Reasoning with Reasons: A Verified Formalisation of Lewis's theory of common knowledge

Huub Vromen
[Radboud University]
[huub.vromen@ru.nl]

November 3, 2025

Abstract

David Lewis's theory of common knowledge and convention has been influential in philosophy, economics, and game theory. While Lewis provided informal arguments, subsequent work has attempted to formalise his theory rigorously. This paper presents a complete mechanically verified formalisation of Lewis's theory in the Lean 4 proof assistant, using justification logic with explicit reason terms.

The formalisation demonstrates that Lewis's crucial axioms A1 and A6 can be *proven as theorems* rather than assumed as axioms, vindicating Lewis's intuitions about the logic of indication. We systematically compare three approaches to formalising Lewis: Cubitt and Sugden's syntactic approach (which works but leaves A1 and A6 unexplained), Sillari's modal logic approach (which fails—A1 is actually false in standard modal logic), and our justification logic approach (which succeeds by making reasons explicit and composable).

The complete formalisation (2,874 lines of Lean code with 1,450+ lines of documentation) is available as supplementary material. All proofs have been mechanically verified, providing the highest level of assurance that Lewis's argument is sound.

Keywords: Common knowledge, convention, David Lewis, justification logic, formal verification, Lean 4, mechanised proof

1 Introduction

David Lewis's theory of common knowledge and convention has profoundly influenced philosophy, economics, and game theory since its publication in 1969. At the heart of Lewis's account lies a subtle but crucial argument: that certain situations—what he calls a "basis for common knowledge"—generate an infinite hierarchy of higher-order reasons to believe. Lewis's informal argument for this claim left gaps that subsequent scholars have attempted to fill through rigorous formalisation.

Recent work by Vromen [2024] argued that the most prominent formalisation attempts face fundamental problems. Specifically, Vromen showed that Sillari [2005]'s widely-cited modal logic approach fails—Lewis's key axiom A1 is actually *false* in Sillari's framework. Moreover, Vromen demonstrated that Cubitt and Sugden [2003]'s syntactic approach, while mathematically correct, leaves unexplained why Lewis's crucial assumptions (axioms A1 and A6) should hold. As an alternative, Vromen proposed that justification logic—a framework developed by Artemov [2001], Artemov and Fitting [2019] that makes reasons for belief explicit—provides the correct foundation. In justification logic, Lewis's axioms A1 and A6 become *theorems* rather than stipulated axioms, vindicating Lewis's intuitions about how indication and reasoning work.

However, Vromen's arguments, like those of Cubitt-Sugden and Sillari before them, were developed informally using standard mathematical notation and prose. While philosophically compelling, informal arguments cannot provide the same level of assurance as mechanically verified proofs. Gaps can remain hidden, unstated assumptions can slip through, and subtle errors can persist even after careful peer review.

1.1 Contribution: Machine-Checked Verification

This paper provides *machine-checked verification* of the key arguments from Vromen [2024] by formalizing them in the Lean 4 proof assistant [de Moura et al., 2015, de Moura and Ullrich, 2021]. We present three complete formalisations totaling 2,874 lines of verified code:

1. **Cubitt-Sugden baseline** (574 lines): A direct formalisation of the syntactic approach where R ("has reason to believe") and Ind ("indication") are primitive relations, and axioms A1–A6 must be assumed. This proves Lewis's theorem but leaves the crucial assumptions unexplained.

2. **Sillari modal logic critique** (1,050+ lines): A formalisation of Sillari's approach using Kripke semantics, together with *machine-checked counterexamples* showing that axiom B3 (the modal analog of A1) fails, that axiom C4 fails, and that Lewis's theorem is either false or trivially vacuous depending on interpretation.

3. **Justification logic solution** (1,250+ lines): A formalisation using explicit reason terms where R is *defined* as the existence of a reason, not taken as primitive. In this framework, we *prove* A1 and A6 as theorems from more basic principles about reasoning, and we prove Lewis's main theorem with a non-trivial inductive argument.

All three formalisations have been mechanically verified by Lean 4's type checker, providing mathematical certainty that the proofs are correct and that no hidden assumptions have been introduced.

1.2 Why Formal Verification Matters for Philosophy

The use of proof assistants is standard practice in mathematics and computer science, where they provide assurance that complex proofs are correct [**??**]. But philosophical arguments have traditionally relied on informal reasoning, checked only by human readers. This raises a natural question: what does mechanical verification add to philosophical inquiry?

Certainty vs. plausibility. Informal philosophical arguments can be persuasive without being rigorously correct. They may contain gaps that readers charitably fill in, or they may rely on principles that seem obvious but turn out to fail in edge cases. Mechanical verification eliminates this gap between persuasiveness and correctness. When Lean accepts a proof, we know with mathematical certainty that the conclusion follows from the stated premises—no gaps, no hidden assumptions, no subtle errors.

Explicit assumptions. formalisation forces us to make every assumption explicit. In informal work, we might write "by standard logical reasoning..." or "it is clear that..." without fully specifying what principles we're invoking. A proof assistant demands complete precision. This paper shows exactly which logical principles are needed for Lewis's argument: not "all tautologies" (as in Cubitt-Sugden) but only three specific principles (T1, T2, T3) about conjunction, transitivity, and meta-reasoning.

Reusability and reproducibility. The Lean formalisation can be checked by anyone with the Lean 4 toolchain. Future researchers can build on this foundation, extending the formalisation to Lewis's full theory of convention or applying it to related problems in social epistemology and game theory. The code serves as a precise, executable specification of the theory.

Discovery through formalisation. While this paper primarily verifies existing arguments, the formalisation process itself can reveal insights. For instance, the explicit construction of reason terms in the justification logic approach makes vivid *why* indication works as Lewis said—we can literally see the composite reasons being built up through the application operator. This concreteness aids philosophical understanding.

1.3 Relationship to Vromen (2024)

It is crucial to understand what this paper contributes beyond Vromen [2024]. The philosophical insights—that Sillari's approach fails, that justification logic is the correct framework, that A1 and A6 should be theorems rather than axioms—were established in that earlier work. This paper does not claim to *discover* these results.

Rather, this paper provides:

- **Machine-checked verification** that the informal arguments are correct
- **Constructive proofs** showing exactly how to derive A1 and A6
- **Explicit counterexamples** to Sillari's approach, verified by Lean
- **Complete formal specifications** of all three approaches
- **Methodological demonstration** of proof assistants for philosophy

The relationship is analogous to that between a mathematical conjecture and its formal proof in a theorem prover. The conjecture (here: justification logic is the right framework) is intellectually prior, but the formal verification provides a different kind of confidence—not philosophical persuasion but mathematical certainty.

Throughout this paper, we explicitly acknowledge which results originated in Vromen [2024] and which are novel contributions of the formalisation. Our aim is not to claim credit for philosophical insights we did not originate, but to demonstrate how those insights can be rigorously verified and made available for future research.

1.4 Technical and Philosophical Innovations

Beyond verification, this formalisation makes several novel contributions:

Defeasible reasoning fragment. Standard justification logic includes an operator + for combining evidence and axioms that make reasoning monotonic. However, reasons to believe are *defeasible*—they can be defeated by new information. The classic example is the Nixon diamond [Horty, 2012]: Nixon is a Quaker (reason to believe he's a pacifist) and Nixon is a Republican (reason to believe he's not a pacifist). Both reasons coexist.

Our formalisation uses only the application operator $\cdot$ (for modus ponens), omitting + and its monotonicity axioms. This makes the logic *non-monotonic*, better modeling practical reasoning. We show that this defeasible fragment suffices for Lewis's purposes—we need only reason composition via application, not arbitrary evidence combination. This is a technical contribution: demonstrating that Lewis's argument works in a minimal, defeasible fragment of justification logic.

Explicit reason construction. The Lean proofs don't just assert that reasons exist; they *construct* them. For example, the proof of A1 explicitly shows how to build a reason for φ from a reason for $(A \to \varphi)$ and a reason for A by using the multiplication operator. This constructiveness provides additional insight into the mechanism of reasoning that abstract existence proofs cannot.

Pedagogical resource. The formalisation includes 1,450+ lines of documentation, explaining each definition, axiom, and theorem in detail. It includes comparison tables, concrete examples, and references to the philosophical literature. This makes it not just a verification artifact but an educational resource for understanding Lewis's theory and the three formalisation approaches.

1.5 Roadmap

The remainder of the paper proceeds as follows:

Section 2 provides background on Lewis's theory, the gaps in his informal argument, and the three previous formalisation attempts (Cubitt-Sugden, Sillari, Vromen 2024).

Section 3 presents the Lean 4 formalisation in detail, organized around the three approaches: Cubitt-Sugden's baseline (§**??**), Sillari's failed modal logic approach (§**??**), and the justification logic solution (§**??**). For each approach, we explain the core definitions, key theorems, and what the formalisation reveals.

Section 4 discusses what mechanical verification guarantees, the formalisation process, and the confidence we can have in the results.

Section 5 analyses *why* justification logic succeeds where the other approaches fail or remain incomplete, drawing out the philosophical implications of the formal results.

Section 6 surveys related work on common knowledge, justification logic applications, and formal methods in philosophy.

Section 7 reflects on methodological lessons, strategies for making formal work accessible to philosophers, and future directions.

Section 8 summarizes the contributions and their broader implications.

All Lean code is publicly available at [GitHub URL to be added upon publication], enabling readers to verify the results independently and build on this foundation.

1.6 Reading Guide for Philosophers

This paper is written for philosophers interested in common knowledge, convention, and formal epistemology, not for Lean experts. We have structured it to be accessible to readers without prior exposure to proof assistants:

- **Lean code snippets** are minimal and carefully explained
- **Mathematical notation** is used where possible instead of raw code
- **Philosophical interpretation** accompanies every technical result
- **Complete proofs** are in the supplementary Lean files, not the main text
- **Concrete examples** (like the candle scenario, traffic conventions) ground abstract definitions

Readers who wish to understand the philosophical arguments without engaging with the formal details can focus on:

- The comparison tables (Tables 1, 2, 3, **??**)
- The philosophical significance paragraphs following each technical result
- Section 5, which synthesizes the lessons

Readers interested in the formal details are encouraged to consult the Lean files directly, which include extensive documentation and references to the philosophical literature.

1.7 A Note on Terminology

Following Lewis [1969] and Vromen [2024], we use "reason to believe" rather than "knowledge" or "belief." This is deliberate: Lewis's theory is about normative relations (what agents *should* believe given their evidence), not psychological states (what they *actually* believe). The distinction matters for his account of convention, where coordination requires that agents act on their reasons even if they haven't fully formed the corresponding beliefs.

We also use "indication" as Lewis did, meaning roughly "provides reason to infer." This differs from standard logical implication because it's sensitive to the epistemic situation of the agent—what evidence they have access to.

When we say "justification logic," we mean Artemov's framework where reasons (or "justifications") are explicit terms in the logic, not implicit accessibility relations. We use "reason" rather than "justification" to maintain continuity with Lewis's terminology, but the technical framework is the same.

2 Background: Lewis and Common Knowledge

This section provides necessary background for understanding the formalisation. We begin with Lewis's original informal theory (§2.1), then examine the gap in his argument that motivated subsequent formalisation attempts (§2.2), and finally survey the three major approaches to filling that gap (§2.3).

2.1 Lewis's Original Theory (1969)

Lewis [1969] developed his theory of common knowledge to explain how conventions arise and persist in populations. His central question was: what makes a regularity of behaviour R a convention in a population P? His answer invoked the notion of *common knowledge*—not just that everyone follows R, but that everyone knows everyone follows R, everyone knows that everyone knows everyone follows R, and so on ad infinitum.

2.1.1 Reasons to Believe vs. Knowledge

A crucial feature of Lewis's account is that he couches it in terms of *reasons to believe* rather than *knowledge* or *actual beliefs*. This is deliberate: knowledge is factive (you can only know what's true), but Lewis needs to account for situations where people coordinate based on shared but potentially false information.

As Lewis later acknowledged [Lewis, 1978], the term "common knowledge" is a misnomer—his theory is really about *common reason to believe*. We follow Vromen [2024] in using "reason to believe" throughout, staying closer to Lewis's intent even when departing from his terminology.

2.1.2 The Definition of Common Knowledge

Lewis's definition is deceptively simple [Lewis, 1969, p. 56]:

> Let us say that it is common knowledge in a population P that __ if and only if some state of affairs A holds such that:
>
> (1) Everyone in P has reason to believe that A holds.
>
> (2) A indicates to everyone in P that everyone in P has reason to believe that A holds.
>
> (3) A indicates to everyone in P that __.
>
> We can call any such state of affairs A a *basis for common knowledge* in P that __.

The notion of *indication* is defined as:

> 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 __. [Lewis, 1969, pp. 52–53]

2.1.3 Concrete Examples

To make this abstract definition concrete, consider Lewis [1969, p. 56]'s own example:

Example 2.1 (The Town Square Meeting)**.** Suppose you and I have agreed to meet at noon in the town square, and we both arrive and see each other there. Let A be the state of affairs: "You and I are facing each other in the town square at noon." Let φ be the proposition: "You and I are meeting as agreed."

Then:

- **C1**: Each of us can see the other, so we each have reason to believe A holds.
- **C2**: Since I can see you seeing me, A gives me reason to believe you have reason to believe A. Symmetrically for you.
- **C3**: The state of affairs of our facing each other in the square gives each of us reason to believe we're meeting as agreed.

Lewis's theorem says: under these conditions, we achieve *common reason to believe* φ—not just that we each believe we're meeting, but that I believe you believe we're meeting, you believe I believe you believe we're meeting, and so on indefinitely.

Lewis's key insight: certain states of affairs are *self-evident* in a strong sense. They provide publicly accessible, self-reinforcing evidence that generates an infinite hierarchy of higher-order reasons to believe. The traffic convention provides another example: years of observing everyone drive on the right creates a public state of affairs indicating the regularity will continue.

2.2 The Gap in Lewis's Argument

Lewis's informal argument for his main theorem—that conditions C1–C3 generate common reason to believe—proceeds by iteration. He shows:

- **First-order:** From C3 and C1, everyone has reason to believe φ.
- **Second-order:** From C2 and C3, everyone has reason to believe that everyone has reason to believe φ.
- **Third-order:** By similar reasoning, everyone has reason to believe that everyone has reason to believe that everyone has reason to believe φ.
- **And so on...**

The pattern seems clear, but Lewis relies on a principle he states informally [Lewis, 1969, p. 53]:

> Therefore, if A indicates to x that y has reason to believe that A holds, and if A indicates to x that __, and if x has reason to believe that y shares x's inductive standards and background information, then A indicates to x that y has reason to believe that __ (this reason being y's reason to believe that A holds).

This principle—call it **axiom S**—is needed to move from the n-th level to the $(n+1)$-th level in the iteration. But Lewis provides no proof of S. He argues informally that it should follow from shared inductive standards and background information, but the argument has gaps. This is the central technical problem that subsequent formalisations attempt to solve.

2.3 Three Approaches to Filling the Gap

Three major attempts have been made to formalize Lewis's theory and prove his main theorem rigorously. Each represents a different paradigm for understanding "reason to believe" and "indication."

2.3.1 Cubitt and Sugden (2003): The Syntactic Approach

Cubitt and Sugden [2003] were the first to provide a rigorous formalisation. Their approach is *syntactic*: they treat "i has reason to believe φ" and "A indicates φ to i" as primitive predicates in a formal language, and axiomatize their behaviour.

Key definitions. $R_i(\varphi)$ means individual i has reason to believe proposition φ; $\mathrm{Ind}_{A,i}(\varphi)$ means state of affairs A indicates φ to individual i.

Key axioms.

- **A1** (Detachment): $\mathrm{Ind}_{A,i}(\varphi) \wedge R_i(A) \implies R_i(\varphi)$
- **A6** (Transitivity): If A indicates that j has reason to believe A, and i has reason to believe that A indicates φ to j, then A indicates that j has reason to believe φ.
- **C4** (Shared reasoning): If A indicates φ to i, then i has reason to believe that A indicates φ to any other agent j.

With these axioms, Cubitt and Sugden prove Lewis's main theorem rigorously. The proof is valid and fills the gap. However, the axioms A1, A6, and C4 are *assumed*—they work, but we don't know *why* they're true. They're stipulations about how indication behaves, not theorems derived from more fundamental principles.

2.3.2 Sillari (2005): The Modal Logic Approach

Sillari [2005] attempted a different approach using modal epistemic logic with Kripke semantics. This is the *semantic* paradigm: interpret "reason to believe" in terms of possible worlds and accessibility relations. Sillari uses standard modal logic: $\Box_i \varphi$ is true at world w iff φ holds at all worlds accessible from w via agent i's accessibility relation.

Sillari's definition of indication. Sillari defines:

$$\mathrm{Ind}_{A,i}(\varphi) := \Box_i A \wedge (A \to \varphi)$$

That is, A indicates φ to i at world w if i has reason to believe A at w, and the material conditional $A \to \varphi$ holds at w.

The attempted proof. Sillari defines a modal operator $\mathrm{CRG}(\varphi)$ for "common reason to believe φ in group G" in terms of reachability: φ is common reason to believe if it holds at all worlds reachable from the actual world via any finite sequence of agents' accessibility relations. He then attempts to prove Lewis's theorem by induction on path length.

The problems. Vromen [2024] identified three critical problems:

1. **Axiom B3 fails.** Sillari's axiom B3 states: $\Box_i A \wedge \mathrm{Ind}_{A,i}(\varphi) \implies \Box_i \varphi$. Vromen constructed a counterexample showing this is *false* in Kripke frames.

2. **Axiom C4 fails.** This axiom also fails in Kripke semantics.

3. **Lewis's theorem is false or trivial.** Depending on whether we interpret Lewis's conditions C1–C3 as holding only at the actual world (local interpretation) or at all worlds (global interpretation), Lewis's theorem is either false (counterexamples exist) or trivially true (the proof is vacuous).

The conclusion is stark: modal logic is *the wrong framework* for Lewis's theory. The failure is not a technicality that can be fixed with different frame conditions; it's fundamental to the modal approach.

2.3.3 Vromen (2024): The Justification Logic Approach

Vromen [2024] proposed a radically different approach using Artemov's justification logic [Artemov, 2001, Artemov and Fitting, 2019]. This is the *proof-theoretic* paradigm: make reasons explicit as terms in the logic.

Explicit reasons. Instead of saying "i has reason to believe φ" (a black box), we say "reason r justifies i in believing φ." Reasons are objects in the logic that we can quantify over, compose, and track.

The basic setup. The primitive is a ternary relation:

$$r :_i \varphi \quad \text{"reason } r \text{ justifies agent } i \text{ in believing } \varphi\text{"}$$

We then *define*:

$$R_i(\varphi) := \exists r.\, r :_i \varphi \quad \text{"there exists a reason for } i \text{ to believe } \varphi\text{"}$$

$$\mathrm{Ind}_{A,i}(\varphi) := \exists r.\, r :_i (A \to \varphi) \quad \text{"} i \text{ has a reason to believe } A \to \varphi\text{"}$$

The key difference: in Cubitt-Sugden, R and Ind are primitive. In Sillari, they're defined via modal operators. Here, they're defined via *explicit reasons.*

Reason composition. Reasons form a multiplicative structure: if r and s are reasons, then $r \cdot s$ is also a reason. The multiplication represents *application*—applying a reason for an implication to a reason for the antecedent to get a reason for the consequent.

The fundamental axiom is:

Axiom 2.2 (Application Rule (AR))**.** If reason s justifies i in believing $\alpha \to \beta$, and reason t justifies i in believing α, then $s \cdot t$ justifies i in believing β:

$$(s :_i (\alpha \to \beta)) \wedge (t :_i \alpha) \implies (s \cdot t) :_i \beta$$

A1 becomes a theorem. With this framework, axiom A1 is no longer an axiom—it's a *theorem*:

Theorem 2.3 (A1 is provable in justification logic)**.** $\mathit{Ind}_{A,i}(\varphi) \wedge R_i(A) \implies R_i(\varphi)$

Proof sketch.

1. Assume $\mathrm{Ind}_{A,i}(\varphi)$, so $\exists s.\, s :_i (A \to \varphi)$ for some reason s.
2. Assume $R_i(A)$, so $\exists t.\, t :_i A$ for some reason t.
3. By the application rule (AR), $s \cdot t :_i \varphi$.
4. Therefore $\exists r.\, r :_i \varphi$ (namely, $r = s \cdot t$), hence $R_i(\varphi)$. □

The proof is simple but profound. It shows that A1 follows from the logical structure of reasoning with explicit justifications. Lewis's intuition (captured in the word "thereby") is vindicated: having a reason for $A \to \varphi$ and a reason for A *thereby* gives you a reason for φ because you can literally apply the first reason to the second.

A6 and minimal assumptions. Similarly, axiom A6 becomes a theorem in justification logic. Importantly, the justification logic approach requires only minimal logical assumptions. Unlike Cubitt-Sugden (who assume agents reason to all tautologies) or Sillari (who assume the necessitation rule), Vromen needs only three specific axioms (T1, T2, T3) about conjunction, transitivity, and meta-reasoning—basic inferences real agents can perform.

The significance. The justification logic formalisation shows that Lewis's theory is *correct as stated*. When we use the right conceptual framework (explicit reasons), the key assumptions (A1, A6) that seemed to require stipulation turn out to be theorems. This vindicates Lewis's informal argument: his intuitions about "thereby" and indication captured something fundamental about the logic of reasoning.

2.4 What This Paper Adds: Mechanical Verification

All three approaches described above—Cubitt-Sugden's syntactic approach, Sillari's modal logic approach, and Vromen's justification logic approach—were developed using informal mathematics. While Vromen [2024] provided arguments for why Sillari's approach fails and why justification logic succeeds, those arguments were presented in prose and standard mathematical notation.

This paper takes the next step: *machine-checked verification* of these results in Lean 4. We formalize all three approaches, prove (or disprove) the key theorems, and verify the entire argument mechanically. The result is not just philosophical persuasion but mathematical certainty.

In the next section, we present the Lean formalisation in detail.

3 Formalisation in Lean 4

3.1 The Lean 4 Proof Assistant

Lean 4 [de Moura and Ullrich, 2021] is an interactive theorem prover based on dependent type theory. Its expressive type system allows natural formalisation of philosophical arguments, while its strong verification guarantees ensure logical correctness. Mechanical verification means every proof step is checked by Lean's kernel, ensuring no gaps, unstated assumptions, or logical errors exist.

3.2 Core Definitions

The formalisation comprises three files totaling 2,874 lines: `Cubitt_Sugden.lean` (574 lines) demonstrates the syntactic baseline where A1–A6 are axioms;

Table 1: Comparison of R operator across three approaches

Approach	**Definition of $R\,i\,\varphi$**	**Internal Structure?**
Cubitt-Sugden	Primitive relation	No – black box
Sillari	$\Box_i\varphi$ (at all accessible worlds)	No – implicit
Vromen (this work)	$\exists r.\, \texttt{rb}\, r\, i\, \varphi$	**Yes – explicit reason**

`Sillari_improved.lean` (1,050+ lines) proves modal logic fails with verified counterexamples; and `reasons_improved.lean` (1,250+ lines) provides the solution via justification logic where A1 and A6 become theorems. We focus on the third file, which contains the main contribution.

3.2.1 Reasons and Beliefs

The fundamental innovation of justification logic is making reasons *explicit*. Rather than saying "individual i believes φ" or "individual i knows φ", we say "reason r justifies individual i in believing φ."

Definition 3.1 (Justification relation)**.** The primitive relation is:

$$\texttt{rb} : \text{reason} \to \text{indiv} \to \text{Prop} \to \text{Prop}$$

We read $\texttt{rb}\, r\, i\, \varphi$ as: "reason r justifies individual i in believing proposition φ."

This is the atomic notion in our formalisation. Everything else is built from this and the composition operation on reasons.

Definition 3.2 (Reason composition)**.** Reasons form a multiplicative structure. The multiplication $r * s$ represents the *application* of reason r to reason s. This will be crucial for deriving A1.

Definition 3.3 (Having reason to believe)**.** Individual i has reason to believe φ if there exists at least one reason justifying this belief:

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

Key insight. This existential quantification makes the difference. In Cubitt-Sugden, R is a primitive relation with no internal structure. In Sillari, R corresponds to $\Box_i$ with accessibility relations. Here, R is *defined* as the existence of an explicit reason term.

This seemingly small change—making reasons explicit—is what enables proving A1 and A6 rather than assuming them.

Table 2: Definition of indication across three approaches

Approach	Definition of $\texttt{Ind}\,A\,i\,\varphi$
Cubitt-Sugden	Primitive relation (undefined)
Sillari	$R\,i\,A \wedge (A \to \varphi)$ (conjunction)
Vromen	$R\,i\,(A \to \varphi) = \exists r.\, \texttt{rb}\,r\,i\,(A \to \varphi)$

3.2.2 Indication

Lewis's crucial notion of "indication" can now be defined precisely.

Definition 3.4 (Indication)**.** Individual i's reason for A indicates φ when i has reason to believe the conditional $A \to \varphi$:

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

Expanded: $\texttt{Ind}\,\texttt{rb}\,A\,i\,\varphi$ means $\exists r.\,\texttt{rb}\,r\,i\,(A \to \varphi)$ – there exists a reason r that justifies i in believing "if A then φ."

Why this captures Lewis's "thereby". Lewis [1969, pp. 52–53] writes that A indicates B for someone means "the person in question has reason to believe the conditional, if A then B." The key word is "thereby"—when i has reason for A and reason for $(A \to \varphi)$, i thereby has reason for φ. This follows from the structure of reasoning with explicit justifications: the reason r for $(A \to \varphi)$ is a function that can be *applied* to a reason s for A, producing a composite reason $r \cdot s$ for φ.

Sillari's definition might look similar, but the crucial difference is that in modal logic, having $\Box_i(A \to \varphi)$ does not give you a function that can be applied to $\Box_i A$. The modal operators lack the compositional structure that reasons possess.

3.3 Axioms

Our formalisation requires seven axioms—significantly fewer than the nine axioms (A1–A6, CS1–CS3) in Cubitt-Sugden. These are more fundamental: they describe the logic of reasoning with reasons rather than specific properties of common knowledge.

3.3.1 The Application Rule (AR)

Axiom 3.5 (Application Rule)**.**

$$\forall s, t, i, \alpha, \beta.\quad \texttt{rb}(s, i, \alpha \to \beta) \wedge \texttt{rb}(t, i, \alpha) \implies \texttt{rb}(s * t, i, \beta)$$

Interpretation. If reason s justifies i in believing $\alpha \to \beta$, and reason t justifies i in believing α, then their composition $s * t$ justifies i in believing β.

This is the heart of justification logic. It says that reasons compose via application: applying a reason for an implication to a reason for the antecedent yields a reason for the consequent. This is modus ponens at the level of reasons, not just propositions.

3.3.2 Basic Reasoning Axioms (T1–T3)

Three axioms encode basic principles about reasoning:

Axiom 3.6 (T1: Conjunction introduction)**.**

$$\exists a. \forall i, \alpha, \beta. \quad \mathtt{rb}(a, i, \alpha \to \beta \to (\alpha \wedge \beta))$$

There exists a reason a that justifies believing: if you have α and β, then you have $\alpha \wedge \beta$.

Axiom 3.7 (T2: Transitivity of implication)**.**

$$\exists b. \forall i, \alpha, \beta, \gamma. \quad \mathtt{rb}(b, i, ((\alpha \to \beta) \wedge (\beta \to \gamma)) \to (\alpha \to \gamma))$$

There exists a reason b that justifies believing: if $\alpha \to \beta$ and $\beta \to \gamma$, then $\alpha \to \gamma$.

Axiom 3.8 (T3: Meta-reasoning)**.**

$$\exists c. \forall i, j, \alpha, \beta. \quad \mathtt{rb}(c, i, R(j, \alpha \to \beta) \to (R(j, \alpha) \to R(j, \beta)))$$

There exists a reason c that justifies i in believing: if j has reason to believe $\alpha \to \beta$, and j has reason to believe α, then j has reason to believe β. This allows agents to reason about other agents' reasoning.

Minimality. Unlike Cubitt-Sugden (who assume agents have reasons to believe all tautologies), we only assume these three specific reasoning principles—more realistic for actual agents who can perform basic inferences without immediately grasping all logical truths.

3.4 Main Results

3.4.1 Helper Lemmas (E1–E3)

Three lemmas follow from the axioms:

Lemma 3.9 (E1: Modus ponens)**.**

$$\forall i, \alpha, \beta. \quad R(i, \alpha \to \beta) \wedge R(i, \alpha) \implies R(i, \beta)$$

Proof sketch. Extract witnesses $s :_i (\alpha \to \beta)$ and $t :_i \alpha$. Apply AR to get $(s * t) :_i \beta$. Therefore $\exists r.\, r :_i \beta$, so $R\, i\, \beta$.

Lemma 3.10 (E2: Conjunction and transitivity)**.**

$$\forall i, \alpha, \beta, \gamma. \quad R(i, \alpha \to \beta) \wedge R(i, \beta \to \gamma) \implies R(i, \alpha \to \gamma)$$

Proof sketch. Use T1 to form conjunction of the two implications, then apply T2 for transitivity, then use E1.

Lemma 3.11 (E3: Meta-level reasoning)**.**

$$\forall i, j, \alpha, \beta. \quad R(i, R(j, \alpha \to \beta)) \implies R(i, R(j, \alpha) \to R(j, \beta))$$

Proof sketch. Direct application of T3 and E1.

3.4.2 Theorem: A1 is Provable

Theorem 3.12 (A1 – Lewis's detachment principle)**.**

$$\forall i, A, \varphi. \quad \mathit{Ind}(A, i, \varphi) \wedge R(i, A) \implies R(i, \varphi)$$

Proof (in Lean): *Unfold definitions. From $\mathit{Ind}(A, i, \varphi)$ we get $\exists s.\, \texttt{rb}(s, i, A \to \varphi)$. From $R(i, A)$ we get $\exists t.\, \texttt{rb}(t, i, A)$. By the application rule AR, we have $\texttt{rb}(s * t, i, \varphi)$. Therefore $\exists r.\, \texttt{rb}(r, i, \varphi)$, which is $R(i, \varphi)$.* □

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

Status in different approaches:

- Cubitt-Sugden: **AXIOM** (A1) – must be assumed
- Sillari: **FALSE** – counterexample exists in Kripke frames
- This work: **THEOREM** – proven from AR in 6 lines

Key insight. The proof works because AR gives us reason *composition.* When we have a reason s for $(A \to \varphi)$ and a reason t for A, we can literally apply s to t (via multiplication) to get a composite reason $s \cdot t$ for φ. This is what "thereby" means formally. In Cubitt-Sugden, there was no explanation for *why* indication works—it was simply axiomatized. Here we see it follows from the compositional structure of reasoning with explicit justifications.

3.4.3 Lewis's Main Theorem

Theorem 3.13 (Lewis's Convention Theorem)**.** *Given:*

$$\begin{aligned} &\textit{C1:} && \forall i.\, R(i, A) \\ &\textit{C2:} && \forall i, j.\, \mathit{Ind}(A, i, R(j, A)) \\ &\textit{C3:} && \forall i.\, \mathit{Ind}(A, i, \varphi) \\ &\textit{C4:} && \forall \alpha, i, j.\, \mathit{Ind}(A, i, \alpha) \implies R(i, \mathit{Ind}(A, j, \alpha)) \end{aligned}$$

Then for any proposition p in the R-closure $G(\varphi, p)$:

$$\forall i.\, R(i, p)$$

Proof (in Lean): *By induction on the R-closure structure. The base case uses C3. The inductive step uses C4 to lift indication, A6 to compose indication, and A1 to extract the reason.* □

where `Grb` $\varphi\, p$ represents the R-closure: propositions reachable from φ by iteratively applying "i has reason to believe" for various individuals.

Statement. If everyone in P has reason to believe A, and the indication structure satisfies C2–C4, then for any proposition p in the R-closure of φ, everyone has reason to believe p.

Why this proof is non-trivial. Unlike in Cubitt-Sugden (where it follows straightforwardly from axioms) or Sillari (where it's either false or vacuous), this proof shows a genuine inductive structure. We use A1 (which we proved!) and A6 (which we proved!) at each step, demonstrating that Lewis's conditions genuinely suffice for common knowledge.

Significance. This is Lewis's main result—common knowledge (iterated knowledge to all finite levels) arises from the initial conditions (everyone believes A) plus the indication structure. The formalisation shows that Lewis's argument is completely rigorous and the conclusion follows necessarily from the premises.

4 Verification and Validation

Having presented the three formalisations, we examine what mechanical verification provides and what confidence we can have in the results.

4.1 What Gets Mechanically Checked

When Lean 4 accepts a proof, its type checker verifies:

Type correctness. Every expression must have a well-defined type. Type errors that would make definitions meaningless are impossible. We cannot accidentally quantify over propositions when we meant to quantify over reasons, or apply a function to the wrong number of arguments.

Logical validity. Every proof must be logically valid according to Lean's underlying type theory. This means every step is justified by axioms or previously proven lemmas, no gaps exist (unlike informal proofs that might say "clearly..."), no circular reasoning occurs (Lean's termination checker prevents this), and no inconsistent axioms are introduced.

Completeness. Our formalisation contains **zero uses of `sorry`** (Lean's admitted proof mechanism)—every proof is complete. Every theorem statement has a verified proof, every lemma used has itself been proven, and all proof obligations are satisfied.

Explicitness of assumptions. Perhaps most valuable: *all assumptions are explicit.* When we state `A1_theorem`, the type signature makes clear exactly what the theorem depends on—a justification relation, a proposition serving as the basis, the definitions of `Ind` and `R`, and nothing else. No hidden global assumptions exist.

4.1.1 What Lean Does Not Check

Understanding the limits is important:

Correctness of the formalisation. Lean verifies that our *formal* statements follow from our *formal* axioms. It does not verify that our formalisation correctly captures Lewis's *intended* meaning. If we misinterpret Lewis, Lean won't catch that error. This is the "formalisation gap"—bridging it requires philosophical judgment, not just technical skill.

Adequacy of axioms. Lean checks that our theorems follow from axioms, but not whether axioms are true or philosophically adequate. We address this by keeping axioms minimal (7, down from 9 in Cubitt-Sugden), using well-established principles (AR is standard in justification logic), and making axioms explicit for scrutiny.

Philosophical interpretation. Lean verifies that A1 is a theorem, not that this is *philosophically significant.* That's an argument we make in Section 5, and readers must judge its merits.

4.2 The Formalisation Process

The formalisation followed an iterative process across three phases:

Phase 1: Cubitt-Sugden. We began with direct transcription of their semi-formal presentation: choosing Lean types, translating axioms A1–A6, formalizing the R-closure as an inductive type, and proving Lewis's theorem.

This was relatively straightforward, with main challenges being structural decisions about definitions and choosing appropriate proof strategies.

Phase 2: Sillari. Formalizing the modal approach required defining Kripke frames and accessibility relations, implementing modal operators semantically, constructing specific counterexample frames, and proving axioms fail on those frames. Counterexamples could be constructive: we build explicit frames (e.g., two worlds with specific accessibility and proposition values) and prove by computation that axioms fail. Mechanical verification provided assurance that our constructions were correct.

Phase 3: Vromen. The justification logic formalisation required most innovation because Vromen's informal presentation left some details implicit. We formalized reason terms as a type with multiplication, translated AR into Lean, derived intermediate lemmas (E1–E3), proved A1 and A6 as theorems, and proved Lewis's theorem using derived results. The main challenge was determining exactly which axioms were needed—through experimentation, we found only T1–T3 (plus AR) suffice, without full logical omniscience.

Several technical decisions shaped the formalisation. We use Lean's `inductive` mechanism to define the R-closure, providing natural induction principles. We parameterize definitions rather than using global constants, making dependencies explicit and enabling reuse. We balance automation (using tactics for routine steps) with explicit proofs (for philosophically significant steps where the reasoning matters).

4.3 Confidence Levels

What confidence do we have in the results?

Very high confidence in logical correctness. If the formalisation compiles, the proofs are valid. Lean's kernel is small (few thousand lines) and has been extensively reviewed. The probability of a kernel bug affecting our results is negligible.

High confidence in axiom choice. Our axioms (AR, T1–T3, E1–E3) are either standard (AR) or straightforward logical principles. They're weaker than assuming full logical omniscience, making the results more robust.

Moderate confidence in formalisation adequacy. We've been careful to stay close to Lewis's text and document all interpretive choices. But other readings are possible. We mitigate this by comparing three different formalisations—agreement across approaches increases confidence.

Lower confidence in some details. For instance, whether the existential rules (E1–E3) are the *most natural* way to handle quantification is debatable. But the key results (A1 and A6 as theorems) don't depend sensitively on these choices.

4.4 Comparison with Peer Review

Mechanical verification complements traditional peer review.

Peer review evaluates: philosophical adequacy, novelty and significance, whether formalisation captures intended meaning, and writing quality.

Mechanical verification guarantees: logical correctness (no proof gaps), explicit assumptions, independent verification by anyone with tools, and catching subtle errors humans might miss.

Complementary roles. Peer reviewers must still evaluate whether this is the right formalisation of Lewis, whether axioms are philosophically justified, whether results are philosophically significant, and whether presentation is clear. But reviewers need not check whether all proof steps are valid (Lean guarantees this), whether any cases were missed (completeness is verified), or whether hidden assumptions crept in (all assumptions are explicit). This division of labour makes peer review more effective—reviewers can focus on philosophical substance rather than checking technical details.

4.5 Reusability and Reproducibility

A key benefit is that the formalisation can be reused and extended.

Reproducibility. Anyone with Lean 4 can verify our results independently: download the code, install Lean 4, run `lake build`, and Lean will check all proofs (takes ∼15 minutes). This is *algorithmic reproducibility*—verification by deterministic process, not human judgment.

Extensibility. Future researchers can build on this formalisation to extend to Lewis's full theory of convention (coordination problems, equilibria, salience, precedent), apply to other domains (game theory, communication, distributed computing, social epistemology), explore variants (different notions of indication, alternative axiom systems, connections to other logics), or use pedagogically (teaching material for formal epistemology, example of philosophical formalisation).

4.6 Methodological Lessons

Formalisation is most valuable when: (1) the argument is complex (Lewis's iterative argument with nested quantifiers is hard to track informally), (2) subtle errors are possible (Sillari's approach seemed plausible but had hidden problems), (3) assumptions are implicit (Lewis's argument left unstated what logical principles were needed), (4) multiple interpretations exist (different formalisations illuminate different aspects), and (5) reuse is anticipated (foundation for future work).

Mechanical verification does not make informal arguments obsolete. Vromen [2024]'s informal arguments were essential for identifying problems, proposing solutions, explaining why A1 and A6 should be theorems, and motivating

Table 3: Fundamental design choices across three approaches

Feature	**Cubitt-Sugden**	**Sillari**	**Vromen**
Paradigm	Syntactic	Semantic (modal)	Proof-theoretic
R operator	Primitive relation	Modal $\Box_i$	$\exists r.\, \mathrm{rb}\, r\, i\, \varphi$
What R means	"In i's logic"	"At accessible worlds"	"Explicit reason exists"
Indication	Primitive relation	$R\, i\, A \wedge (A \to \varphi)$	$R\, i\, (A \to \varphi)$
Reason structure	None (black box)	Implicit (accessibility)	**Explicit (terms)**
Compose reasons?	No	No	**Yes (multiplication)**
A1 status	AXIOM	**FAILS**	**THEOREM**
A6 status	AXIOM	N/A	**THEOREM**
Lewis theorem	From axioms	False/vacuous	**Non-trivial proof**
Omniscience	Yes (all tautologies)	Yes (modal K)	**No (minimal)**
# of axioms	9 (A1–A6, CS1–CS3)	12+ (modal axioms)	**7 (AR, T1–T3, E1–E3)**

philosophical significance. The formalisation *verified* those arguments but didn't originate them.

The gap between informal concepts and formal representations is unavoidable. Bridging it requires careful reading of source material, comparing multiple formalisation attempts, extensive documentation explaining design choices, and philosophical judgment about adequacy. Mechanical verification ensures that *given* our formalisation, the results are correct, but choosing the *right* formalisation remains a philosophical task.

4.7 Summary

Mechanical verification provides: certainty that proofs are logically valid and complete, explicitness about assumptions and dependencies, reproducibility through algorithmic checking, reusability as foundation for future research, and assurance that subtle errors have not crept in. It does not provide: philosophical insight (that comes from informal reasoning), guarantee of correct formalisation (that requires judgment), or proof that premises are true (that requires empirical work). The value lies in the *combination* of philosophical insight (from Vromen's 2024 paper) and mechanical verification (this paper).

5 Why This Approach Succeeds

5.1 Three Paradigms for Formalizing Lewis

We have seen three attempts to formalize Lewis's theory, each representing a different paradigm for understanding "having reason to believe" and indication.

The table reveals a fundamental progression: Cubitt-Sugden works but is opaque; Sillari adds structure but the wrong kind; this work adds the *right* kind of structure—explicit, composable reason terms.

5.2 Why the Syntactic Approach is Limited

Cubitt and Sugden's [2003] approach treats "having reason to believe" as a primitive relation with no internal structure. This syntactic paradigm defines R and `Ind` as basic predicates, postulates axioms, and derives consequences.

What works. The syntactic approach proves Lewis's theorem. If we accept axioms A1–A6 (plus CS1–CS3), the conclusion follows. This is mathematically correct, and Cubitt-Sugden deserve credit for providing the first rigorous formalisation.

What's missing: Explanation. By making A1 and A6 axioms, the approach leaves unexplained why indication works as intended and why indication propagates through existentials. Lewis [1969, p. 53] says having "reason to believe the conditional, if A then that A." But why does having reason to believe $(A \to \varphi)$, together with reason to believe A, give you reason to believe φ? The syntactic approach simply *stipulates* this via axiom A1 without explaining why.

The black box problem. The fundamental limitation is that R is a black box. We cannot look inside to see why someone has reason to believe something, track which reasons justify which beliefs, or compose reasons to create new reasons. This is philosophically unsatisfying—Lewis's theory is about *reasoning*, but the syntactic approach gives no insight into the reasoning process itself.

Assessment. Strengths: first rigorous formalisation, mathematically correct, proves Lewis's theorem, clear and accessible. Weaknesses: A1 and A6 unexplained (axioms not theorems), no insight into reasoning process, R is a black box, logical omniscience required, 9 axioms (not minimal). Verdict: A valuable baseline, but philosophically incomplete. Shows that Lewis's theorem *can* be proven, but doesn't explain *why* the key assumptions hold.

5.3 Why Modal Logic Fails

Sillari [2005] attempted to formalize Lewis using modal epistemic logic, interpreting "reason to believe" as necessity in a Kripke frame with accessibility relations. This semantic paradigm has been successful for many epistemic problems, making it natural to try for Lewis.

In modal logic, each agent i has an accessibility relation R_i on possible worlds. We define: $\Box_i\varphi$ holds at world w iff φ holds at all worlds w' that i considers possible from w. Sillari defines indication as $\texttt{Ind}_{\text{modal}}(A, i, \varphi) := R_{\text{modal}}(i, A) \wedge (A \to \varphi)$.

The fatal flaw: A1 (B3) fails. The formalisation proves (in `SillariCritique.lean`, theorem `B3_fails`) that the modal logic analog of A1 is **false**. Consider a Kripke frame with worlds $\{w_0, w_1, w_2\}$, agent i accessing both w_1 and w_2

from w_0, and propositions A and φ both true at w_1 and false at w_2. At w_0: $(A \to \varphi)$ is true at both accessible worlds (either both true or both false), so $\Box_i(A \to \varphi)$ is TRUE. But $\Box_i A$ is FALSE (A doesn't hold at w_2) and $\Box_i \varphi$ is FALSE (φ doesn't hold at w_2). Thus $\Box_i(A \to \varphi) \wedge \Box_i A \to \Box_i \varphi$ is FALSE. This is exactly B3 (Sillari's A1), and it fails.

Why it fails. In modal logic, $\Box_i(A \to \varphi)$ can be true "vacuously"—the conditional holds at accessible worlds because A is false at some, not because there's a genuine connection between A and φ. There's no mechanism to "apply" the box operator to evidence. Modal logic reasons about sets of worlds, not justifications. The formula $\Box_i(A \to \varphi)$ says "at all worlds i considers possible, if A then φ"—but this doesn't give agent i a procedure or reason that can be applied to evidence for A to conclude φ.

The two interpretations problem. Vromen [2024, pp. 406–407] identifies that Sillari's formalisation admits two interpretations, and under both, the approach fails. Without transitivity (no S2 axiom), B3 actually FAILS. With transitivity (S2 axiom: $\Box_i \varphi \to \Box_i \Box_i \varphi$), S2 makes the base case of Lewis's theorem (E) FALSE, rendering Lewis's theorem vacuously true. The file `SillariCritique.lean` proves these results.

Assessment. Modal logic is the wrong framework because: (1) no compositional structure—$\Box_i$ operators don't apply to each other, (2) reasoning about worlds not justifications—wrong level of analysis, (3) vacuous truth problem—conditionals true for wrong reasons, (4) missing Lewis's insight—"thereby" requires applying reasons. Verdict: Modal logic fundamentally cannot capture Lewis's theory. This is not a technical problem to be solved with more axioms—it's a mismatch between framework and phenomenon.

5.4 Why Justification Logic Succeeds

Justification logic provides exactly what's needed: explicit, composable reason terms that can be tracked through inferences.

The key innovation: Explicit reasons. The fundamental move is making reasons first-class entities. Instead of $R\,i\,\varphi$ (black box) or $\Box_i \varphi$ (sets of worlds), we have $\exists r.\, \texttt{rb}\, r\, i\, \varphi$ (explicit justifications). We can talk about specific reasons, compose them via multiplication, and track them through proofs.

Reason composition via the Application Rule. AR is what makes "thereby" work:

$$\text{AR}: \quad \text{rb}\ s\ i\ (\alpha \to \beta) \to \text{rb}\ t\ i\ \alpha \to \text{rb}\ (s * t)\ i\ \beta \tag{1}$$

If s is a reason for $(\alpha \to \beta)$ and t is a reason for α, then the application of s to t (written $s * t$) is a reason for β. This is exactly what Lewis needed. When Lewis says "thereby," he means: apply the reason for the conditional to the reason for the antecedent.

Deriving A1: The payoff. With explicit reasons and AR, we can prove A1:

```
lemma A1 (rb : reason -> indiv -> Prop -> Prop) (A : Prop) :
    forall {i : indiv} {alpha : Prop},
    Ind rb A i alpha -> R rb i A -> R rb i alpha := by
  intro i alpha h1 h2
  obtain <t, ht> := h2  -- Extract reason t for A
  obtain <s, hs> := h1  -- Extract reason s for (A -> alpha)
  use s * t              -- Construct reason s*t for alpha
  exact AR rb hs ht      -- Apply application rule
```

The crucial step is applying s to t via multiplication, getting a composite reason $s * t$ that justifies α. Modal logic couldn't do this because boxes don't compose. Cubitt-Sugden couldn't do this because R has no internal structure.

Deriving A6: The full power. A6 is crucial for Lewis's main theorem—it ensures that indication relations propagate correctly through the population P. The proof uses E2 (transitivity) and E3 (meta-reasoning):

```
lemma A6 (rb : reason -> indiv -> Prop -> Prop) (A : Prop)
    (a b c : reason) :
    forall alpha {i j : indiv},
    Ind rb A i (R rb j A) ->
    R rb i (Ind rb A j alpha) ->
    Ind rb A i (R rb j alpha) := by
  intro p i j h1 h2
  -- h1: R i (A -> R j A)
  -- h2: R i (R j (A -> p))
  -- Goal: R i (A -> R j p)
  have h3 : R rb i (R rb j A -> R rb j p) := E3 rb c h2
  have h4 : R rb i (A -> R rb j p) := E2 rb a b h1 h3
  exact h4
```

The existential rules (E1, E2, E3) are sound in justification logic semantics. Cubitt-Sugden had to assume A6; we prove it.

No logical omniscience. We only assume specific logical rules needed (T1, T2, T3), not all tautologies. This is more realistic—real agents don't immediately see all tautological consequences but must perform inferences using specific reasons.

Minimal axioms. We need only 7 axioms: AR (fundamental to all justification logic), T1–T3 (basic logical reasoning), E1–E3 (existential reasoning). Compare to Cubitt-Sugden (9 axioms) and Sillari (12+). Fewer axioms mean simpler logical foundation, less we're assuming, more confidence in conclusions. Crucially: the axioms we have are *general principles about reasoning*, not *specific claims about common knowledge.* A1 and A6 (specific to Lewis) are *derived*, not assumed.

Philosophical vindication. Lewis [1969] had an intuition that having reason for a conditional and reason for the antecedent *thereby* gives you reason for the consequent. He built his theory of convention on this intuition. But is the intuition correct? The justification logic formalisation answers: No, it's not arbitrary. The behaviour Lewis described for indication *follows from the logical structure of reasoning with explicit justifications.* When we formalize what it means to have reasons for beliefs and apply one reason to another, we discover that Lewis's A1 is a *theorem*, not an axiom. This vindicates Lewis's approach—his intuitions about "thereby" and indication captured something fundamental about the logic of reasoning.

5.5 Summary

What each approach shows. Cubitt-Sugden: Lewis's theorem *can* be proven from suitable axioms (contribution: first rigorous formalisation, feasibility). Sillari: Standard modal logic is inadequate for Lewis (contribution: identifies what doesn't work). Vromen: Lewis's theorem with A1, A6 as *theorems* (contribution: correct formalisation with explanation, vindicates Lewis's intuitions).

The progression. (1) Can we prove it? Yes (Cubitt-Sugden). (2) Can we prove it in modal logic? No (Sillari). (3) Can we prove it *and* explain why the key steps work? Yes (this work).

Why success matters. For Lewis scholarship: complete verified formalisation showing his theory works exactly as described. For formal epistemology: concrete example where justification logic succeeds where modal logic fails, suggesting justification logic may be the right framework for reasoning processes (not just knowledge states). For philosophical formalisation: model of how formalisation can do more than verify—it can *explain*. By deriving A1 and A6 rather than assuming them, we understand *why* indication works as Lewis said. For the research community: 2,874 lines of fully documented, mechanically verified code available as reusable resource.

The three approaches tell a complete story: Cubitt-Sugden proved feasibility, Sillari showed what doesn't work, this work provides the correct formalisation with explanation. Together, they demonstrate the value and challenges of bringing formal methods to philosophical problems.

6 Related Work

This formalisation sits at the intersection of common knowledge in formal epistemology, justification logic, and proof assistants in philosophy. We position our contribution within each area.

6.1 Formalisations of Common Knowledge

6.1.1 Modal Logic Approaches

The dominant approach uses modal epistemic logic with Kripke semantics [Fagin et al., 1995, van Ditmarsch et al., 2008]. Common knowledge C_G is defined as the infinite conjunction $K_G\varphi \wedge K_G K_G\varphi \wedge K_G K_G K_G\varphi \wedge \ldots$, where $K_G\varphi := \bigwedge_{i\in G} K_i\varphi$. This framework has been successful in distributed computing [?], game theory [Aumann, 1976], and multi-agent systems [?].

However, modal formalisations differ fundamentally from Lewis. First, they use *knowledge* (factive) rather than Lewis's *reasons to believe* (non-factive, evidence-based). As Paternotte [2011] notes, conventions can be based on shared false beliefs. Second, modal operators provide no mechanism for tracking *which evidence* justifies beliefs. Lewis's "indication" depends on tracking evidential relationships, which modal operators obscure.

Our formalisation shows precisely why modal logic is inadequate: axiom B3 (the modal analog of A1) is *false* in Kripke semantics. This reveals a fundamental mismatch between the framework and Lewis's concept.

6.1.2 Fixed-Point and Dynamic Approaches

Fixed-point definitions [Barwise, 1981, Heifetz, 1999] characterize common knowledge as the largest proposition ψ such that $\psi \implies \varphi \wedge E_G\psi$. While elegant, this still uses knowledge operators and doesn't capture Lewis's notion of a "basis" for common knowledge.

Dynamic epistemic logic [van Ditmarsch et al., 2008] models how knowledge changes through announcements. However, DEL still uses knowledge operators rather than reasons. A future direction would be developing *dynamic justification logic* modeling how public events make reasons available.

Game-theoretic accounts [Aumann, 1976, ?] typically assume knowledge modeled via partitions rather than reasons to believe. They don't address how common knowledge arises from observable situations. Our formalisation complements this work by providing foundations for how common reason to believe emerges.

6.2 Justification Logic and Applications

6.2.1 Artemov's Program

Artemov [2001] introduced justification logic to provide provability semantics for intuitionistic logic. Instead of $\Box\varphi$, write $t : \varphi$ ("term t is a proof of φ"). Extended to epistemic logic [Artemov, 2006], this addresses logical omniscience: agents need not believe all logical consequences, only those they have explicit justifications for.

Standard justification logic includes application $(s \cdot t)$, sum $(s + t)$, and proof checker $(!t)$ operators. We use only application, omitting $+$ and $!$

to create a *defeasible fragment* suitable for non-monotonic reasoning. As Vromen [2024] argues, reasons to believe can be defeated by new evidence, unlike mathematical proofs.

6.2.2 Applications to Common Knowledge

Artemov [2006] first applied justification logic to common knowledge, introducing *justified common knowledge*. Bucheli et al. [2011] extended this, modeling both implicit and explicit common knowledge.

Our work differs: we formalize *Lewis's specific theory* using a minimal defeasible fragment, prove A1/A6 are theorems, and explicitly compare with failed approaches. The relationship is complementary: Artemov and colleagues develop general theory; we show how a fragment applies to Lewis's problem.

Work on defeasible justification logic [Renne, 2009, **?**] supports non-monotonic reasoning through evidence elimination and updates. Our approach is simpler—we just omit the sum operator—though relating our fragment to these frameworks remains open.

6.3 Proof Assistants in Philosophy

Proof assistants remain uncommon in philosophy. Notable exceptions include **?**'s formalisation of Gödel's ontological argument in Isabelle/HOL, **?**'s formalisation of Anselm's argument in Coq, and **?**'s discussion of proof assistants for philosophical logic.

We chose Lean 4 for its readable syntax (closer to mathematical notation), powerful type system (dependent types express quantification naturally), active community, and good tooling (VSCode integration). Other proof assistants could work, but Lean 4 provides the best balance of power and accessibility.

6.4 Convention and Coordination

Our formalisation addresses only Lewis's common knowledge theory (Chapter 2 of *Convention*), not his full theory involving coordination problems, salience, preference structures, and alternative regularities. Cubitt and Sugden [2003] formalize Lewis's full account, modeling coordination games and convention emergence. Vanderschraaf [1998], **?** develop evolutionary accounts partially departing from Lewis.

Our formalisation provides the foundation for formalizing Lewis's full theory. One could extend our Lean code to model coordination problems, formalize salience, and prove claims about convention stability.

6.5 Philosophical Methodology

Philosophers disagree about formalisation's value. Cappelen [2018] argues formal methods can obscure; Williamson [2020] counters that rigor requires precision. Our view: formalisation is valuable when arguments are complex, informal presentations have gaps, different interpretations exist, and precision reveals substantive issues. Lewis's theory satisfies all criteria.

Experimental philosophy [Knobe and Nichols, 2008] tests philosophical intuitions empirically. Our contribution is orthogonal: we show Lewis's *argument* is valid, not that *premises* are true. Empirical work could test whether real situations satisfy C1–C4. Interestingly, experimental work [**??**] suggests folk concepts may be more defeasible than traditional epistemology assumes, supporting our defeasible fragment choice.

6.6 Positioning This Work

Our contribution combines: (1) Lewis's specific theory, (2) justification logic with explicit reasons, (3) defeasible fragment, (4) mechanical verification in Lean 4 (first for common knowledge), (5) comparative analysis of three approaches, and (6) verified counterexamples proving Sillari's approach fails.

No prior work combines these elements. The closest is Vromen [2024], which developed philosophical arguments informally; we provide mechanical verification. This formalisation serves as foundation for formalizing Lewis's full theory, example of proof assistant use in philosophy, starting point for dynamic justification logic, and case study in defeasible reasoning with explicit justifications. The broader lesson: mechanizing philosophical arguments is feasible, valuable, and reveals insights not visible in informal presentations.

7 Discussion

This section reflects on methodological lessons and future research directions.

7.1 When is Formalisation Valuable?

Based on this case study, mechanical verification is most valuable when:

The argument has complex logical structure. Lewis's theorem involves nested quantifiers, multiple iteration levels, and subtle dependencies between conditions. Formalisation makes this structure explicit and checkable.

Different interpretations exist. By formalizing all three approaches (syntactic, modal, proof-theoretic), we can compare them precisely and identify where they succeed or fail.

Subtle errors are possible. Sillari's approach seemed plausible and was published in *Synthese*, yet contained a subtle flaw that persisted for nearly two decades. Mechanical verification makes such errors impossible to miss.

Precision matters philosophically. Making everything explicit reveals insights. Our formalisation shows that Lewis needs only three specific logical principles (T1–T3), not full logical omniscience—making his theory more realistic and robust than previously thought.

Conversely, formalisation may not be worthwhile when arguments are simple and obviously valid, key concepts resist formalisation, the goal is exploration rather than verification, or resources are limited.

7.2 The Role of Informal Arguments

Mechanical verification does not make informal arguments obsolete. The creative philosophical work—identifying problems with Sillari's approach, proposing justification logic as a solution—happened at the informal level in Vromen [2024]. The formalisation *verified* those insights but didn't originate them. This suggests a productive division of labour: informal work explores concepts and identifies problems; formal work verifies arguments and ensures correctness. Both are valuable; neither replaces the other.

7.3 Strategies for Accessibility

Making formal work accessible to philosophical audiences requires several strategies. We minimize code in the main text, using it only when illustrating key points. Instead of raw Lean syntax, we often translate to standard mathematical notation familiar to philosophers. After every technical result, we explain its philosophical significance—for example, why proving A1 as a theorem matters. Comparison tables (Tables 1, 2, 3, **??**) make differences between approaches visually clear. Concrete examples—the town square meeting, traffic conventions, counterexamples to Sillari—bridge the gap between formal definitions and intuitive understanding. The Lean files contain 1,450+ lines of documentation (50% of total code), making the formalisation a pedagogical resource.

The learning curve for proof assistants remains a barrier. As the community grows and resources improve—through better tutorials, collabouration between philosophers and computer scientists, shared libraries, and workshops—this barrier will lower.

7.4 Future Directions

Several research directions emerge from this work:

Completeness. Our defeasible fragment is sound, but is it complete? Standard justification logic (with sum and verification operators) is complete [Artemov and Fitting, 2019], but our fragment omits these. A definitive answer requires further work.

Dynamic extensions. Our formalisation is static, but real situations are dynamic. Following van Ditmarsch et al. [2008]'s dynamic epistemic logic, we could develop dynamic justification logic modeling how public announcements make reasons available, how agents update reasons given new evidence, and how conventions emerge through repeated interaction.

Full formalisation of convention. We formalized only Lewis's theory of common knowledge. The full theory involves coordination problems, salience, precedent, and alternative regularity conditions. These could be formalized using game-theoretic tools alongside our epistemic framework.

Other applications. Justification logic might address other philosophical problems: scientific reasoning (how evidence accumulates), legal reasoning (chains of evidence and standards of proof), testimony (how justification transmits through testimonial chains), and argumentation theory (argument structure and inference patterns). Each application would require domain-specific extensions while preserving core ideas.

7.5 Limitations

Lean verifies that our formal claims follow from our formal axioms, but doesn't verify that our formalisation correctly captures Lewis's intended meaning. We've stayed close to Lewis's text and documented interpretive choices, but other interpretations are possible. We formalized three specific approaches but not all possible approaches (fixed-point definitions, game-theoretic models, category-theoretic treatments). Some features are Lean-specific; other proof assistants might require different encodings. Finally, mechanical verification guarantees correctness but doesn't automatically confer understanding—that's why we've emphasized philosophical interpretation throughout.

7.6 Conclusion

This formalisation demonstrates that proof assistants can contribute meaningfully to philosophy through verification (confirming informal arguments), precision (making assumptions explicit), discovery (revealing insights like minimal axiom requirements), comparison (enabling precise approach evaluation), and foundation-building (providing verified results for future work). These benefits come at a cost—learning curve and time investment—but for complex arguments with subtle structure, the benefits outweigh the costs. As proof assistants become more accessible, we expect mechanization to become a standard tool in the philosopher's toolkit.

8 Conclusion

David Lewis's theory of common knowledge, developed informally in 1969, has profoundly influenced philosophy, economics, and game theory. But Lewis's informal argument contained gaps, and subsequent attempts to formalize his theory revealed fundamental challenges. This paper has provided machine-checked verification of the key arguments in Vromen [2024], using the Lean 4 proof assistant to settle questions that remained open for decades.

8.1 Summary of Contributions

We have formalized three approaches to Lewis's theory and proven the following results with mechanical verification:

8.1.1 The Cubitt-Sugden Baseline

Cubitt and Sugden [2003] provided the first rigorous formalisation of Lewis, treating "has reason to believe" and "indication" as primitive relations and axiomatizing their behaviour. We have verified:

- Lewis's theorem follows from axioms A1–A6
- The formalisation is mathematically correct
- But A1 and A6 remain unexplained (assumed, not proven)

This establishes that Lewis's argument *can* be made rigorous, though questions about why his key assumptions hold remain open.

8.1.2 Sillari's Failed Modal Logic Approach

Sillari [2005] attempted to formalize Lewis using standard modal epistemic logic with Kripke semantics. We have verified:

- **Axiom B3 (modal analog of A1) fails**: Explicit counterexample constructed and verified
- **Axiom C4 (Cubitt-Sugden) also fails**: Verified counterexample
- **Semantic inconsistency**: Axiom B3 contradicts the semantic clause S2
- **Lewis's theorem**: Either false (local interpretation) or trivially vacuous (global interpretation)

This definitively shows that modal logic is inadequate for Lewis's theory. The failure is not a technicality but reveals a fundamental mismatch.

8.1.3 The Justification Logic Solution

Following Vromen [2024], we formalized Lewis's theory using justification logic with explicit reason terms. We have verified:

- **A1 is a theorem**, not an axiom (proven from the application rule AR)
- **A6 is a theorem**, not an axiom (proven from lemmas E2 and E3)
- **Lewis's main theorem** holds with a non-trivial inductive proof
- **Minimal assumptions**: Only three logical axioms (T1–T3) needed, not full logical omniscience
- **Defeasible fragment**: The logic omits the sum operator, supporting non-monotonic reasoning

This vindicates Lewis's informal intuitions: when we use the right conceptual framework (explicit reasons that can be composed), his crucial assumptions become theorems rather than stipulations.

8.2 Philosophical Significance

Beyond the technical results, this formalisation has philosophical implications:

Lewis was right. The justification logic formalisation shows that Lewis's intuitions about "indication" and "thereby" were correct. His claim that having reason for $A \to \varphi$ and having reason for A *thereby* gives you reason for φ is not a stipulation—it follows from the logical structure of reasoning with explicit justifications.

Framework matters. The three approaches aren't just different notations for the same content. They represent genuinely different conceptual frameworks, and the choice of framework determines what can be proven. Modal logic (treating reasons as implicit in accessibility relations) cannot prove A1. Justification logic (treating reasons as explicit, composable terms) can.

Formalisation reveals structure. By making everything explicit, the formalisation revealed that Lewis needs only minimal logical assumptions (T1–T3), not full logical omniscience. This makes his theory more realistic and shows it's more robust than previously thought.

Explicit reasons are philosophically illuminating. The justification logic approach doesn't just verify Lewis—it explains *why* his argument works. We can see the composite reasons being constructed, track which evidence justifies which beliefs, and understand the mechanism of coordination through shared reasons.

8.3 Methodological Significance

This work also contributes to methodology:

Proof assistants for philosophy. This appears to be the first mechanically verified formalisation of common knowledge in any proof assistant. It demonstrates that philosophical arguments—even subtle, complex ones about epistemology and social coordination—can be mechanized. This opens possibilities for applying formal methods to other philosophical problems.

Verification vs. discovery. The formalisation verified arguments developed informally in Vromen [2024] but didn't originate them. This illustrates the complementary roles of informal and formal reasoning: informal work explores and proposes; formal work verifies and makes precise.

Making assumptions explicit. A key benefit of formalisation is that *all* assumptions are explicit. We now know exactly which principles Lewis's argument requires (AR, T1–T3, C1–C4), with no hidden dependencies.

Reusable foundation. The verified formalisation serves as a foundation for future work on Lewis's full theory of convention, on extensions to dynamic settings, and on applications to game theory and social epistemology.

8.4 Looking Forward

Several directions for future research emerge:

Complete Lewis's theory. We formalized Chapter 2 of *Convention* (common knowledge). The rest of the book (coordination problems, salience, convention stability) could be formalized, building on our foundation.

Dynamic extensions. Develop a dynamic justification logic modeling how public events make reasons available and how agents update their beliefs over time.

Empirical testing. Use experimental methods to test whether Lewis's premises (C1–C4) hold in real situations and whether people reason as Lewis predicts.

Other applications. Apply justification logic to other philosophical problems: scientific reasoning, legal evidence, testimony, argumentation theory.

Semantics and completeness. Investigate the semantic foundations of our defeasible fragment and determine whether completeness holds.

Each of these directions could yield new insights and demonstrate the value of formal methods in philosophy.

8.5 Final Remarks

David Lewis's theory of common knowledge is a cornerstone of social epistemology and game theory. By providing the first mechanically verified formalisation of this theory, we have:

- Established with certainty that Lewis's argument is valid
- Definitively refuted the modal logic approach
- Vindicated the justification logic framework
- Made all assumptions explicit and minimal
- Created a foundation for future formal work

More broadly, we have demonstrated that proof assistants can contribute meaningfully to philosophy. Mechanical verification provides a level of assurance that informal arguments—no matter how carefully developed and peer-reviewed—cannot match. For suitable problems, the investment in formalisation is worthwhile.

As proof assistants become more accessible and the community of formal philosophers grows, we expect this kind of work to become more common. This paper serves as both a specific contribution (verified Lewis) and a methodological example (how to use Lean for philosophy).

The formalisation (2,874 lines of verified code with extensive documentation) is publicly available at [GitHub URL]. We encourage readers to examine it, verify our results independently, and build upon it. The future of rigorous philosophy may well involve a partnership between human insight and machine verification—each contributing what it does best.

Lewis showed us that common knowledge arises from publicly available evidence. Our formalisation shows us that Lewis was right—and now we can be certain.

Acknowledgments

The author thanks [names] for helpful comments and discussions. The formalisation was developed with assistance from Claude (Anthropic) for documentation and presentation. All technical content and proofs are the work of the author.

References

Sergei N. Artemov. Explicit provability and constructive semantics. *The Bulletin of Symbolic Logic*, 7(1):1–36, 2001. doi: 10.2307/2687821.

Sergei N. Artemov. Justified common knowledge. *Theoretical Computer Science*, 357(1-3):4–22, 2006. doi: 10.1016/j.tcs.2006.03.009.

Sergei N. Artemov and Melvin Fitting. *Justification Logic: Reasoning with Reasons*. Cambridge University Press, Cambridge, 2019. ISBN 9781108348263. doi: 10.1017/9781108348263.

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

Jon Barwise. Scenes and other situations. *The Journal of Philosophy*, 78(7): 369–397, 1981. doi: 10.2307/2026481.

Samuel Bucheli, Roman Kuznets, and Thomas Studer. Justifications for common knowledge. *Journal of Applied Non-Classical Logics*, 21(1):35–60, 2011. doi: 10.3166/jancl.21.35-60.

Herman Cappelen. Fixing language: An essay on conceptual engineering. *Inquiry*, 61(1):109–115, 2018. doi: 10.1080/0020174X.2017.1385522.

Robin P. Cubitt and Robert Sugden. Common knowledge, salience and convention: A reconstruction of David Lewis' game theory. *Economics and Philosophy*, 19(2):175–210, 2003. doi: 10.1017/S0266267103001123.

Leonardo de Moura and Sebastian Ullrich. The Lean 4 theorem prover and programming language. In André Platzer and Geoff Sutcliffe, editors, *Automated Deduction – CADE 28*, volume 12699 of *Lecture Notes in Computer Science*, pages 625–635, Cham, 2021. Springer. doi: 10.1007/978-3-030-79876-5_37.

Leonardo de Moura, Soonho Kong, Jeremy Avigad, Floris van Doorn, and Jakob von Raumer. The Lean theorem prover (system description). In Amy P. Felty and Aart Middeldorp, editors, *Automated Deduction – CADE-25*, volume 9195 of *Lecture Notes in Computer Science*, pages 378–388, Cham, 2015. Springer. doi: 10.1007/978-3-319-21401-6_26.

Ronald Fagin, Joseph Y. Halpern, Yoram Moses, and Moshe Y. Vardi. *Reasoning about Knowledge*. MIT Press, Cambridge, MA, 1995. ISBN 9780262562003.

Aviad Heifetz. Iterative and fixed point common belief. *Journal of Philosophical Logic*, 28(1):61–79, 1999. doi: 10.1023/A:1004357300525.

John F. Horty. *Reasons as Defaults*. Oxford University Press, Oxford, 2012. doi: 10.1093/acprof:oso/9780199744077.001.0001.

Joshua Knobe and Shaun Nichols. An experimental philosophy manifesto. pages 3–14, 2008.

David K. Lewis. *Convention: A Philosophical Study*. Harvard University Press, Cambridge, MA, 1969. ISBN 9780674165250.

David K. Lewis. Truth in fiction. *American Philosophical Quarterly*, 15(1): 37–46, 1978.

Cédric Paternotte. Being realistic about common knowledge: A Lewisian approach. *Synthese*, 183(2):249–276, 2011. doi: 10.1007/s11229-010-9770-y.

Bryan Renne. Evidence elimination in multi-agent justification logic. In *Proceedings of the 11th Conference on Theoretical Aspects of Rationality and Knowledge (TARK 2009)*, pages 227–236, California, 2009. ACM Press. doi: 10.1145/1562814.1562845.

Giacomo Sillari. A logical framework for convention. *Synthese*, 147(2):379–400, 2005. doi: 10.1007/s11229-005-1352-z.

Hans van Ditmarsch, Wiebe van der Hoek, and Barteld Kooi. *Dynamic Epistemic Logic*. Springer, Dordrecht, 2008. ISBN 9781402058387. doi: 10.1007/978-1-4020-5839-4.

Peter Vanderschraaf. Knowledge, equilibrium and convention. *Erkenntnis*, 49(3):337–369, 1998. doi: 10.1023/A:1005461514200.

Huub Vromen. Reasoning with reasons: Lewis on common knowledge. *Economics & Philosophy*, 40(2):397–418, 2024. doi: 10.1017/S0266267123000238.

Timothy Williamson. *Suppose and Tell: The Semantics and Heuristics of Conditionals*. Oxford University Press, Oxford, 2020. doi: 10.1093/oso/9780198860662.001.0001.