A Formal Analysis of Hollis' Paradox

Thomas Ågotnes[1,2] and Chiaki Sakama[3]

[1] University of Bergen, Norway thomas.agotnes@uib.no
[2] Shanxi University, China
[3] Wakayama University, Japan sakama@wakayama-u.ac.jp

Abstract. In Hollis' paradox, A and B each chose a positive integer and whisper their number to C. C then informs them, jointly, that they have chosen different numbers and, moreover, that neither of them are able to work out who has the greatest number. A then reasons as follows: B cannot have 1, otherwise he would know that my number is greater, and by the same reasoning B knows that I don't have 1. But then B also cannot have 2, otherwise he would know that my number is greater (since he knows I don't have 1). This line of reasoning can be repeated indefinitely, effectively forming an inductive proof, ruling out any number – an apparent paradox. In this paper we formalise Hollis' paradox using public announcement logic, and argue that the root cause of the paradox is the wrongful assumption that A and B assumes that C's announcement necessarily is *successful*. This resolves the paradox without assuming that C can be untruthful, or that A and B are not perfect reasoners, like other solutions do. There are similarities to the surprise examination paradox. In addition to a semantic analysis in the tradition of epistemic logic, we provide a syntactic one, deriving conclusions from a set of premises describing the initial situation – more in the spirit of the literature on Hollis' paradox. The latter allows us to pinpoint which assumptions are actually necessary for the conclusions resolving the paradox.

Keywords: Epistemic Logic · Hollis' Paradox · Public Announcement Logic

1 Introduction

In *A paradoxical train of thought* [9], Martin Hollis describes the following situation.

> *A thinks of a number and whispers it privately to C. B does the same. C tells them, 'You have each thought of a different positive whole number. Neither of you can work out whose is the greater'. . . . Sitting alone in his homebound train, A muses as follows. 'I picked 157 and have no idea what B picked. So, assuming that he indeed chose a different positive whole number, C is right. . . . Well, clearly B did not choose 1, as he would then be able to work out that mine is greater; and by the same token he knows that I did not choose 1. So he did not choose 2, since he could then use the previous reasoning to prove that my number is greater. Similarly, he can know that I did not choose 2 either. With 2 out of the way, I infer that he did not choose 3; and he can infer that I did not choose 3. . . . I can keep this up for ever. But that is absurd. It means that I cannot have picked 157, which I certainly did.*

Several solutions attempting to resolve the apparent paradox have been proposed [14, 11, 17] (see also Hollis' response to the two first in [10]). What they have in common is that they argue that the announcement by C might not be truthful, and even if it were A and B might not have justified belief in that. Like most well known epistemic puzzles, Hollis' paradox leaves many assumptions implicit or ambiguous, so let us in this paper assume the following: (a) all agents always tell the truth (if they say something it is true and they know that it is true) and (b) this is common knowledge among all agents. Thus, we will be modelling knowledge rather than belief, and at any point in time an agent's knowledge is a result of the information she has received. We also assume that it is common knowledge that everyone is a perfect reasoner[4].

As far as we are aware, no *formal* analysis of Hollis' paradox appears in the literature, unlike most other well known epistemic or doxastic puzzles or paradoxes which have been studied using dynamic epistemic logic – see [19, 18] for an overview and references. Indeed, the precision and clarity of formal logic has been crucial in understanding these puzzles and clarify hidden premises (and these puzzles have again been a driving force as case studies in the development of dynamic epistemic logic).

In this paper we use public announcement logic [15] to model and analyse Hollis' paradox. This allows us to untangle subtleties in the alleged paradox, and in particular to be precise about the distinction between truth *before* an announcement and *after*, a distinction often lost in other analyses of the paradox. We argue that the root cause of the paradoxical situation is a wrongful assumption that the announcements by C always are *successful*, i.e., that they always remain true after they are announced. In Hollis' argument, this assumption is used as a premise in the inductive "proof". This has, as far as we know, not been pointed out in other studies of the paradox, and we believe this is the first solution to the paradox that does not rely on weakening the assumptions outlined above. However, it should come as no surprise. As pointed out already in [14], Hollis' paradox is similar[5] to *the surprise examination paradox*[6] which was first analysed using dynamic epistemic logic by Gerbrandy [7, 8]. Gerbrandy pointed out that the root cause of that paradox is the same phenomenon that lies behind many other epistemic puzzles with counter-intuitive solutions, the *muddy children (or three wise men) problem* [5] being the most well known, namely that announcements can become false as a result of being announced [7] – they are not necessarily successful. Olin [14] also points out that there are still "important differences" between the two paradoxes. We discuss the connection further in the last section of the paper.

In addition to arguing why, under the assumptions outlined above, Hollis' paradox is actually not a paradox, we shed light on other epistemic aspects of the puzzle, such as whether common knowledge must be assumed (it must not) or how many layers of nested knowledge are relevant (two). We provide two alternative and complementary analyses: a semantic analysis (in the style of Gerbrandy) where we give a single

[4] Hollis [9] already hints at this assumption: "...each of us has to assume that the other is not stupid...".

[5] Olin [14] claims that it is "a version of surprise examination"; Hollis [10] on the other hand argues that his paradox is "wider".

[6] See [12] for an overview of different variants and a discussion of historic origins.

[7] In muddy children, that happens in the last joint announcement by the children.

model of the initial situation described in the story and show that it has certain logical properties (Section 3), as well as a syntactic analysis (more in the style of Hollis and his respondents in *Analysis*, but formalised) where we describe the situation using a set of logical formulas and show that the same properties can be derived (Section 4). First, we give a brief technical introduction to epistemic logic and the logic of public announcements (see [19] for more details).

2 Background

2.1 Epistemic Logic

The most popular *epistemic logic* (i.e., logic for reasoning about knowledge) is modal propositional epistemic logic [5]. It extends propositional logic over a set of primitive propositions P with modalities K_a, where a is one of the agents in a given finite set Ag of agents. Intuitively, $K_a\phi$ means that agent a knows ϕ. Formally, the language is defined by the following grammar:

$$\varphi ::= p \mid \neg\varphi \mid \varphi_1 \wedge \varphi_2 \mid K_a\varphi$$

where $p \in P$ and $a \in Ag$. It is interpreted in *(epistemic) models* $M = (S, \sim, V)$ where S is a non-empty set of *states* (or worlds); $\sim$ gives an equivalence relation $\sim_a$ on S for each $a \in Ag$, a's *accessibility relation*; and $V : P \to \wp(S)$ is a *valuation function*, saying which primitive propositions are true in which states. Intuitively, $s \sim_a t$ models that agent a cannot discern between the states s and t; if the state of the world is s she considers it possible that it is actually t, and vice versa.

We write $M, s \models \phi$ to denote the fact that formula ϕ is true in state s of model M, defined recursively as follows:

$$\begin{array}{ll} M, s \models p \quad \Leftrightarrow s \in V(p) & M, s \models K_a\varphi \;\; \Leftrightarrow (\forall t \in S)(s \sim_a t \Rightarrow M, t \models \varphi) \\ M, s \models \neg\varphi \Leftrightarrow M, s \not\models \varphi & M, s \models \varphi \wedge \psi \Leftrightarrow M, s \models \varphi \;\&\; M, s \models \psi \end{array}$$

Thus, $K_a\varphi$ is true if and only if φ is true in all indiscernible (for a) states. We use the usual derived propositional connectives, in addition to $\hat{K}_a\phi$ for $\neg K_a\neg\phi$, intuitively meaning that agent a considers that ϕ *possible*, i.e., that φ is true in at least one indiscernible state.

2.2 Public Announcement Logic

Public announcement logic (PAL) [15] extends epistemic logic in order to be able to reason about *change* in agents' knowledge and ignorance, resulting from a specific type of events: public announcements (such as the ones made by C in Hollis' paradox). Syntactically PAL extends epistemic logic with modalities of the form $[\phi]$ where ϕ is a formula. A formula $[\phi]\psi$ intuitively means that *after ϕ is truthfully*[8] *and publicly announced, ψ becomes true*. Formally, the language is defined by the following grammar:

$$\varphi ::= p \mid \neg\varphi \mid \varphi_1 \wedge \varphi_2 \mid K_a\varphi \mid [\varphi_1]\varphi_2$$

[8] Here and in the following we mean "truthful" in the strong sense that the announcement is in fact true (rather than only believed to be true).

where $p \in P$ and $a \in Ag$. This language is also interpreted in epistemic models, extending the interpretation of the epistemic language with a clause for the public announcement operators. Informally, $[\phi]\psi$ is true in a state s in a model M ($M, s \models [\phi]\psi$) if ψ is true in state s ($M', s \models \psi$) in the model (call it M') resulting from *removing all states t in M where ϕ is false* ($M, t \models \neg\phi$). This captures the epistemic effects of a public announcement of ϕ: after the announcement, no-one considers it possible that $\neg\phi$ was[9] true, no-one considers it possible that anyone considered $\neg\phi$ possible, and so on (it becomes *common knowledge* that ϕ was true, capturing the word "publicly" above). Formally:

$$M, s \models [\varphi]\psi \quad \iff \quad (M, s \models \varphi \Rightarrow M|\varphi, s \models \psi)$$

where $M|\varphi = (S', \sim', V')$ is a model such that for any $a \in Ag$ and $p \in P$, $S' = \{t \in S \mid M, t \models \varphi\}$, $\sim'_a = \sim_a \cap (S' \times S')$, and $V'(p) = V(p) \cap S'$.

The precondition $M, s \models \phi$ in the interpretation of $[\phi]\psi$ is needed because without it the definition would not be well-defined: if ϕ is false in s then s itself would be removed in the model update. This captures the "truthful" in the informal reading "*after ϕ is truthfully and publicly announced, ψ becomes true*" - or, alternatively, "*if ϕ is true then ψ will become true after ϕ is publicly announced*". The *dual*, $\langle\phi\rangle\psi = \neg[\phi]\neg\psi$, means that ϕ *is true and ψ will become true after ϕ is publicly announced.*

We write $M \models \phi$ to denote the fact that ϕ is true in all states in model M. A formula ϕ is *valid* if $M \models \phi$ for all models M. When Γ is a set of formulas, $\Gamma \models \phi$ means that for all M, s, if $M, s \models \Gamma$ then $M, s \models \phi$ (ϕ is logically entailed by Γ).

We say that a formula ϕ is an *(un)successful update* in M, s iff $M, s \models \langle\phi\rangle\phi$ ($M, s \models \langle\phi\rangle\neg\phi$); ϕ is a *successful formula* iff $[\phi]\phi$ is valid and an *unsuccessful formula* if not.

2.3 Axioms

Axiomatisations of epistemic logic and Public Announcement Logic are shown in Table 1. These axiomatisations are sound and complete [15, 19], in the sense that any formula is valid if and only if it can be derived using these axioms and rules.

We write $\vdash \phi$ to denote that formula ϕ is derivable (is a theorem), i.e., that there is a finite sequence of formulas ending with ϕ where every formula is either an instance of an axiom schema or the result of applying an inference rule to formulas earlier in the sequence. When Γ is a set of formulas, $\Gamma \vdash \phi$ ("ϕ can be derived from Γ") means that there is a finite subset $\{\gamma_1, \ldots, \gamma_k\}$ of Γ such that $\vdash \bigwedge_{1 \leq i \leq k} \gamma_i \to \phi$.

3 A Semantic Analysis

Hollis' paradox is well suited to a semantic (model theoretic) analysis, because the story intuitively and implicitly completely describes a single epistemic model. Figure 1 shows

[9] If ϕ is, e.g., a primitive proposition, then "was true" is the same as "is true". However, this is not the case in general: it could be that ϕ was true in a certain state before the announcement, but became false in the same state *as a result of the announcement*. The canonical example of the latter is the so-called Moore sentence $\phi = p \wedge \neg K_a p$.

Propositional tautology instances	Prop		
$K_a(\phi \to \psi) \to (K_a\phi \to K_a\psi)$	KD	$[\phi]p \leftrightarrow (\phi \to p)$	APerm
$K_a\phi \to \phi$	T	$[\phi]\neg\psi \leftrightarrow (\phi \to \neg[\phi]\psi)$	ANeg
$K_a\phi \to K_aK_a\phi$	4	$[\phi](\psi \wedge \chi) \leftrightarrow ([\phi]\psi \wedge [\phi]\chi)$	AConj
$\neg K_a\phi \to K_a\neg K_a\phi$	5	$[\phi]K_a\psi \leftrightarrow (\phi \to K_a[\phi]\psi)$	AKnow
From ϕ and $\phi \to \psi$, infer ψ	MP	$[\phi][\psi]\chi \leftrightarrow [\phi \wedge [\phi]\psi]\chi$	AComp
From ϕ, infer $K_a\phi$	Nec		

Table 1. Axiomatisation of epistemic logic (left) and PAL (left and right).

the epistemic model of the agents' knowledge after they have chosen their numbers but *before* C makes any announcement. A state corresponds to each agent having selected a number, we will refer to the combination as a *selection*. We only model A and B (as agents a and b respectively); C's knowledge is not relevant for the paradox beyond the assumption that his two announcements are actually true when they are made. Let 7_a be an atomic proposition meaning that agent a has chosen the number 7, and similarly for other numbers and for agent b. Also, let p_a mean that agent a's number is strictly greater than agent b's, and p_b that agent b's number is strictly greater than agent a's. We can now formalise the two announcements "you have each thought of a different number" and "neither of you can work out whose is the greater", respectively as:

$$ann1 = p_a \vee p_b \qquad ann2 = \neg K_a p_a \wedge \neg K_a p_b \wedge \neg K_b p_a \wedge \neg K_b p_b$$

While $ann1$ is straightforward, the formalisation $ann2$ of the second announcement deserves comment. In this formalisation we interpret "work out" as "deduce". "Work out" doesn't seem to imply, e.g., waiting for further information, asking questions, or guessing. Indeed, this is a common interpretation: informal descriptions of Hollis' paradox that have appeared after the original statement [9] explicitly use "deduce" instead of "work out"; e.g., [16] ("Neither of you can deduce which number is greatest"). It is worth noting that this formalisation is similar to Gerbrandy's formalisation of the the announcement in the surprise exam paradox [8], and that it has been argued [2] that the latter does not capture the intended meaning and that a stronger *self-referential* proposition is needed. In Section 5 we discuss why the same argument does not apply to our case. Also note that this formalisation is made in the context of the assumptions made in the introduction (common knowledge of truthfulness, perfect reasoners). A formula $K_a\phi$ holds iff ϕ follows from the information agent a currently has, and can thus be deduced by a perfect reasoner. $\neg K_a\phi$ holds if a cannot deduce ϕ (work out that ϕ holds).

In this initial model, agents a and b each only know their own number and consider any possibility for the other agent's number. They have no additional information (yet). For example, we have that $M_1, (2,3) \models \neg K_a p_b$: if A has selected 2 and B has selected 3, then A does not know that B's number is highest. In fact, in *all* states, i.e., no matter what the selection is, it holds that none of the agents know which number is greatest: $M_1 \models ann2$. However, note that if the selection, e.g., is $(1,1)$, A knows that her number cannot be strictly greater than B's: $M_1, (1,1) \models K_a\neg p_a$.

Let us now consider the situation immediately after C makes the announcement $ann1$. This announcement is *informative* for A and B; they learn something from it.

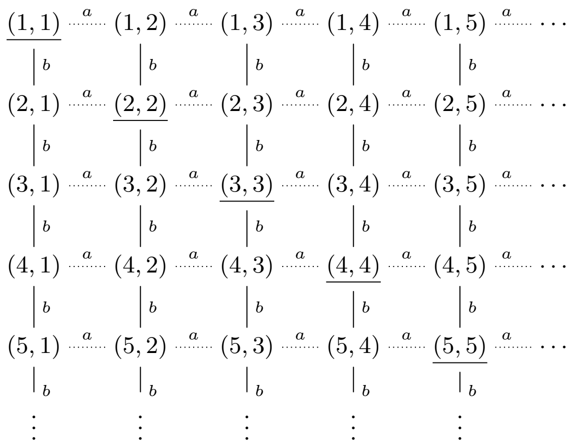
(1, 1) a (1, 2) a (1, 3) a (1, 4) a (1, 5) a ···
b b b b b
(2, 1) a (2, 2) a (2, 3) a (2, 4) a (2, 5) a ···
b b b b b
(3, 1) a (3, 2) a (3, 3) a (3, 4) a (3, 5) a ···
b b b b b
(4, 1) a (4, 2) a (4, 3) a (4, 4) a (4, 5) a ···
b b b b b
(5, 1) a (5, 2) a (5, 3) a (5, 4) a (5, 5) a ···
b b b b b
⋮ ⋮ ⋮ ⋮ ⋮

Fig. 1. Initial model M_1. In state $(2, 3)$ agent a has chosen the number 2 and agent b has chosen the number 3, and so on for the other states. The accessibility relation for agent a is depicted using dotted lines. Reflexive loops and transitive "jumps" are not shown; the actual accessibility relation is the reflexive, transitive closure of the relation in the picture. More intuitively: agent a cannot discern between states on the same row. Similarly for agent b, solid lines, and the same column. Atom $p_a$is true in all states to the left of the underlined diagonal; p_b in all states to the right of the diagonal. States where $ann1$ is *false* are underlined. $ann2$ is true in all states.

Thus we have to update the model M_1 with the new information $ann1$ which is jointly received by a and b. We do that by removing the states in model M_1 where $ann1$ is false. The resulting model, M_2, is shown in Figure 2.

As mentioned, the agents' knowledge has now changed, and in particular we have that $M_2 \models 1_a \rightarrow K_a p_b$. Similarly, $M_2 \models 1_b \rightarrow K_b p_a$. In words: if A has chosen the number 1, she *now* knows that she has a strictly lower number than B. Written another way: $M_1 \models 1_a \rightarrow [ann1]K_a p_b$.

As a consequence, we now (after the first announcement) have that the statement $ann2$ is not true in, e.g., state $(1, 3)$: $M_2, (1, 3) \models \neg ann2$.

Consider now the announcement of $ann2$ by C. The consequence of this announcement is that no one no longer considers states where $ann2$ was false (at the moment the announcement was made) possible (i.e., the bold states in the figure), and we update model M_2 by removing those states. The resulting model, M_3, is also illustrated in Figure 2. Observe that we now have that, e.g., $M_3, (2, y) \models K_a p_b$ for all $y > 2$, and $M_3, (x, 2) \models K_b p_a$ for all $x > 2$. In other words, $M_2 \models 2_a \rightarrow [ann2]K_a p_b$, or: $M_1 \models 2_a \rightarrow [ann1][ann2]K_a p_b$ – no matter what the selection is, if A's number is 2 then she will know that B's number is highest after both announcements.

Let us consider the claims in the statement of the paradox. "clearly B did not choose 1": this is true; $M_1 \models [ann1][ann2]K_a \neg 1_b$."... and by the same token he knows that I did not choose 1": also true; $M_1 \models [ann1][ann2]K_b \neg 1_a$. "So he did not choose 2, since he could then use the previous reasoning to prove that my number is greater": no, this is in fact not true. In fact, no matter what the selection is, each of the two agents

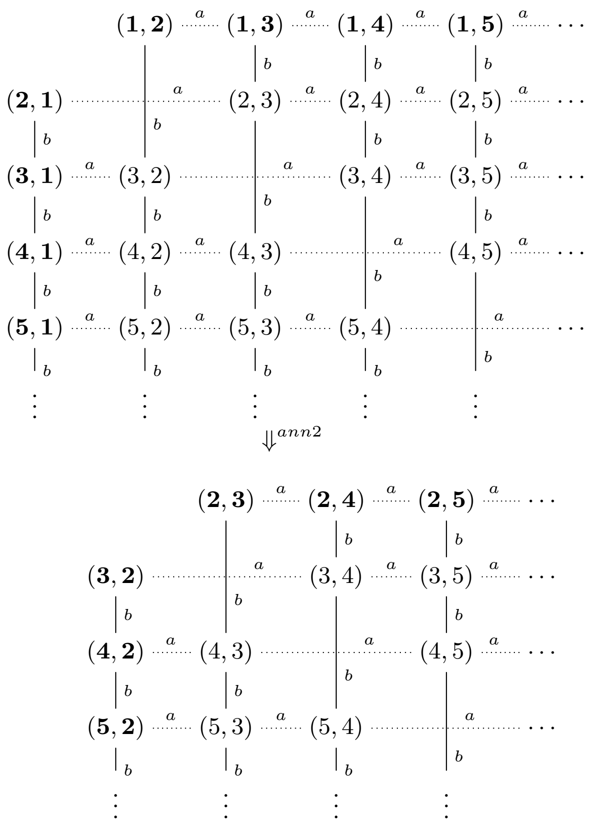

Fig. 2. M_2 (top), the result of announcing $ann1$ in M_1. M_3 (bottom), the result of announcing $ann2$ in M_2. States where $ann2$ is *false* are in **bold**. $ann1$ is true in all states.

considers it possible that the other agent has 2 unless she has it herself:

$$M_1 \models \neg 2_a \rightarrow [ann1][ann2]\neg K_a \neg 2_b$$

and similarly with a and b swapped. This shows that the inductive argument in the "proof" of the paradox halts. The reason that the argument in the "proof" of the paradox doesn't work is that while the announcement $ann2$ might have been successful, A and B cannot know that: $M_1 \models (\neg 2_a \wedge \neg 2_b) \rightarrow [ann1][ann2](ann_2 \wedge \neg K_a ann_2 \wedge \neg K_b ann_2)$.

So why do we still have states $(2, y)$ and $(x, 2)$ in the model after the second announcement? Observe that the second announcement removed all states $(1, y)$ and $(x, 1)$. What enabled this was that $ann1$ was announced first – without that $ann2$ would not have removed those states. $ann2$ plays a similar role for the states $(2, y)$ and $(x, 2)$: after the announcement of $ann2$, $ann2$ becomes *false* in those states. However, it was true in the same states before the announcement, which is why they are not removed.

It could perhaps be argued that C implicitly meant something stronger than just that $ann2$ was true at the moment it was announced, for example that it would also stay true after the announcement (and that this was clear to A and B)[10]. That would be modelled explicitly by the announcement $ann2' = ann2 \wedge [ann2]ann2 = \langle ann2 \rangle ann2$. The effect of that announcement would in fact be identical to the effect of announcing $ann2$ twice in a row. As argued above, the third announcement (announcing $ann2$ a second time) would remove the $(2, y)$ and $(x, 2)$ states. Now, after this annoucement, $ann2$ becomes false in all $(3, y)$ and $(x, 3)$ states. Formally: $M_1 \models 3_a \rightarrow [ann1][ann2][ann2]K_a p_b$ (while $M_1 \models 3_a \rightarrow [ann1][ann2]\neg K_a p_b$). We can continue this argument: repeating the announcement "Neither of you can deduce which number is greatest" removes more and more states. It is only in this sense that "you can extend this line of reasoning to include any number you like" is true: extending this line of reasoning implies that the announcement has to be made *again* to exclude the number 2, and again for the number 3, and so on. If repeated enough times, we will reach a point where either A or B has learned who has the greatest number, and the announcement is unsuccessful and cannot be repeated any more[11]. In the statement of the paradox, the announcement is only made once, which explains why the reasoning cannot be extended beyond the number 1. This resolves the paradox.

4 A Syntactic Analysis

We now turn to analyse the paradox *syntactically*, by describing the situation as a set of formulas Γ, and deriving conclusions from them. In particular, we will show, similarly to in the model theoretic analysis, that

$$\Gamma \vdash [ann1][ann2]K_a \neg 1_b$$

[10] Gerbrandy [8, pp. 27–29] discusses the same point in the context of surprise examination.

[11] This can be expressed elegantly by the iterated announcement operator in [13]: $M_1 \models \langle ann2^* \rangle \neg ann2$, which is true iff $M_1 \models \underbrace{\langle ann2 \rangle \cdots \langle ann2 \rangle}_{n} \neg ann2$ for some $n \geq 1$. See also [20] for a further disucssion of this and related operators.

– after the two announcements A knows that B does not have 1, but

$$\Gamma \vdash 157_a \to [ann1][ann2]\neg K_a \neg 2_b$$

– she does *not* know that B does not have 2 (in the case that A has 157 as in the description of the paradox), stopping the inductive train of thought in its tracks.

4.1 Describing the initial situation

We start by defining Γ, describing A's and B's initial knowledge and ignorance. For the purpose of the two derivations mentioned above we basically only need two premises (more discussion on this perhaps surprising fact below).

The first is that *everyone knows their own number*. For any $i \in \{a, b\}$:

$$x_i \to K_i x_i \tag{A0}$$

and furthermore that *this is known by both A and B*. For any $i, j \in \{a, b\}$:

$$K_j(x_i \to K_i x_i) \tag{A1}$$

Since (A1) implies (A0) (see epistemic logic axiom T), we actually only need (A1). We will use axiom T in the same way implicitly in the following.

The second is that initially (before any announcements) *each agent considers it possible that the other has chosen any number* (and this is known by both). For any $i, j \in \{a, b\}$ and any number y, we write $\bar{i}$ for "the other agent", i.e., $\bar{a} = b$ and $\bar{b} = a$:

$$K_j \hat{K}_i y_{\bar{i}} \tag{A2}$$

In addition to these two[12] premises we need some bookkeeping: the logic of the linear order of the natural numbers and the agents' knowledge of that. This is captured by the following three premises.

First, the relationship betwen p_b and p_a. *If i's number is greatest, then the other agent's number is not* (and this is known). For any $i, j \in \{a, b\}$:

$$K_j(p_i \to \neg p_{\bar{i}}) \tag{A3}$$

Second, we need two premises describing the relationship between atoms of the form 156_a and p_a. The first says that *one is the lowest number* (and anyone knows this, and anyone knows that anyone knows this[13]). The second is that *if agent i has the greatest number then p_i holds* (and this is known). For any $i, j, k \in \{a, b\}$ and numbers $x > y$:

$$K_j K_k(1_i \to \neg p_i) \tag{A4}$$

$$K_j K_k((x_i \wedge y_{\bar{i}}) \to p_i) \tag{A5}$$

[12] There are two *schemas* but actually infinitely many formulas.

[13] We could assume that these premises are *common knowledge*, writing e.g., $C_{\{a,b\}}(1_i \to \neg p_i)$. However, it turns out that assuming common knowledge is *not needed*, and it is of interest to illucidate exactly how many levels of nested knowledge is sufficient: e.g., two levels for (A4).

Thus, we let Γ be all instances of (A1)–(A5):

$$\Gamma = \left\{ \begin{array}{ll} K_j(x_i \rightarrow K_i x_i), & \\ K_j \hat{K}_i y_{\bar{i}}, & \\ K_j(p_i \rightarrow \neg p_{\bar{i}}), & : i, j, k \in \{a, b\}, x, y \in \mathbb{N}, x > y \\ K_j K_k(1_i \rightarrow \neg p_i), & \\ K_j K_k((x_i \wedge y_{\bar{i}}) \rightarrow p_i) & \end{array} \right\}$$

where $\mathbb{N}$ is the set of natural numbers. Note that, while Γ is an infinite set of premises, any derivation $\Gamma \vdash \phi$ of ϕ from Γ can only use a finite number of those premises.

4.2 Simplifying announcements

It is a straightforward exercise in PAL to show that, for any ϕ,

$$\vdash [ann1][ann2]\phi \leftrightarrow [\beta]\phi \tag{1}$$

where

$$\beta = (p_b \vee p_a) \wedge \\ \neg K_b(p_a \rightarrow p_b) \wedge \neg K_b(p_b \rightarrow p_a) \wedge \neg K_a(p_a \rightarrow p_b) \wedge \neg K_a(p_b \rightarrow p_a)$$

From $K_b(p_b \rightarrow \neg p_a) \in \Gamma$ (A3) and similarly for the other combinations, we also have[14]:

$$\Gamma \vdash \alpha \leftrightarrow \beta \tag{2}$$

where

$$\alpha = (p_b \vee p_a) \wedge \hat{K}_b p_a \wedge \hat{K}_b p_b \wedge \hat{K}_a p_a \wedge \hat{K}_a p_b$$

4.3 I know that she does not have 1

We now show that $\Gamma \vdash [ann1][ann2]K_a\neg 1_b$. Here and in the following we often combine several proof steps. In particular, we liberally use known epistemic logic and PAL theorems – referred to as "S5" and "PAL" respectively.

1	$\Gamma \vdash K_a(1_b \rightarrow K_b 1_b)$	(A1)
2	$\Gamma \vdash K_a K_b(1_b \rightarrow \neg p_b)$	(A4)
3	$\Gamma \vdash K_a(K_b 1_b \rightarrow K_b \neg p_b)$	2, S5
4	$\Gamma \vdash K_a(1_b \rightarrow K_b \neg p_b)$	1, 3, S5
5	$\Gamma \vdash K_a(1_b \rightarrow \neg\alpha)$	4, Prop
6	$\Gamma \vdash K_a(\alpha \rightarrow \neg(\alpha \rightarrow 1_b))$	5, Prop
7	$\Gamma \vdash K_a(\alpha \rightarrow \neg[\alpha]1_b)$	6, APerm
8	$\Gamma \vdash K_a[\alpha]\neg 1_b$	7, ANeg
9	$\Gamma \vdash \alpha \rightarrow K_a[\alpha]\neg 1_b$	8, Prop
10	$\Gamma \vdash [\alpha]K_a\neg 1_b$	9, AKnow
11	$\Gamma \vdash [ann1][ann2]K_a\neg 1_b$	10, Eq. (1), Eq. (2), Prop

[14] Observe that α expresses that (1) the two numbers are different, and (2) both agents consider each of the numbers to be the greatest (α implies $ann1 \wedge ann2$ but not the other way around).

4.4 But I don't know that she does not have 2

We show that$\mathit{\Gamma} \vdash 157_a \rightarrow [ann1][ann2]\neg K_a\neg 2_b$. "$x/y$" means "replace x with y".

1	$\mathit{\Gamma} \vdash \hat{K}_a 2_b$	(A2), S5
2	$\mathit{\Gamma} \vdash 157_a \rightarrow K_a 157_a$	(A1), S5
3	$\mathit{\Gamma} \vdash K_a((157_a \wedge 2_b) \rightarrow p_a)$	(A5), S5
4	$\mathit{\Gamma} \vdash K_a 157_a \rightarrow K_a(2_b \rightarrow p_a)$	3, S5, Prop
5	$\mathit{\Gamma} \vdash 157_a \rightarrow \hat{K}_a(2_b \wedge (2_b \rightarrow p_a))$	1, 2, 4, S5
6	$\mathit{\Gamma} \vdash 157_a \rightarrow \hat{K}_a(2_b \wedge (p_b \vee p_a))$	5, Prop
7	$\mathit{\Gamma} \vdash K_a \hat{K}_a 2_b$	(A2)
8	$\mathit{\Gamma} \vdash K_a 157_a \rightarrow K_a \hat{K}_a(2_b \wedge 157_a)$	7, S5
9	$\mathit{\Gamma} \vdash K_a((2_b \wedge 157_a) \rightarrow p_a)$	(A5), S5
10	$\mathit{\Gamma} \vdash K_a 157_a \rightarrow K_a \hat{K}_a p_a$	8, 9, S5
11	$\mathit{\Gamma} \vdash 157_a \rightarrow K_a \hat{K}_a p_a$	2, 10, Prop
12	$\mathit{\Gamma} \vdash 157_a \rightarrow K_a \hat{K}_a p_b$	Like 7-11: 2_b/158_b and p_a/p_b
13	$\mathit{\Gamma} \vdash K_a \hat{K}_b 3_a$	(A2)
14	$\mathit{\Gamma} \vdash K_a(2_b \rightarrow K_b 2_b)$	(A1)
15	$\mathit{\Gamma} \vdash K_a(2_b \rightarrow \hat{K}_b(3_a \wedge 2_b))$	13, 14, S5
16	$\mathit{\Gamma} \vdash K_a K_b(3_a \wedge 2_b \rightarrow p_a)$	(A5)
17	$\mathit{\Gamma} \vdash K_a(2_b \rightarrow \hat{K}_b p_a)$	15, 16, S5
18	$\mathit{\Gamma} \vdash K_a(2_b \rightarrow \hat{K}_b p_b)$	Like 13-17: 3_a/1_a and p_a/p_b
19	$\mathit{\Gamma} \vdash 157_a \rightarrow \hat{K}_a(2_b \wedge \alpha)$	6, 11, 12, 17, 18, S5
20	$\mathit{\Gamma} \vdash 157_a \rightarrow (\alpha \rightarrow \neg K_a(\alpha \rightarrow \neg 2_b))$	19, Prop
21	$\mathit{\Gamma} \vdash 157_a \rightarrow (\alpha \rightarrow \neg K_a(\alpha \rightarrow \neg(\alpha \rightarrow 2_b)))$	20, Prop
22	$\mathit{\Gamma} \vdash 157_a \rightarrow (\alpha \rightarrow \neg K_a(\alpha \rightarrow \neg[\alpha]2_b))$	21, APerm
23	$\mathit{\Gamma} \vdash 157_a \rightarrow (\alpha \rightarrow \neg K_a[\alpha]\neg 2_b)$	22, ANeg
24	$\mathit{\Gamma} \vdash 157_a \rightarrow (\alpha \rightarrow (\alpha \wedge \neg K_a[\alpha]\neg 2_b))$	23, Prop
25	$\mathit{\Gamma} \vdash 157_a \rightarrow (\alpha \rightarrow \neg(\alpha \rightarrow K_a[\alpha]\neg 2_b))$	24, Prop
26	$\mathit{\Gamma} \vdash 157_a \rightarrow (\alpha \rightarrow \neg[\alpha]K_a\neg 2_b)$	25, AKnow
27	$\mathit{\Gamma} \vdash 157_a \rightarrow [\alpha]\neg K_a\neg 2_b$	26, ANeg
28	$\mathit{\Gamma} \vdash 157_a \rightarrow [ann1][ann2]\neg K_a\neg 2_b$	29, Eq. (1), Eq. (2)

4.5 Dealing with infinite disjunction

In the previous section we showed how to derive $\mathit{\Gamma} \vdash 157_a \rightarrow [ann1][ann2]\neg K_a\neg 2_b$, and which assumptions were sufficient for that derivation. The number 157, taken from the original formulation of the paradox, is of course arbitrary – it could be replaced with 15 or 1570 or indeed any number different from 2 itself. So we get $\mathit{\Gamma} \vdash 15_a \rightarrow [ann1][ann2]\neg K_a\neg 2_b$ and so on in the same way. By this reasoning, it seems that we should be able to get the more general $\mathit{\Gamma} \vdash \neg 2_a \rightarrow [ann1][ann2]\neg K_a\neg 2_b$. However, this does in fact not hold – the assumptions in $\mathit{\Gamma}$ turn out to not be strong enough to make $\neg 2_a \rightarrow [ann1][ann2]\neg K_a\neg 2_b$ derivable. To see this, consider the model M_4 and its transformations as a result of the two announcements in Figure 3. It is easy to see that $M_4,(-,3) \models \mathit{\Gamma}$, but since $M_6,(-,3) \models K_a\neg 2_b$ we have that $M_4,(-,3) \not\models$

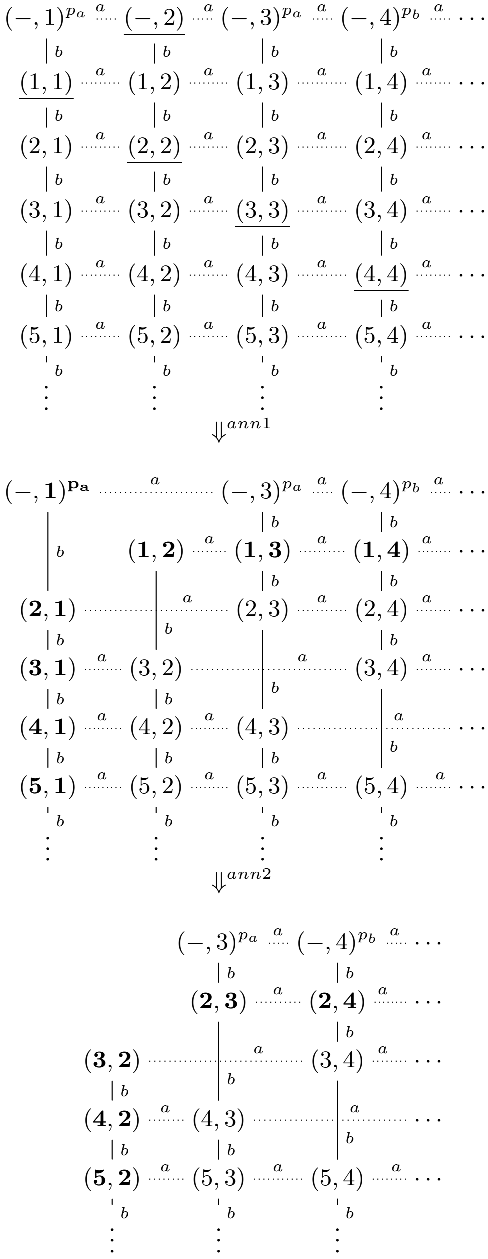

Fig. 3. Models M_4 (top), as well as M_5 (middle) and M_6 (bottom) – the results of announcing $ann1$ in M_4 and $ann2$ in M_5, respectively. States where $ann1$/$ann2$ is false are underlined/in bold. The valuation is the same as in M_1 for corresponding states. For the "new" states (first row), the valuation is as follows: x_b is given by the state, e.g., 3_b is true in state $(-, 3)$; x_a is *false* in all these states; the truth values of p_a and p_b are indicated in the figure.

$[ann1][ann2]\neg K_a\neg 2_b$. In other words, $\Gamma \not\models \neg 2_a \rightarrow [ann1][ann2]\neg K_a\neg 2_b$, and thus $\Gamma \not\vdash \neg 2_a \rightarrow [ann1][ann2]\neg K_a\neg 2_b$.

So, Γ must be strengthened if we want to derive $\neg 2_a \rightarrow [ann1][ann2]\neg K_a\neg 2_b$, so that models like M_4 are ruled out. That model contains states where one agent (a) has not chosen any number, clearly conflicting with the description of the puzzle[15]. However, the assumption that A has chosen some number corresponds to an infinite disjunction of the form $\bigvee_{x\geq 1} x_a$, which cannot be written as a formula.

It turns out, however, that a weaker assumption is sufficient. Notice that if we have that $\neg 1_a$ and $\neg 2_a$ and 2_b, it follows that p_a – if A doesn't have 1 or 2 she must have a number greater than B's number 2. $\neg 1_a \wedge \neg 2_a \wedge 2_b \rightarrow p_a$ does not follow from Γ (to see this observe that it is false in state $(-,2)$ in M_4). We now strengthen Γ with a generalisation of that assumption, namely, for any $i,j \in \{a,b\}$, $k \geq 1$ and $m \leq k$:

$$K_j(\neg 1_i \wedge \neg 2_i \wedge \cdots \wedge \neg k_i \wedge m_{\bar{i}} \rightarrow p_i) \tag{A6}$$

We will also need a negative variant of A1 (everyone knows their own number), saying that if I have *not* chosen x then I know that. For any $i,j \in \{a,b\}$:

$$K_j(\neg x_i \rightarrow K_i\neg x_i) \tag{A1'}$$

Finally, we will need to assume the following as a first principle (any $i \in \{a,b\}$):

$$K_i(\hat{K}_a p_b \wedge \hat{K}_b p_a) \tag{A7}$$

– in the initial situation (before any announcements), A considers it possible that B has chosen a greater number, and conversely for B (note that we cannot assume, e.g., $\hat{K}_a p_a$ – because if A has 1 she does not consider it possible that her number is greater than B's).

Let Γ' be Γ extended with premises (A6), (A1') and (A7), i.e., $\Gamma' = \Gamma \cup \{K_j(\neg 1_i \wedge \neg 2_i \wedge \cdots \wedge \neg k_i \wedge m_{\bar{i}} \rightarrow p_i), K_j(\neg x_i \rightarrow K_i\neg x_i), K_i(\hat{K}_a p_b \wedge \hat{K}_b p_a) : i,j \in \{a,b\}, k \geq 1, m \leq k\}$. We now show that $\Gamma' \vdash \neg 2_a \rightarrow [ann1][ann2]\neg K_a\neg 2_b$.

In the following, by "L. 4.4:x-y" we mean "like in lines x to y in the proof in Sec. 4.4".

[15] Nevertheless, that was no problem for $157_a \rightarrow [ann1][ann2]\neg K_a\neg 2_b$, which happens to hold in those models too.

1	$\Gamma' \vdash (\neg 1_a \wedge \neg 2_a) \to \hat{K}_a(2_b \wedge (p_b \vee p_a))$	L. 4.4:1-6; $157_a/(\neg 1_a \wedge \neg 2_a)$, (A1)/(A1’), (A5)/(A6)
2	$\Gamma' \vdash (\neg 1_a \wedge \neg 2_a) \to K_a \hat{K}_a p_a$	L. 4.4:7-11; $157_a/(\neg 1_a \wedge \neg 2_a)$ (A5)/(A6)
3	$\Gamma' \vdash \hat{K}_a p_b$	(A7)
4	$\Gamma' \vdash (\neg 1_a \wedge \neg 2_a) \to K_a \hat{K}_a p_b$	3, Prop
5	$\Gamma' \vdash K_a \hat{K}_b p_a$	(A7)
6	$\Gamma' \vdash K_a(2_b \to \hat{K}_b p_a)$	5, S5
7	$\Gamma' \vdash K_a(2_b \to \hat{K}_b p_b)$	L. 4.4:13-18; $3_a/1_a$, p_a/p_b
8	$\Gamma' \vdash (\neg 1_a \wedge \neg 2_a) \to \hat{K}_a(2_b \wedge \alpha)$	1, 2, 4, 6, 7, S5
9	$\Gamma' \vdash (\neg 1_a \wedge \neg 2_a) \to [\alpha]\neg K_a \neg 2_b$	L. 4.4:19-27
10	$\Gamma' \vdash (\neg 1_a \wedge \neg 2_a) \to (\alpha \to \langle\alpha\rangle \neg K_a \neg 2_b)$	9, PAL
11	$\Gamma' \vdash (\neg 1_a \wedge \neg 2_a \wedge \alpha) \to \langle\alpha\rangle \neg K_a \neg 2_b$	10, Prop
12	$\Gamma' \vdash \alpha \to \neg 1_a$	as in Sec. 4.3
13	$\Gamma' \vdash (\neg 2_a \wedge \alpha) \to \langle\alpha\rangle \neg K_a \neg 2_b$	11, 12, Prop
14	$\Gamma' \vdash \neg 2_a \to (\alpha \to \langle\alpha\rangle \neg K_a \neg 2_b$	13, Prop
15	$\Gamma' \vdash \neg 2_a \to [\alpha]\neg K_a \neg 2_b$	14, PAL ([19, Prop. 4.13])
16	$\Gamma' \vdash \neg 2_a \to [ann1][ann2]\neg K_a \neg 2_b$	15, Eq. (1), Eq. (2)

5 Discussion

We have argued that under assumptions about common knowledge of truthfulness and perfect reasoners, Hollis’ paradox can be resolved by observing that the second announcement is not neccessarily successful. Note that it will *actually* be a *successful update* – except in the cases that either A or B has chosen 1 or 2. Thus, a more precise explanation is that the agents *don’t know whether the announcement was successful*. As is well known in dynamic epistemic logic, an announcement can be unsuccessful yet informative, a likely source of the confusion behind the so-called paradox.

As mentioned in the introduction, there are similarities between Hollis’ paradox and the surprise examination paradox. In particular, they are built on the same fallacy: that announcements always are successful. This was first pointed out for the surprise examination paradox by Gerbrandy [8], using a variant of public announcement logic, in a similar way to the semantic analysis in this paper. Several other logical analyses have since appeared [4, 3, 12, 2]. While it can be argued that the root cause behind the two “paradoxes” is the same (unsuccessful formulas), the logical modelling is quite different. Gerbrandy’s formalisation has in common with our formalisation of Hollis’ paradox that there is an initial announcement that eliminates some states in the model and that the (false) assumption that the initial announcement would stay true after that initial elimination would eliminate yet more states and that this can be repeated in several steps eventually leading to a paradoxical situation where all states have been eliminated. In both cases the "paradox" can be seen as an inductive "proof" that actually fails after the first step due to the false premise that the initial announcement is successful. A significant difference is that in the surprise examination paradox the state space is finite, while in Hollis’ paradox the state space is infinite and an inductive argument is crucial.

Other significant differences is that the state space in Gerbrandy's solution is very simple, consisting of only three states, and our model of Hollis' paradox is more complex, while on the other hand the announcement of "the exam date will be a surprise" in the former[16], is more complex than $ann2$. The reason for the latter is the iterative opening of the doors which has no correspondent in Hollis' paradox. In fact, from a modelling perspective Hollis' paradox has more in common with *Sum and Product* [6], with a state space that is a (in that case finite) subset of the cartesian product of the natural numbers and where states are eliminated in a sequence of announcements. In that case the announcements are given explicitly and there is no paradox.

It has however, been forcefully argued [2] that Gerbrandy's *non-self-referential* formalisation of the announcement is not a very natural interpretation of the sentence "the exam date will be a surprise" nor is it indeed the interpretation most commenters on the paradox agree with. This argument hinges on the word "will" which refers to the future and in particular, it is argued, to the actual future immediately after the announcement is made, and thus that a *self-referential* interpretation of the statement to mean "you will not know in advance the exam day (i.e., after hearing *this* very announcement") [17]. This is indeed convincing, but we argue that the same argument does not apply to Hollis' paradox where the announcement is "neither of you can work it out" (or "neither of you can deduce it" [16]). Granted, "can work out" (or "deduce") seem to refer to the future as well, but a perfect reasoner has at any point already "worked out" (deduced) all possible consequences of her knowledge. The operative word here is "*can*", referring to the present, the announcement is not "neither of you *will* be able to work out".

Our formalisation hinges on the two assumptions of common knowledge of truthfulness and perfect reasoners, both of which it would be interesting to relax in future work on formalisations. Modeling non-perfect reasoners (see, e.g., [1]) might seem particularly relevant since it gives more meaning to the phrase "can work out", but there are no clues in the description of the paradox how the agents abilities to "work out" things are limited (indeed, on the contrary, as mentioned in the introduction Hollis hinted at joint knowledge of good reasoning abilities).

While the semantic modelling of the initial situation in Hollis' paradox allowed us to pinpoint exactly where the inductive argument breaks down, existing discourse on Hollis' paradox [9, 14, 11, 10, 17] typically employ (informal) derivations of conclusions from premises in some implicit epistemic/doxastic logic. In keeping with this tradition we also provided a "syntactic" analysis where we modelled the initial situation as a set of premises and derived our conclusions from them – albeit in a more detailed, formal way. This furthermore allowed us to pinpoint which of the facts in the initial situation were sufficient for the conclusions. It turned out that we did not need to completely describe the grid model from the semantic analysis. Furthermore, while it can clearly be argued that it is implicitly assumed that it is common knowledge that A and B each know their own number, the derivation of the fact that none of the agents can rule out

[16] $(we \wedge \neg K we) \vee (th \wedge [\neg we]\neg K th) \vee (fr \wedge [\neg we][\neg th]\neg K fr) \vee K\bot$. Note that Gerbrandy assumes that the knowledge modalities are K45 rather than S5.

[17] Note that this kind of self-reference is not the same as saying that "you don't know it now and you still don't know it after it is announced that you don't know it now" as briefly discussed at the end of Section 3.

that the other one has 2 only relies on *general* knowledge (everybody-knows) of that fact. That conclusion only relies on up to 2 levels of nested knowledge of any of the premises (everybody knows that everybody knows).

The fact that we don't need to assume common knowledge of the premises has an interesting corollary. Intuitively, a "static" epistemic or doxastic logic seems to be insufficient to deal with the paradox, because we need to be able to reason about knowledge/beliefs at different time points – in particular "before" and "after" announcements. Indeed, failure to make that distinction is exactly what lies behind the original paradox as well as other attempts to resolve it. However, the fact that we don't need common knowledge means that the premises, conclusions and the whole derivation can be translated into pure (static!) epistemic logic [15]! So, Hollis' paradox can be resolved by pure "static" epistemic reasoning about the initial situation after all.

Acknowledgments The first author was supported by L. Meltzers Høyskolefond.

References

1. Thomas Ågotnes. *A logic of Finite Syntactic Epistemic States*. Ph.D. thesis, Department of Informatics, University of Bergen, Norway, 2004.
2. Alexandru Baltag, Nick Bezhanishvili, and David Fernández-Duque. The topology of surprise. In *Proceedings of the International Conference on Principles of Knowledge Representation and Reasoning*, volume 19, pages 33–42, 2022.
3. Alexandru Baltag and S Smets. Surprise?! an answer to the hangman, or how to avoid unexpected exams! In *Logic and Interactive Rationality Seminar (LIRA), slides*, 2009.
4. Hans Van Ditmarsch and Barteld Kooi. The secret of my success. *Synthese*, 151(2):201–232, 2006.
5. R. Fagin, J.Y. Halpern, Y. Moses, and M.Y. Vardi. *Reasoning about Knowledge*. MIT, 1995.
6. H. Freudenthal. Solution to problem no. 223. *Nieuw Archief voor Wiskunde*, 3(18):102–106, 1970.
7. Jelle Gerbrandy. *Bisimulations on planet Kripke*. University of Amsterdam, 1999.
8. Jelle Gerbrandy. The surprise examination in dynamic epistemic logic. *Synthese*, 155:21–33, 2007.
9. Martin Hollis. A paradoxical train of thought. *Analysis*, 44(4):205–206, 1984.
10. Martin Hollis. More paradoxical epistemics. *Analysis*, 46(4):217–218, 1986.
11. Michael Kinghan. A paradox derailed: reply to hollis. *Analysis*, 46(1):20–24, 1986.
12. Alexandru Marcoci. The surprise examination paradox in dynamic epistemic logic. Master's thesis, Universiteit van Amsterdam, 2010.
13. Joseph S Miller and Lawrence S Moss. The undecidability of iterated modal relativization. *Studia Logica*, 79:373–407, 2005.
14. Doris Olin. On a paradoxical train of though. *Analysis*, 46(1):18–20, 1986.
15. J.A. Plaza. Logics of public communications. In *Proc. of ISMIS '89*, pages 201–216, 1989.
16. William Poundstone. *Labyrinths of reason: Paradox, puzzles, and the frailty of knowledge*. Anchor, 2011.
17. George Rea. A variation of hollis' paradox. *Analysis*, 47(4):218–220, 1987.
18. H. van Ditmarsch and B. Kooi. *One Hundred Prisoners and a Light Bulb*. Springer International Publishing, 2015.
19. H. van Ditmarsch, W. van der Hoek, and B. Kooi. *Dynamic Epistemic Logic*, volume 337 of *Synthese Library*. Springer, 2007.
20. Hans van Ditmarsch. To be announced. *Information and Computation*, 292:105026, 2023.