Inferring Lewisian Common Knowledge Using Theory of Mind Reasoning in a Forward-Chaining Rule Engine

Extended Abstract

Stephen Cranefield
University of Otago
Dunedin, New Zealand
stephen.cranefield@otago.ac.nz

Sriashalya Srivathsan*
Eastern University, Sri Lanka
Vantharumoolai, Sri Lanka
sriashalyas@esn.ac.lk

Jeremy Pitt
Imperial College London
London, United Kingdom
j.pitt@imperial.ac.uk

ABSTRACT

This paper presents a practical technique for inferring common knowledge based on the approach of David Lewis, who identified three conditions that are sufficient for information about the world and other agents' reasoning mechanisms to lead to chains of iterated mutual knowledge. We consider agents with theory-of-mind rules that model other agents' beliefs. We prove that only two levels of nested models of other agents are necessary to achieve common knowledge. We illustrate this approach with an implemented scenario involving information on monuments in a public forum.

KEYWORDS

Common knowledge; David Lewis; Theory of mind

ACM Reference Format:
Stephen Cranefield, Sriashalya Srivathsan, and Jeremy Pitt. 2024. Inferring Lewisian Common Knowledge Using Theory of Mind Reasoning in a Forward-Chaining Rule Engine: Extended Abstract. In *Proc. of the 23rd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2024), Auckland, New Zealand, May 6 – 10, 2024*, IFAAMAS, 3 pages.

1 INTRODUCTION

Inferring common knowledge is a crucial component of practical reasoning to allow agents to coordinate efficiently with others. However, while common knowledge has been the focus of many logical theories, these generally focus on defining whether a proposition is said to be common knowledge (usually using infinitely nested knowledge operators or fixed points) rather than explaining how a proposition comes to be common knowledge. Consequently, there is an absence of both algorithms and software for practical reasoning with common knowledge.

In this paper, we extend philosopher David Lewis's informal analysis of common knowledge [7], as formalised by Cubitt and Sugden [5]. Lewis considered situations where an observed state of affairs A "indicates" that a certain proposition P holds. He proposed three conditions that are sufficient, given shared inductive standards and background knowledge relevant to A and P, for A to lead to common knowledge of P in the traditional logical sense. However, neither Lewis nor Cubitt and Sugden explained how these conditions and the existence of sufficient shared background knowledge and reasoning can be inferred to hold in concrete practical settings. We fill this gap by adapting and grounding the theory of C&S for situated agents with a reasoning mechanism based on a forward-chaining rule engine and theory-of-mind rules, which enable other agents' beliefs and rules to be modelled. Our agents maintain models of nested beliefs and received percepts, e.g., what the agent believes *any fool* (inspired by a theory of McCarthy [8]) will perceive and believe.

*The second author was at the University of Otago when this work was done.

Proc. of the 23rd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2024), N. Alechina, V. Dignum, M. Dastani, J.S. Sichman (eds.), May 6 – 10, 2024, Auckland, New Zealand. © 2024 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org).

2 THEORIES OF COMMON KNOWLEDGE

Common knowledge is traditionally expressed by the infinite conjunction $C\varphi = \varphi \wedge E\varphi \wedge E^2\varphi \wedge E^3\varphi \wedge \cdots$, where $K_i\,\varphi$ denotes "agent i knows φ", $E\varphi \equiv \wedge_i K_i\,\varphi$ denotes "everyone knows φ", and E^n abbreviates a nested sequence of n E operators [9]. It can be axiomatised using the Fixed-Point Axiom $C\varphi \leftrightarrow E(\varphi \wedge C\varphi)$ alongside the following induction rule: *From* $\varphi \rightarrow E(\psi \wedge \varphi)$ *infer* $\varphi \rightarrow C\psi$ [6].

Artemov [2] notes that "there is no conventional cut-elimination in [such] common-knowledge systems [1]", which "practically rules out automated proof search and severely limits the usage of formal methods in analyzing knowledge". Therefore, in this work we turn to the analysis of common knowledge by Lewis [7].

There are three key concepts in Lewis's theory. As Lewis only gives an informal presentation, our account is informed by the formalisation by Cubitt and Sugden (henceforth C&S) [5].

A state of affairs. According to C&S [5], Lewis considers states of affairs to be "alternative specifications of how the world, as seen by the modeller, really might be", but drop Lewis's distinction between a state of affairs A and the proposition 'A holds'. We take an agent perspective and consider states of affairs to be set of percepts that our agent of interest has actually perceived.

A reason to believe. Lewis's conception of common knowledge involves reasoning about what other agents have a *reason to believe* rather than what they actually believe, which cannot be known. C&S call this *warranted belief*. We assume that a software agent will always have some reason for creating a belief of its own or within its model of another agent's beliefs. Thus, all our agent's beliefs are warranted. We consider that reasoning about what agents will perceive and believe is a theory of mind (ToM) problem and encode knowledge about shared background facts and reasoning rules using explicit theory-of-mind rules.

An indication relation. Lewis defines *A indicates to someone i that a proposition x holds* to mean that if i has reason to believe that A holds, i would *thereby* have reason to believe x. C&S consider that Lewis intends this to be stronger than material implication and that the reason for believing that A holds must *provide i's reason* for

believing that x is true. Neither Lewis nor C&S provide semantics for this notion, but C&S provide six properties they believe any indication relation should satisfy. In contrast, we provide specific semantics for indication in terms of a proof tree for x created from the theory-of-mind rules.[1]

Lewis identified the following sufficient, but not necessary, conditions for common knowledge to arise from observations and common standards for inductive reasoning [7]:[2]

> *Let us say that it is common knowledge in a population $\mathcal{P}$ that P if and only if some state of affairs A holds such that:* **(L1)** *Everyone in $\mathcal{P}$ has reason to believe that A holds;* **(L2)** *A indicates to everyone in $\mathcal{P}$ that everyone in $\mathcal{P}$ has reason to believe that A holds;* **(L3)** *A indicates to everyone in $\mathcal{P}$ that P holds.*

Lewis argues informally that when these conditions hold for some $\mathcal{P}$, A and P, an infinite chain of nested beliefs about P can be shown to hold, in the style of traditional logical definitions of common knowledge. C&S formalised these three conditions, added a fourth condition that was implicit in Lewis's text as the existence of "suitable ancillary premises regarding our [shared] rationality, inductive standards, and background information" and provided an iterative proof pattern that the four conditions result in common knowledge. We provide an inductive proof and define an alternative to the fourth condition that allows us to prove that only two levels of nested knowledge about other agents are necessary for common knowledge to be inferred when using our ToM approach.

3 OUR THEORY OF MIND APPROACH

We assume an agent maintains a chain of nested models recording its own percepts and beliefs, its beliefs of the percepts and beliefs of other agents, and so on. Percepts and beliefs are denoted by $percept(M, \phi)$ and $bel(M, \phi)$, where M names the model: either the symbol $\odot$, representing the agent's own percepts and beliefs, or a term $M{\gg}Ag$, where M is a model name, $\gg$ is a left-associative binary operator and Ag is an agent name or the reserved constant af. Following the notion of "any fool knows" [8], we use af to represent *any fool*. For example, $bel(\odot{\gg}af{\gg}af, blue(sky))$ means the agent believes that any fool believes any (other) fool believes the sky is blue. The scope of *any fool* may be defined by a specified set of *af scope percepts*, e.g., af may represent only individuals who are perceived to be citizens located in a city's public square.

An agent has theory-of-mind rules that model shared beliefs and reasoning rules. These can create new beliefs (or in one special case, percepts) within the current model M or create percepts and/or beliefs in a nested model $M{\gg}Ag$.

We model states of affairs as sets of percepts. Given a set of propositional atoms $A = \{A_1, \ldots, A_n\}$, we write $percepts(M, A)$ to denote the set of percepts $\{percept(M, A_1), \ldots, percept(M, A_n)\}$ that are all in the same model M. In logical formulas we overload this notation and write $percepts(M, A)$ to mean $\bigwedge_{p \in percepts(M,A)} p$.

We write $percepts(M, A)$ ind ψ to denote indication, where ψ can be another set of $percepts(M', A')$, a single percept $percept(M', B)$, or a single belief $bel(M', B)$. The full paper [4] defines semantics for indication based on proof trees created from the ToM rules.

[1] For details, see the full paper [4].
[2] We have changed the variables and added the labels.

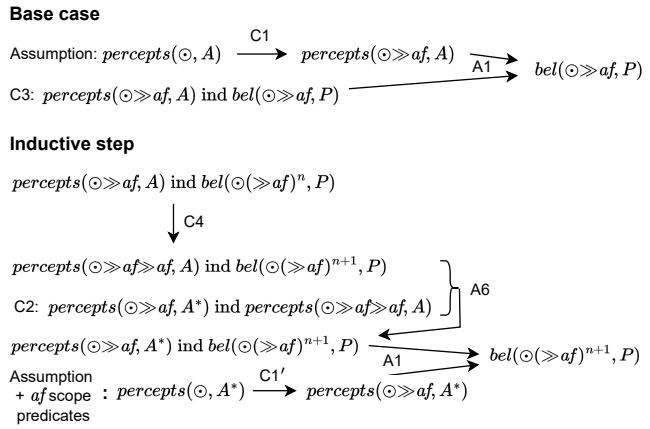

Figure 1: Our inductive proof that perceiving A leads to common knowledge of P when C1 to C4 hold.

We now present our adaptations of C&S's conditions for a state of affairs A to be a basis for common knowledge of a proposition P.

$$percepts(\odot, A) \rightarrow percepts(\odot{\gg}af, A) \quad (C1)$$

$$percepts(\odot, A^*) \rightarrow percepts(\odot{\gg}af, A^*) \quad (C1')$$

$$percepts(\odot{\gg}af, A^*) \text{ ind } percepts(\odot{\gg}af{\gg}af, A) \quad (C2)$$

$$percepts(\odot{\gg}af, A) \text{ ind } bel(\odot{\gg}af, P) \quad (C3)$$

$$\forall n{\geq}1\colon percepts(\odot{\gg}af, A) \text{ ind } bel(\odot(\gg af)^n, P) \rightarrow percepts(\odot{\gg}af{\gg}af, A) \text{ ind } bel(\odot(\gg af)^{n+1}, P) \quad (C4)$$

where A^* augments A with the set of af scope percepts and $\odot(\gg af)^n$ denotes the model $\odot{\gg}af{\gg}\cdots{\gg}af$ with n occurrences of af.

In the words of C&S, C1 and C2 hold if A is *self-revealing* and *public*. C3 holds if A indicates P to everyone. C4 is a special case of C&S's condition about shared knowledge and reasoning standards.

Figure 1 shows our inductive proof that our versions of C&S's properties of indication A1 and A6 (which our semantics satisfy) and the conditions above lead to common knowledge of P when A is perceived, because $bel(\odot(\gg af)^n, P)$ holds for all $n \geq 1$.

However, C4 cannot be verified with a finite set of nested models. The full paper [4] presents an alternative *isomorphism* test comparing $\odot{\gg}af$ and $\odot{\gg}af{\gg}af$, and proves that C4 follows from this test. Using this test together with C1 to C3, common knowledge can be inferred using only these two models.

4 IMPLEMENTATION AND SCENARIO

The full paper describes an implementation [3] that defines the conditions for common knowledge using SWI-Prolog [11] and the Pfc forward-chaining rule library [10]. It presents ToM rules for a scenario involving information on a public monument. We show that an agent that maintains only two levels of nested knowledge is able to evaluate the isomorphism test and conditions C1 to C3 to infer that the information on the monument is common knowledge.

ACKNOWLEDGMENTS

This work was supported by the Marsden Fund Council from New Zealand Government funding, managed by Royal Society Te Apārangi.

REFERENCES

[1] Luca Alberucci and Gerhard Jäger. 2005. About cut elimination for logics of common knowledge. *Annals of Pure and Applied Logic* 133, 1 (2005), 73–99. https://doi.org/10.1016/j.apal.2004.10.004

[2] Sergei Artemov. 2006. Justified common knowledge. *Theoretical Computer Science* 357, 1 (2006), 4–22. https://doi.org/10.1016/j.tcs.2006.03.009

[3] Stephen Cranefield, Sriashalya Srivathsan, and Jeremy Pitt. 2024. An implementation of Lewisian common knowledge through theory-of-mind rules (version 1.0). (2024). https://doi.org/10.5281/zenodo.10566378

[4] Stephen Cranefield, Sriashalya Srivathsan, and Jeremy Pitt. 2024. *A practical approach to recognising Lewisian common knowledge using theory-of-mind rules*. Technical Report. University of Otago. http://hdl.handle.net/10523/16567

[5] Robin P Cubitt and Robert Sugden. 2003. Common knowledge, Salience and convention: A reconstruction of David Lewis's game theory. *Economics & Philosophy* 19, 2 (2003), 175–210.

[6] Ronald Fagin, Joseph Y. Halpern, Yoram Moses, and Moshe Y. Vardi. 1995. *Reasoning about knowledge*. MIT Press.

[7] David Lewis. 1969. *Convention: A philosophical study*. Harvard University Press.

[8] John McCarthy, Masahiko Sato, Takeshi Hayashi, and Shigeru Igarashi. 1978. *On the Model Theory of Knowledge*. Technical Report STAN-CS-78-667. Stanford University, Stanford, CA, USA.

[9] J.-J. Ch. Meyer and W. van der Hoek. 1995. *Epistemic Logic for AI and Computer Science*. Cambridge University Press. https://doi.org/10.1017/CBO9780511569852

[10] Richard Fritzson Tim Finin and David Matuszek. 1989. Adding forward chaining and truth maintenance to Prolog. In *Proceedings of the Fifth IEEE Conference on Artificial Intelligence Applications*. IEEE Computer Society, 123–130. https://doi.org/10.1109/CAIA.1989.49145

[11] Jan Wielemaker, Tom Schrijvers, Markus Triska, and Torbjörn Lager. 2012. SWI-Prolog. *Theory and Practice of Logic Programming* 12, 1-2 (2012), 67–96.