Automating Public Announcement Logic and the Wise Men Puzzle in Isabelle/HOL

Christoph Benzmüller and Sebastian Reiche

March 17, 2025

Abstract

We present a shallow embedding of public announcement logic (PAL) with relativized general knowledge in HOL. We then use PAL to obtain an elegant encoding of the wise men puzzle, which we solve automatically using sledgehammer.

Contents

1 Public Announcement Logic (PAL) in HOL

An earlier encoding and automation of the wise men puzzle, utilizing a shallow embedding of higher-order (multi-)modal logic in HOL, has been presented in [1, 2]. However, this work did not convincingly address the interaction dynamics between the involved agents. Here we therefore extend and adapt the universal (meta-)logical reasoning approach of [1] for public announcement logic (PAL) and we demonstrate how it can be utilized to achieve a convincing encoding and automation of the wise men puzzle in HOL, so that also the interaction dynamics as given in the scenario is adequately addressed. For further background information on the work presented here we refer to [3, 4].

theory *PAL* **imports** *Main* **begin**
nitpick-params[*user-axioms*,*expect*=*genuine*]

Type i is associated with possible worlds

typedecl i
type-synonym $\sigma = i \Rightarrow bool$
type-synonym $\tau = \sigma \Rightarrow i \Rightarrow bool$
type-synonym $\alpha = i \Rightarrow i \Rightarrow bool$

type-synonym $\varrho = \alpha \Rightarrow bool$

Some useful relations (for constraining accessibility relations)

definition $reflexive{::}\alpha \Rightarrow bool$
where $reflexive\ R \equiv \forall x.\ R\ x\ x$
definition $symmetric{::}\alpha \Rightarrow bool$
where $symmetric\ R \equiv \forall x\ y.\ R\ x\ y \longrightarrow R\ y\ x$
definition $transitive{::}\alpha \Rightarrow bool$
where $transitive\ R \equiv \forall x\ y\ z.\ R\ x\ y \wedge R\ y\ z \longrightarrow R\ x\ z$
definition $euclidean{::}\alpha \Rightarrow bool$
where $euclidean\ R \equiv \forall x\ y\ z.\ R\ x\ y \wedge R\ x\ z \longrightarrow R\ y\ z$
definition $intersection\text{-}rel{::}\alpha \Rightarrow \alpha \Rightarrow \alpha$
where $intersection\text{-}rel\ R\ Q \equiv \lambda u\ v.\ R\ u\ v \wedge Q\ u\ v$
definition $union\text{-}rel{::}\alpha \Rightarrow \alpha \Rightarrow \alpha$
where $union\text{-}rel\ R\ Q \equiv \lambda u\ v.\ R\ u\ v \vee Q\ u\ v$
definition $sub\text{-}rel{::}\alpha \Rightarrow \alpha \Rightarrow bool$
where $sub\text{-}rel\ R\ Q \equiv \forall u\ v.\ R\ u\ v \longrightarrow Q\ u\ v$
definition $inverse\text{-}rel{::}\alpha \Rightarrow \alpha$
where $inverse\text{-}rel\ R \equiv \lambda u\ v.\ R\ v\ u$
definition $big\text{-}union\text{-}rel{::}\varrho \Rightarrow \alpha$
where $big\text{-}union\text{-}rel\ X \equiv \lambda u\ v.\ \exists R.\ (X\ R) \wedge (R\ u\ v)$
definition $big\text{-}intersection\text{-}rel{::}\varrho \Rightarrow \alpha$
where $big\text{-}intersection\text{-}rel\ X \equiv \lambda u\ v.\ \forall R.\ (X\ R) \longrightarrow (R\ u\ v)$

In HOL the transitive closure of a relation can be defined in a single line.

definition $tc{::}\alpha \Rightarrow \alpha$
where $tc\ R \equiv \lambda x\ y. \forall Q.\ transitive\ Q \longrightarrow (sub\text{-}rel\ R\ Q \longrightarrow Q\ x\ y)$

Logical connectives for PAL

abbreviation $patom{::}\sigma \Rightarrow \tau$ $(\langle {}^{A}\text{-}\rangle [79]80)$
where ${}^{A}p \equiv \lambda W\ w.\ W\ w \wedge p\ w$
abbreviation $ptop{::}\tau$ $(\langle \top \rangle)$
where $\top \equiv \lambda W\ w.\ True$
abbreviation $pneg{::}\tau \Rightarrow \tau$ $(\langle \neg\text{-}\rangle [52]53)$
where $\neg\varphi \equiv \lambda W\ w.\ \neg(\varphi\ W\ w)$
abbreviation $pand{::}\tau \Rightarrow \tau \Rightarrow \tau$ (**infixr**$\langle \wedge \rangle 51$)
where $\varphi \wedge \psi \equiv \lambda W\ w.\ (\varphi\ W\ w) \wedge (\psi\ W\ w)$
abbreviation $por{::}\tau \Rightarrow \tau \Rightarrow \tau$ (**infixr**$\langle \vee \rangle 50$)
where $\varphi \vee \psi \equiv \lambda W\ w.\ (\varphi\ W\ w) \vee (\psi\ W\ w)$
abbreviation $pimp{::}\tau \Rightarrow \tau \Rightarrow \tau$ (**infixr**$\langle \rightarrow \rangle 49$)
where $\varphi \rightarrow \psi \equiv \lambda W\ w.\ (\varphi\ W\ w) \longrightarrow (\psi\ W\ w)$
abbreviation $pequ{::}\tau \Rightarrow \tau \Rightarrow \tau$ (**infixr**$\langle \leftrightarrow \rangle 48$)
where $\varphi \leftrightarrow \psi \equiv \lambda W\ w.\ (\varphi\ W\ w) \longleftrightarrow (\psi\ W\ w)$
abbreviation $pknow{::}\alpha \Rightarrow \tau \Rightarrow \tau$ $(\langle \mathbf{K}\text{-}\ \text{-}\rangle)$
where $\mathbf{K}\ r\ \varphi \equiv \lambda W\ w. \forall v.\ (W\ v \wedge r\ w\ v) \longrightarrow (\varphi\ W\ v)$
abbreviation $ppal{::}\tau \Rightarrow \tau \Rightarrow \tau$ $(\langle [\mathbf{!}\text{-}]\text{-}\rangle)$
where $[\mathbf{!}\varphi]\psi \equiv \lambda W\ w.\ (\varphi\ W\ w) \longrightarrow (\psi\ (\lambda z.\ W\ z \wedge \varphi\ W\ z)\ w)$

Glogal validity of PAL formulas

abbreviation $pvalid{::}\tau \Rightarrow bool$ $(\langle \lfloor\text{-}\rfloor \rangle [7]8)$

where $\lfloor\varphi\rfloor \equiv \forall\, W.\forall\, w.\ W\ w \longrightarrow \varphi\ W\ w$

Introducing agent knowledge (K), mutual knowledge (E), distributed knowledge (D) and common knowledge (C).

abbreviation *EVR*::$\varrho\Rightarrow\alpha$
  where $EVR\ G \equiv$ *big-union-rel* G
abbreviation *DIS*::$\varrho\Rightarrow\alpha$
  where $DIS\ G \equiv$ *big-intersection-rel* G
abbreviation *agttknows*::$\alpha\Rightarrow\tau\Rightarrow\tau$ (‹$\mathbf{K}_{\_}$ -›)
  where $\mathbf{K}_r\ \varphi \equiv\ \mathbf{K}\ r\ \varphi$
abbreviation *evrknows*::$\varrho\Rightarrow\tau\Rightarrow\tau$ (‹$\mathbf{E}_{\_}$ -›)
  where $\mathbf{E}_G\ \varphi \equiv\ \mathbf{K}\ (EVR\ G)\ \varphi$
abbreviation *disknows* :: $\varrho\Rightarrow\tau\Rightarrow\tau$ (‹$\mathbf{D}_{\_}$ -›)
  where $\mathbf{D}_G\ \varphi \equiv \mathbf{K}\ (DIS\ G)\ \varphi$
abbreviation *prck*::$\varrho\Rightarrow\tau\Rightarrow\tau\Rightarrow\tau$ (‹$\mathbf{C}_{\_}$⦇-|-⦈›)
  where $\mathbf{C}_G$⦇$\varphi|\psi$⦈ $\equiv \lambda W\ w.\ \forall\, v.\ (tc\ ($*intersection-rel* $(EVR\ G)\ (\lambda u\ v.\ W\ v \wedge \varphi\ W\ v))\ w\ v) \longrightarrow (\psi\ W\ v)$
abbreviation *pcmn*::$\varrho\Rightarrow\tau\Rightarrow\tau$ (‹$\mathbf{C}_{\_}$ -›)
  where $\mathbf{C}_G\ \varphi \equiv\ \mathbf{C}_G$⦇$\top|\varphi$⦈

Postulating S5 principles for the agent's accessibility relations.

abbreviation *S5Agent*::$\alpha\Rightarrow$*bool*
  where *S5Agent* $i \equiv$ *reflexive* $i \wedge$ *transitive* $i \wedge$ *euclidean* i
abbreviation *S5Agents*::$\varrho\Rightarrow$*bool*
  where *S5Agents* $A \equiv \forall\, i.\ (A\ i \longrightarrow$ *S5Agent* $i)$

Introducing "Defs" as the set of the above definitions; useful for convenient unfolding.

named-theorems *Defs*
declare *reflexive-def*[*Defs*] *symmetric-def*[*Defs*] *transitive-def*[*Defs*]
  euclidean-def[*Defs*] *intersection-rel-def*[*Defs*] *union-rel-def*[*Defs*]
  sub-rel-def[*Defs*] *inverse-rel-def*[*Defs*] *big-union-rel-def*[*Defs*]
  big-intersection-rel-def[*Defs*] *tc-def*[*Defs*]

Consistency: nitpick reports a model.

lemma *True* **nitpick** [*satisfy*] **oops**

2 Automating the Wise Men Puzzle

Agents are modeled as accessibility relations.

consts a::α b::α c::α
abbreviation *Agent*::$\alpha\Rightarrow$*bool* (‹$\mathcal{A}$›) **where** $\mathcal{A}\ x \equiv x = a \vee x = b \vee x = c$
axiomatization where *group-S5*: *S5Agents* $\mathcal{A}$

Common knowledge: At least one of a, b and c has a white spot.

consts *ws*::$\alpha\Rightarrow\sigma$
axiomatization where *WM1*: $\lfloor\mathbf{C}_{\mathcal{A}}\ ({}^{A}ws\ a \vee {}^{A}ws\ b \vee {}^{A}ws\ c)\rfloor$

Common knowledge: If x does not have a white spot then y knows this.

axiomatization where
WM2ab: $\lfloor \mathbf{C}_{\mathcal{A}}\ (\neg({}^{A}ws\ a) \rightarrow (\mathbf{K}_b\ (\neg({}^{A}ws\ a))))\rfloor$ **and**
WM2ac: $\lfloor \mathbf{C}_{\mathcal{A}}\ (\neg({}^{A}ws\ a) \rightarrow (\mathbf{K}_c\ (\neg({}^{A}ws\ a))))\rfloor$ **and**
WM2ba: $\lfloor \mathbf{C}_{\mathcal{A}}\ (\neg({}^{A}ws\ b) \rightarrow (\mathbf{K}_a\ (\neg({}^{A}ws\ b))))\rfloor$ **and**
WM2bc: $\lfloor \mathbf{C}_{\mathcal{A}}\ (\neg({}^{A}ws\ b) \rightarrow (\mathbf{K}_c\ (\neg({}^{A}ws\ b))))\rfloor$ **and**
WM2ca: $\lfloor \mathbf{C}_{\mathcal{A}}\ (\neg({}^{A}ws\ c) \rightarrow (\mathbf{K}_a\ (\neg({}^{A}ws\ c))))\rfloor$ **and**
WM2cb: $\lfloor \mathbf{C}_{\mathcal{A}}\ (\neg({}^{A}ws\ c) \rightarrow (\mathbf{K}_b\ (\neg({}^{A}ws\ c))))\rfloor$

Positive introspection principles are implied.

lemma *WM2ab'*: $\lfloor \mathbf{C}_{\mathcal{A}}\ (({}^{A}ws\ a) \rightarrow \mathbf{K}_b\ ({}^{A}ws\ a))\rfloor$
using *WM2ab group-S5* **unfolding** *Defs* **by** *metis*
lemma *WM2ac'*: $\lfloor \mathbf{C}_{\mathcal{A}}\ (({}^{A}ws\ a) \rightarrow \mathbf{K}_c\ ({}^{A}ws\ a))\rfloor$
using *WM2ac group-S5* **unfolding** *Defs* **by** *metis*
lemma *WM2ba'*: $\lfloor \mathbf{C}_{\mathcal{A}}\ (({}^{A}ws\ b) \rightarrow \mathbf{K}_a\ ({}^{A}ws\ b))\rfloor$
using *WM2ba group-S5* **unfolding** *Defs* **by** *metis*
lemma *WM2bc'*: $\lfloor \mathbf{C}_{\mathcal{A}}\ (({}^{A}ws\ b) \rightarrow \mathbf{K}_c\ ({}^{A}ws\ b))\rfloor$
using *WM2bc group-S5* **unfolding** *Defs* **by** *metis*
lemma *WM2ca'*: $\lfloor \mathbf{C}_{\mathcal{A}}\ (({}^{A}ws\ c) \rightarrow \mathbf{K}_a\ ({}^{A}ws\ c))\rfloor$
using *WM2ca group-S5* **unfolding** *Defs* **by** *metis*
lemma *WM2cb'*: $\lfloor \mathbf{C}_{\mathcal{A}}\ (({}^{A}ws\ c) \rightarrow \mathbf{K}_b\ ({}^{A}ws\ c))\rfloor$
using *WM2cb group-S5* **unfolding** *Defs* **by** *metis*

Automated solutions of the Wise Men Puzzle.

theorem *whitespot-c*: $\lfloor [\mathbf{!}\neg\mathbf{K}_a({}^{A}ws\ a)]([\mathbf{!}\neg\mathbf{K}_b({}^{A}ws\ b)](\mathbf{K}_c\ ({}^{A}ws\ c)))\rfloor$
using *WM1 WM2ba WM2ca WM2cb* **unfolding** *Defs* **by** (*smt* (*verit*))

For the following, alternative formulation a proof is found by sledgehammer, while the reconstruction of this proof using trusted methods (often) fails; this hints at further opportunities to improve the reasoning tools in Isabelle/HOL.

theorem *whitespot-c'*:
$\lfloor [\mathbf{!}\neg((\mathbf{K}_a\ ({}^{A}ws\ a)) \vee (\mathbf{K}_a\ (\neg{}^{A}ws\ a)))]([\mathbf{!}\neg((\mathbf{K}_b\ ({}^{A}ws\ b)) \vee (\mathbf{K}_b\ (\neg{}^{A}ws\ b)))](\mathbf{K}_c\ ({}^{A}ws\ c)))\rfloor$
using *WM1 WM2ab WM2ac WM2ba WM2bc WM2ca WM2cb* **unfolding** *Defs*
— sledgehammer by (smt (verit))
oops

Consistency: nitpick reports a model.

lemma *True* **nitpick** [*satisfy*] **oops**
end

References

[1] C. Benzmüller. Universal (meta-)logical reasoning: Recent successes. *Science of Computer Programming*, 172:48–62, 2019.

[2] C. Benzmüller. Universal (meta-)logical reasoning: The wise men puzzle (Isabelle/HOL Dataset). *Data in Brief*, 24(103823):1–5, 2019.

[3] C. Benzmüller and S. Reiche. Modeling and automating public announcement logic with relativized common knowledge as a fragment of HOL in LogiKEy. Technical Report arXiv:2111.01654, CoRR, 2021.

[4] S. Reiche and C. Benzmüller. Public announcement logic in HOL. In M. A. Martins and S. Igor, editors, *Dynamic Logic. New Trends and Applications. DaLi 2020*, volume 12569 of *Lecture Notes in Computer Science*. Springer, Cham, 2020.