\}.$$

Then $\models TK\varphi \to \forall\bar{x}(TK\psi(\bar{x}) \leftrightarrow \pi(\bar{x}))$.

Proof. If $\models TK\varphi \to TK\psi(\bar{a})$ then, since $\Delta_C(\bar{x})$ is a partition, $\bar{a}$ satisfies some disjunct of $\pi(\bar{x})$, so that $\models \pi(\bar{a})$. Conversely, if $\models \pi(\bar{a})$, then we can find $\bar{a}'$ such that $\bar{a}$ and $\bar{a}'$ satisfy the same disjunct of $\pi(\bar{x})$ and $\models TK\varphi \to TK\psi(\bar{a}')$. Since $\bar{a}$ and $\bar{a}'$ satisfy the same disjunct of $\pi(\bar{x})$, we can find a permutation of individuals f such that $f : \bar{a}' \mapsto \bar{a}$ and f is the identity on C. Thus, $f(TK\varphi) = TK\varphi$ and $f(TK\psi(\bar{a}')) = TK\psi(\bar{a})$. By Lemma 6, $\models TK\varphi \to TK\psi(\bar{a})$. Thus, for any tuple $\bar{a}$, $\models \pi(\bar{a})$ if and only if $\models TK\varphi \to TK\psi(\bar{a})$.

Let W be any set of interpretations and let $\bar{a}$ be any tuple such that $W \models TK\varphi \wedge TK\psi(\bar{a})$. By Corollary 4, $\models TK\varphi \to TK\psi(\bar{a})$. By the result of the previous paragraph, $\models \pi(\bar{a})$. Hence $\models TK\varphi \to \forall\bar{x}(TK\psi(\bar{x}) \to \pi(\bar{x}))$. Conversely, let W be any set of interpretations and $\bar{a}$ any tuple such that $W \models TK\varphi \wedge \pi(\bar{a})$. By the result of the previous paragraph, $\models TK\varphi \to TK\psi(\bar{a})$, so $W \models TK\psi(\bar{a})$. Hence $\models TK\varphi \to \forall\bar{x}(\pi(\bar{x}) \to TK\psi(\bar{x}))$. □

Before we proceed to the proof of Theorem 4, we should note how it is used by Corollary 4 and Lemma 12. Suppose that the sentence ψ in the statement of Corollary 4 has TK-rank less than N, for some integer N. Then, for the proof to go through, we need only assume that Theorem 4 holds for sentences ψ with TK-rank less than N. Likewise, suppose that the formula $\psi(\bar{x})$ in the statement of Lemma 12 has TK-rank less than N, for some integer N. Again, for the proof to go through, we need only assume that Corollary 4 holds for sentences $\psi(\bar{a})$ with TK-rank less than N, and hence we need only assume that Theorem 4 holds for sentences with TK-rank less than N. This observation helps us to present the proof of Theorem 4 more succinctly.

Proof of Theorem 4: We proceed by induction on the TK-rank N of ψ. The case $N = 0$ is dealt with by Corollary 1.

Suppose $N > 0$ and that the result holds for all φ whatever, and for all ψ with TK-rank less than N. Let $TK\psi'(\bar{x})$ be a maximal TK-subformula of ψ, so that the TK-rank of of $\psi'(\bar{x})$ is less than N. Hence we can apply Lemma 12 by inductive hypothesis, obtaining a rigid formula $\pi(\bar{x})$ such that

$$\models TK\varphi \to \forall\bar{x}(TK\psi'(\bar{x}) \leftrightarrow \pi(\bar{x})).$$

Now let ψ_1 be the result of substituting $\pi(\bar{x})$ for $TK\psi'(\bar{x})$ in ψ. Since $TK\psi'(\bar{x})$ is a maximal TK-formula in ψ—that is, it does not occur embedded in any TK-operator, we infer, by Lemma 10, that $\models TK\varphi \to (K\psi \leftrightarrow K\psi_1)$. Proceeding in the same way, we eliminate the remaining maximal TK-subformulas of ψ_1 obtaining, say, the basic sentence ψ_m with $\models TK\varphi \to (K\psi \leftrightarrow K\psi_m)$. The result then follows from Corollary 1. □

Our final objective is to generalize Theorem 3.

Lemma 13. *Let φ, χ be basic sentences and let $\psi_1(\bar{x}_1), \ldots, \psi_n(\bar{x}_n)$ be any formulas. Then $\models \varphi \to \chi$ if and only if*

$$\models (\varphi \wedge \bigwedge_{1 \le i \le n} \forall \bar{x}_i \neg TK\psi_i(\bar{x}_i)) \to \chi.$$

Proof. For the non-trivial direction, suppose that $W \models_w \varphi \wedge \neg\chi$. Let $p_1, p_2, \ldots$ be the proposition letters not occurring in φ or χ, and let the arity of $\bar{x}_i$ be m_i. For all i ($1 \le i \le n$), let $\bar{a}_{i,1}, \bar{a}_{i,2}, \ldots$ be any enumeration of all the m_i-tuples of names. Now modify the interpretations in W according to the following recipe:

> **for** $j = 1, 2, \ldots$ **do**
>   **for** $i = 1, 2, \ldots, n$ **do**
>   **if** $\models TK\psi_i(\bar{a}_{i,j}) \to Kp_{jn+i}$ **then**
>   make p_{jn+i} false at some interpretation in W
>   **else**
>   make p_{jn+i} true at all interpretations in W.

Let the result of this process be W^*. We claim that, for all i ($1 \le i \le n$), $W^* \models \forall \bar{x}_i \neg TK\psi_i(\bar{x}_i)$. For let $\bar{a}$ be any m_i-tuple of names. Then, for some j, $\bar{a} = \bar{a}_{i,j}$. Since p_{jn+i} is objective, either $\models TK\psi_i(\bar{a}_{i,j}) \to Kp_{jn+i}$ or $\models TK\psi_i(\bar{a}_{i,j}) \to \neg Kp_{jn+i}$. In the former case, by construction of W^*, $W^* \models \neg Kp_{jn+i}$; in the latter, $W^* \models Kp_{jn+i}$. Either way, $W^* \not\models TK\psi_i(\bar{a}_{i,j})$. It follows that $W^* \models \forall \bar{x}_i \neg TK\psi_i(\bar{x}_i)$, as required. On the other hand, since φ and χ are basic and do not mention $p_1, p_2, \ldots$, we still have $W^* \models_{w^*} \varphi \wedge \neg\chi$ for some $w^* \in W^*$. Thus, $\not\models (\varphi \wedge \bigwedge_{1 \le i \le n} \forall \bar{x}_i \neg TK\psi_i(\bar{x}_i)) \to \chi$. □

Lemma 14. *Let φ be a basic sentence and let $\psi_1(\bar{x}_1), \ldots, \psi_n(\bar{x}_n)$ be any formulas. Define*

$$\theta \ := \ \varphi \wedge \bigwedge_{1 \le i \le n} \forall \bar{x}_i \neg TK\psi_i(\bar{x}_i).$$

Then $\models TK\theta \to TK\varphi$.

Proof. Observe that

$$K\theta \leftrightarrow K\varphi \wedge \bigwedge_{1 \le i \le n} \forall \bar{x}_i \neg TK\psi_i(\bar{x}_i). \tag{3}$$

Now suppose $W \models TK\theta$. Certainly, then, $W \models K\varphi$. Now let χ be objective with $\not\models K\varphi \to \chi$. By Lemma 13 (using $K\varphi$ in pace of φ), and by (3), we have $\not\models K\theta \to \chi$, and so $W \models \neg K\chi$. Thus, $W \models TK\varphi$. □

Lemma 15. *Let φ be a basic sentence and let $\psi_1(\bar{x}_1), \ldots, \psi_n(\bar{x}_n)$ be any formulas. Suppose $TK\theta$ is satisfiable, where*

$$\theta \;:=\; \varphi \wedge \bigwedge_{1 \leq i \leq n} \forall \bar{x}_i \neg TK\psi_i(\bar{x}_i).$$

Then $\models TK\varphi \leftrightarrow TK\theta$.

Proof. By Lemma 14, $\models TK\theta \rightarrow TK\varphi$. It follows that $TK\varphi$ is satisfiable.

Certainly, $\models TK\varphi \rightarrow K\varphi$. We claim further that $TK\varphi \rightarrow \forall \bar{x}_i \neg TK\psi_i(\bar{x}_i)$ for all i ($1 \leq i \leq n$). For suppose otherwise: let $\bar{a}$ be a tuple of names such that $TK\varphi \wedge TK\psi_i(\bar{a})$ is satisfiable. By Corollary 4, $\models TK\varphi \rightarrow TK\psi_i(\bar{a})$, and since $\models TK\theta \rightarrow TK\varphi$, we have both $\models TK\theta \rightarrow TK\psi_i(\bar{a})$ and $\models TK\theta \rightarrow \neg TK\psi_i(\bar{a})$, contradicting the satisfiability of $TK\theta$. This establishes the claim.

Thus, we have $\models TK\varphi \rightarrow K\theta$; and trivially, $\models K\theta \rightarrow K\varphi$. Since $TK\varphi$ is satisfiable, Lemma 11 yields $TK\varphi \leftrightarrow TK\theta$. □

Theorem 5. *Let φ be any sentence. Then there exists an objective sentence φ^* such that $\models TK\varphi \leftrightarrow TK\varphi^*$.*

Proof. We proceed by induction on the TK-rank N of φ. The case $N = 0$ is dealt with by Theorem 3.

Suppose that $N > 0$ and that the result holds for all φ with TK-rank less than N. We may suppose that $TK\varphi$ is satisfiable, since otherwise we need only set $\varphi^* := \bot$. Let $TK\psi_1(\bar{x}_1), \ldots, TK\psi_n(\bar{x}_n)$ be the maximal TK-subformulas of φ (i.e. those not occurring within the scope of any other TK-operator). By Lemma 12, we can find rigid formulas $\pi_i(\bar{x}_i)$ ($1 \leq i \leq n$) such that

$$\models TK\varphi \rightarrow \bigwedge_{1 \leq i \leq n} \forall \bar{x}_i(\pi_i(\bar{x}_i) \leftrightarrow TK\psi_i(\bar{x}_i)). \tag{4}$$

Therefore, $\models TK\varphi \rightarrow K(\varphi \wedge \bigwedge_{1 \leq i \leq n} \forall \bar{x}_i(\pi_i(\bar{x}_i) \leftrightarrow TK\psi_i(\bar{x}_i)))$. Trivially, $\models K(\varphi \wedge \bigwedge_{1 \leq i \leq n} \forall \bar{x}_i(\pi_i(\bar{x}_i) \leftrightarrow TK\psi_i(\bar{x}_i))) \rightarrow K\varphi$; and we are assuming that $TK\varphi$ is satisfiable. Hence, by Lemma 11,

$$\models TK\varphi \leftrightarrow TK\left(\varphi \wedge \bigwedge_{1 \leq i \leq n} \forall \bar{x}_i(\pi_i(\bar{x}_i) \leftrightarrow TK\psi_i(\bar{x}_i))\right).$$

Let $\tilde{\varphi}$ be the result of substituting $\pi_i(\bar{x}_i)$ for $TK\psi_i(\bar{x}_i)$ in φ, for every i ($1 \leq i \leq n$). Define

$$\theta \;:=\; \tilde{\varphi} \wedge \bigwedge_{1 \leq i \leq n} \forall \bar{x}_i(\pi(\bar{x}_i) \leftrightarrow TK\psi_i(\bar{x}_i)).$$

Since $\models \theta \rightarrow \forall \bar{x}_i(\pi_i(\bar{x}_i) \leftrightarrow TK\psi_i(\bar{x}_i))$ for all i ($1 \leq i \leq n$), and since the $TK\psi_i(\bar{x}_i)$ do not occur within the scope of any TK-operators in φ, Lemma 10 yields $\models \theta \rightarrow (\varphi \leftrightarrow \tilde{\varphi})$, whence

$$\models K\theta \rightarrow K\varphi. \tag{5}$$

Similarly, from (4), Lemma 10 yields $\models TK\varphi \rightarrow (\tilde{\varphi} \leftrightarrow \varphi)$. Combining this again with (4), we easily obtain

$$\models TK\varphi \rightarrow K\theta. \tag{6}$$

From (5) and (6), and the satisfiability of $TK\varphi$, we obtain, by Lemma 11, $TK\varphi \leftrightarrow TK\theta$. It follows that $TK\theta$ is satisfiable.

We remark that, by construction, the TK-rank of each of the $\psi_i(\bar{x}_i)$ is less than N, and that $\tilde{\varphi}$ is basic. We consider two cases:

Case 1: $\pi_i(\bar{x}_i)$ is satisfiable for some i $(1 \leq i \leq n)$. Let $\bar{a}$ be a tuple such that $\models \pi_i(\bar{a})$. Then, since $TK\theta$ is satisfiable, so is $TK\theta \wedge TK\psi_i(\bar{a})$. By Corollary 4, $\models TK\theta \leftrightarrow TK\psi_i(\bar{a})$, and since the TK-rank of $\psi_i(\bar{a})$ is less than N, by inductive hypothesis, we can find an objective φ^* such that $\models TK\psi_i(\bar{a}) \leftrightarrow TK\varphi^*$. Hence, $\models TK\varphi \leftrightarrow TK\varphi^*$ and we are done.

Case 2: $\pi_i(\bar{x}_i)$ is unsatisfiable for every i $(1 \leq i \leq n)$. Then

$$\models TK\theta \leftrightarrow TK\left(\tilde{\varphi} \wedge \bigwedge_{1\leq i\leq n} \forall \bar{x}_i \neg TK\psi_i(\bar{x}_i)\right).$$

Moreover, since $TK\theta$ is satisfiable, Lemma 15 yields $\models TK\theta \leftrightarrow TK\tilde{\varphi}$. Since $\tilde{\varphi}$ is basic, Theorem 3 guarantees the existence of an objective φ^* such that $\models TK\tilde{\varphi} \leftrightarrow TK\varphi^*$. Hence $\models TK\varphi \leftrightarrow TK\varphi^*$ and we are done. □

6 Introducing time

So far, we have finessed the issue of time in the language $\mathcal{FOLTK}$, treating the temporal reference in $K\varphi$, $TK\varphi$, in the context of an agent's evolving degrees of belief, as implicit. However, it is straightforward to introduce temporal arguments to these operators; that is the subject of this section.

For simplicity, let us assume that time is discrete, and that, after some finite time, all knowledge ceases. The former assumption is reasonable for artificial agents; the latter is reasonable for all agents, natural or artificial. Fix some integer N, then, and consider two families of operators: $\{K_n\}_{n\leq N}$, and $\{TK_n\}_{n\leq N}$, where $K_n\varphi$ has the interpretation "the agent knows at time n that φ", and similarly for $TK_n\varphi$. In addition, we shall assume that the agent never forgets. Again, this assumption is reasonable for some kinds of artificial agent. In this context, it is sufficient to restrict attention to formulas which are 'forward-looking'—that is, which feature no temporal operator with index m embedded within the scope of a temporal operator with a larger index n. After all, at the time $n-1$, we may assume all formulas of the forms $K_m\psi$ and $TK_m\psi$ to have probability 1 or 0, and therefore not to be any longer relevant for the purpose of conditionalization. Accordingly, we define a series of languages $\mathcal{FOLTK}_n^N$ $(1 \leq n \leq N)$ as follows:

if r is an m-ary predicate letter and $t_1, \ldots, t_m$ are terms, then $r(t_1, \ldots, t_m)$ is a formula of $\mathcal{FOLTK}_n^N$;

if $1 \leq n < n' \leq N$, then any formula of $\mathcal{FOLTK}_{n'}^N$ is a formula of $\mathcal{FOLTK}_n^N$;

if φ and ψ are formulas of $\mathcal{FOLTK}_n^N$ and x is a variable, then $\varphi \wedge \psi$, $\varphi \vee \psi$, $\neg\varphi$, $\exists x\varphi$, $\forall x\varphi$, $K_n\varphi$ and $TK_n\varphi$ are formulas of $\mathcal{FOLTK}_n^N$.

Finally, we let $\mathcal{FOLTK}^N = \mathcal{FOLTK}_1^N$. Thus, for example, $TK_1\neg K_2 p$ and $TK_2\neg K_2 p$ are $\mathcal{FOLTK}^2$-formulas, but $K_2\neg TK_1 p$ (which is not 'forward-looking') is not. It helps to think of the languages $\mathcal{FOLTK}_n^N$ as being constructed in order of decreasing values of n. Evidently, $\mathcal{FOLTK}_N^N$ is just a notational variant of $\mathcal{FOLTK}$, with the operators K and TK replaced by K_N and TK_N, respectively; and $\mathcal{FOLTK}_n^N$ is the fragment of $\mathcal{FOLTK}^N$ involving no modal operators with indices less than n.

For convenience of expression, we call a formula $\varphi \in \mathcal{FOLTK}^N$ *n-objective* if it involves no operators K_m or TK_m with $m \leq n$. Thus, the notion of an n-objective formula is the natural generalization of the notion of an objective formula in the context of $\mathcal{FOLTK}$. In particular, an N-objective $\mathcal{FOLTK}^N$-formula involves no occurrences of modal operators at all; and an n-objective $\mathcal{FOLTK}^N$-formula ($n < N$) is simply an $\mathcal{FOLTK}_{n+1}^N$-formula.

The semantics of $\mathcal{FOLTK}^N$ can then be given along the same lines as for $\mathcal{FOLTK}$; the chief difference is that formulas must now be evaluated with respect to a structure modelling the evolving knowledge of the agent. Accordingly, let us define a structure to be a pair $M = \langle W, \{\sim_n\}_{1 \leq n \leq N}\rangle$, where W is a set of interpretations, and each $\sim_n$ $(1 \leq n \leq N)$ an equivalence relation on W. To capture the assumption that our agent does not forget previously acquired information, we take $m \leq n$ to imply $\sim_m \supseteq \sim_n$. We can now define the truth-conditions for $\mathcal{FOLTK}^N$-formulas recursively in much the same way as for $\mathcal{FOLTK}$. We give here only the clauses governing the modal operators, since the remaining clauses are essentially identical to the non-temporal case. Taking M to be the structure $\langle W, \{\sim_n\}_{1 \leq n \leq N}\rangle$, we define:

$M \models_w K_n\varphi$ if and only if, for all $w' \in W$ such that $w \sim_n w'$, $M \models_{w'} \varphi$;

$M \models_w TK_n\varphi$ if and only if $M \models_w K_n\varphi$ and $M \models_w \neg K_n\chi$ for all n-objective sentences χ such that $\not\models K_n\varphi \rightarrow \chi$.

It is easy to see that, for all $N \geq 1$, the recursion in in these clauses terminates properly.

The theorems established in Sections 3–5 then transfer unproblematically to the temporalized case. In particular, we have the following two facts for all n $(1 \leq n \leq N)$. (i) Let φ and ψ be any sentences of $\mathcal{FOLTK}_n^N$. Then $\models TK_n\varphi \rightarrow K_n\psi$ or $\models TK_n\varphi \rightarrow \neg K_n\psi$. (ii) Let φ be any sentence of $\mathcal{FOLTK}_n^N$. Then there exists an n-objective sentence φ^* such that $\models TK_n\varphi \leftrightarrow TK_n\varphi^*$.

Finally, we mention the problem of whether the assumption of mortality could be relaxed. That is: could we devise a single (finitary) language featuring operators $K_1, K_2, \ldots$ and $TK_1, TK_2, \ldots$? Formally, the most obvious relaxation of the semantics given above for $\mathcal{FOLTK}^N$ would yield non-terminating recursive definitions. Moreover, informally, it is unclear that the sought-after operators would make sense. For consider the set of formulas

$$TK_1(\neg TK_2\top), \;\; TK_2(\neg TK_3\top), \;\; TK_3(\neg TK_4\top), \;\; \ldots$$

Could all these propositions be true together? On the one hand, they contain nothing we would wish to regard as an obvious consistency; on the other, it is disquieting that an agent should be regarded as having, as his only knowledge at any given time, the knowledge that, at the next time, he will know more. We should regard this question as, at present, unsolved. However, as we remarked above, the assumption of a finite life-span is reasonable for all agents, natural or artificial.

7 Conclusion

We began this paper by recounting the well-known diachronic Dutch book argument in support of Bayesian inference—the policy of revising degrees of belief by conditionalization. We showed that this argument is most persuasive in the special case of an artificial agent whose language of thought contains *epistemically categorical* propositions—that is, propositions the mere truth of which determine the agent's knowledge, and hence its degrees of belief. We set about constructing such a language. Specifically, we defined a first-order language, $\mathcal{FOLTK}$, containing modal operators K and TK having the interpretations "The agent knows that ..." and "The agent's total knowledge is that ...", respectively. We showed that all propositions of the form $TK\psi$ are epistemically categorical in the required sense. We further showed that, for such an agent, total knowledge is always total knowledge of an objective proposition. We indicated briefly how these operators could be transferred to a (bounded) temporal setting. For artificial agents employing such a language, Bayesian inference is the correct belief-revision policy: if the agent's degrees of belief are represented by the probability distribution p, and its new total knowledge is the objective proposition ψ, then the agent should revise its degrees of belief by conditionalizing on $TK\psi$. Put another way: for such agents, the language $\mathcal{FOLTK}$ represents the right degree of epistemic ascent to save the diachronic Dutch book argument.

References

[1] Maya Bar-Hillel and Ruma Falk. Some teasers concerning conditional probabilities. *Cognition*, 11:109–122, 1982.

[2] Joseph Y. Halpern. *Reasoning about Uncertainty*. MIT Press, Cambridge, MA, 2003.

[3] R. Jeffrey. Conditioning, kinematics and exchangeability. In B. Skyrms and W. Harper, editors, *Causation, Chance and Credence*, volume 1. Kluwer, Dordrecht, 1988.

[4] Richard C. Jeffrey. Probability and the art of judgment. In P.A. Chinstein and O. Hannaway, editors, *Observation, Experiment and Hypothesis in Modern Physical Science*. Cambridge University Press, MIT Press, 1985.

[5] Hector J. Levesque. All I know: a study in autoepistemic logic. *Artificial Intelligence*, 42:263–309, 1990.

[6] Hector K. Levesque and Gerhard Lakemeyer. *The Logic of Knowledge Bases*. MIT Press, Cambridge, MA, 2000.

[7] J.B. Paris. *The Uncertain Reasoner's Companion: a Mathematical Perspective*. Cambridge Tracts in Theoretical Computer Science. Cambridge University Press, Cambridge, 1995.

[8] Ian Pratt-Hartmann. Total knowledge. In *Proceedings of the National Conference of the American Association for Artificial Intelligence (AAAI-00)*, pages 423–428, Cambridge, MA, 2000. MIT Press/AAAI Press.

[9] Paul Teller. Conditionalization and observation. *Synthese*, 26:218–258, 1973.