paper 8

Formalizing Lewis's Theory of Common Knowledge in Justification Logic

Huub Vromen

Abstract

David Lewis's theory of common knowledge, presented in *Convention* (1969), provides a foundational account of how shared knowledge arises in populations. However, previous attempts to formalize this theory have encountered significant difficulties. Cubitt and Sugden (2003) required key principles as unexplained axioms, while Sillari (2005) showed that a modal logic formalization leads to counterexamples. This paper presents a formalization in justification logic that resolves these problems. By using explicit reason terms, we prove Lewis's axioms A1 and A6 as theorems rather than assumptions, and provide a constructive proof of Lewis's main theorem on common knowledge. The formalization has been fully verified in the Lean 4 proof assistant.

1 Introduction

Common knowledge plays a fundamental role in social interaction, coordination, and convention. When a proposition is commonly known in a population, everyone knows it, everyone knows that everyone knows it, and so on ad infinitum. David Lewis's *Convention* (1969) provided the first systematic philosophical analysis of how such infinite hierarchies of knowledge can arise from finite conditions.

Lewis's key insight was that common knowledge can be grounded in a *basis*—a publicly observable state of affairs that *indicates* the target proposition to everyone, and whose indicative power is itself commonly known. However, formalizing this intuitive account has proven difficult.

Cubitt and Sugden (2003) provided the first rigorous formalization but required two crucial principles—A1 and A6—as unexplained axioms. Sillari (2005) attempted a formalization in modal logic but discovered that A1 fails in general Kripke frames, leading to counterexamples where the conditions for common knowledge are satisfied but the conclusion fails.

This paper presents a formalization in justification logic that resolves these difficulties. The key innovation is using *explicit reason terms* that track the evidential basis for beliefs. This allows us to:

1. Prove A1 and A6 as theorems from minimal assumptions
2. Provide a constructive proof of Lewis's main theorem
3. Capture Lewis's notion of "thereby" in indication
4. Avoid logical omniscience assumptions

The formalization has been fully verified in the Lean 4 proof assistant with no unproven assumptions.

2 Background: Lewis's Theory

2.1 The Problem of Common Knowledge

Lewis was concerned with explaining how conventions arise and persist. A convention, roughly, is a regularity in behavior that solves a coordination problem and is sustained by common knowledge of

the regularity and the preferences of the agents involved.

The challenge is explaining how common knowledge—an infinite hierarchy of iterated knowledge—can arise from finite cognitive resources. Lewis's solution involves the notion of *indication*.

2.2 Indication and Bases for Common Knowledge

Lewis defines indication as follows (1969, p. 52–53):

> "A indicates to someone i that ___ if and only if, if i had reason to believe that A held, i would **thereby** have reason to believe that ___"

The word "thereby" is crucial—it suggests that the reason to believe the conclusion is *based on* the reason to believe the premise. This evidential dependence is central to our formalization.

A state of affairs A is a *basis for common knowledge* of φ in a population P if:

C1 Everyone in P has reason to believe A holds

C2 A indicates to everyone in P that everyone has reason to believe A

C3 A indicates to everyone in P that φ

C4 Indications are transparent: if A indicates α to i, then i has reason to believe that A indicates α to every j

Lewis's theorem states that if these conditions hold, then every proposition in the *G-closure* of φ—the closure of φ under the operation "everyone has reason to believe that ___"—is believed by everyone.

3 Previous Formalizations

3.1 Cubitt and Sugden (2003)

Cubitt and Sugden provided the first rigorous formalization of Lewis's theory. They took the "reason to believe" operator R and the indication relation as primitives, and required two key principles as axioms:

A1 If A indicates φ to i, and i has reason to believe A, then i has reason to believe φ

A6 If A indicates to i that j has reason to believe A, and i has reason to believe that A indicates φ to j, then A indicates to i that j has reason to believe φ

While this formalization is correct, it is philosophically unsatisfying because it provides no explanation of *why* A1 and A6 hold. They appear as brute axioms rather than consequences of the nature of reasons and indication.

3.2 Sillari (2005)

Sillari attempted to remedy this by formalizing the theory in modal logic. He interpreted "i has reason to believe φ" as $\Box_i\varphi$ (the modal necessity operator for agent i) and defined indication as:

$$\mathrm{Ind}_i(\varphi, \psi) := \mathrm{R}_i\varphi \wedge (\varphi \to \psi)$$

However, Sillari discovered that A1 fails in general Kripke frames. The problem is that modal logic's implicit treatment of knowledge allows frames where $\Box(\varphi \to \psi)$ and $\Box\varphi$ hold at a world but $\Box\psi$ fails at accessible worlds. This creates counterexamples to Lewis's theorem.

4 The Justification Logic Approach

4.1 Core Ideas

Justification logic, developed by Artemov (2001), extends modal logic with explicit reason terms. Instead of simply asserting $\Box\varphi$ ("φ is known"), we write $t : \varphi$ ("t is a reason for φ"). This allows us to track the evidential basis for beliefs and to compose reasons using explicit operations.

The key insight for formalizing Lewis is that his use of "thereby" in the definition of indication suggests reasons are *objects* that can be tracked and combined. When A indicates φ to i, it's not just that i would come to believe φ—i would have a *specific reason* for φ that is *based on* their reason for A.

4.2 Formal Definitions

We work with a reason-belief relation $\text{rb}(r, i, \varphi)$ meaning "r is for agent i a reason to believe proposition φ". Reasons form a multiplicative structure where $s * t$ represents applying reason s (for an implication) to reason t (for the antecedent).

Definition 1 (Reason to Believe)**.** *Agent i has reason to believe φ if there exists some reason r such that $rb(r, i, \varphi)$:*

$$R(i, \varphi) := \exists r.\, rb(r, i, \varphi)$$

Definition 2 (Indication)**.** *A indicates φ to agent i if i has reason to believe the implication $A \to \varphi$:*

$$Ind(A, i, \varphi) := R(i, A \to \varphi)$$

This definition captures the "thereby" in Lewis's formulation. If i has reason t for $A \to \varphi$ and reason s for A, then by the application rule, i has composite reason $t * s$ for φ. Crucially, $t * s$ *contains* s as a component—the reason to believe φ is based on the reason to believe A.

4.3 Axioms

We require one fundamental rule and three tautologies:

Axiom 3 (Application Rule (AR))**.** *If agent i has reason s for $\alpha \to \beta$ and reason t for α, then i has reason $s * t$ for β:*

$$rb(s, i, \alpha \to \beta) \land rb(t, i, \alpha) \to rb(s * t, i, \beta)$$

Axiom 4 (T1: Conjunction)**.** *There exists a distinguished reason `conjConst` that justifies conjunction formation:*

$$rb(\texttt{conjConst}, i, \alpha \to \beta \to (\alpha \land \beta))$$

Axiom 5 (T2: Transitivity)**.** *There exists a distinguished reason `transConst` that justifies transitivity of implication:*

$$rb(\texttt{transConst}, i, ((\alpha \to \beta) \land (\beta \to \gamma)) \to (\alpha \to \gamma))$$

Axiom 6 (T3: Distribution)**.** *There exists a distinguished reason `distConst` that justifies distribution of reasons:*

$$rb(\texttt{distConst}, i, R(j, \alpha \to \beta) \to (R(j, \alpha) \to R(j, \beta)))$$

Note that we do *not* assume agents have reasons for all tautologies (logical omniscience). Only these three specific forms are required, matching empirical evidence about human reasoning capabilities.

5 Main Results

5.1 Derived Rules

From the axioms, we derive the following rules:

Lemma 7 (E1: Modus Ponens for Reasons)**.**

$$R(i, \alpha \to \beta) \to R(i, \alpha) \to R(i, \beta)$$

Proof. Given reasons s for $\alpha \to \beta$ and t for α, apply AR to get reason $s * t$ for β. □

Lemma 8 (L1: Conjunction for Reasons)**.**

$$R(i, \alpha) \to R(i, \beta) \to R(i, \alpha \wedge \beta)$$

Proof. Apply T1 twice using AR to combine the reasons. □

Lemma 9 (E2: Transitivity for Reasons)**.**

$$R(i, \alpha \to \beta) \to R(i, \beta \to \gamma) \to R(i, \alpha \to \gamma)$$

Proof. Use L1 to form the conjunction, then apply T2 via AR. □

Lemma 10 (E3: Distribution Rule)**.**

$$R(i, R(j, \alpha \to \beta)) \to R(i, R(j, \alpha) \to R(j, \beta))$$

Proof. Direct application of T3 via AR. □

5.2 Lewis's Axioms as Theorems

The crucial result is that A1 and A6 are now *theorems*, not axioms:

Lemma 11 (A1 as Theorem)**.** *If A indicates φ to i, and i has reason to believe A, then i has reason to believe φ:*

$$Ind(A, i, \varphi) \to R(i, A) \to R(i, \varphi)$$

Proof. Immediate from E1, since $\mathrm{Ind}(A, i, \varphi) = \mathrm{R}(i, A \to \varphi)$. □

Lemma 12 (A6 as Theorem)**.** *If A indicates to i that j has reason to believe A, and i has reason to believe that A indicates φ to j, then A indicates to i that j has reason to believe φ:*

$$Ind(A, i, R(j, A)) \to R(i, Ind(A, j, \varphi)) \to Ind(A, i, R(j, \varphi))$$

Proof. Use E3 to get $\mathrm{R}(i, \mathrm{R}(j, A) \to \mathrm{R}(j, \varphi))$, then use E2 to chain with the first premise. □

The fact that these proofs are trivial—A1 is literally three lines—suggests that justification logic is the natural framework for Lewis's theory.

5.3 G-Closure and Lewis's Main Theorem

Definition 13 (G-Closure)**.** *The G-closure of φ is the smallest set containing φ and closed under the operation "i has reason to believe ___" for all agents i:*

- *φ is in the G-closure*
- *If p is in the G-closure, so is $R(i, p)$ for any i*

Theorem 14 (Lewis's Theorem)**.** *If A is a basis for common knowledge of φ (satisfying C1–C4), then every agent has reason to believe every proposition in the G-closure of φ.*

Proof. By induction on the G-closure. We prove the stronger claim that every agent i has *indication* that A implies p, for all p in the closure.

Base case: $p = \varphi$. By C3, A indicates φ to i.

Inductive case: $p = \mathrm{R}(j, u)$ where A indicates u to i (by IH). By C4, i has reason to believe that A indicates u to j. By A6 with C2, A indicates $\mathrm{R}(j, u)$ to i.

Finally, apply A1 with C1 to convert indication to actual reason. □

6 Discussion

6.1 Why Justification Logic Succeeds

The justification logic formalization succeeds where modal logic fails because it makes explicit the evidential structure that Lewis's theory requires. The "thereby" in his definition of indication is captured by the composition of reason terms: when $t * s$ is a reason for φ, it contains s (the reason for A) as a component.

This explicit structure ensures that indications can always be "cashed out"—if you have a justified belief in an implication and a justified belief in the antecedent, you necessarily get a justified belief in the consequent. Modal logic's implicit treatment of knowledge allows frames where this fails.

6.2 Minimal Assumptions

The formalization requires only three tautologies (T1, T2, T3) beyond the application rule. This is philosophically significant because:

1. It avoids logical omniscience—agents need not have reasons for all tautologies
2. It matches empirical evidence about human reasoning capabilities
3. It shows exactly what inferential capacities are needed for common knowledge

6.3 Comparison of Approaches

Feature	Cubitt-Sugden	Sillari	This paper
R operator	Primitive	Modal $\Box$	$\exists r.\, \mathrm{rb}(r, i, \varphi)$
Indication	Primitive	$\mathrm{R} \wedge (\varphi \to \psi)$	$\mathrm{R}(i, A \to \varphi)$
A1 status	Axiom	Fails	Theorem
A6 status	Axiom	N/A	Theorem
Logical omniscience	Yes	Yes	No

7 Conclusion

This paper has presented a formalization of Lewis's theory of common knowledge in justification logic that resolves the difficulties encountered by previous approaches. By using explicit reason terms, we capture the evidential structure implicit in Lewis's notion of indication and prove his key axioms as theorems from minimal assumptions.

The formalization has been fully verified in the Lean 4 proof assistant, ensuring the correctness of all proofs. The code is available at [repository URL].

The success of this formalization suggests that Lewis's theory is best understood in terms of explicit justifications and evidence rather than modal operators. Common knowledge arises when there is a public, shared structure of reasons that everyone can access, inspect, and combine to reach conclusions.

References

- Artemov, S. (2001). Explicit provability and constructive semantics. *Bulletin of Symbolic Logic*, 7(1), 1–36.
- Artemov, S. (2006). Justified common knowledge. *Theoretical Computer Science*, 357, 4–22.
- Artemov, S. & Fitting, M. (2019). *Justification Logic: Reasoning with Reasons*. Cambridge University Press.
- Cubitt, R. & Sugden, R. (2003). Common knowledge, salience and convention. *Economics & Philosophy*, 19, 175–210.
- Lewis, D. (1969). *Convention: A Philosophical Study*. Harvard University Press.
- Sillari, G. (2005). A logical framework for convention. *Synthese*, 147, 379–400.
- Vromen, H. (2024). Reasoning with reasons: Lewis on common knowledge. *Economics & Philosophy*, 40, 397–418.