A Formalisation of Lewis's theory of Common Knwoledge

Huub Vromen
Radboud University
huub.vromen@ru.nl

October 20, 2025

Abstract

We present a complete formalization in Lean 4 of David Lewis's theory of conventions and common knowledge, along with a critical analysis of Sillari's attempted modal logic formalization. Our development consists of three main components: (1) a formalization of Lewis's original argument using a justification logic framework with explicit reasons, (2) an implementation of Sillari's Kripke semantics approach, and (3) formal proofs showing fundamental problems with Sillari's formalization. Specifically, we demonstrate that crucial axioms (B3/A1 and C4) fail under Sillari's definitions, and that Lewis's main theorem either becomes false or trivial depending on how it is interpreted. Our formalization comprises approximately [X] lines of Lean code and includes [Y] theorems and lemmas.

1 Introduction

David Lewis's 1969 theory of conventions provides a foundational account of how coordination problems are solved through common knowledge. His framework has been influential in philosophy, game theory, linguistics, and social epistemology. However, the precise logical structure of Lewis's argument has been debated.

Sillari (2005, 2008) attempted to formalize Lewis's account using standard modal logic with Kripke semantics. While this approach has pedagogical appeal, we show through formal verification in Lean 4 that it suffers from fundamental problems:

- Lewis's crucial axiom A1 (Sillari's B3) fails under the Kripke semantics

- Cubitt and Sugden's axiom C4 also fails
- Lewis's main theorem either has counterexamples or becomes vacuous

We also provide an alternative formalization using justification logic with explicit reasons to believe, showing that Lewis's original argument can be made rigorous without the problems that plague Sillari's approach.

1.1 Contributions

- First complete formalization of Lewis's convention argument in a proof assistant
- Formal verification that Sillari's modal logic formalization is flawed
- Explicit counterexamples demonstrating the failure of key axioms
- Alternative formalization using justification logic
- Machine-checked proofs of all claims

1.2 Related Work

Prior work includes Cubitt and Sugden's (2003) informal analysis, Sillari's modal formalization, and various game-theoretic treatments. To our knowledge, this is the first mechanized verification of these results.

2 Lewis's Original Framework

Lewis's account centers on two key concepts: *reasoning* (R) and *indication* (Ind).

2.1 Core Definitions

Definition 1 (Reasoning). $\mathsf{R}_i\phi$ *means individual i has reason to believe proposition ϕ.*

Definition 2 (Indication). $\mathsf{Ind}^i_A\psi$ *means that given common knowledge A, individual i's reasons for A indicate ψ.*

In Lean, we represent these using explicit reason terms:

```
-- R: having a reason to believe
def R (rb : reason → indiv → Prop → Prop) (i : indiv) (  : Prop) :
 ↪ Prop :=
   r, rb r i

-- Indication: having reason to believe the implication
def Ind (rb : reason → indiv → Prop → Prop)
    (  : Prop) (i : indiv) (  : Prop) : Prop :=
  R rb i (  →  )
```

2.2 Justification Logic Framework

Our formalization uses a logic of reasons based on Artemov's justification logic. The key axiom is the *application rule*:

```
-- Application rule: if i has reason s for (  →  )
-- and reason t for  , then i has reason (s * t) for
axiom AR (rb : reason → indiv → Prop → Prop) :
   {s t : reason} {i : indiv} {   : Prop},
  rb s i (  →  ) → rb t i   → rb (s * t) i
```

This allows us to track *how* individuals reason, not just *that* they have beliefs.

2.3 The R-Closure

Lewis's main construction is the R-closure: the set of propositions reachable from ϕ by iteratively applying reasoning operators.

```
inductive RC (R : indiv → Prop → Prop) (  : Prop) : Prop → Prop
| base : RC R
| step {u : Prop} (j : indiv) (hu : RC R   u) :
    RC R   (R j u)
```

This captures nested reasoning like: ϕ, $\mathsf{R}_j\phi$, $\mathsf{R}_i(\mathsf{R}_j\phi)$, etc.

2.4 Lewis's Main Theorem

Theorem 3 (Lewis)**.** *Under conditions C1-C4, if p is in the R-closure of ϕ, then every individual reasons to p.*

Proof. The proof proceeds by induction on the R-closure structure:

```
lemma everyone_reason_of_rc
  (A1 :   {i : indiv} {p : Prop}, Ind A i p → R i A → R i p)
  (A6 :   {i j : indiv} {u : Prop},
    Ind A i (R j A)   R i (Ind A j u) → Ind A i (R j u))
  (C1 :   i : indiv, R i A)
  (C2 :   i j : indiv, Ind A i (R j A))
  (C3 :   i : indiv, Ind A i  )
  (C4 :   {i j : indiv} {u : Prop},
    Ind A i u → R i (Ind A j u))
  (hp : RC R   p) :   i : indiv, R i p := by
  intro i
  have hInd : Ind A i p :=
    ind_of_rc A6 C2 C3 C4 hp
  exact A1 hInd (C1 i)
```

The key insight is that indication propagates through the closure: if q is in the R-closure, then every individual indicates q (lemma `ind_of_rc`). □

2.5 Derived Results

We verify specific instances showing elements at various depths in the R-closure:

```
-- Depth 1: R j
lemma L1 [...] (i j : indiv) : R i (R j  )

-- Depth 2: R j (R k  )
lemma L2 [...] (i j k : indiv) : R i (R j (R k  ))

-- Depth 3: R j (R k (R    ))
lemma L3 [...] (i j k   : indiv) : R i (R j (R k (R    )))
```

3 Sillari's Modal Formalization

Sillari attempts to formalize Lewis using standard Kripke semantics for epistemic logic. While conceptually simpler, we show this approach is fundamentally flawed.

3.1 Kripke Semantics Approach

```
-- Multi-agent Kripke frame
structure MultiAgentFrame (Agent : Type) where
  World : Type
  rel : Agent → World → World → Prop

-- R: knowledge operator (box modality)
def R (i : Agent) (  : frame.World → Prop) :
    frame.World → Prop :=
  fun w =>   v, frame.rel i w v →   v

-- Ind: knowledge plus material implication
def Ind (i : Agent) (    : frame.World → Prop) :
    frame.World → Prop :=
  fun w => R i   w   (  w →   w)
```

3.2 Common Reason to Believe via Reachability

```
-- Reachability via transitive closure
inductive trcl (r : frame.World → frame.World → Prop) :
    frame.World → frame.World → Prop
| base {x y} : r x y → trcl r x y
| step {x y z} : r x y → trcl r y z → trcl r x z

-- CRB: common reason to believe
def CRB (  : frame.World → Prop) (s : frame.World) : Prop :=
    w, trcl connected s w →   w
```

4 Failure of Sillari's Formalization

4.1 Axiom B3 Fails (Lewis's A1)

Theorem 4 (B3 Counterexample)**.** *There exists a Kripke frame where* $\mathsf{R}_i\phi \wedge \mathsf{Ind}_i^{\phi}\psi$ *holds but* $\mathsf{R}_i\psi$ *fails.*

Proof. Consider a two-world frame with worlds s and t, where agent i relates s to t but not s to itself. Let $\phi(w) := (w \neq s)$ and $\psi(w) := (w \neq t)$.

```
lemma B3_fails
  (h1 : two_worlds s t)
  (h2a : ¬ frame.rel i s s)
  (h2b : frame.rel i s t) :
    ¬ (  w (   : frame.World → Prop),
       R i   w → Ind i     w → R i   w) := by
  let   := fun w => w   t
  let   := fun w => w   s
  push_neg
  -- R i   s holds: t is the only successor, and   t
  have h4 : R i   s := by rw [R]; aesop
  -- Ind i     s holds
  have h5 : Ind i     s := by rw [Ind]; aesop
  -- But R i   s fails: at t, ¬  t
  have h6 : ¬ R i   s := by
    intro hR
    exact (hR t h2b) rfl
  refine  s,  ,  , ?_, ?_, ?_
  · exact h4
  · exact h5
  · exact h6
```

At world s: $\mathsf{R}_i\phi$ holds (the only accessible world t satisfies ϕ), and $\mathsf{Ind}_i^{\phi}\psi$ holds (since $\phi \to \psi$ at s). But $\mathsf{R}_i\psi$ fails because at accessible world t, we have $\neg\psi(t)$. □

This is devastating for Sillari's approach since Lewis's original argument crucially depends on axiom A1.

4.2 Axiom C4 Also Fails

Cubitt and Sugden proposed axiom C4 as essential for Lewis's account:

Theorem 5 (C4 Counterexample). $\mathsf{Ind}_i^{\phi}\psi \not\Rightarrow \mathsf{R}_i(\mathsf{Ind}_j^{\phi}\psi)$

```
lemma C4_fails
  (h2a : ¬ frame.rel i s s)
  (h2b : frame.rel i s t) :
    ¬   w (    : frame.World → Prop),
      (Ind i     w → R i (Ind j    ) w) := by
  let   := fun _ : frame.World => True
  let   := fun w : frame.World => w = s
  push_neg
  have h3 : Ind i     s := by
    constructor
    { intro w _; aesop }
    { aesop }
  have h3a : ¬ R i (Ind j    ) s := by
    rw [R]
    push_neg
    use t
    constructor
    { exact h2b }
    { intro hn
      have hphi :   t := by aesop
      have hp :   t := hn.2 hphi
      have h3b : ¬   t := by aesop
      aesop }
  use s,  ,
```

4.3 Lewis's Theorem: Two Failed Interpretations

Sillari's formulation of Lewis's theorem is ambiguous. We examine both interpretations:

4.3.1 Local Interpretation (FALSE)

If assumptions C1-C3 hold only at world s:

Theorem 6 (Counterexample: One Agent)**.** *There exists a frame with one agent and three worlds where all local conditions hold but* CRB *fails.*

```
lemma Lewis_fails_1i
  (h3w : three_worlds s u v)
  (hrel : frame.rel = fun (_ : Agent) (w1 w2 : frame.World) =>
    (w1 = s   w2 = u)   (w1 = u   w2 = v)) :
    (   : frame.World → Prop),
    R i1   s
    Ind i1 (R i1  ) (R i1 (R i1  )) s
    Ind i1 (R i1  ) (R i1 (fun w => w = u)) s
    ¬ CRB (fun w => w = u) s
```

Frame structure: $s \to u \to v$ (linear chain). With $\phi := (\cdot \neq s)$ and $\psi := (\cdot = u)$, all local conditions hold at s, but CRB ψ s fails because v is reachable and $\psi(v)$ is false.

4.3.2 Global Interpretation (TRIVIAL)

If assumptions hold at all worlds, the proof becomes vacuous:

```
lemma lewis_s_2
  (C1 :   w, Rg   w)
  (C3 :   w, Ind i (Rg  ) (Rg  ) w) :
    CRB   s := by
  have hRg _all :   w, Rg   w :=
    fun w => (C3 w).2 (C1 w)
  intro v hv
  induction hv with
  | base h_edge =>
      rcases h_edge with  j, hj
      exact (hRg _all x) j y hj
  | step _ _ ih => exact ih
```

Note: Assumption C2 is completely unused! This suggests the global interpretation misses Lewis's intended logical structure.

5 Implementation Statistics

Our formalization consists of three main files:

- `Neeley.lean`: Lewis's original argument (justification logic) - [X] lines
- `Neeley_closure.lean`: R-closure and inductive proofs - [Y] lines
- `Sillari.lean`: Kripke semantics and counterexamples - [Z] lines

Total statistics:

- Lines of code: [Total]
- Definitions: [Count]
- Theorems/Lemmas: [Count]
- Axioms (justification logic): 4
- Counterexamples: 4

6 Discussion

6.1 Why Sillari's Approach Fails

The fundamental problem is that Kripke semantics conflates *having* a belief with *having a reason* for that belief. Lewis's original argument tracks explicit reasons and their composition, which cannot be captured by simple accessibility relations.

The failure of B3/A1 shows that the material implication $\phi \to \psi$ at a world does not suffice to transmit reasons from ϕ to ψ. Lewis needs something stronger: a *reason for the implication*.

6.2 Advantages of Justification Logic

Our alternative formalization using justification logic:

- Makes the reasoning structure explicit
- Avoids the failed axioms (A1 and A6 become provable)
- Matches Lewis's original informal presentation
- Provides constructive proofs (not just semantic validity)

6.3 Implications for Convention Theory

These results have philosophical significance:

- Standard epistemic logic may be inadequate for analyzing conventions
- The structure of common knowledge in coordination problems is more subtle than previously recognized
- Formalization reveals hidden assumptions in informal arguments

7 Conclusion

We have presented the first complete mechanized verification of Lewis's convention theory and demonstrated fundamental flaws in Sillari's modal logic formalization. Our results show that:

1. Lewis's original argument can be made rigorous using justification logic
2. Sillari's Kripke semantics approach fails on multiple axioms
3. Machine verification reveals problems invisible to informal analysis

Future work includes extending to full game-theoretic conventions, analyzing alternative modal approaches, and exploring other applications of justification logic to social epistemology.

7.1 Code Availability

Complete source code is available at: `https://github.com/hjvromen/lewis-lean`

References

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

Lewis, David. 1969. *Convention: a philosophical study.* Cambridge, MA: Harvard University Press.

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