2006

TR-2006004: Justified Knowledge Is Sufficient

Evangelia Antonakos

Justified Knowledge is Sufficient

Evangelia Antonakos
Eva@Antonakos.net

April 20, 2006

Abstract

Three formal approaches to public knowledge are "any fool" knowledge by McCarthy (1970), Common Knowledge by Halpern and Moses (1990), and Justified Knowledge by Artemov (2004). We compare them to mathematically address the observation that the light-weight systems of Justified Knowledge and 'any fool knows' suffice to solve standard epistemic puzzles for which heavier solutions based on Common Knowledge are offered by standard textbooks. Specifically we show that epistemic systems with Common Knowledge modality C are conservative with respect to Justified Knowledge systems on formulas $\chi \wedge C\varphi \to \psi$, where χ, φ, and ψ are C-free. We then notice that formalization of standard epistemic puzzles can be made in the aforementioned form, hence each time there is a solution within a Common Knowledge system, there is a solution in the corresponding Justified Knowledge system.

1 Multi-agent Logics

The logics T_n, $\mathsf{S4}_n$, and $\mathsf{S5}_n$ are logics in which each of the finitely many (n) agents has a knowledge operator K_i which is T, or $\mathsf{S4}$, or $\mathsf{S5}$ respectively. We only consider cases where all agents' modalities are of the same logical strength.

Definition 1. *The formal systems for* T_n, $\mathsf{S4}_n$, *and* $\mathsf{S5}_n$ *are as follows: propositional logic plus for* K_i, $i = 1, 2, \ldots, n$ *we have*

Axioms for $\mathsf{S4}_n$:

K : $K_i(\varphi \to \psi) \to (K_i\varphi \to K_i\psi)$ — each agent can do *modus ponens*

T : $K_i\varphi \to \varphi$ — agents can know only true propositions

$\mathsf{4}$: $K_i\varphi \to K_iK_i\varphi$ — agents have positive introspection

Rules:

Necessitation: $\vdash \varphi \Rightarrow \vdash K_i\varphi$

For T_n, omit the final axiom.

For $\mathsf{S5}_n$, add negative introspection: $\neg K_i\varphi \to K_i\neg K_i\varphi$.

Definition 2. *Kripke models for* $\mathsf{S4}_n$*:* $M = \langle W,\ R_1,\ R_2, \ldots,\ R_n,\ \Vdash \rangle$ *where*

- W is a non-empty set of worlds
- $R_i \subseteq W \times W$ is agent i's accessability relation. R_i is reflexive and transitive.
- $\Vdash \subseteq W \times Var$ where Var is the set of propositional variables. The forcing relation $\Vdash$ is naturally extended to all formulas so that R_i corresponds to K_i:

$$M, u \Vdash K_i\varphi \Leftrightarrow \forall v \in M\ uR_iv \to M, v \Vdash \varphi\ .$$

For T_n-models, each R_i is reflexive while for $\mathsf{S5}_n$-models, each R_i is an equivalence relation.

Theorem 1. T_n, $\mathsf{S4}_n$, *and* $\mathsf{S5}_n$ *are sound and complete with respect to their models, as shown in* [FHMV95].

Multi-agent systems are enhanced by the addition of modalities which take into account the public knowledge of agents. Three such modalities C, J, and O will be discussed, all of which model variations of public information. We will compare their logical strengths, semantics, and complexity and will see why justified knowledge (J) systems are sufficient to solve classical epistemic puzzles, a role usually designated for common knowledge (C).

2 Common Knowledge

The most recognized conception of public knowledge is common knowledge, and the literature addressing it, both philosophical (notably [L69]) and mathematical, is vast. Informally, the epistemic operator $C\varphi$, to be read 'φ is common knowledge,' can be given as infinite conjunction:

$$C\varphi \leftrightarrow \varphi \wedge E\varphi \wedge E^2\varphi \wedge E^3\varphi \wedge \cdots \wedge E^n\varphi \ldots$$

where $E\varphi = K_1\varphi \wedge K_2\varphi \wedge \ldots \wedge K_n\varphi$ (*everyone knows* φ) and K_i is an individual agent's knowledge operator corresponding to T, $\mathsf{S4}$ or $\mathsf{S5}$ as appropriate. One formal characterization which [FHMV95] and [vBS04] take is via a Fixed Point Axiom

$$C\varphi \leftrightarrow E\,(\varphi \wedge C\varphi)$$

and the Induction Rule

$$\frac{\varphi \rightarrow E\,(\varphi \wedge \psi)}{\varphi \rightarrow C\psi}$$

yielding common knowledge to be the greatest fixed point solution [FHMV95]. Common knowledge does not take into account the means by which the knowledge is acquired. As we will see, this is in contrast to justified knowledge. The distinction between the infinite conjunction, the fixed point axiom, and how common knowledge is achieved is addressed in [Bar88]. [G92] too, provides a useful survey with several nice examples but does not include a distinct formalism. There is also an equivalent axiomatic formulation of common knowledge which replaces the induction rule with the induction axiom in [MvdH95], which, for technical convenience, we will use.

Definition 3. T_n^C, $\mathsf{S4}_n^C$, *and* $\mathsf{S5}_n^C$ *axiom systems:*
Propositional Logic *plus*
Axioms:
T, $\mathsf{S4}$, *or* $\mathsf{S5}$ *axioms for* K_i, $i = 1, 2, \ldots, n$, *respectively;*
K: $C(\varphi \rightarrow \psi) \rightarrow (C\varphi \rightarrow C\psi)$;
T: $C\varphi \rightarrow \varphi$;
$C\varphi \rightarrow E(C\varphi)$, *where* $E\varphi = K_1\varphi \wedge K_2\varphi \wedge \cdots \wedge K_n\varphi$;
Induction Axiom: $\varphi \wedge C(\varphi \rightarrow E\varphi) \rightarrow C\varphi$.
Rules:
Necessitation for all K_i: $\vdash \varphi \Rightarrow \vdash K_i\varphi$;
Necessitation for C: $\vdash \varphi \Rightarrow \vdash K_i\varphi$.

Definition 4. *Models for* T_n^C, $\mathsf{S4}_n^C$, *and* $\mathsf{S5}_n^C$: $M = \langle W, R_1, R_2, \ldots, R_n, R_C, \Vdash\rangle$ *where*
$\bullet M = \langle W, R_1, R_2, \ldots, R_n \Vdash\rangle$ *is a* T_n, $\mathsf{S4}_n$, *or* $\mathsf{S5}_n$ *model, respectively*
$\bullet$ R_C *is a reflexive and transitive relation such that*

$$\textit{Transitive Closure}\,\big(\bigcup_{i=1}^{n} R_i\big) = R_C.$$

• *The forcing relation* $\Vdash$ *is naturally extended to all formulas so that* R_C *corresponds to* C*:* $M, u \Vdash C\varphi \Leftrightarrow \forall v \in M\ uR_C v \rightarrow M, v \Vdash \varphi$.

Theorem 2. T_n^C, $\mathsf{S4}_n^C$, *and* $\mathsf{S5}_n^C$ *are sound and complete with respect to their models.*

Proof. One can be found in [FHMV95], beginning on p. 70. □

The agents' logic plays a role in determining the strength of the common knowledge operator C. In the systems defined above, C is always at least as strong as K_i. Showing that in T_n^C, $\mathsf{S4}_n^C$, and $\mathsf{S5}_n^C$, C satisfies the T, $\mathsf{S4}$, and $\mathsf{S5}$ axioms, respectively, is given as an exercise in [FHMV95], p. 93.

3 Justified Knowledge

Justified knowledge was introduced by Artemov in [A04] as the forgetful projection of the *evidence-based* knowledge represented by an appropriate adaptation of LP (Logic of Proofs). In LP systems ($\mathsf{T}_n\mathsf{LP}$, $\mathsf{S4}_n\mathsf{LP}$, $\mathsf{S5}_n\mathsf{LP}$), each formula / subformula carries with it a proof term representing a particular proof of the formula / subformula from the axioms. Justified knowledge systems are ones in which all proofs are identified as one. Whereas $C\varphi$ asserts that φ is common knowledge, $J\varphi$ asserts that φ is common knowledge arising from a proof of φ or some other agreed-upon acceptable set of evidences. Though the proof of φ is not explicitly presented with the assertion $J\varphi$, it is reproducible. This is the important Realization Theorem which provides an algorithm to reconstruct proof terms. For more details on this, the reader should consult [A04, A06].

As with the common knowledge logics, the construction of the justified knowledge logics T_n^J, $\mathsf{S4}_n^J$, and $\mathsf{S5}_n^J$ builds on the multi-agent logics. In C systems the agents' logic determines the strength of C while in J systems the strength of J is chosen independently to be weaker, stronger, or the same as that of the agents'. In the aforementioned logics, the modality J will be assumed to be $\mathsf{S4}$ unless otherwise specified.

Definition 5. T_n^J, $\mathsf{S4}_n^J$, *and* $\mathsf{S5}_n^J$ *axiom systems:*
Propositional Logic *plus*
Axioms:
T, $\mathsf{S4}$, *or* $\mathsf{S5}$ *axioms for* K_i, $i = 1, 2, \ldots, n$;

$\mathsf{S4}$ *axioms for* J *;*
Connection Principle: $J\varphi \rightarrow \ K_i\varphi$ *.*
Rules:
Necessitation for all K_i*:* $\vdash \varphi \Rightarrow \vdash K_i\varphi$ *;*
Necessitation for J*:* $\vdash \varphi \Rightarrow \vdash J\varphi$ *.*

Definition 6. *Models for* T_n^J*,* $\mathsf{S4}_n^J$*, and* $\mathsf{S5}_n^J$*:* $M = \langle W, R_1, R_2, \ldots, R_n, R_J, \Vdash\rangle$ *where*
• $M = \langle W, R_1, R_2, \ldots, R_n \Vdash\rangle$ *is a* T_n*,* $\mathsf{S4}_n$*, or* $\mathsf{S5}_n$ *model, respectively*
• R_J *is a reflexive and transitive relation such that*

$$\textit{Transitive Closure}\ \big(\bigcup_{i=1}^{n} R_i\big) \subseteq R_J.$$

• *The forcing relation* $\Vdash$ *is naturally extended to all formulas so that* R_J *corresponds to* J*:* $M, u \Vdash J\varphi \Leftrightarrow \forall v \in M\ uR_Jv \rightarrow M, v \Vdash \varphi$*.*

Theorem 3. T_n^J*,* $\mathsf{S4}_n^J$*, and* $\mathsf{S5}_n^J$ *are sound and complete with respect to their models, as shown in* [A04, A06]*.*

Recall that in common knowledge models, $R_C = \text{Transitive Closure}(\bigcup_{i=1}^{n} R_i)$ and so $R_C \subseteq R_J$. Thus in a context where we can compare the two, *i.e.* a hybrid model with both R_C and R_J, it seems $J\varphi \Rightarrow C\varphi$ but not vice versa. More formally, we have the following propositions.

Definition 7. *Let* φ^* *be* φ *with each instance of a* J *replaced by a* C*.*

Proposition 1. $(\mathsf{S4}_n^J)^* \subset \mathsf{S4}_n^C$ *but* $(\mathsf{S4}_n^J)^* \neq \mathsf{S4}_n^C$ *.*

Proof. It needs to be shown that the $*$-translation of each each rule and axiom of $\mathsf{S4}_n^J$ is provable in $\mathsf{S4}_n^C$. Artemov shows this in [A04, A06] using the equivalent axiomatization of $\mathsf{S4}_n^C$ from [FHMV95]. □

The proof shows it is only the Induction Axiom of $\mathsf{S4}_n^C$ which is not provable in $(\mathsf{S4}_n^J)^*$, yielding the strict inclusion.

Corollary 1. *Let* $\mathsf{I.A.}$ *be the induction axiom. Then*
$\mathsf{S4}_n^C \equiv (\mathsf{S4}_n^J)^* + \mathsf{I.A.}$ *and* $\mathsf{T}_n^C \equiv (\mathsf{T}_n^J)^* + \mathsf{I.A.}$ *where* J *is an* $\mathsf{S4}$ *modality.*
$\mathsf{S5}_n^C \equiv (\mathsf{S5}_n^J)^* + \mathsf{I.A.}$ *when* J *is an* $\mathsf{S5}$ *modality.*

$\mathsf{S5}_n^J$ where J is a $\mathsf{S5}$ modality was considered in [R06].

Proof. The strict inclusion of the J systems follows from the previous proposition and noticing that C satisfies the $\mathsf{4}$ axiom in T_n^C and the $\mathsf{5}$ axiom in $\mathsf{S5}_n^C$. When the induction axiom is added, the equivalence is clear. □

Thus indeed, in any J-model, $J\varphi \Rightarrow C\varphi$. The evidence-based common knowledge semantics for J systems are further enriched by Artemov's Realization Theorem mentioned at the start of the section. This gives a constructive approach to recovering or realizing the full proof terms of the evidence-based knowledge systems.

Theorem 4 (Realization Theorem). *There is an algorithm that given an* $\mathsf{S4}_n^J$*-derivation of a formula* φ*, retrieves an* $\mathsf{S4}_n\mathsf{LP}$*-formula* ψ*, a* realization of φ*, such that* φ *coincides with* ψ°*, where* $\circ$ *replaces all proof terms with* J*, and* $\mathsf{S4}_n\mathsf{LP}$ *proves* ψ*.*

This theorem and a realization theorem for $\mathsf{S5}_n^J$ where J is an $\mathsf{S4}$-modality is established in [A04, A06] while a realization theorem for $\mathsf{S5}_n^J$ where J is an $\mathsf{S5}$-modality is given in [R06]. Other major advantages to justified knowledge are

- proofs in $\mathsf{S4}_n^J$ are normalizable ([A04, A06]), but not in $\mathsf{S4}_n^C$
- $\mathsf{S4}_2^J$ is *PSPACE*-complete [D05], whereas for $n \geq 2$, $\mathsf{S4}_n^C$ is *EXPTIME*-complete [FHMV95].

These features have been exploited by Bryukhov in [B05] to develop an automated theorem prover for $\mathsf{S4}_n^J$. Justified Knowledge offers simpler, more constructive, and more automation-friendly approach to common knowledge.

4 Any Fool's Knowledge

McCarthy's model of common knowledge via "any fool knows" apparently goes back to roughly 1970 ([FHMV95], p. 13), though its first published appearance is in [MSTI78]. In this epistemic multi-agent system, the modality for each agent is denoted by S, and the additional virtual agent, "any fool," is denoted by O. In [MSTI78] p. 2, whatever any fool knows, "everyone knows that everyone else knows," and so someone knows. Thus we may add an additional axiom linking the fool to the other people: $O\varphi \rightarrow S\varphi$. Call this the linking axiom. This corresponds exactly to Artemov's connection principle: $J\varphi \rightarrow K_i\varphi$. When McCarthy et al. use subscripted modals, S_i, to specify individual agents, S_0 is the distinguished any fool operator O. Thus

we see that the "fool" is a particular agent, hence in any axiom, we may replace all S modals by Os, though not vice versa.

Definition 8. *The McCarthy et al axioms are based on propositional logic plus:*

linking axiom: $O\varphi \to S\varphi$
K0: $S\varphi \to \varphi$
K1: $O(S\varphi \to \varphi)$
K2: $O(O\varphi \to OS\varphi)$
K3: $O(S\varphi \wedge S(\varphi \to \psi) \to S\psi)$
K4: $O(S\varphi \to SS\varphi)$
K5: $O(\neg S\varphi \to S\neg S\varphi)$.

We will look at three systems identified in [MSTI78] given by axioms K0-K3, K0-K4, and K0-K5.[1] These will be referred to as MT, $\mathsf{M4}$, and $\mathsf{M5}$ respectively. Model semantics and completeness results for a variant of $\mathsf{M5}$ is stated in [MSTI78]. These logics were introduced explicitly to formalize and solve epsitemic puzzles, of which Wise Men and Unfaithful Wives are addressed in [MSTI78]. However, rather than borrowing from standard formulations of modal logics, the authors seem to have developed these particular axioms by considering just what would be needed for these particular examples. Despite this, we can see that Artemov's justified knowledge operator J plays a role equivalent to McCarthy's any fool operator O. In particular, we have the following theorem.

Definition 9. $\star$ *replaces* J *by* O *and* K_i *by* S_i.

Theorem 5. $(\mathsf{T}_n^J)^\star \equiv \mathsf{MT}$ *and* $(\mathsf{S4}_n^J)^\star \equiv \mathsf{M4}$ *when* J *is an* $\mathsf{S4}$ *modality and* $(\mathsf{S5}_n^J)^\star \equiv \mathsf{M5}$ *when* J *is an* $\mathsf{S5}$ *modality.*

Proof. Immediate from the following three lemmas. □

Lemma 1. $(\mathsf{T}_n^J)^\star \equiv \mathsf{MT}$.

Proof. Recall that J is an $\mathsf{S4}$ modality while the K_i are T modalities.

($\Leftarrow$) To show $(\mathsf{T}_n^J)^\star \supset \mathsf{MT}$, T_n^J must satisfy MT axioms (K0-K3 and the linking axiom), where Os are Js and S_is are K_is.

[1] In [MSTI78], $\mathsf{K0}$ is omitted from these lists. Given other statements in the paper, this clearly is just an oversight.

Linking axiom: $\mathsf{T}_n^J \vdash J\varphi \to K_i\varphi$; the connection principle
K0: $\mathsf{T}_n^J \vdash K_i\varphi \to \varphi$; T axiom for K_i
K1: $\mathsf{T}_n^J \vdash J(K_i\varphi \to \varphi)$; J necessitation of T axiom of K_i
K2: $\mathsf{T}_n^J \vdash J(J\varphi \to JK_i\varphi)$;
1. $\mathsf{T}_n^J \vdash J\varphi \to JJ\varphi$ $\mathsf{4}$ axiom for J
2. $\mathsf{T}_n^J \vdash J\varphi \to K_i\varphi$ the connection principle
3. $\mathsf{T}_n^J \vdash J(J\varphi \to K_i\varphi)$ from 2. by J necessitation
4. $\mathsf{T}_n^J \vdash JJ\varphi \to JK_i\varphi$ from 3. by K axiom for J
5. $\mathsf{T}_n^J \vdash J\varphi \to JK_i\varphi$ from 1. and 4.
6. $\mathsf{T}_n^J \vdash J(J\varphi \to JK_i\varphi)$ from 5. by J necesitation

K3: $\mathsf{T}_n^J \vdash J(K_i\varphi \wedge K_i(\varphi \to \psi) \to K_i\psi)$; J necessitation of K axiom for K_i.

As mentioned above, "any fool" is a particular agent and so in any axiom all the Ss may be replaced by Os. Consider K0′-K3′ and the linking axiom′ where we do just that:

Linking axiom′: $\mathsf{T}_n^J \vdash J\varphi \to J\varphi$; propositional tautology
K0′: $\mathsf{T}_n^J \vdash J\varphi \to \varphi$; T axiom for J
K1′: $\mathsf{T}_n^J \vdash J(J\varphi \to \varphi)$; J necessitation of T axiom of J
K2′: $\mathsf{T}_n^J \vdash J(J\varphi \to JJ\varphi)$; J necessitaton of $\mathsf{4}$ axiom for J
K3′: $\mathsf{T}_n^J \vdash J(J\varphi \wedge J(\varphi \to \psi) \to J\psi)$; J necessitation of K axiom for J.

($\Rightarrow$) $(\mathsf{T}_n^J)^\star \subset \mathsf{MT}$. We must show that MT satisfies the $(\mathsf{T}_n^J)^\star$ axioms and rules. Remember that "any fool" O is a particular S agent.

S axioms:
K: $\mathsf{MT} \vdash S\varphi \wedge S(\varphi \to \psi) \to S\psi$; by K3, linking axiom, K0
T: $\mathsf{MT} \vdash S\varphi \to \varphi$; K0
O axioms:
T: $\mathsf{MT} \vdash O\varphi \to \varphi$; by K0, O is a particular S
K: $\mathsf{MT} \vdash O\varphi \wedge O(\varphi \to \psi) \to O\psi$; by K axiom for S, O is a an S
$\mathsf{4}$: $\mathsf{MT} \vdash O\varphi \to OO\varphi$; by K2, T axiom for O, O is an S
Connection axiom: $\mathsf{MT} \vdash O\varphi \to S\varphi$; the linking axiom

O necessitation: This follows from the fact that each S and O axiom is nesessitated.

K axiom for S and O is necessitated by K3.

T axiom for S and O is necessitated by K1.

$\mathsf{4}$ axiom for O is necessitated by K2.

S necessitation: This follows from O necessitation and the linking axiom. □

Lemma 2. $(\mathsf{S4}_n^J)^\star \equiv \mathsf{M4}$.

Proof. Recall that J and K_i are $\mathsf{S4}$ modalities.
($\Leftarrow$) $(\mathsf{S4}_n^J)^\star \supset \mathsf{M4}$ follows from previous lemma and
K4: $\mathsf{S4}_n^J \vdash J(K_i\varphi \to K_iK_i\varphi)$; by J necessitation of $\mathsf{4}$ axiom for K_i.
K4′ = K2′

($\Rightarrow$) $(\mathsf{S4}_n^J)^\star \subset \mathsf{M4}$ follows from previous lemma and
S axioms:
$\mathsf{4}$: $\mathsf{M4} \vdash S\varphi \to SS\varphi$; by K4, T axiom for O.
O necessitation: $\mathsf{4}$ axiom for S is necessitated by K4. □

Lemma 3. $(\mathsf{S5}_n^J)^\star \equiv \mathsf{M5}$.

Proof. Recall that J and K_i are $\mathsf{S5}$ modalities.
($\Leftarrow$) $(\mathsf{S5}_n^J)^\star \supset \mathsf{M5}$ follows from previous lemma and
K5: $\mathsf{S4}_n^J \vdash J(\neg K_i\varphi \to K_i\neg K_i\varphi)$; by J necessitation of $\mathsf{5}$ axiom for K_i.
K5′: $\mathsf{S4}_n^J \vdash J(\neg J\varphi \to J\neg J\varphi)$; by J necessitation of $\mathsf{5}$ axiom for J.

($\Rightarrow$) $(\mathsf{S5}_n^J)^\star \subset \mathsf{M5}$ follows from previous lemma and
S axioms:
$\mathsf{5}$: $\mathsf{M4} \vdash \neg S\varphi \to S\neg S\varphi$; by K5, T axiom for O
O axioms:
$\mathsf{5}$: $\mathsf{M5} \vdash \neg O\varphi \to O\neg O\varphi$; by K5, T axiom for O, O is an S.
O necessitation: $\mathsf{5}$ axiom for O and S necessitated by K5. □

Despite quite different motivations and technical backgrounds, McCarthy's "any fool" and Artemov's justified knowledge approaches lead to the same multi-modal logics.

Corollary 2. *There is a Realization Theorem for* MT, $\mathsf{M4}$, *and* $\mathsf{M5}$ *providing evidence-based semantics for McCarthy's "any fool" knowledge operator* O.

Also, with the addition of an induction rule, these 'fool' logics provide another alternative normal modal axiomatization for [FHMV95] common knowledge systems, where the common knowledge operator is represented by an additional axiom schema on the top of the multi-modal version of the corresponding basic modal logic.

5 Epistemic Puzzles

Standard epistemic puzzles in the literature such as Wise Men, Muddy Children, and Unfaithful Wives, have traditionally been solved using strong common knowledge systems like $\mathsf{S5}_n^C$. By noting examples provided in [MSTI78] and [A04, A06], we see that justified knowledge can be used to formulate the same puzzles. In addition, $\mathsf{S5}$ is an unnecessarily strong assumption as multi-agent T systems are often sufficient. Though [F06] offers common knowledge solutions via tidy tableau proofs, common knowledge solutions such as ones offered in [FHMV95], seem to make unwarranted or hidden assumptions, most notably that each agent in these puzzles has the same Kripke model in mind. The formalizations, and hence deductive solutions, we can get with justified knowledge are more satisfying on this front.

These puzzles are structured so that we are given some facts about the world and some common knowledge assumptions, from which we are to conclude some fact about the world or an agent's knowledge. In particular, a solution to these puzzles is not a conclusion about whether a proposition is common knowledge. Schematically these puzzles are of the form $\chi \wedge C\varphi \to \psi$, where χ, φ, and ψ are C-free. In the following section we will see that common knowledge systems are conservative with respect to justified knowledge systems on formulas of this form, so that any solution in a common knowledge system can be obtained in the corresponding justified knowledge system.

We motivated this by two examples: muddy children and wise men. For classical common knowledge solutions the reader may consult [FHMV95] and [MvdH95].

Muddy Children

> Assume there are $n > 1$ children playing outside. While they have been playing, some of them have gotten muddy foreheads. Each child can see whether the others have mud on their foreheads, but no one has mentioned it. At some point the father comes out and announces for all to hear, "at least one of you has mud on your forehead." He asks, "Do you know whether your own forehead is dirty?" The kids answer simultaneously. The question and answer rounds continue. We will assume that all hear and understand the father and that all, including the father, speak truthfully.

It is well known that if $k \geq 1$ of them are muddy, all the children will

answer that they don't know whether they are muddy the first $k-1$ times they are asked. The k^{th} time they are asked, each muddy child will reply that she knows that she is muddy. The $k{+}1^{st}$ time they are asked, each clean child will reply that she knows that she is not muddy.

Consider an instance of this puzzle with three children called 1, 2, and 3, where 1 and 2 are muddy and 3 is clean. We can give the specifications of this puzzle using facts about the world and explicit justified knowledge assumptions. There will then be a solution consisting of a formal derivation in T_3^J using to conclude J-free facts. This approach follows one presented in [A04].

Let the atomic propositions p_i stand for 'child i has a muddy forehead.' Binary triples are shorthand notation for conjunctions of p_is, *e.g.* $\mathbf{101} = p_1 \wedge \neg p_2 \wedge p_3$. $Kw_i\varphi$ is shorthand for $K_i\varphi \vee K_i\neg\varphi$, i.e., '$i$ knows whether φ.' *'Knowing about others'* is formalized as '$K.A.O.$' $= \bigwedge_{j\neq i} Kw_jp_i$. As the children can see the others' foreheads, we add $K.A.O.$ as an axiom.

The specifications for this problem can be made by describing the initial state of the world as

$$\mathsf{L}(0) = \mathbf{110} + J(K.A.O.) + J(\neg\mathbf{000}).$$

After the children answer the first question (with a chorus of "No" in this case), the state of the world can be represented by

$$\mathsf{L}(1) = \mathsf{L}(0) + J(\neg Kw_1p_1) + J(\neg Kw_2p_2) + J(\neg Kw_3p_3).$$

Loosely, after m^{th} round of question and answers,

$$\mathsf{L}(\mathsf{m}+1) = \mathsf{L}(\mathsf{m}) + J(\text{children's answers to the } m^{th} \text{ question}).$$

Now, straight forward derivations, left to the reader, show that $\mathsf{T}_3^J \vdash \mathsf{L}(1) \to K_1p_1$, $\mathsf{T}_3^J \vdash \mathsf{L}(1) \to K_2p_2$, and $\mathsf{T}_3^J \vdash \mathsf{L}(2) \to K_3\neg p_3$ so that children 1 and 2 will answer "yes, I know" to the second question, and 3 will answer, "yes, I know" to the third question. Note that these derivations follow the format $\chi \wedge J\varphi \to \psi$ mentioned above.

In the specifications above, the state of the world must be calculated by hand after each question is answered. A related formalization for muddy

children due to Fitting can be used for automatic updating. This is a slight reformulation for consistency of notation, the original is in [F06], p. 41. The reader is referred to [F06] for details.

$$\mathsf{T}_3^J \vdash I \wedge J(K.A.O.) \rightarrow (J(\neg(J(I) \rightarrow S)) \rightarrow K_i p_i) \text{ for some } i$$

Here I is the initial (current) state of the world and $S = Kw_1p_1 \vee Kw_2p_2 \vee Kw_3p_3$, *i.e.* someone knows whether they're muddy.

By this simple derivation we see that this too, satisfies the paradigm of $\chi \wedge J\varphi \rightarrow \psi$:

1. $I \wedge J(K.A.O.) \rightarrow (J(\neg(J(I) \rightarrow S)) \rightarrow K_i p_i)$	Fitting's Axiom
2. $I \wedge J(K.A.O.) \wedge J(\neg(J(I) \rightarrow S)) \rightarrow K_i p_i$	by importation
3. $I \wedge J(K.A.O.) \wedge J(J(I) \wedge \neg S) \rightarrow K_i p_i$	propositional reasoning
4. $I \wedge J(K.A.O.) \wedge JJ(I) \wedge J(\neg S) \rightarrow K_i p_i$	J distributivity
5. $I \wedge J(K.A.O.) \wedge J(I) \wedge J(\neg S) \rightarrow K_i p_i$	T axiom for J
6. $I \wedge J(K.A.O. \wedge I \wedge \neg S) \rightarrow K_i p_i$	J distributivity.

Wise Men

> A king has three wise men. It is common knowledge that there are three red hats and two white hats. The king places a hat on the head of each wise man so that none sees the color of his own hat. The king then hides the remaining two hats. He asks the wise men, sequentially, if they know the color of their own hat. The first wise man says that he does not know; then the second wise man says that he does not know. Does the third wise man know what color his hat is? If so, what color is it?

The third wise man's hat is red, and he will know this after hearing the answers of the first two wise men.

As with muddy children, we can formulate the situation and derive a solution in T_3^J. The set-up and derivations in the following lemma are lifted directly from [A06], section 8. We will use $K.A.O.$ as in muddy children. Here, p_i will stand for 'i's hat is white'. The initial situation can be described by

$$\mathsf{L}(0) = J(K.A.O.) + J(\neg \mathbf{000}).$$

Note that here we don't include the actual state of world as we did with muddy children. The situation after the first and second wise men said they

didn't know is represented by

$$\mathsf{L}(2) = \mathsf{L}(0) + J(\neg Kw_1p_1) + J(\neg Kw_2p_2) \ .$$

Lemma 4. $\mathsf{T}_3^J \vdash \mathsf{L}(2) \to K_3p_3$

Proof. First, we prove $\mathsf{L}(2) \to K_3(\neg\mathbf{100})$.

1. $\mathbf{100} \to K_1(\neg p_2 \wedge \neg p_3)$, from $J(K.A.O.)$
2. $K_1(\neg p_2 \wedge \neg p_3) \to K_1p_1$, from $J(\neg\mathbf{000})$
3. $\mathbf{100} \to K_1p_1$, from 1. and 2.
4. $\neg Kw_1p_1 \to \neg\mathbf{100}$, from 3.
5. $J(\neg Kw_1p_1) \to J(\neg\mathbf{100})$, from 4., by T_3^J-reasoning
6. $J(\neg\mathbf{100})$, from $J(\neg Kw_1p_1)$ and 5.
7. $K_3(\neg\mathbf{100})$, by T_3^J-reasoning.

Likewise, using $J(\neg Kw_2p_2)$ we obtain $K_3(\neg\mathbf{010})$.

Next, we prove that $\mathsf{L}(2) \to K_3(\neg\mathbf{110})$.

1. $\mathbf{110} \to K_2(\mathbf{110} \vee \mathbf{100})$, from $J(K.A.O.)$
2. $\mathbf{110} \to K_2(\mathbf{110})$, since $J(\neg\mathbf{100})$
3. $\mathbf{110} \to Kw_2p_2$, by T_3^J-reasoning
4. $\neg Kw_2p_2 \to \neg\mathbf{110}$, by propositional logic
5. $J(\neg Kw_2p_2) \to J(\neg\mathbf{110})$, by T_3^J-reasoning
6. $J(\neg\mathbf{110})$, from $J(\neg Kw_2p_2)$ and 5.
7. $K_3(\neg\mathbf{110})$, by T_3^J-reasoning.

Finally, we conclude K_3p_3, since $K_3\neg(\mathbf{000} \vee \mathbf{100} \vee \mathbf{010} \vee \mathbf{110})$, that is, all combinations with $\neg p_3$ have been ruled out. □

From the above proof is it clear not only that the third wise man knows he wears a red hat, but that he would know this even if he were blind or otherwise can not see the others' hats. Note that the above reasoning does not use Kw_3p_1 and Kw_3p_2.

We have seen that both these puzzles man be solved in relatively weak multi-agent T systems with justified knowledge, moreover, the specifications for these problems can be given in the form $\chi \wedge J\varphi \to \psi$ where χ, φ, and ψ contain no J operators. In the following section we have a conservativity result which will show that for any solution in a C system, there will be one in the corresponding J system.

6 Limited Conservativity

A logic T with language $\mathcal{L}$ is a conservative extension of a logic T' with language $\mathcal{L}' \subseteq \mathcal{L}$ if for sentences φ of $\mathcal{L}'$, T proves φ only if T' proves φ.

Recall the definition for $*$ from Section 3 which renames J to C. As the logics $(\mathsf{S4}_n^J)^*$ and $\mathsf{S4}_n^C$ have the same language and yet are not equal, it is clear that $\mathsf{S4}_n^C$ can not be a conservative extension of $(\mathsf{S4}_n^J)^*$, it is however a conservative extension over all formulas in which C occurs only negatively.

Theorem 6. *If φ is a formula of $\mathsf{S4}_n^J$ such that all occurrences of J in φ are negative, then $\mathsf{S4}_n^J \vdash \varphi \quad \Leftrightarrow \quad \mathsf{S4}_n^C \vdash (\varphi)^*$.*

In some sense this result is tight as the induction axiom $(\varphi \wedge J(\varphi \rightarrow E\varphi) \rightarrow J\varphi)$ which distinguishes $(\mathsf{S4}_n^J)^*$ from $\mathsf{S4}_n^C$ has, along with a negative occurrence of J, a single positive occurrence of J.

Proof. $(\Rightarrow)$ is secured by the inclusion $(\mathsf{S4}_n^J)^* \subset \mathsf{S4}_n^C$.

$(\Leftarrow)$ This direction is a consequence of the Main Lemma which follows. We show this direction by proving the contrapositive. Suppose φ is a formula of $\mathsf{S4}_n^J$ such that all occurrences of J in φ are negative and $\mathsf{S4}_n^J \nvdash \varphi$. By completeness, there is a model M and a world x such that $M, x \Vdash \neg\varphi$. By the Main Lemma, $M^C, x \Vdash^C \neg(\varphi)^*$, hence $\mathsf{S4}_n^C \nvdash (\varphi)^*$, since M^C (with R_J ignored) is a model for $\mathsf{S4}_n^C$. $\square$

Lemma 5 (Main Lemma). *Let M be a $\mathsf{S4}_n^J$-model. Add the relation R_C of reachability along $R_1, \ldots R_n$ to M and get the augmented model M^C, where $\Vdash^C$ coincides with $\Vdash$ on variables, and the modality C corresponds to R_C. Let φ be a formula of $\mathsf{S4}_n^J$. Then*

if all occurrences of J in φ are positive, then $x \Vdash \varphi \quad \Rightarrow \quad x \Vdash^C (\varphi)^*$;

if all occurrences of J in φ are negative, then $x \Vdash \neg\varphi \quad \Rightarrow \quad x \Vdash^C \neg(\varphi)^*$.

Proof. By induction on φ.

Base case is secured by the definition of $\Vdash^C$.

Boolean case: $\varphi \equiv \psi \rightarrow \theta$.

Subcase: all occurrences of J in φ are positive and $x \Vdash \varphi$. Then $x \Vdash \neg\psi$ or $x \Vdash \theta$. In the former case all occurrences of J in ψ are negative and, by the induction hypothesis, $x \Vdash^C \neg(\psi)^*$. In the latter case all occurrences of J

in θ are positive and, by the induction hypothesis, $x \Vdash^C (\theta)^*$. In either case, $x \Vdash^C (\varphi)^*$.

Subcase: all occurrences of J in φ are negative and $x \Vdash \neg\varphi$. Then $x \Vdash \psi$ and $x \Vdash \neg\theta$. Since all occurrences of J in ψ are positive and all occurrences of J in θ are negative, by the induction hypothesis, $x \Vdash^C (\psi)^*$ and $x \Vdash^C \neg(\theta)^*$, hence $x \Vdash^C \neg(\varphi)^*$.

Case: $\varphi \equiv K_i\psi$.

Subcase: all occurrences of J in φ are positive and $x \Vdash \varphi$. Then all occurrences of J in ψ are positive and $y \Vdash \psi$, for all y such that xR_iy. By the induction hypothesis, $y \Vdash^C (\psi)^*$, for all y such that xR_iy, hence $x \Vdash^C (K_i\psi)^*$, *i.e.*, $x \Vdash^C (\varphi)^*$.

Subcase: all occurrences of J in φ are negative and $x \Vdash \neg\varphi$. Then for some y such that xR_iy, $y \Vdash \neg\psi$. Since all occurrences of J in ψ are also negative, by the induction hypothesis, $y \Vdash^C \neg(\psi)^*$, hence $x \Vdash^C \neg(K_i\psi)^*$, *i.e.*, $x \Vdash^C \neg(\varphi)^*$.

Case: $\varphi \equiv J\psi$.

Subcase: all occurrences of J in φ are positive and $x \Vdash \varphi$. Then all occurrences of J in ψ are also positive and $y \Vdash \psi$, for all y such that xR_Jy. Since $R_C \subseteq R_J$, $y \Vdash \psi$, for all y such that xR_Cy. By the induction hypothesis, $y \Vdash^C (\psi)^*$, for all y such that xR_Cy. Hence $x \Vdash^C C(\psi)^*$, *i.e.*, $x \Vdash^C (J\psi)^*$, *i.e.*, $x \Vdash^C (\varphi)^*$.

Subcase: 'all occurrences of J in φ are negative and $x \Vdash \neg\varphi$' is impossible, since $\varphi \equiv J\psi$ and the displayed occurrence of J is positive in $J\psi$. □

Corollary 3. *If χ, φ and ψ are formulas in the language of $\mathsf{S4}_n$, then $\mathsf{S4}_n^C \vdash \chi \wedge C\varphi \to \psi \Leftrightarrow \mathsf{S4}_n^J \vdash \chi \wedge J\varphi \to \psi$.*

Proof. As per the theorem, $\chi \wedge J\varphi \to \psi$ has J only in negative position. □

Corollary 4. *If χ, φ and ψ are formulas in the language of T_n and J is an $\mathsf{S4}$ modality, then $\mathsf{T}_n^C \vdash \chi \wedge C\varphi \to \psi \Leftrightarrow \mathsf{T}_n^J \vdash \chi \wedge J\varphi \to \psi$.*

Proof. Direct consequence of the theorem if main lemma starts with T_n^J-models and completeness for T_n^J. □

Corollary 5. *If χ, φ and ψ are formulas in the language of $\mathsf{S5}_n$ and J is an $\mathsf{S5}$ modality, then $\mathsf{S5}_n^C \vdash \chi \wedge C\varphi \to \psi \Leftrightarrow \mathsf{S5}_n^J \vdash \chi \wedge J\varphi \to \psi$.*

Proof. Direct consequence of the theorem if main lemma starts with $\mathsf{S5}_n^J$-models and completeness for $\mathsf{S5}_n^J$. □

7 Conclusions

Informally, we can consider to C-, J-free formulas as representing facts about states of knowledge of real epistemic agents $K_1, \ldots, K_n$. A standard setup of a epistemic puzzle consists of certain assumptions about the state of the world and agents' knowledge (χ) and common knowledge ($C\varphi$) whereas the conclusions are normally made about states of knowledge of real agents (ψ).

It seems that formalization of standard epistemic puzzles can be made in the form of $\chi \wedge C\varphi \rightarrow \psi$, hence each time there is a solution within a Common Knowledge system, there is a solution in the corresponding Justified Knowledge system. In employing justified knowledge, we make our assumptions explicit (by giving $\mathsf{L}(\mathsf{0})$) and are able to work with systems of lower complexity and a nicer proof theoretical behavior (justified knowledge systems). In practical terms, induction is redundant.

We have also seen that Artemov's justified knowledge systems are equivalent to McCarthy's 'any fool' systems which were built particularly to address these epistemic puzzles. This lends credence to applications of J systems to these puzzles and endows the O systems with a constructive, evidence-based semantics via the Realization Theorem.

We may also care to consider whether there is there a wider class of puzzles to which these observations or conservativity may apply and whether there is a benefit to considering a logic which contains both J and C modalities.

References

[A01] S. Artemov. Explicit provability and constructive semantics. *Bulletin of Symbolic Logic,* 7(1):1-36, 2001.

[A04] S. Artemov. Evidence-Based Common Knowledge. Technical Report TR-2004018, CUNY Ph.D. Program in Computer Science, 2004.

[A06] S. Artemov. Justified Common Knowledge. To appear in *Theoretical Computer Science*, 2006.

[Bar88] J. Barwise. Three Views of Common Knowledge. TARK 1988: 365-379.

[vBS04] J. van Benthem, D Sarenac. The Geometry of Knowledge. ILLC Report PP-2004-21, University of Amsterdam, 2004.

[B05] Y. Bryukhov. *Integration of decision procedures into high-order interactive provers*. PhD thesis, CUNY Graduate School, 2005.

[D05] S. Demri. Complexity of Simple Dependent Bimodal Logics. Laboratoire LEIBNIZ-CNRS, U.M.R. 5522, manuscript, 2005.

[FHMV95] R. Fagin, J. Halpern, Y. Moses, and M. Vardi. *Reasoning About Knowledge.* MIT Press, 1995.

[F06] M. Fitting. Modal Proof Theory. To appear in *Handbook of Modal Logic*, P. Blackburn, J. van Benthem, and F. Wolter, editors, Elsevier, 2006.

[G92] J. Geanakoplos. Common Knowledge. TARK 1992: 254-315.

[L69] D. Lewis. *Convention*. Harvard University Press, 1969.

[MSTI78] J. McCarthy, M. Sato, T. Hayashi, and S. Igarishi. On the Model Theory of Knowledge. Technical Report STAN-CS-78-657, Stanford University, 1978.

[MvdH95] J.-J. Ch. Meyer and W. van der Hoek. *Epistemic Logic of AI and Computer Science.* Cambridge Tracts in Theoretical Computer Science, 41. Cambridge University Press, 1995.

[R06] N. Rubtsova. Evidence Reconstruction of Epistemic Modal Logic S5. To appear in *Computer Science in Russia 2006, Proceedings*, Lecture Notes in Computer Science, Springer 2006.