Dynamics of Lying

Hans van Ditmarsch, LORIA — CNRS / Université de Lorraine

Abstract

We propose a dynamic logic of lying, wherein a 'lie that φ' (where φ is a formula in the logic) is an action in the sense of dynamic modal logic, that is interpreted as a state transformer relative to the formula φ. The states that are being transformed are pointed Kripke models encoding the uncertainty of agents about their beliefs. Lies can be about factual propositions but also about modal formulas, such as the beliefs of other agents or the belief consequences of the lies of other agents. We distinguish two speaker perspectives: (Obs) an outside *observer* who is lying to an agent that is modelled in the system, and (Ag) an *agent* who is lying to another agent, and where both are modelled in the system. We distinguish three addressee perspectives: (Cred) the *credulous* agent who believes everything that it is told (even at the price of inconsistency), (Skep) the *skeptical* agent who only believes what it is told if that is consistent with its current beliefs, and (Rev) the belief *revising* agent who believes everything that it is told by consistently revising its current, possibly conflicting, beliefs. The logics have complete axiomatizations, which can most elegantly be shown by way of their embedding in what is known as action model logic or in the extension of that logic to belief revision.

1 Introduction

A Grimm brothers fairytale called 'A Liar's Tale' [17] contains the passage "A crab was chasing a hare which was running away at full speed; and high up on the roof lay a cow which had climbed up there." This contains very obvious lies. Nobody considers it possible that this is true. Crabs are reputedly slow, hares are reputedly fast.

In 'The Liar's Tale', none of the lies are believed.

In the movie 'The Invention of Lying' the main character Mark goes to a bank counter and finds out he only has \$300 in his account. But he needs \$800. Lying has not yet been invented in the 20th-century fairytale country of this movie. Then and there, Mark invents lying. He tells the bank employee assisting him that there must be a mistake: he has \$800 in his account. He is lying. She responds, oh well, then there must be a mistake with your account data, because on my screen it says you only have \$300. And she gives him \$800! In the remainder of the movie, Mark gets very rich.

In the movie 'The Invention of Lying', all lies are believed.

Sometimes other people believe you when you lie and sometimes they don't. When can you get away with a lie? In Section 3.1 we present the consecutive numbers riddle, wherein two agents Anne and Bill are each told a natural number, and where they know that their numbers are one apart. If Anne is told 2 and Bill is told 3, Anne is uncertain between Bill having 1 or 3, and Bill is uncertain between Anne having 2 or 4. So both Anne and Bill do not know their number.

Suppose that Anne says to Bill: "I know your number." Anne is lying. Bill does not consider it possible that Anne knows his number: he will not believe the lie. Alternatively, suppose that Anne says to Bill: "I don't know your number." She is telling the truth. Bill, in response, says to Anne: "I know your number." Now Bill is lying. If Anne believes that, she could then say "I know your number," as she believes Bill to have 1. Her announcement is truthful (i.e., she believes it to be true) but an honest mistake, because her belief is incorrect. However, as a result of Anne's announcement Bill will learn Anne's number, so that his announcement "I know your number," that was a lie at the time, now has become true. That is, if you are still following us.

In more realistic scenarios, some lies are believed and some are not.

It seems not so clear how all this should be formalized in a logic interpreted on epistemic modal structures, and this is the topic of our paper.

1.1 The modal dynamics of lying

What is a lie? Let p be an atomic proposition (propositional variable). You lie to me that p, if you believe that p is false while you say that p, and with the intention that I believe p. The thing you say, we call the announcement. If you succeed in your intention, then I believe p, and I also believe that your announcement of p was truthful, i.e., that you believed that p when you said that p. In this investigation we abstract from the intentional aspect of lying (with the exception of Section 7). We model lying as a dynamic operation, and in that sense we only model the realization of the intention: the successful lie. This is an abstraction. Such an abstraction is similar to that in AGM belief revision [2], wherein one models how to incorporate new information in an agent's belief set, but abstracts from the process that made the new information acceptable to the agent. Our proposal is firmly grounded in modal logic. We employ dynamic epistemic logic [47].

What are the modal preconditions and postconditions of a lie? Let us for now assume that p is a Boolean proposition. Given two agents a (female) and b (male), in our exposition a will typically be the speaker or sender and b will then be the receiver or addressee. However, a and b are not agent roles but agent names. We also model dialogue wherein agents speak in turn; so these roles may swap. Formula $B_a p$ stands for 'agent a believes that p'.

The precondition of 'a is lying that p to b' is $B_a \neg p$ ($\neg$ is negation). Stronger preconditions are conceivable, e.g., that the addressee considers it possible that the lie is true,

$\neg B_b \neg p$, or that the speaker believes that, $B_a \neg B_b \neg p$. These conditions may not always hold while we still call the announcement a lie, because the speaker may not know whether the additional conditions are satisfied. We therefore will (initially) only require precondition $B_a \neg p$.

We should contrast the announcement that p by a lying agent, with other forms of announcement. Just as a lying agent believes that p is false when it announces p, a truthful agent believes that p is true when it announces p. The precondition for a lying announcement by a is $B_a \neg p$, and so the precondition for a truthful announcement by a is $B_a p$. Besides the truthful and the lying announcement there is yet another form of announcement, because in modal logic there are always three instead of two possibilities: either you believe p, or you believe $\neg p$, or you are uncertain whether p. The last corresponds to the precondition $\neg(B_a p \vee B_a \neg p)$ ($\vee$ is disjunction). An announcement wherein agent a announces p while she is uncertain about p we propose to call a bluffing announcement (we do not know of accepted standard terminology for this form of announcement).

We conclude that to the three mutually exclusive and complete preconditions $B_a p$, $B_a \neg p$, and $\neg(B_a p \vee B_a \neg p)$ we associate the truthful*,* lying*, and* bluffing *announcement that p.*[1]

We now consider the postconditions of 'a is lying that p to b'. If a's intention to deceive b is successful, b believes p after the lie. Therefore, $B_b p$ should be a postcondition of a successful execution of the action of lying. Also, the precondition should be preserved: $B_a \neg p$ should still be true after the lie. In the first place, we propose logics to achieve this. However, this comes at a price. In case the agent b already believed the opposite, $B_b \neg p$, then b's beliefs are inconsistent afterwards. (This merely means that b's accessibility relation is empty, not that the logic is inconsistent.) There are two different solutions for this: either b does not change his beliefs, so that $B_b \neg p$ still holds after the lie, or the belief $B_b \neg p$ is given up in order to consistently incorporate $B_b p$. The three alternative postconditions after the lie that p are therefore: (Cred) always make $B_b p$ true after the lie (even at the price of inconsistency), (Skep) only make $B_b p$ true if agent b considered p possible before the lie ($\neg B_b \neg p$), and (Rev) always make $B_b p$ true by a consistency preserving process of belief revision. These are all modelled. We consider these alternatives 'agent

[1] Strictly binding formal belief preconditions to these terms is intended as a strong barrier to avoid pitfalls and digressions into philosophy and epistemology. This is necessary because their everyday usage is ambiguous. We have tried to stay close to dictionary meanings and reported usage. This footnote reports our findings. 'Truthful' is synonymous with 'honest'. Dictionaries do not make a difference between an agent telling the truth and an agent believing that it is telling the truth. A modal logician has to make a choice. We mean the latter, exclusively. A truthful announcement may therefore not be a true announcement. It is tempting, when looking for a single term, to call a truthful agent a truthteller (a somewhat archaic usage) but that would imply that we would not require a truthteller to tell the truth, maybe stretching usage too far. So we did not, and stick to 'truthful agent'. The dictionary meaning for the verb bluff is 'to cause to believe what is untrue' or 'to deceive or to feign'. Feigning belief in p means suggesting belief in p, for example by saying that you believe it, even if this is not the case. That corresponds to $\neg B_a p$ as precondition. However, this would make lying a form of bluffing, as $B_a \neg p$ implies $\neg B_a p$. It is common and according to Gricean conversational norms that saying something that you believe to be false is worse than (or, at least, different from) saying something that you do not believe to be true. This brings us to $\neg B_a p \wedge \neg B_a \neg p$ ($\wedge$ is conjunction), equivalent to $\neg(B_a p \vee B_a \neg p)$.

types': the credulous agent Cred, the skeptical agent Skep, and the belief revising agent Rev. Sometimes we obtain $B_b B_a p$ instead of $B_b p$ as a postcondition.

The action of lying is modelled as a dynamic modal operator. The dynamic modal operator for 'lying that p' is interpreted as an epistemic state transformer. An epistemic state is a pointed Kripke model (a model with a designated state) that encodes the beliefs of the agents. In this dynamic epistemic setting we can distinguish two different lying agents: (Obs) the case of an external observer (an agent who is not explicitly modelled in the structures and in the logical language), who is lying to an agent modelled in the system, and (Ag) the case of one agent lying to another agent, where both are explicitly modelled. Obs and Ag are considered agent types as well.

The belief operators B_a do not merely apply to Boolean propositions p but to any proposition φ with belief modalities. This is known as higher-order belief. The preconditions and postconditions of lying may be such higher-order formulas. In the semantics, the generalization from 'lying that p' to 'lying that φ' for any proposition, does not present any problem. But in the syntax it presents a problem: we can no longer require, for example, that the beliefs of the speaker a remain unchanged, or that, after the lie, the addressee b believes the formula of the lie. For a typical example, suppose that p is false, that a knows the truth about p, and that b is uncertain whether p. Consider the lie by a to b that $p \wedge \neg B_b p$ (a Moorean sentence). This is a lie: a knows (believes correctly) that p is false, and therefore that the sentence $p \wedge \neg B_b p$ is false. Clearly, we wish to say that the lie was successful if $B_b p$ holds afterwards, not if $B_b(p \wedge \neg B_b p)$ holds ('belief of the announced formula'), as this is an inconsistency. Our proposed modelling allows us to elegantly explain lying about such modal formulas. Lying in the consecutive numbers riddle provides another example of that.

1.2 A short history of lying

We conclude this introduction with a review of literature on lying.

Philosophy Lying has been a thriving topic in the philosophical community for a long, long time [39, 9, 29, 30]. Almost any analysis starts with quoting Augustine on lying:

> "that man lies, who has one thing in his mind and utters another in words"
> "the fault of him who lies, is the desire of deceiving in the uttering of his mind"

In other words: lying amounts to saying that p while believing that $\neg p$, with the intention to make believe p, our starting assumption. The requirements for the belief preconditions and postconditions in such works are illuminating [30]. For example, the addressee should not merely believe the lie but believe it to be believed by the speaker. Indeed, ... and even believe it to be commonly believed by speaker and addressee, would the modal logician say. Scenarios involving eavesdroppers (can you lie to an agent who is not the addressee?) are relevant for logic and multi-agent system design, and also relevant are claims that you can only lie if you really say something: an omission is not a lie [30]. Wrong, says

the computer scientist: if the protocol is common knowledge, you can lie by not acting when you should have; say, by not stepping forward in the muddy children problem when you should have because you know that you are muddy. The philosophical literature also clearly distinguishes between false propositions and propositions believed to be false by the speaker but in fact true, so that when you lie about them, you actually tell the truth. Gettier-like scenarios are presented, including delayed justification [34].[2] Much is said on the morality of lying [9] and on its intentional aspect [30]. As said, we abstract from the intentional aspect of lying except in Section 7. We also abstract from its moral aspect.

Cognitive science Lying excites great interest in the general public. Examples of popular science books on lying are [41, 43]. In psychology, biology, and other experimental sciences, lying and deception are related. A cuckoo is 'lying' if it is laying eggs in another bird's nest. Relevant for our investigation is, that it is typical to be believed, and that lying is therefore the exception. We model the 'successful' lie that is indeed believed, unless there is evidence to the contrary, namely prior belief in the opposite. Also relevant in cognitive science and biology is that the detection of lying is costly, and that this is a reason to be typically believed. In logic, the cost of reasoning is computational complexity. In cognitive science, the cost of reasoning is experimentally determined response time. A proved relation between computational complexity and response time in the presence of lying would be highly relevant for our investigation. The cost of reasoning as response time is investigated in 'Theory of Mind' (ToM) studies, such as [12, 21] in this special issue. A false belief in the ToM sense is not necessarily a lie, it is false in the sense of being a *mistaken belief*. The interplay between reasoning capabilities and embedded abilities (spontaneous response [12], implicit knowledge about false beliefs [21]) seems crucial in ToM studies. In this special issue, [13] is closer to the aims of our investigation. They observe that "Deception demands the ability to simulate another's reaction in order to determine if a lie will be believable. It is even harder if the goal of the lie is to manipulate." Here we see the intentional aspect of the lie again, translated into computational cost.

Economics In economics, 'cheap talk' is making false promises. Your talk is cheap if you do not intend to execute an action that you publicly announced to plan. It is therefore a lie, it is deception [16, 23]. Our focus is different. We do not model lying about planned actions but lying about propositions, and in particular we model the belief consequences of such lies. Economists postulate probabilities for lying strategies and truthful strategies, to be tested experimentally. We only distinguish lies that are always believed from lies that (in the face of contradictory prior belief) are never believed.

Logic Papers that model lying as an epistemic action, inducing a transformation of an epistemic model, include [5, 40, 7, 45, 25, 48, 37]. Lying by an external observer has been

[2]Suppose that you believe that $\neg p$ and that you lie that p. Later, you find out that your belief was mistaken because p was really true. You can then with some justification say "Ah, so I was not really lying."

discussed by Baltag and collaborators from the inception of dynamic epistemic logic onward [5]; the later [7] also discusses lying in logics with plausible belief, as does [45]. In [48] the conscious update in [15] is applied to model lying by an external observer. In [38] the authors give a modal logic of lying and bluffing, including intentions. Instead of bluffing they call this bullshit, after [14]. Strangely they do not model lying as a dynamic modality; it is therefore unclear how they can deal with the contradictory Moorean phenomena [20] discussed above. The recent [37] lists varous modal logics combining knowledge and intention, where the philosophically familiar themes of how to stack belief preconditions and belief postconditions reappear. In [40, 25] the unbelievable update is considered; this is the issue of consistency preservation for belief, as in our treatment of unbelievable lies (rejecting the lie that p if you already believe $\neg p$). The promising [27] allows explicit reference in the logical language to truthful, lying and bluffing agents, thus enabling some form of self-reference—this work also inspired us to call our similar distinctions *agent types*.

1.3 Contributions and overview

The novel contribution of our paper is a precise model of the informative consequences of two agents lying to each other, and a logic for that, including a treatment of bluffing. This agent-to-agent-lying, in the logic called agent announcement logic, is presented in Section 3 (the lying agent has type Ag). A special, simpler, case is that of an outside observer (Obs) who is lying to an agent that is modelled in the system. This (truthful and lying) public announcement logic is treated in Section 2. Section 4 on action models is an alternative perspective on the frameworks presented in Section 2 and Section 3. Section 5 adapts the logics of the Sections 2 and 3 (where the 'credulous' addressee has type Cred) to the requirement that unbelievable lies should not be incorporated (type Skep). Subsequently, Section 6 adapts these logics to the requirement that unbelievable lies, on the contrary, should be incorporated, but consistently so (type Rev). This involves structures with plausibility relations. These sections are also original contributions. Section 7 reviews the literature on how to model the intentional aspect of a lie in dynamic epistemic logic.

2 Truthful and lying public announcements

In this section we model lying by an outside observer (Obs) to a credulous addressee (Cred).

The logic of truthful and lying public announcements [48] is a version of the logic of so-called 'arrow elimination' public announcements [15], which is an alternative for the better known 'state elimination' logic of truthful public announcements [32, 6]. Its language, structures, and semantics are as follows. Given are a finite set of agents A and a countable set of propositional variables P (let $a \in A$ and $p \in P$).

Definition 1 (Language)

$$\mathcal{L}(!, ¡) \ni \varphi ::= p \mid \neg\varphi \mid (\varphi \wedge \psi) \mid B_a\varphi \mid [!\varphi]\psi \mid [¡\varphi]\psi \qquad \dashv$$

Other propositional connectives are defined by abbreviation. For $B_a\varphi$, read 'agent a believes formula φ'. Agent variables are $a, b, c, \ldots$. For $[!\varphi]\psi$, read 'after truthful public announcement of φ, formula ψ (is true)'. For $[¡\varphi]\psi$, read 'after lying public announcement of φ, formula ψ (is true)' (after the lie that φ, ψ is true). The dual operator for the necessity-type announcement operator is by abbreviation defined as $\langle !\varphi\rangle\psi := \neg[!\varphi]\neg\psi$. If $B_a\neg\varphi$, we say that φ is *unbelievable* (for a) and, consequently, if $\neg B_a\neg\varphi$, we say that φ is *believable* (for a). This is also read as 'agent a considers it possible that φ'. Without the announcement operators we get the language $\mathcal{L}$ of epistemic logic.

Definition 2 An *epistemic model* $M = (S, R, V)$ consists of a *domain* S of *states* (or 'worlds'), an *accessibility function* $R : A \to \mathcal{P}(S \times S)$, where each $R(a)$, for which we write R_a, is an accessibility relation, and a *valuation* $V : P \to \mathcal{P}(S)$, where each $V(p)$ represents the set of states where p is true. For $s \in S$, (M, s) is an *epistemic state*. ⊣

An epistemic state is also known as a pointed Kripke model. We often omit the parentheses in (M, s). Four model classes will appear in this work. Without any restrictions we call the model class $\mathcal{K}$. The class of models where all accessibility relations are transitive and euclidean is called $\mathcal{K}45$, and if they are also serial it is called $\mathcal{K}D45$. The class of models where all accessibility relations are equivalence relations is $\mathcal{S}5$. Class $\mathcal{K}D45$ is said to have the *properties of belief*, and $\mathcal{S}5$ to have the *properties of knowledge*.

Definition 3 Assume an epistemic model $M = (S, R, V)$.

$$
\begin{array}{lll}
M, s \models p & \text{iff} & s \in V_p \\
M, s \models \neg\varphi & \text{iff} & M, s \not\models \varphi \\
M, s \models \varphi \wedge \psi & \text{iff} & M, s \models \varphi \text{ and } M, s \models \psi \\
M, s \models B_a\varphi & \text{iff} & \text{for all } t \in S : R_a(s,t) \text{ implies } M, t \models \varphi \\
M, s \models [!\varphi]\psi & \text{iff} & M, s \models \varphi \text{ implies } M^\varphi, s \models \psi \\
M, s \models [¡\varphi]\psi & \text{iff} & M, s \models \neg\varphi \text{ implies } M^\varphi, s \models \psi
\end{array}
$$

where epistemic model M^φ is as M except that $R_a^\varphi := R_a \cap (S \times [\![\varphi]\!]_M)$ (and where $[\![\varphi]\!]_M := \{s \in S \mid M, s \models \varphi\}$). ⊣

The announcing agent is not modelled in public announcement logic, but only the effect of its announcements on the audience, the set of all agents. Truthful public announcement logic is the logic to model the revelations of a benevolent god, taken as the truth without questioning. Lying public announcements can be seen as made by a malevolent entity, the devil. Everything he says is false. Everything is a lie. The addressed agents always assume that god is talking to them: new information is accepted by the agents independent from the truth of that information.

If we define $[\varphi]\psi$ by abbreviation as $[!\varphi]\psi \wedge [¡\varphi]\psi$, then $[\varphi]\psi$ means 'after public announcement of φ (independently from the truth of φ), ψ is true'; and therefore, $[\varphi]B_a\psi$ means 'after public announcement of φ, agent a believes ψ'. We therefore call it the *believed public announcement* of φ (in contrast to the truthful public announcement of φ *with state elimination semantics* [32]). This dynamic operator is the primitive in the original

[15], where it is called *conscious update*. In [48], the believed public announcement of φ is called *manipulative update* with φ. The original proposal there is to view believed public announcement of φ as non-deterministic choice (as in action logics and PDL-style logics) between truthful public announcement of φ and lying public announcement of φ.

The interaction between announcement and other operators than belief we assume known [47]. It changes predictably in all logics we present. We will only vary the dynamic part of the logic. The axioms for belief after truthful and after lying public announcement are as follows.

Definition 4 (Axioms for belief after truthful and lying public announcement [48])

$$\begin{array}{lcl} [!\varphi]B_a\psi & \leftrightarrow & \varphi \to B_a[!\varphi]\psi \\ [¡\varphi]B_a\psi & \leftrightarrow & \neg\varphi \to B_a[!\varphi]\psi. \end{array}$$ ⊣

After the lying public announcement that φ, agent a believes that ψ, if and only if, on condition that φ is false, agent a believes that ψ after truthful public announcement that φ. To the credulous person who believes the lie, the lie appears to be the truth.

The reduction principle in [15, 24] for the interaction between belief and believed announcement is, in terms of our language, $[\varphi]B_a\psi \leftrightarrow B_a(\varphi \to [\varphi]\psi)$. This seems to have a different shape, as the modal belief operator binds the entire implication. But it is indeed valid in our semantics (a technical detail we did not find elsewhere).

Proposition 5 $\models [\varphi]B_a\psi \leftrightarrow B_a(\varphi \to [\varphi]\psi)$ ⊣

Proof

$$\begin{array}{lcl} [\varphi]B_a\psi & \Leftrightarrow & [!\varphi]B_a\psi \wedge [¡\varphi]B_a\psi \\ & \Leftrightarrow & (\varphi \to B_a[!\varphi]\psi) \wedge (\neg\varphi \to B_a[!\varphi]\psi) \\ & \Leftrightarrow & B_a[!\varphi]\psi \\ & \Leftrightarrow^{\sharp} & B_a(\varphi \to [!\varphi]\psi) \\ & \Leftrightarrow^{*} & B_a((\varphi \to [!\varphi]\psi) \wedge (\varphi \to [¡\varphi]\psi)) \\ & \Leftrightarrow & B_a(\varphi \to ([!\varphi]\psi \wedge [¡\varphi]\psi)) \\ & \Leftrightarrow & B_a(\varphi \to [\varphi]\psi) \end{array}$$

The $\sharp$-ed equivalence holds because from the semantics of truthful and lying announcement directly follows that $[!\varphi]\psi \leftrightarrow (\varphi \to [!\varphi]\psi)$. The *-ed equivalence holds because

$$\begin{array}{lcl} \varphi \to [¡\varphi]\psi & \Leftrightarrow & \varphi \to (\neg\varphi \to [¡\varphi]\psi) \\ & \Leftrightarrow & (\varphi \wedge \neg\varphi) \to [¡\varphi]\psi \\ & \Leftrightarrow & \bot \to [¡\varphi]\psi \\ & \Leftrightarrow & \top \end{array}$$

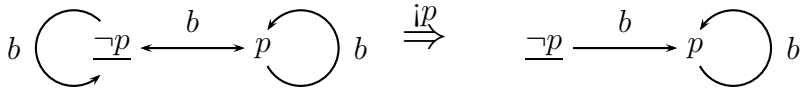
□

Proposition 6 ([48]) The axiomatization of the logic of truthful and lying public announcements is complete (for the model class $\mathcal{K}$ and for the model class $\mathcal{K}45$). ⊣

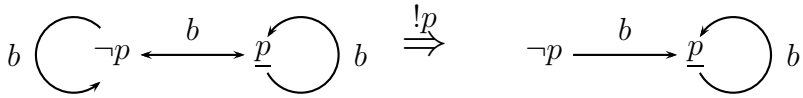
Proof Completeness is shown by a reduction argument. All formulas in $\mathcal{L}(!, ¡)$ are equivalent to formulas in $\mathcal{L}$ (epistemic logic). By means of equivalences such as in the axiom for belief after lying one can rewrite each formula to an equivalent one without announcement operators. □

The logic of truthful and lying public announcement satisfies the property 'substitution of equivalents' (substitution of formulas by logically equivalent formulas preserves validity), but the logic does not satisfy the property 'substitution of variables' (substitution of propositional variables by formulas preserves validity). For example, $[!p]p$ is valid but, clearly, $[!(p \wedge \neg B_a p)](p \wedge \neg B_a p)$ is invalid.

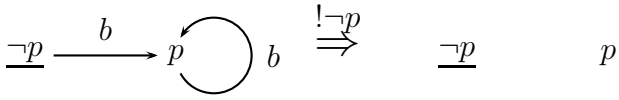
For an example of this logic we show the effect of truthful and lying announcement of p to an agent b in an epistemic model where b is uncertain whether p. The actual state (underlined) must be different in these models: when lying that p, p is in fact false, whereas when truthfully announcing that p, p is in fact true. For lying we get



whereas for truthfulness we get



The class $\mathcal{K}D45$ is not closed under truthful public announcements. Suppose agent b incorrectly believes p and processes the truthful public announcement that $\neg p$. We get



On the left, we have that $\neg p \wedge B_b p$ is true. The new information $!\neg p$ results in eliminating arrows to the p-state, and consequently agent b's accessibility relation becomes empty. He believes everything. The model on the left is in $\mathcal{K}D45$, but the model on the right is not in $\mathcal{K}D45$, because it is not serial.

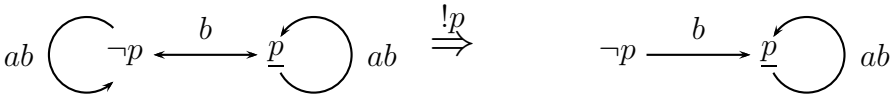
3 Agent announcements

In this section we model lying by an agent in the system (Ag) to a credulous addressee (Cred). We start by motivating our modelling choices, after which we formally introduce the language and semantics.

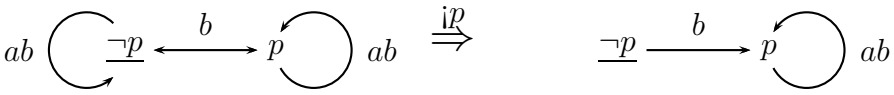
In the logic of lying and truthful public announcements, the announcing agent is an outside observer and is implicit. Therefore, it is also implicit that it believes that the announcement is false or that it believes that the announcement is true. In dynamic

epistemic logic, it is common to formalize 'agent a truthfully announces φ' as 'the outside observer truthfully announces $B_a\varphi$'. However, 'agent a lies that φ' cannot be modelled as 'the outside observer lies that $B_a\varphi$'. It is important to understand why this is the case and therefore we will explain it in detail.

Consider an epistemic state (with equivalence relations, encoding knowledge) where b does not know whether p, a knows whether p, and p is true. Agent a is in the position to tell b the truth about p. A truthful public announcement of B_ap (arrow elimination semantics, Definition 3) indeed simulates that a truthfully and publicly announces p.



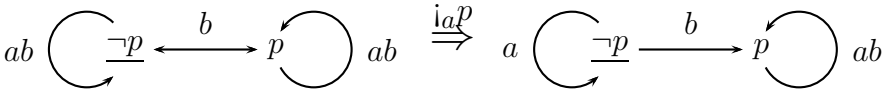
Given the same model, now suppose p is false, and that a lies that p. A lying public announcement of B_ap (it is lying, because it satisfies the required precondition $\neg B_ap$) does not result in the desired information state, because this makes agent a believe her own lie. And as she already knew $\neg p$, this makes a's beliefs inconsistent.



The problem here is that, no matter how complex the formula of the announcement, both lying and truthful public announcements will always similarly affect the (accessibility relations and thus) beliefs of the speaker and the addressee. This is because the announcements are public actions.

To model one agent lying to another agent we need a more complex form of model change than uniform restriction of the accessibility relation for all agents. We need to differentiate between the relation encoding the uncertainty of the speaker and the relation encoding the uncertainty of the addressee. This observation is the main reason for the logic of agent-to-agent lying that we propose in this section.

Consider the following adaptation of the running example. A lie by a to b that p should have the following effect:



After this lie we have that a still believes that $\neg p$, but that b believes that p. (We even have that b believes that a and b have common belief of p.) We satisfied the requirements of a lying agent announcement.

The precondition for agent a lying that φ is $B_a\neg\varphi$. The precondition for agent a truthfully announcing that φ is $B_a\varphi$. Another form of agent announcement is *bluffing*. You are bluffing that φ, if you say that φ but are uncertain whether φ. The precondition for agent a bluffing is therefore $\neg(B_a\varphi \vee B_a\neg\varphi)$. These are three mutually exclusive preconditions and they are also complete (the disjunction of all three is the triviality $\top$). No matter

what a actually believes, she will say φ, and the addressed agent b believes that he is told the truth, namely that a believes φ.

We recall that for public announcement, where the beliefs of the speaker were implicit, we had only two preconditions for announcing φ: φ and $\neg\varphi$, for truthful and lying public announcement. The 'third' would have been be $\neg(\varphi \vee \neg\varphi)$ but that is a contradiction, equivalent to $\bot$. The devil can lie, but the devil cannot bluff.

The accessibility relations in the example are treated differently in that: lying does not affect the accessibility relation of the speaker but only of that of the addressee. This implements the intuition that the speaker a can say φ *no matter what*, whether she believes what she says, believes the opposite, or is uncertain. Her announcement that φ does not inform the speaker about new facts (but typically changes her beliefs in the beliefs of the addressee). Whereas the addressee b takes φ to be the truth *no matter what*. We even require that the addressee takes *the speaker's belief in* φ to be the truth no matter what (otherwise an addressee already believing $\neg\varphi$ would not be able to infer that the speaker is mistaken). The generalization of the example is therefore that: the accessibility relation for the speaker a remains the same; states where φ is believed by speaker a remain accessible to addressee b; and states where φ is not believed by speaker a are no longer accessible to addressee b.

The requirements for agent lying are embodied by the following syntax and semantics.

Definition 7 The *language of agent announcement logic* is defined as

$$\mathcal{L}(!_a, ¡_a, !¡_a) \ \ni \ \varphi ::= p \mid \neg\varphi \mid (\varphi \wedge \psi) \mid B_a\varphi \mid [!_a\varphi]\psi \mid [¡_a\varphi]\psi \mid [!¡_a\varphi]\psi \qquad \dashv$$

The inductive constructs $[!_a\varphi]\psi$, $[¡_a\varphi]\psi$, and $[!¡_a\varphi]\psi$ stand for, respectively, a truthfully announces φ, a is lying that φ, and a is bluffing that φ; where agent a addresses all other agents b. We also define by abbreviation $[_a\varphi]\psi$ as $[!_a\varphi]\psi \wedge [¡_a\varphi]\psi \wedge [!¡_a\varphi]\psi$ (the dual may be more intuitive to the reader).

Definition 8

$$\begin{array}{lll} M, s \models [!_a\varphi]\psi & \text{iff} & M, s \models B_a\varphi \text{ implies } M_a^\varphi, s \models \psi \\ M, s \models [¡_a\varphi]\psi & \text{iff} & M, s \models B_a\neg\varphi \text{ implies } M_a^\varphi, s \models \psi \\ M, s \models [!¡_a\varphi]\psi & \text{iff} & M, s \models \neg(B_a\varphi \vee B_a\neg\varphi) \text{ implies } M_a^\varphi, s \models \psi \end{array}$$

where M_a^φ is as M except for the accessibility relation R' defined as (S is the domain of M, and $a \neq b$)

$$\begin{array}{lcl} R'_a & := & R_a \\ R'_b & := & R_b \cap \ (S \times [\![B_a\varphi]\!]_M). \end{array} \qquad \dashv$$

The principles for a making a truthful, lying, or bluffing announcement to b are as follows. The belief consequences for the speaker a are different from the belief consequences for the addressee(s) b.

Definition 9 (Axioms for the belief consequences of agent announcements)

$$
\begin{array}{lcll}
[!_a\varphi]B_b\psi & \leftrightarrow & B_a\varphi \rightarrow B_b[!_a\varphi]\psi & \\
[!_a\varphi]B_a\psi & \leftrightarrow & B_a\varphi \rightarrow B_a[_a\varphi]\psi & (ii) \\
[¡_a\varphi]B_b\psi & \leftrightarrow & B_a\neg\varphi \rightarrow B_b[!_a\varphi]\psi & (iii) \\
[¡_a\varphi]B_a\psi & \leftrightarrow & B_a\neg\varphi \rightarrow B_a[_a\varphi]\psi & \\
[!¡_a\varphi]B_b\psi & \leftrightarrow & \neg(B_a\varphi \vee B_a\neg\varphi) \rightarrow B_b[!_a\varphi]\psi & \\
[!¡_a\varphi]B_a\psi & \leftrightarrow & \neg(B_a\varphi \vee B_a\neg\varphi) \rightarrow B_a[_a\varphi]\psi & \dashv
\end{array}
$$

Axioms (ii) and (iii) are only outlined for illustrative purposes. To the addressee, the announcement always appears to be the truth. Therefore, after the lie that φ the addressee b believes ψ, iff, on the condition $B_a\neg\varphi$ that it is a lie, the addressee believes that ψ holds after the truthful announcement $B_a\varphi$ (iii). To grasp the intuition behind the axioms for the speaker we have to realize that the setting is for general modal accessibility, not for $\mathcal{K}45$ or $\mathcal{S}45$. Therefore $B_a\neg\varphi$ may be true in a given state, but not $B_aB_a\neg\varphi$: if a is lying she may not know that, she may consider states possible wherein, while announcing φ, she would have been truthful or she would have been bluffing. Therefore, after truthfully announcing φ the speaker a believes ψ, iff, on the condition $B_a\varphi$ that it is truthful, the speaker believes that ψ holds after *any* (either truthful or lying or bluffing) announcement $B_a\varphi$ (ii).[3]

For $\mathcal{K}45$ or $\mathcal{S}45$ agents we get a description closer to our intuition. This is because the liar then knows that it is lying, the bluffer that it is bluffing, and the truthful agent that it is truthful. Then we can even derive $[!_a\varphi]B_a\psi \leftrightarrow (B_a\varphi \rightarrow B_a[!_a\varphi]\psi)$ (ii), etc. for the other two cases. In case (ii) both the speaker and the addressee then believe that the consequences are those of truthful announcement.

Proposition 10 The axiomatization of the logic of agent announcements is sound. $\dashv$

Proof We show (ii) and (iii), the others are similar. The soundness of all axioms also, more succinctly, follows from modelling agent announcements as action models and using the reduction principle for 'belief after action', see Section 4.

$(ii \Rightarrow)$ Let $M, s \models [!_a\varphi]B_a\psi$, let $M, s \models B_a\varphi$, and let $(s,t) \in R_a$. We have to show that $M_a^\varphi, t \models \psi$. From $M, s \models [!_a\varphi]B_a\psi$ and $M, s \models B_a\varphi$ follows $M_a^\varphi, s \models B_a\psi$. From $(s,t) \in R_a$ (in M) and $R'_a = R_a$ (where R'_a in M_a^φ) follows $(s,t) \in R'_a$. From $M_a^\varphi, s \models B_a\psi$ and $(s,t) \in R'_a$ follows $M_a^\varphi, t \models \psi$.

$(ii \Leftarrow)$ Let $M, s \models B_a\varphi \rightarrow B_a[_a\varphi]\psi$. We now have to show that $M, s \models [!_a\varphi]B_a\psi$. To show that, let $M, s \models B_a\varphi$, and $(s,t) \in R'_a$. It remains to show that $M_a^\varphi, t \models \psi$. From $(s,t) \in R'_a$ and $R_a = R'_a$ follows $(s,t) \in R_a$. From $M, s \models B_a\varphi \rightarrow B_a[_a\varphi]\psi$, $M, s \models B_a\varphi$ and $(s,t) \in R_a$ follows $M, t \models [_a\varphi]\psi$, and therefore ($[_a\varphi]$ is unconditional!) $M_a^\varphi, t \models \psi$.

[3] Alternatively, we can take $[_a\varphi]$ as a primitive operator of the logic and define truthtelling, lying and bluffing by abbrevation as, respectively, $[!_a\varphi]\psi \leftrightarrow (B_a\varphi \rightarrow [_a\varphi]\psi)$, $[¡_a\varphi]\psi \leftrightarrow (B_a\neg\varphi \rightarrow [_a\varphi]\psi)$, and $[!¡_a\varphi]\psi \leftrightarrow (\neg(B_a\varphi \vee B_a\neg\varphi) \rightarrow [_a\varphi]\psi)$. We then need two axioms only, namely $[_a\varphi]B_a\psi \leftrightarrow B_a[_a\varphi]\psi$ and $[_a\varphi]B_b\psi \leftrightarrow B_b(B_a\varphi \rightarrow [_a\varphi]\psi)$. This suggestion is from Emiliano Lorini.

(*iii* $\Rightarrow$) Let $M, s \models [¡_a\varphi]B_b\psi$, let $M, s \models B_a\neg\varphi$, and let $(s,t) \in R_b$. There are two cases. If $M, t \not\models B_a\varphi$, then $M, t \models [!_a\varphi]\psi$ holds trivially. Otherwise, we have to show that $M^\varphi_a, t \models \psi$. From $(s,t) \in R_b$ and $M, t \models B_a\varphi$ follows $(s,t) \in R'_b$ (in M^φ_a). From $M, s \models [¡_a\varphi]B_b\psi$ and $M, s \models B_a\neg\varphi$ follows $M^\varphi_a, s \models B_b\psi$. From that and $(s,t) \in R'_b$ follows $M^\varphi_a, t \models \psi$.

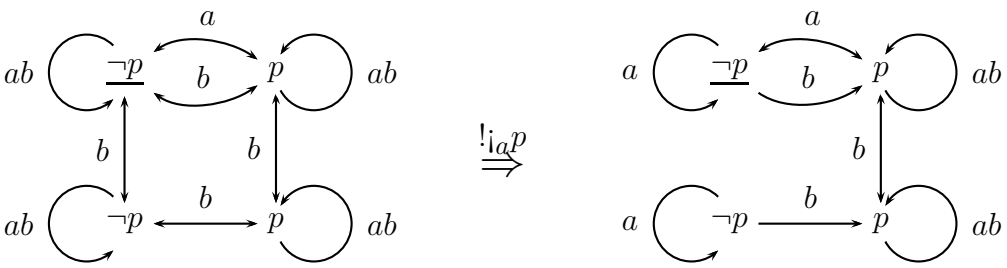
(*iii* $\Leftarrow$) Now assume $M, s \models B_a\neg\varphi \rightarrow B_b[!_a\varphi]\psi$. We have to show that $M, s \models [¡_a\varphi]B_b\psi$. To show that, let $M, s \models B_a\neg\varphi$, and $(s,t) \in R'_b$. We then also have $(s,t) \in R_b$. From that, the assumption and $M, s \models B_a\neg\varphi$ follows $M, t \models [!_a\varphi]\psi$. As $(s,t) \in R'_b$ we have that $M, t \models B_a\varphi$. From $M, t \models B_a\varphi$ and $M, t \models [!_a\varphi]\psi$ follows $M^\varphi_a, t \models \psi$, as required.

□

Proposition 11 The axiomatization of the logic of agent announcements is complete. ⊣

Proof Soundness was shown in Prop. 10. Just as in the previous logic (Prop. 6) completeness is shown by a reduction argument. All formulas in $\mathcal{L}(!_a, ¡_a, !_{¡_a})$ are equivalent to formulas in epistemic logic. In the axioms above, the announcement operator is (on the right) always pushed further down into any given formula. As before, the result holds for model classes $\mathcal{K}$, $\mathcal{K}45$ and $\mathcal{S}5$. An alternative, indirect, completeness proof is that agent announcement logic is action model logic for a given action model (Section 4). □

As an example illustrating the difference between a truthful, lying and bluffing agent announcement we present the following model wherein the addressee b, hearing the announcement of p by agent a, considers all three possible. In fact, a is bluffing, and a's announcement that p is false. After the announcement, b incorrectly believes that p, but a is still uncertain whether p. If the bottom-left state had been the actual state, a would have been lying that p, and if the bottom-right state had been the actual state, it would have been a truthful announcement by a that p. (In the figure we assume transitivity of the accessibility relation.)



Unbelievable announcements The axiomatization of the logic of agent announcements is incomplete for $\mathcal{K}D45$. This is because of the problem of unbelievable announcements. In Sections 5 and 6 we present alternative logics wherein believable announcements (announcements of φ to addressee b such that $\neg B_b\neg B_a\varphi$ is true) are treated differently from unbelievable announcements (such that $B_b\neg B_a\varphi$ is true). These logics are complete for class $\mathcal{K}D45$.

Public announcements are agent announcements Consider a depiction of an epistemic model. The outside observer is the guy or girl looking at the picture: you, the reader. She can see all different states. She has no uncertainty and her beliefs are correct. It is therefore that her truthful announcements are true and that her lying announcements are false. It is also therefore that 'truthful public announcement logic' is not a misnomer, it is indeed the logic of how to process new information that is true. We can model the outside observer as an agent gd, for '*g*od or the *d*evil'.

Proposition 12 Given an epistemic model M, let $gd \in A$ be an agent with an accessibility relation that is the identity on M. Then $M^{\varphi}_{gd} = M^{\varphi}$. Let φ, ψ not contain announcement operators, then $[!_{gd}\varphi]\psi$ is equivalent to $[!\varphi]\psi$, and $[¡_{gd}\varphi]\psi$ is equivalent to $[¡\varphi]\psi$. ⊣

Proof In Definition 8, the accessibility of the addressees is adjusted to $R'_b := R_b \cap (S \times [\![B_{gd}\varphi]\!]_M)$. As R_{gd} is the identity, $B_{gd}\varphi$ is equivalent to φ. So, $R'_b := R_b \cap (S \times [\![\varphi]\!]_M)$, as in Definition 8. □

Therefore, public announcements are a special form of agent announcements. The two logics complement each other. They do not really represent different approaches to lying. The outside observer can justifiably be considered another agent type Obs, a special case of agent type Ag, the agent making the announcement.

The postconditions of lying For a lying to b that p (a propositional variable), we required postcondition $B_b p$ (or stronger: $B_b B_a p$) and persistence of precondition $B_a \neg p$. (For formulas that are not variables such realizability objectives are hard to achieve in dynamic epistemic logics.) This is indeed the case.

Proposition 13 All valid are: (i) $[¡_a p]B_b B_a p$, (ii) $[¡_a p]B_a \neg p$, (iii) $[¡p]B_b p$, and (iv) $[¡p]\neg p$.⊣

Proof We show (i) and (ii); (iii) and (iv) are the special case for the omniscient agent gd for whom $B_{gd}\varphi \leftrightarrow \varphi$. Equivalence $[_a\varphi]p \leftrightarrow p$ in step ♯ holds because the postcondition is a variable and because $[_a\varphi]$ has precondition $\top$.

$$
\begin{array}{lcl}
[¡_a p]B_b B_a p & \Leftrightarrow & B_a\neg p \to B_b[!_a p]B_a p \\
& \Leftrightarrow & B_a\neg p \to B_b(B_a p \to B_a[_a p]p) \\
& \Leftrightarrow^{\sharp} & B_a\neg p \to B_b(B_a p \to B_a p) \\
& \Leftrightarrow & B_a\neg p \to B_b\top \\
& \Leftrightarrow & B_a\neg p \to \top \\
& \Leftrightarrow & \top
\end{array}
\qquad
\begin{array}{lcl}
[¡_a p]B_a\neg p & \Leftrightarrow & B_a\neg p \to B_a[_a p]\neg p \\
& \Leftrightarrow & B_a\neg p \to B_a\neg p \\
& \Leftrightarrow & \top
\end{array}
$$

□

In agent lying, $[¡_a p]B_b p$ is not valid! We can only achieve that the addressee believes that the speaker believes p, not that the addressee believes p. The source of the new information cannot be discarded. If the addressee already believed to the contrary, he will now believe that the speaker is mistaken: $B_b(\neg p \wedge B_a p)$, see below.

Lying about beliefs Agents may announce factual propositions (Boolean formulas) but also modal propositions, and thus be lying and bluffing about them. In the consecutive numbers riddle (Section 3.1) the announcements 'I know your number' and 'I do not know your number' are modal propositions, and the agents may be lying about those.

For our target agents, that satisfy introspection (so that $B_aB_a\varphi \leftrightarrow B_a\varphi$ and $B_a\neg B_a\varphi \leftrightarrow \neg B_a\varphi$ are validities), the distinction between bluffing and lying seems to become blurred. If I am uncertain whether p, I would be bluffing if I told you that p, but I would be lying if I told you that I believe that p. The announcement that p satisfies the precondition $\neg(B_ap \vee B_a\neg p)$. It is bluffing that p (it is $!_{¡a}p$). But the announcement that B_ap satisfies the precondition $B_a\neg B_ap$[4], the negation of the announcement. It is lying that B_ap (it is $¡_aB_ap$). We would prefer to call both bluffing, and that 'a announces that B_ap' is *strictly* 'a announces that p'. A general solution to avoid such ambiguity involves more than merely stripping a formula of an outer B_a operator: a announcing that B_aB_ap should also strictly be a announcing that p, and a announcing that $B_ap \wedge B_aq$ should strictly be a announcing that that $p \wedge q$. We need recursion.

Definition 14 (Strictly lying) Let a be an agent with consistent beliefs ($KD45$). An announcement by a that φ is *strict* iff φ is equivalent to $\top$ or φ is equivalent to $\bigvee_i \psi_i^a \wedge \bigwedge_i \neg B_a\neg\psi_i^a$ where all ψ_i^a are alternating disjunctive forms (see proof below) for agents other than a. An agent announcement $¡_a\varphi$ is *strictly lying* iff there is a φ' equivalent to φ such that φ' is strict and $¡_a\varphi'$ is a lying agent announcement. (Similarly for *strictly bluffing*.) $\dashv$

Proposition 15 For each $\varphi \in \mathcal{L}(!_a, ¡_a, !_{¡a})$ there is an equivalent $\psi \in \mathcal{L}$ that is strict. $\dashv$

Proof We first define the *alternating disjunctive form* (*adf*). An *adf* is a disjunction of formulas of shape $\psi_0 \wedge \bigwedge_{b\in A'}(B_b \bigvee_i \psi_i^b \wedge \bigwedge_i \neg B_b\neg\psi_i^b)$, where $\psi_0 \in \mathcal{L}$, $A' \subseteq A$, and where ψ_i^b is an *adf* for a group $A'' \subset A$ of agents other than b (the 'alternating' in alternating disjunctive forms). The indices i range over $1, \ldots, n$ ($n \geq 1$). We may assume that for each b some ψ_i^b is non-trivial (i.e., not equivalent to $\top$), because otherwise we can take $A' \setminus b$ in the conjunction above. Each formula in multi-agent $\mathcal{K}D45$ is equivalent to an *adf* [18].[5]

Now for the proof. Let $\varphi \in \mathcal{L}(!_a, ¡_a, !_{¡a})$. Determine an equivalent $\varphi' \in \mathcal{L}$ (i.e., rewrite φ into a multi-agent epistemic formula, without announcement operators). Let ψ be an *adf* equivalent to $B_a\varphi'$. It is elementary to see that ψ must be either equivalent to $\top$ or have the form $B_a \bigvee_i \psi_i^a \wedge \bigwedge_i \neg B_a\neg\psi_i^a$ and where some ψ_i^a is non-trivial. We now observe that

$$\begin{array}{rcl} \psi & \Leftrightarrow & B_a \bigvee_i \psi_i^a \wedge \bigwedge_i \neg B_a\neg\psi_i^a \\ & \Leftrightarrow & B_a \bigvee_i \psi_i^a \wedge \bigwedge_i B_a\neg B_a\neg\psi_i^a) \\ & \Leftrightarrow & B_a(\bigvee_i \psi_i^a \wedge \bigwedge_i \neg B_a\neg\psi_i^a) \end{array}$$

[4]From $\neg(B_ap \vee B_a\neg p)$ follows $\neg B_ap$, and with negative introspection we get $B_a\neg B_ap$.

[5]Reported in [18], an *adf* is a multi-agent generalization of the (single-agent) disjunctive form as in [31, p.35] (where it is called $\mathcal{S}5$ normal form), and a special case of the disjunctive form as in [11]. An *adf* contains no stacks of B_a operators without an intermediary B_b operator for another agent, i.e., if $B_a\chi$ is a subformula of an *adf* and $B_a\chi''$ is a subformula of χ then there is an agent $b \neq a$ and a χ' such that $B_b\chi'$ is a subformula of χ and $B_a\chi''$ is a subformula of χ'.

and that $\bigvee_i \psi_i^a \wedge \bigwedge_i \neg B_a \neg \psi_i^a$ is strict. □

Mistakes and lies There are two sorts of mistaken beliefs.

The first kind of mistaken belief is when I believe $p \to \neg q$ and I believe p, and, when questioned about my belief in q, confidently state that I believe that q. This is a mistaken belief. The formula $B_a q$ should be false, because $B_a \neg q$ is a deductive consequence of $B_a(p \to \neg q)$ and $B_a p$. To explain this sort of mistaken belief we need to model non-omniscient agents. We do not do that here.

The second kind of mistaken belief is when I believe that p but when in fact p is false. Even omniscient/logical agents can make such mistakes. This, we can address. It amounts to the truth of $\neg p \wedge B_a p$.

What is the difference between a lie and a mistake? I am lying that φ if I say φ and believe $\neg\varphi$ (independently from the truth of φ), whereas I am mistaken about φ (I mistakenly but truthfully announce φ) if I say φ and believe φ, but φ is false. (Let us rule out the dual form of mistake, when I am lying that φ while actually φ is true.) Let a be the speaker, then the precondition of lying that φ is $B_a \neg\varphi$ and the precondition of a mistaken truthful announcement that φ is $\neg\varphi \wedge B_a\varphi$. The speaker can therefore distinguish a lie from a mistake.

How about the addressee? Clearly, a $\mathcal{K}D45$ agent cannot distinguish another agent lying from being mistaken, because it can itself be mistaken about the announced proposition and about the speaker's beliefs of that proposition. We can then only observe that if a says φ given that $B_b(\neg\varphi \wedge B_a\varphi)$, then b believes a to be mistaken, whereas if a says φ given that $B_b B_a \neg\varphi$, then b believes a to be lying.

But if both knowledge and belief play a role, then the addressee can distinguish a lie from a mistake. A standard assumption in many games (therefore called 'fair games') and in many other distributed systems, is initial common knowledge of all agents of the uncertainty about the system. (I *know* — I can assume to know — that you do not know my card. If not, it's not fair, then you are cheating.) In such a system, beliefs that are not knowledge can only result from incorrect public updates (public announcements whose preconditions are not considered possible by some agents, such as lies) or private updates (telling other players about your cards in your absence, exchanging cards under the table out of view of the other players, etc.). For such an addressee, a mistake is that a says φ when $K_b(\neg\varphi \wedge B_a\varphi)$, whereas a lie is that a says φ when $K_b K_a \neg\varphi$. (For 'a knows that φ', the $S5$ notion, we write $K_a\varphi$.) We find this observation important enough to make it into a proposition.

Proposition 16 In fair games players can distinguish lies from mistakes. ⊣

The consecutive numbers riddle gives an example of such a lie and such a mistake.

3.1 Lying in the consecutive numbers riddle

> *Anne and Bill are each going to be told a natural number. Their numbers will be one apart. The numbers are now being whispered in their respective ears.*

They have common knowledge of this scenario. Suppose Anne is told 2 and Bill is told 3.

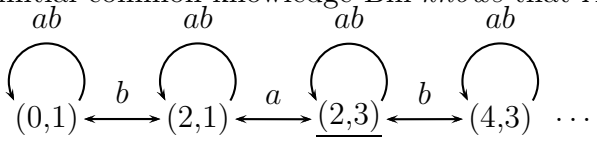
The following truthful conversation between Anne and Bill now takes place:

- *Anne: "I do not know your number."*
- *Bill: "I do not know your number."*
- *Anne: "I know your number."*
- *Bill: "I know your number."*

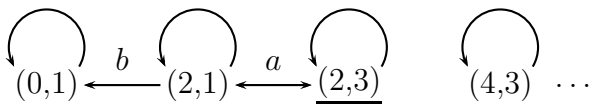
Explain why is this possible.

This consecutive numbers riddle [26] is an example of agent announcements. For a standard analysis, in terms of $S5$ knowledge, see [49]. We show two different scenarios for the consecutive numbers riddle with lying. With lying, the riddle involves a speaker feigning knowledge and consequently an addressee incorrectly believing something to be knowledge, so we move from knowledge to belief. In the communication, only the sentences 'I know your number' and 'I do not know your number' occur. Bluffing can therefore not be demonstrated: introspective agents believe their uncertainty and believe their beliefs.

The first scenario consists of Anne lying in her first announcement. Bill does not believe Anne's announcement: his accessibility relation from actual state $(2,3)$ has become empty. Bill's beliefs are therefore no longer consistent. Here, the analysis stops. We do not treat Bill's 'announcement' "That's a lie" as a permitted move in the game. On the assumption of initial common knowledge Bill *knows* that Anne was lying and not mistaken.



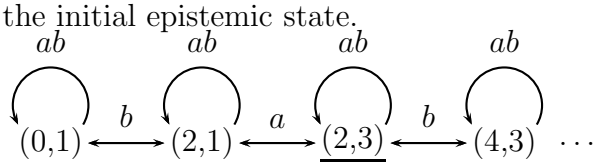
- Anne: "I know your number." *Anne is lying*



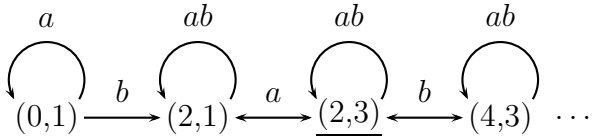
- Bill: "That's a lie."

In the second scenario Anne initially tells the truth, after which Bill is lying, resulting in Anne mistakenly concluding (and truthfully announcing) that she knows Bill's number: she believes it to be 1. This mistaken announcement by Anne is informative to Bill. He learns from it (correctly) that Anne's number is 2, something he didn't know before.

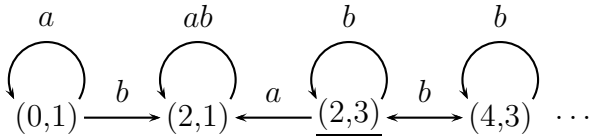
In the second scenario, Bill gets away with lying, because Anne considered it possible that he told the truth. Bill knows (believes correctly, and justifiably) that Anne's second announcement was a mistake and not a lie. The justification is, again, common knowledge in the initial epistemic state.



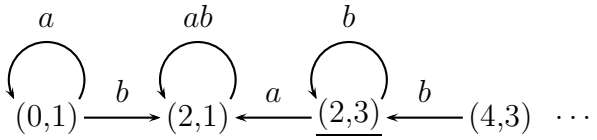
- Anne: "I do not know your number."



- Bill: "I know your number." *Bill is lying*



- Anne: "I know your number." *Anne is mistaken.*



4 Action models and lying

This section introduces action models, an alternative formalization of (truthful and lying) public announcements and agent announcements.

Whether I am truthfully announcing φ to you, or am lying that φ, or am bluffing that φ, to you it all appears as the truthful announcement of φ. Whereas to me, the speaker, they all appear to me as the announcement of φ. (They are truthful, lying, or bluffing, but unless in the $\mathcal{K}D45$ or $\mathcal{S}5$ case I cannot know that.) Different agents have different perspectives on this action. Action models [6] are a familiar way to formalize uncertainty about actions in the form of such different perspectives on 'what the real action is'. The action model for truthful public announcement (state elimination semantics) can be viewed as a singleton action model. This is well-known. We can view truthful and lying public announcement ('believed announcement', the arrow elimination semantics) as the different points, respectively, of a two-point action model. This is somewhat less well-known. (The modelling of believed announcements [15] as an action model was suggested in [42, 24].) We can also view truthful, lying and bluffing agent announcement as the respective different points of a three-point action model. These should be seen as alternative descriptions of the logics of lying in terms of a well-known framework. It has the additional advantage of independently validating the various axioms for lying and bluffing, and providing alternative completeness proofs (action model logic is complete). The following can be found in a standard introduction to action models [47].

An action model is a structure like a Kripke model but with a precondition function instead of a valuation function.

Definition 17 (Action model) An *action model* $\mathsf{M} = (\mathsf{S}, \mathsf{R}, \mathsf{pre})$ consists of a *domain* S of *actions*, an *accessibility function* $\mathsf{R} : A \to \mathcal{P}(\mathsf{S} \times \mathsf{S})$, where each R_a is an accessibility relation, and a *precondition function* $\mathsf{pre} : \mathsf{S} \to \mathcal{L}_X$, where $\mathcal{L}_X$ is a logical language. A pointed action model is an *epistemic action*. ⊣

A truthful public announcement of φ (state elimination semantics) is a singleton action model with precondition φ and with the single action accessible to all agents.

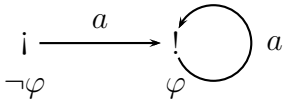
Performing an epistemic action in an epistemic state means computing their restricted modal product. This product encodes the new state of information.

Definition 18 (Update of an epistemic state with an action model) Given an epistemic state (M, s) where $M = (S, R, V)$ and an epistemic action (M, s) where $\mathsf{M} = (\mathsf{S}, \mathsf{R}, \mathsf{pre})$. Let $M, s \models \mathsf{pre}(\mathsf{s})$. The update $(M \otimes \mathsf{M}, (s, \mathsf{s}))$ is an epistemic state where $M \otimes \mathsf{M} = (S', R', V')$ such that

$$\begin{array}{lll} S' & = & \{(t, \mathsf{t}) \mid M, t \models \mathsf{pre}(\mathsf{t})\} \\ ((t, \mathsf{t}), (t', \mathsf{t}')) \in R'_a & \text{iff} & (t, t') \in R_a \text{ and } (\mathsf{t}, \mathsf{t}') \in \mathsf{R}_a \\ (t, \mathsf{t}) \in V'(p) & \text{iff} & t \in V(p) \end{array}$$ ⊣

The domain of $M \otimes \mathsf{M}$ is the product of the domains of M and M, but restricted to state/action pairs (t, t) such that $M, t \models \mathsf{pre}(\mathsf{t})$, i.e., such that the action can be executed in that state. (Note that a state in the resulting domain is no longer an abstract object, as before, but such a state/action-pair.) An agent considers a pair (t, t) possible in the next information state if she considered the previous state t possible, and the execution of action t in that state. And the valuations do not change after action execution.

Definition 19 (Action model for truthful and lying public announcement) The action model M' for truthful and lying public announcement that φ, where $\varphi \in \mathcal{L}(!, ¡)$, consists of two actions (suggestively) named ! and ¡, with $\mathsf{pre}(!) = \varphi$ and $\mathsf{pre}(¡) = \neg\varphi$, and such that for all agents only action ! is accessible. Truthful public announcement of φ is the epistemic action $(\mathsf{M}', !)$. Lying public announcement of φ is the epistemic action $(\mathsf{M}', ¡)$. The action model can be depicted as follows. Preconditions are shown below the actions.



⊣

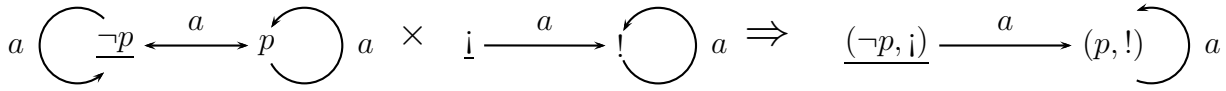
This terminology is not ambiguous in view of our earlier results, as we have the following.

Proposition 20 ([42, 24])

$$\begin{array}{lll} M, s \models [!\varphi]\psi & \text{iff} & M \otimes \mathsf{M}', (s, !) \models \psi \\ M, s \models [¡\varphi]\psi & \text{iff} & M \otimes \mathsf{M}', (s, ¡) \models \psi \ . \end{array}$$ ⊣

Proof Elementary, by induction on ψ. □

For an example, we show the execution of $(\mathsf{M}', ¡)$ for 'the outside observer is lying to a that p', in the epistemic state where the agent a is uncertain whether p but where p is false.



Consider this epistemic model (S, R, V). As before, the states in the epistemic model are named after their valuation, so that $S = \{\neg p, p\}$; the state named $\neg p$ is simply the state where p is false, etc. The modal product (S', R', V') consists of two states; $(\neg p, ¡) \in S'$ because $M, \neg p \models \mathsf{pre}(¡)$, as $\mathsf{pre}(¡) = \neg p$, and $(p, !) \in S'$ because $M, p \models p$. Then, $((\neg p, ¡), (p, !)) \in R'_a$ because $(\neg p, p) \in R_a$ (in M) and $(¡, !) \in \mathsf{R}_a$ (in the action model M'), etc. It is an artifact of the example that the shape of the action model is the shape of the next epistemic state. That is merely a consequence of the fact that the initial epistemic state has the universal accessibility relation for the agent on a domain of all valuations of the atoms occurring in the precondition (here: a domain of two states, for the two valuations of p).

Possibly more elegantly, we can also show the correspondence established in Proposition 20 from the perspective of a different logic. With an epistemic action (M, s) we can associate a dynamic modal operator $[\mathsf{M}, \mathsf{s}]$ in a logical language where an enumeration of action model frames is a parameter of the inductive language definition, apart from propositional variables P and agents A.

Definition 21 (Language of action model logic)

$$\mathcal{L}(\otimes) \ni \varphi ::= p \mid \neg\varphi \mid (\varphi \wedge \psi) \mid B_a\varphi \mid [\mathsf{M}, \mathsf{s}]\psi \qquad \dashv$$

The last clause is in fact inductive, if we realize that the preconditions of all actions in M, including s, are also of type formula. For example, truthful public announcement logic is an instantiation of that language for the singleton set of actions $\{!\}$, where we view '!' as an operation with two input formulae φ and ψ and that returns as output the announcement formula $[!\varphi]\psi$.

Definition 22 (Semantics of $[\mathsf{M}, \mathsf{s}]$)

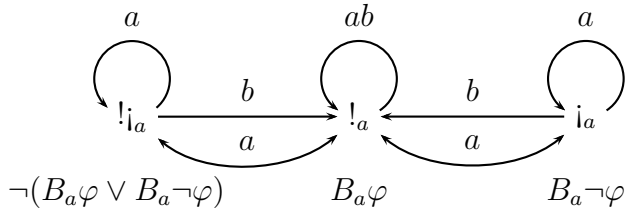
$$M, s \models [\mathsf{M}, \mathsf{s}]\psi \text{ iff } M, s \models \mathsf{pre}(\mathsf{s}) \text{ implies } M \otimes \mathsf{M}, (s, \mathsf{s}) \models \psi \qquad \dashv$$

Given the action model for truthful and lying public announcement that φ of Definition 19, but with $\varphi \in \mathcal{L}(\otimes)$, and given an inductively defined translation $tr : \mathcal{L}(\otimes) \to \mathcal{L}(!, ¡)$ with only nontrivial clauses $tr([\mathsf{M}', !]\psi) := [!tr(\varphi)]tr(\psi)$ and $tr([\mathsf{M}', ¡]\psi) := [¡tr(\varphi)]tr(\psi)$, the correspondence established in Proposition 20 can be viewed from the perspective of the dynamic modal operators for action models (where we suggestively execute the first step of the translation).

$$\begin{array}{lll} M, s \models [!tr(\varphi)]tr(\psi) & \text{iff} & M, s \models [\mathsf{M}', !]\psi \\ M, s \models [¡tr(\varphi)]tr(\psi) & \text{iff} & M, s \models [\mathsf{M}', ¡]\psi \end{array}$$

We now proceed with the presentation of action models for agent announcements.

Definition 23 (Action model for agent announcement) The action model M'' for agent announcement consists of three actions named $!¡_a$, $!_a$, and $¡_a$ with preconditions $\neg(B_a\varphi \vee B_a\neg\varphi)$, $B_a\varphi$, and $B_a\neg\varphi$, respectively, where $\varphi \in \mathcal{L}(!_a, ¡_a, !¡_a)$. The announcing agent a has universal access on the action model. To the other agents b only action $!_a$ is accessible. Agent a truthfully announcing φ to all other agents b is the epistemic action $(\mathsf{M}'', !_a)$ — with precondition $B_a\varphi$, therefore — and similarly lying and bluffing are the action models $(\mathsf{M}'', ¡_a)$ and $(\mathsf{M}'', !¡_a)$.



⊣

Again, we have the desired correspondence (and this can, again, be formulated in action model logic with an inductive translation).

Proposition 24

$$\begin{array}{lll} M, s \models [!_a\varphi]\psi & \text{iff} & M \otimes \mathsf{M}'', (s, !_a) \models \psi \\ M, s \models [¡_a\varphi]\psi & \text{iff} & M \otimes \mathsf{M}'', (s, ¡_a) \models \psi \\ M, s \models [!¡_a\varphi]\psi & \text{iff} & M \otimes \mathsf{M}'', (s, !¡_a) \models \psi \end{array}$$ ⊣

The action model representations validate the axioms for announcement and belief, for all versions shown; and they justify that these axioms form part of complete axiomatizations. These axioms are instantiations of the more general axiom for an epistemic action followed by a belief, in action model logic. This axiom ([6]) is

$$[\mathsf{M}, \mathsf{s}]B_a\psi \quad \leftrightarrow \quad \mathsf{pre}(\mathsf{s}) \rightarrow \bigwedge_{(\mathsf{s},\mathsf{t})\in \mathsf{R}_a} B_a[\mathsf{M}, \mathsf{t}]\psi$$

In other words, an agent believes ψ after a given action, if ψ holds after any action that is for a indistinguishable from it. For example, in the epistemic action $(\mathsf{M}', ¡)$, with $\mathsf{pre}(¡) = \varphi$, for lying public announcement that φ, ! is the only accessible action from action ¡, and we get

$$\begin{array}{lll} [\mathsf{M}', ¡]B_a\psi & \leftrightarrow & \mathsf{pre}(¡) \rightarrow B_a[\mathsf{M}', !]\psi, \quad \text{and so (Def. 4):} \\ [¡\varphi]B_a\psi & \leftrightarrow & \neg\varphi \rightarrow B_a[!\varphi]\psi \end{array}$$

For another example we derive the axioms (ii) and (iii) of Def. 9, see also Def. 23.

$$\begin{array}{lll} [\mathsf{M}'', !_a]B_a\psi & \leftrightarrow & \mathsf{pre}(!_a) \rightarrow (B_a[\mathsf{M}'', !_a]\psi \wedge B_a[\mathsf{M}'', ¡_a]\psi \wedge B_a[\mathsf{M}'', !¡_a]\psi), \quad \text{and so:} \\ [!_a\varphi]B_a\psi & \leftrightarrow & B_a\varphi \rightarrow B_a[_a\varphi]\psi \end{array}$$

$$\begin{array}{lll} [\mathsf{M}'', ¡_a]B_b\psi & \leftrightarrow & \mathsf{pre}(¡_a) \rightarrow B_b[\mathsf{M}'', !_a]\psi, \quad \text{and so:} \\ [¡_a\varphi]B_b\psi & \leftrightarrow & B_a\neg\varphi \rightarrow B_b[!_a\varphi]\psi \end{array}$$

In view of the identification discussed in this section, in the following we will continue to call any dynamic modal operator for belief change an epistemic action, and also standardly represent epistemic actions by their corresponding action models.

5 Unbelievable lies and skeptical agents

In this section we model lying to a skeptical addressee (Skep).

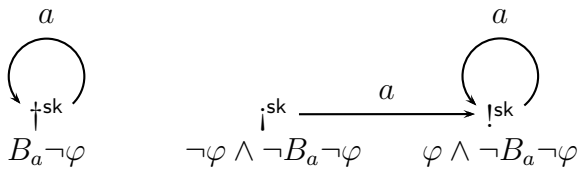
If I tell you φ and you already believe the opposite, accepting this information will make your beliefs inconsistent. This is not merely a problem for lying ('unbelievable lies') but for any form of information update. One way to preserve consistent beliefs is to reject new information if it is inconsistent with your beliefs. (The other way is to accept them, but to remove inconsistent prior beliefs. See Section 6.) Such agents may be called skeptical. In this section we adjust the logics of truthful and lying public announcement and of agent announcement to skeptical agents. This adjustment is elementary. As this topic is alive in the community of dynamic epistemic logicians we incorporate a review of the literature.

Consistency of beliefs is preserved iff seriality is preserved on epistemic models to interpret these beliefs. For the model classes $\mathcal{K}D45$ and $\mathcal{S}5$ that we target, this is the requirement that the class is closed under information update. The perspective on epistemic actions and action models from Section 4 is instructive. The class of $\mathcal{S}5$ epistemic models is closed under update with $\mathcal{S}5$ epistemic actions, such as truthful public announcements, but the class of $\mathcal{K}D45$ models is *not* closed under update with $\mathcal{K}D45$ epistemic actions (see the last paragraph of Section 2). The action models for truthful and lying public announcement, and for agent announcement, are $\mathcal{K}D45$.

Updates that preserve $\mathcal{K}D45$ have been investigated in [40, 4, 25]. Aucher [4] defines a language fragment that makes you go mad ('crazy formulas'). The idea is then to avoid that. Steiner [40] proposes that the agent does not incorporate the new information if it already believes the contrary. In that case, nothing happens. Otherwise, access to states where the information is not believed is eliminated, just as for believed public announcements. This solution to model unbelievable lies (and unbelievable truths!) is similarly proposed in the elegant [25], where it is called 'cautious update' — a suitable term. We will propose to call such agents *skeptical* instead of cautious.

We propose a three-point action model for *truthful and lying public announcement to skeptical agents*, with the semantics motivated by [40, 25].

Definition 25 (Public announcement to skeptical agents) The action model N for truthful and lying public announcement to skeptical agents consists of three actions named $!^{\mathsf{sk}}$, $¡^{\mathsf{sk}}$, and $\dagger^{\mathsf{sk}}$, with preconditions and accessibility relations (for all agents a) as follows.



For $(\mathsf{N}, !^{\mathsf{sk}})$, with $\mathsf{pre}(!^{\mathsf{sk}}) = \varphi \wedge \neg B_a \neg \varphi$, we write $!^{\mathsf{sk}}\varphi$; similarly, we write $¡^{\mathsf{sk}}\varphi$ for $(\mathsf{N}, ¡^{\mathsf{sk}})$ and $\dagger^{\mathsf{sk}}\varphi$ for $(\mathsf{N}, \dagger^{\mathsf{sk}})$. ⊣

The difference with the action model for truthful and lying public announcement is that the alternatives φ and $\neg\varphi$ now have an additional precondition $\neg B_a \neg \varphi$: the announce-

ment should be believable. In the action model there is a separate, disconnected, case for unbelievable announcements: precondition $B_a\neg\varphi$. For unbelievable announcements it does not matter whether φ is a lie or is the truth. The agent chooses to discard it either way. It is skeptical. (We chose the dagger/cross symbol † to indicate that the announcement is discarded / 'killed off'.)

Definition 26 (Axioms) The principles for public announcements to skeptical agents are:

$$\begin{array}{lcl}
[!^{\mathsf{sk}}\varphi]B_a\psi & \leftrightarrow & (\varphi \wedge \neg B_a\neg\varphi) \rightarrow B_a[!^{\mathsf{sk}}\varphi]\psi \\
[¡^{\mathsf{sk}}\varphi]B_a\psi & \leftrightarrow & (\neg\varphi \wedge \neg B_a\neg\varphi) \rightarrow B_a[!^{\mathsf{sk}}\varphi]\psi \\
[\dagger^{\mathsf{sk}}\varphi]B_a\psi & \leftrightarrow & B_a\neg\varphi \rightarrow B_a[\dagger^{\mathsf{sk}}\varphi]\psi
\end{array}$$

⊣

Proposition 27 The axiomatization of the logic for public announcements to skeptical agents is complete. ⊣

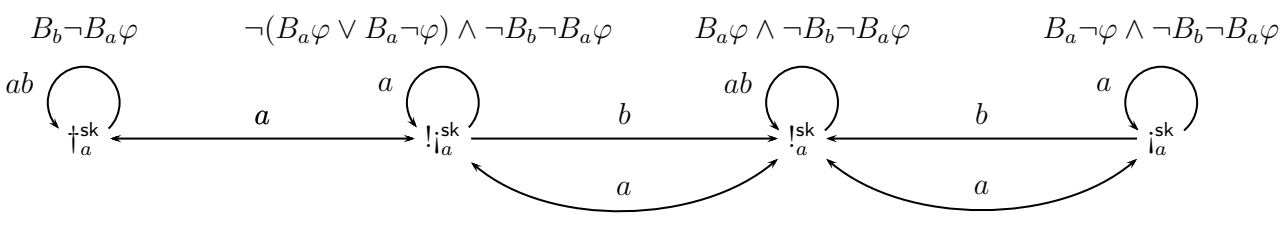
Proof Directly, from the embedding in action model logic. □

In the case of a single agent, the semantics of an unbelievable announcement of φ leaves the structure of the unbelievable part of the model unchanged. This means that agent a does not change its beliefs. An unbelievable public announcement does not make an informative difference for the skeptical agent.

If there are more agents, an announcement can be believable for one agent and unbelievable for another agent. This results in adaptations of the modelling. We consider such matters mere variations, and move on.

The analysis becomes more interesting for agent announcements, namely when the speaker is uncertain whether her lie will be believed by the addressee, as in the consecutive numbers riddle. In (non-skeptical) agent announcements, the addressee incorporates the new information as if the speaker believes the formula of the announcement, disregarding that the speaker may be lying or bluffing. The skeptical addressee *only* incorporates the new information if "the addressee considers it possible that *the speaker believes that* the announced formula is true." This may be different from "the addressee considers it possible that the announced formula is true." Even if the addressee already believes $\neg p$, he may consider it possible that the speaker is truthfully (in the sense of 'honestly') announcing p (i.e., believes p) and is not lying. The addressee b then merely concludes that the speaker a is mistaken in her truthful announcement of p: she says p because she believes p: B_ap. Whereas in fact $B_a\neg p$, because a was lying.

Definition 28 (Action model for agent announcement to skeptics) The action model N' for agent announcements from speaker a to skeptical addressee(s) b is as follows. Assume transitivity of accessibility relations.



For $(\mathsf{N}', ¡_a^{\mathsf{sk}})$ with precondition $B_a\neg\varphi \wedge \neg B_b\neg B_a\varphi$ we write $¡_a^{\mathsf{sk}}\varphi$, etc. (Similarly for public announcement to skeptical agents.) ⊣

This action model encodes the following perspectives. In case addressee b believes that speaker a does not believe φ, he is indifferent between the three alternatives truthful, lying, and bluffing announcement (by a) that φ. But in case b considers it possible that a believes φ, then, as before, he believes a to be truthful about φ. Whereas speaker a, who is saying φ no matter what, is both uncertain (in the case of general accessibility) whether she is lying, truthful, or bluffing, and also whether addressee b is inclined to believe her. Her accessibility relation is the universal relation.

We only give the reduction axioms for lying. In the second axiom we use $[{}_a^{\mathsf{sk}}\varphi]\psi$ as shorthand for $[!_a^{\mathsf{sk}}\varphi]\psi \wedge [¡_a^{\mathsf{sk}}\varphi]\psi \wedge [!¡_a^{\mathsf{sk}}\varphi]\psi \wedge [\dagger_a^{\mathsf{sk}}\varphi]\psi$. It can be said to formalize that the lying agent a is uncertain whether the lie is believable to the addressee b even when this is the case.

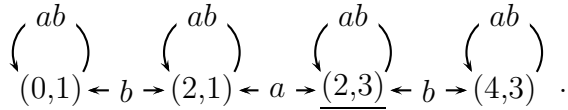
Definition 29 (Axioms) The axioms for a lying to a skeptical agent b, in case a is believed, are as follows.

$$\begin{array}{lcl} [¡_a^{\mathsf{sk}}\varphi]B_b\psi & \leftrightarrow & (B_a\neg\varphi \wedge \neg B_b\neg B_a\varphi) \rightarrow B_b[!_a^{\mathsf{sk}}\varphi]\psi \\ [¡_a^{\mathsf{sk}}\varphi]B_a\psi & \leftrightarrow & (B_a\neg\varphi \wedge \neg B_b\neg B_a\varphi) \rightarrow B_a[{}_a^{\mathsf{sk}}\varphi]\psi \end{array}$$ ⊣

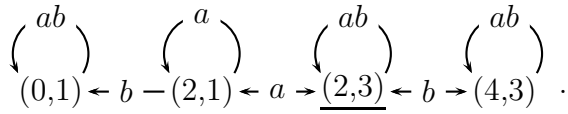
Proposition 30 The axiomatization for agent announcement to skeptical agents is complete (for classes $\mathcal{K}$, $\mathcal{K}45$, $\mathcal{K}D45$, $\mathcal{S}5$). ⊣

Proof For soundness: both axioms are instantiations of the general principle for action models: $[\mathsf{N}', ¡_a^{\mathsf{sk}}]B_b\psi \leftrightarrow (\mathsf{pre}(¡_a^{\mathsf{sk}}) \rightarrow B_b[\mathsf{N}', ¡_a^{\mathsf{sk}}]\psi$, and $[\mathsf{N}', ¡_a^{\mathsf{sk}}]B_a\psi \leftrightarrow (\mathsf{pre}(¡_a^{\mathsf{sk}}) \rightarrow \bigwedge_{\mathsf{s}} B_a[\mathsf{N}', \mathsf{s}]\psi$, where s ranges over the four points of N'. Completeness is again by a reduction argument. Note that this logic is also complete for class $\mathcal{K}D45$: it is seriality preserving. □

The logic of agent announcements to skeptical agents can be applied to model the consecutive numbers riddle. The reader can compare the transition below to the corresponding but slightly different transition in the first scenario in Section 3.1. We observe that, this time, both models are $\mathcal{K}D45$.



- Anne: "I know your number." *Anne is lying*



The skeptical Bill continues to be uncertain about Anne's number. He does not change his factual beliefs. But he does change his beliefs about Anne's beliefs. For example, after Anne's lie, Bill considers it possible that Anne considers it possible that he believed her. Informally, this gives Bill reason to believe that Anne has 2 and not 4. If she had 4, she would know that her lie would not be believed.

6 Lying and plausible belief

We recall the three different attitudes, presented in the introductory Section 1, towards incorporating announcements φ that contradict the beliefs $B_b\neg\varphi$ of an addressee b: (Cred) do it at the price of inconsistent beliefs (public announcements and agent announcements as treated in Sections 2 and 3), (Skep) reject the information (announcements to skeptical agents, treated in Section 5), and (Rev) accept the information by a consistency preserving process removing some old beliefs. This section is devoted to the third way. Going mad is too strong a response, not ever accepting new information seems too weak a response, we now discuss a solution in between. It involves distinguishing stronger from weaker beliefs when revising beliefs. To achieve that, we need to give epistemic models more structure: given a set of states all considered possible by an agent, it may consider some states more plausible than other states, and belief in φ can then be defined as the truth of φ in the most plausible states that are considered possible. We now have more options to change beliefs. We can change the sets of states considered possible by the agent, but we can also change the relative plausibility of states within that set.

Such approaches for belief change involving plausibility have been proposed in [3, 44, 42, 8]. How to model lying with plausibility models is summarily discussed in [7, 45] (as a dialogue, these are different contributions to the same volume), and also in [8, p.54].

We continue in the same vein as in the previous sections and present epistemic actions for plausible (truthful and lying) public announcement, and for plausible (truthful, lying, and bluffing) agent announcement, by an adjustment of the epistemic actions already shown. The epistemic action for plausible public lying is the one in [7, 45, 8]; the epistemic action for plausible agent lying applies the general setup of [8] to a specific (plausibility) action model. Thus it appears we can again present reduction axioms for belief change and complete axiomatizations for these logics. This is so, but it would involve not just modal belief operators $B_a\varphi$ but also conditional belief operators $B_a^\varphi\psi$ (among the states considered possible by a, the most plausible states that satisfy φ also satisfy ψ; so that $B_a\psi$ is $B_a^\top\psi$), and additional axioms for conditional belief. We refer to the cited works by Baltag and Smets [7, 8] for further details.

Definition 31 (Plausibility epistemic models) A *plausibility epistemic model* $M = (S, \sim, <, V)$ has one more parameter than an epistemic model, namely a plausibility function $< : A \to \mathcal{P}(S \times S)$. The accessibility relations for each agent are required to be equivalence relations $\sim_a$. The reflexive clausure of the restriction of $<_a$ to an equivalence class of $\sim_a$ is required to be a *well-preorder*[6]. ⊣

As the accessibility relations are equivalence relations, we write $\sim_a$ instead of R_a, as before. If $s <_a t$ we say that state s is considered more plausible than state t by agent a. The reflexive clausure of $<_a$ is $\leq_a$ and for ($s \leq_a t$ and $t \leq_a s$) we write $s =_a t$ (*equally plausible*).

In this setting we can distinguish knowledge from belief: the agent believes φ, iff φ is true in all most plausible equivalent states; and the agent knows φ, iff φ is true in all

[6] A well-preorder is a well-founded preorder. A preorder is a reflexive and transitive relation. A relation is well-founded if every non-empty subset of the domain has a minimal element.

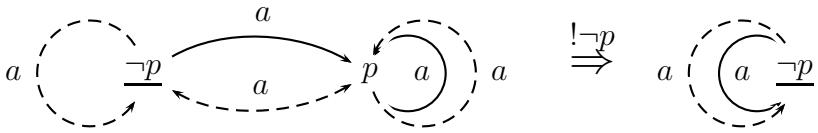
equivalent states.[7] We write $B_a\varphi$ for 'agent a believes φ', as before. There is a natural way to associate an accessibility relation with such belief, based on the equivalence relations and plausibility relations, and we can then define belief in φ as the truth of φ in all accessible states, as usual. Defined in this way, these relations for belief satisfy the $\mathcal{K}D45$ properties.

Definition 32 (Accessibility relation and semantics for belief) Given a plausibility epistemic model M with $\sim_a$ and $<_a$, the accessibility relation R_a is defined as:

$$(s,t) \in R_a \text{ iff } s \sim_a t \text{ and } t \leq_a t' \text{ for all } t' \text{ such that } s \sim_a t'$$

Given a plausibility epistemic model M and a state s in its domain, $M, s \models B_a\varphi$ iff $M, t \models \varphi$ for all t such that $(s,t) \in R_a$. ⊣

An information update that changes the knowledge of the agents, by way of changing the relations $\sim_a$, may also affect their beliefs, by way of changing the derived relations R_a. For example, suppose an agent a is uncertain whether p but considers it more likely that p is true than that p is false; without reason, because in fact p is false. The epistemic state is depicted below, on the left. The arrows in the equivalence relation are dashed lines. The arrows in the accessibility relation R_a for belief are solid lines, as before. In this state B_ap is true, because p is true in the more plausible state. Now agent a is presented with hard evidence that $\neg p$ (state elimination semantics, as in truthful public announcement logic). The state where p is false is eliminated from consideration. The only remaining state has become the most plausible state. In the epistemic state depicted below on the right, $B_a\neg p$ is true. Agent a has revised her belief that p into belief that $\neg p$. In this example, belief changes but knowledge also changes. There are other examples wherein only belief changes.



Belief revision consists of changing the plausibility order between states. This induces an order between deductively closed sets of formulas. The most plausible of these is the set of formulas that are believed. This belief revision is similar to AGM belief revision, seen as changing the plausibility order (partial or total order, or well-preorder) between deductively closed sets of formulas. The contraction that forms part of the revision is with respect to that order. Dynamic epistemic logics for belief revision were developed to model higher-order belief revision and iterated belief revision.

Just as epistemic actions, with underlying action models, generalize public announcements, plausibility epistemic actions generalize simpler forms of public plausibility updates. Accessibility relations for 'considering an action possible' are computed as in the case of plausibility epistemic models.

[7] The former represents weak belief and the latter true strong belief. There are yet other epistemic operators in this setting: safe belief, conditional belief, ... We restrict our presentation to (weak) belief.

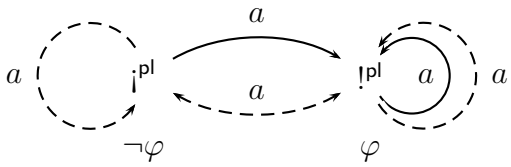
Definition 33 (Plausibility action model) A *plausibility action model* $\mathsf{M} = (\mathsf{S}, \approx, \prec, \mathsf{pre})$ consists of a *domain* S of *actions*, an *accessibility function* $\approx : A \to \mathcal{P}(\mathsf{S} \times \mathsf{S})$, where each $\approx_a$ is an equivalence relation, a plausibility function $\prec : A \to \mathcal{P}(\mathsf{S} \times \mathsf{S})$, and a *precondition function* $\mathsf{pre} : \mathsf{S} \to \mathcal{L}_X$, where $\mathcal{L}_X$ is a logical language. The restriction of $\prec_a$ to an equivalence class of $\approx_a$ must be a well-preorder. A pointed plausibility action model is a *plausibility epistemic action*. ⊣

Definition 34 (Update with a plausibility epistemic action) The update of a plausibility epistemic state with a plausibility epistemic action is computed as the update without plausibilities (see Definition 18), except for the update of the plausibility function that is defined as:

$$(s, \mathsf{s}) <_a (t, \mathsf{t}) \quad \text{iff} \quad \begin{array}{l} \mathsf{s} \prec_a \mathsf{t} \text{ or} \\ \mathsf{s} =_a \mathsf{t} \text{ and } s <_a t \end{array}$$ ⊣

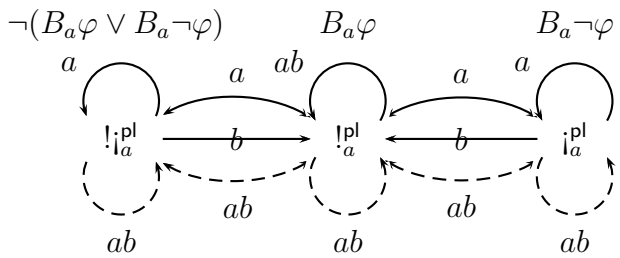
See [8] for details. Using these definitions, we propose the following plausibility epistemic actions for plausible (truthful and lying) public announcement that φ and for plausible (truthful, lying, bluffing) agent announcements that φ (by agent a to agent b). They are like the epistemic actions presented in Section 4, but enriched with plausibilities. As before, the pointed versions of these action models define the appropriate epistemic actions.

Definition 35 (Plausible (truthful and lying) public announcement) Plausible (truthful and lying) public announcements $!^{\mathsf{pl}}\varphi$ and $¡^{\mathsf{pl}}\varphi$ are the epistemic actions defined by corresponding points of the plausibility action model depicted as follows—where $\mathsf{pre}(!^{\mathsf{pl}}) = \varphi$ and $\mathsf{pre}(!^{\mathsf{pl}}) = \neg\varphi$.



⊣

Definition 36 (Plausible agent announcement) Plausible (truthful, lying, and bluffing) agent announcements $!^{\mathsf{pl}}_a\varphi$, $¡^{\mathsf{pl}}_a\varphi$ and $!¡^{\mathsf{pl}}_a\varphi$ are the epistemic actions defined by corresponding points of the plausibility action model consisting of points $!^{\mathsf{pl}}_a$, $¡^{\mathsf{pl}}_a$, and $!¡^{\mathsf{pl}}_a$, and such that $!^{\mathsf{pl}}_a \prec_b !¡^{\mathsf{pl}}_a \prec_b ¡^{\mathsf{pl}}_a$, with universal access for agents a and b, and with preconditions as visualized below.



⊣

Agent b's equivalence relation is the universal relation. He cannot exclude any of the three types of announcement. Agent b's accessibility relation for belief expresses that he considers it most plausible that a was telling the truth. It does not appear from the visualization that he considers it more plausible that a was bluffing than that she was lying. The accessibility relation is the same as in the action model for agent announcement in Section 4. Agent a's accessibility relation is the universal relation on the action model.

As in the previous sections, an addressee rather assumes being told the truth than being told a lie or being bluffed to. But, unlike in the previous sections, we can now also encode more-than-binary preferences between actions. As lying seems worse than bluffing, we make it least plausible to interpret an announcement as lying, more plausible that it is bluffing, and most plausible that it is truthful. That is about as charitable as we can be as an addressee. This seems to be in accordance with pragmatic practice.

Proposition 37 (Axiomatization and completeness) The logics of plausible public lying and plausible agent lying have a complete axiomatization. ⊣

Proof This follows from [8, p.51]. The so-called 'Derived Law of Action-Conditional-Belief' is a reduction axiom for the belief postconditions of plausibility epistemic actions. This axiom involves modalities for belief and for knowledge, as well as conditional modalities for belief and for knowledge. The completeness proof consists of rewriting each formula in the logic to an equivalent formula in the logic of conditional belief (where conditional knowledge is definable as conditional belief).[8] □

As an example of a plausible agent announcement, we take Anne's lying announcement "I know your number" in the first lying scenario in the consecutive numbers riddle. It suffices to refer to the depicted execution for skeptical agents in the final paragraph of Section 5. The accessibility relations for B_a and B_b are as there (after the lie, it remains the same for the speaker a, and not for addressee b). The equivalence relations for a and b are the same before and after the lie. It is a coincidence that in this example a skeptical announcement is a plausible announcement. This is a consequence of the fact that the initial belief accessibility relations are equivalence relations.

7 Intentions

The logical modelling of lies typically involves taking intention into account, where the most basic definition is that "a lies that p if a believes that p is false while a says to b that p is true, with the intention that b believes p" [30]. In this investigation we did not model the intentional part. The reason was that we model lying as an action, and that we model the intention that the addressee believes p as the realization of that belief after successfully

[8] Given a plausibility epistemic model M and a state s in its domain, define R_a^φ as: $(s,t) \in R_a^\varphi$ iff $[s \sim_a t,\ t \leq_a t'$ for all t' such that $s \sim_a t'$, and $M, t \models \varphi]$, and define the semantics of B_a^φ as: $M, s \models B_a^\varphi \psi$ iff $M, t \models \psi$ for all t such that $(s,t) \in R_a^\varphi$. We now have that $[¡^{\mathsf{pl}}\varphi]B_a\varphi \leftrightarrow (\neg\varphi \rightarrow B_a^\varphi[!^{\mathsf{pl}}\varphi]\psi)$. The axioms for plausible agent lying are more complex (they also involve conditional knowledge modalities).

executing the action of lying. Our analysis abstracts from explicitly modelling intention. In this section we explore the integration of intention and belief in action logics.

The modelling of intention and intention change in action logics and in dynamic modal logics has recently received some attention [35, 33, 46, 22]. These works contain well-known principles about the static interaction between knowledge and intention, such as that you are supposed to know your intentions ($I_a\varphi \to K_a I_a\varphi$), but for our purpose of modelling lying as an action their main interest is how they relate intention to action. One can distinguish approaches where intentions are (as standardly) *properties of formulas* and thus interpreted as sets of states [35, 46], from approaches where intentions are *properties of actions* [33, 22]. In [46], formulas are distinguished between good and bad via a choice function; you intend φ if you have a plan (that may consist of several actions) to realize a good φ. The different approaches are related: given choice between action x after which y holds and z after which w holds we can either say that the agent intends x and not z as a property of actions or we can say that the agent intends y and thus should choose x as the only means to realize that. The goal of modelling lying in dynamic epistemic logic with intentions then consists in adapting such (or yet other) approaches to the modelling of lying. A foothold there could be [36] wherein positive and negative outcomes (like good/bad formulas) define offensive and defensive dishonesty in the same sense as offensive and defensive lies in [38].

This direction of research would make the logical language richer and more expressive. However, we see alternatives.

In the first place, it seems to us that when 'a is lying that p' is described as 'a believes that p is false and a says to b that p with the intention that b believes p', this may sometimes *mean* 'a believes that p is false and after a says to b that p, b believes p'. The first formulation simply struggled with a way to describe a dynamic phenomenon in static terms. Maybe 'intention' was a way to point out the discrepancy between the inconsistent prior and posterior belief in p? Although $\neg B_a p$ and $B_b p$ are inconsistent, $\neg B_a p$ and $I_b B_a p$ need not be inconsistent. Otherwise, why is a *truthful* announcement of p to a never described in the literature as something 'intending' to make a believe p?

Secondly, consider generalizing the good/bad distinction between formulas to a totally ordered set of goals. We then have either the tools of AI planning or the tools of game theory at our disposition to continue such investigations. Consider games wherein two players make moves that are epistemic actions, and wherein their goals are formulas describing beliefs. Such imperfect information games with modal logic are presented in [1]. It would be interesting to define *lying games* in this setting, such that each announcement can be either the truth or a lie, and where a high negative payoff is associated with the detection of a lie.

8 Conclusions and further research

We presented various logics for an integrated treatment of lying, bluffing, and truthful announcements, where these are considered epistemic actions inducing transformations of

epistemic models. These logics abstract from the moral and intentional aspect of lying, and only consider fully omniscient agents and flawless and instantaneous information transmission. We presented versions of such logics that treat lies that contradict the beliefs of the addressee differently from those that don't, including a modelling involving plausible belief. Our main result are the various 'agent announcement' logics wherein one agent is lying to another agent and wherein both are explicitly modelled in the system.

There are limitations to our approach in view of the analysis and design of artificially intelligent agents. The intentional aspect of lying is not formalized, explicit agency (as in frameworks like ATL, ATEL, and STIT) is missing, bounded rationality is not modelled, and we also do not model degree of confidence in beliefs.

We envisage future research on common knowledge, complexity, and the liar paradox:

- If agent b believes p as the result of agent a announcing p to b, we not only have $B_b p$ (or $B_b B_a p$) but also $B_b C_{ab} p$: addressee b believes that he and the speaker a now commonly believe that p. Common belief/knowledge operators allow for more refined preconditions. A precondition for agent a successfully lying that φ to agent b seems:

$$B_a \neg\varphi \wedge \neg B_b \neg\varphi \wedge C_{ab}((B_a\varphi \vee B_a\neg\varphi) \wedge \neg(B_b\varphi \vee B_b\neg\varphi))$$

- The computational cost of lying seems a strong incentive against it. In a different context, the computational cost of insincere voting ('lying' about your preference) in social choice theory [10] is intractable in well-designed voting procedures, so that sincere voting is your best strategy. The complexity of model checking and satisfiability of the logics in this paper is unclear. It seems likely that the complexity of satisfiability of truthful and lying public announcement logic is in PSPACE, applying similar results in [28] that builds on [19]. Also, it is unclear how to measure the complexity of a dynamic epistemic communication protocol that may involve lies.

- A philosophical challenge is to model a liar's paradox in dynamic epistemic logic, but this is problematic. Promising progress towards that is reported in [27].

Acknowledgements

I thank the anonymous reviewers of the journal Synthese for their comments, and for their persistence. I gratefully acknowledge comments from Alexandru Baltag, Jan van Eijck, Patrick Girard, Barteld Kooi, Fenrong Liu, Emiliano Lorini, Yoram Moses, Eric Pacuit, Rohit Parikh, Ramanujam, Hans Rott, Sonja Smets, Rineke Verbrugge, and Yanjing Wang. As my work on lying has a long history (from 2008 onward), I am concerned I may have forgotten to credit yet others, for which my apologies. As the editor of the special issue in which this contribution appears, Rineke had many valuable comments in addition to those by the reviewers, and was as always extremely encouraging. Emiliano spared me the embarrassment of including unsound axioms (and an incorrect action model) in the axiomatization of agent announcement logic, for which infinite thanks.

References

[1] T. Ågotnes and H. van Ditmarsch. What will they say? - Public announcement games. *Synthese*, 179(S.1):57–85, 2011.

[2] C.E. Alchourrón, P. Gärdenfors, and D. Makinson. On the logic of theory change: Partial meet contraction and revision functions. *Journal of Symbolic Logic*, 50:510–530, 1985.

[3] G. Aucher. A combined system for update logic and belief revision. In *Proc. of 7th PRIMA*, pages 1–17. Springer, 2005. LNAI 3371.

[4] G. Aucher. Consistency preservation and crazy formulas in BMS. In *Proc. of 11th JELIA*, LNCS 5293, pages 21–33. Springer, 2008.

[5] A. Baltag. A logic for suspicious players: Epistemic actions and belief updates in games. *Bulletin of Economic Research*, 54(1):1–45, 2002.

[6] A. Baltag, L.S. Moss, and S. Solecki. The logic of public announcements, common knowledge, and private suspicions. In *Proc. of 7th TARK*, pages 43–56, 1998.

[7] A. Baltag and S. Smets. The logic of conditional doxastic actions. In *New Perspectives on Games and Interaction*, Texts in Logic and Games 4, pages 9–31. Amsterdam University Press, 2008.

[8] A. Baltag and S. Smets. A qualitative theory of dynamic interactive belief revision. In *Proc. of 7th LOFT*, Texts in Logic and Games 3, pages 13–60. Amsterdam University Press, 2008.

[9] S. Bok. *Lying: Moral Choice in Public and Private Life*. Random House, New York, 1978.

[10] V. Conitzer, J. Lang, and L. Xia. How hard is it to control sequential elections via the agenda? In *Proc. of 21st IJCAI*, pages 103–108. Morgan Kaufmann, 2009.

[11] G. d'Agostino and G. Lenzi. A note on bisimulation quantifiers and fixed points over transitive frames. *J. Log. Comput.*, 18(4):601–614, 2008.

[12] L. de Bruin and A. Newen. The developmental paradox of false belief understanding: A dual-system solution. *Synthese*, 2013. This issue.

[13] C. Dégremont, L. Kurzen, and J. Szymanik. On the tractability of comparing informational structures. *Synthese*, 2013. This issue.

[14] H.G. Frankfurt. *On Bullshit*. Princeton University Press, 2005.

[15] J.D. Gerbrandy and W. Groeneveld. Reasoning about information change. *Journal of Logic, Language, and Information*, 6:147–169, 1997.

[16] U. Gneezy. Deception: The role of consequences. *American Economic Review*, 95(1):384–394, 2005.

[17] J.L.K. Grimm and W.K. Grimm. *Kinder- und Hausmärchen*. Reimer, 1814. Volume 1 (1812) and Volume 2 (1814).

[18] J. Hales. Refinement quantifiers for logics of belief and knowledge. Honours Thesis, University of Western Australia, 2011.

[19] J.Y. Halpern and Y. Moses. A guide to completeness and complexity for modal logics of knowledge and belief. *Artificial Intelligence*, 54:319–379, 1992.

[20] J. Hintikka. *Knowledge and Belief*. Cornell University Press, Ithaca, NY, 1962.

[21] B. Hollebrandse, A. van Hout, and P. Hendriks. First and second-order false-belief reasoning: Does language support reasoning about the beliefs of others? *Synthese*, 2013. This issue.

[22] T.F. Icard, E. Pacuit, and Y. Shoham. Joint revision of beliefs and intention. In *Proceedings of 12th KR*. AAAI Press, 2010.

[23] N. Kartik, M. Ottaviani, and F. Squintani. Credulity, lies, and costly talk. *Journal of Economic Theory*, 134:93–116, 2006.

[24] B. Kooi. Expressivity and completeness for public update logics via reduction axioms. *Journal of Applied Non-Classical Logics*, 17(2):231–254, 2007.

[25] B. Kooi and B. Renne. Arrow update logic. *Review of Symbolic Logic*, 4(4):536–559, 2011.

[26] J.E. Littlewood. *A Mathematician's Miscellany*. Methuen and Company, 1953.

[27] F. Liu and Y. Wang. Reasoning about agent types and the hardest logic puzzle ever. *Minds and Machines*, 2012. Published online first – doi `10.1007/s11023-012-9287-x`.

[28] C. Lutz. Complexity and succinctness of public announcement logic. In *Proc. of the 5th AAMAS*, pages 137–144, 2006.

[29] J.E. Mahon. Two definitions of lying. *Journal of Applied Philosophy*, 22(2):21–230, 2006.

[30] J.E. Mahon. The definition of lying and deception. In *The Stanford Encyclopedia of Philosophy*, 2008. `http://plato.stanford.edu/archives/fall2008/entries/lying-definition/`.

[31] J.-J.Ch. Meyer and W. van der Hoek. *Epistemic Logic for AI and Computer Science*. Cambridge University Press, 1995. Cambridge Tracts in Theoretical Computer Science 41.

[32] J.A. Plaza. Logics of public communications. In *Proc. of the 4th ISMIS*, pages 201–216. Oak Ridge National Laboratory, 1989.

[33] B. Rodenhäuser. Intentions in interaction. Informal Proceedings of 7th LOFT, 2010.

[34] H. Rott. Der Wert der Wahrheit. In M. Mayer, editor, *Kulturen der Lüge*, pages 7–34. Böhlau-Verlag, Köln und Weimar, 2003.

[35] O. Roy. A dynamic-epistemic hybrid logic for intentions and information changes in strategic games. *Synthese*, 171(2):291–320, 2009.

[36] C. Sakama. Dishonest reasoning by abduction. In *Proceedings of 22nd IJCAI*, pages 1063–1064. IJCAI/AAAI, 2011.

[37] C. Sakama. Formal definitions of lying. Proc. of 14th TRUST, 2011.

[38] C. Sakama, M. Caminada, and A. Herzig. A logical account of lying. In *Proc. of JELIA 2010*, LNAI 6341, pages 286–299, 2010.

[39] F.A. Siegler. Lying. *American Philosophical Quarterly*, 3:128–136, 1966.

[40] D. Steiner. A system for consistency preserving belief change. In *Proc. of the ESSLLI Workshop on Rationality and Knowledge*, pages 133–144, 2006.

[41] R. Trivers. *The Folly of Fools – the logic of deceit and self-deception in human life.* Basic Books, 2011.

[42] J. van Benthem. Dynamic logic of belief revision. *Journal of Applied Non-Classical Logics*, 17(2):129–155, 2007.

[43] J.H. van der Velden. *Iedereen liegt maar ik niet.* Bruna, 2011.

[44] H. van Ditmarsch. Prolegomena to dynamic logic for belief revision. *Synthese (Knowledge, Rationality & Action)*, 147:229–275, 2005.

[45] H. van Ditmarsch. Comments on 'The logic of conditional doxastic actions'. In *New Perspectives on Games and Interaction*, Texts in Logic and Games 4, pages 33–44. Amsterdam University Press, 2008.

[46] H. van Ditmarsch, T. de Lima, and E. Lorini. Intention change via local assignments. In *Languages, Methodologies, and Development Tools for Multi-Agent Systems (Proceedings of 3rd LADS)*, pages 136–151. Springer, 2011. LNAI 6822.

[47] H. van Ditmarsch, W. van der Hoek, and B. Kooi. *Dynamic Epistemic Logic*, volume 337 of *Synthese Library.* Springer, 2007.

[48] H. van Ditmarsch, J. van Eijck, F. Sietsma, and Y. Wang. On the logic of lying. In *Games, Actions and Social Software*, LNCS 7010, pages 41–72. Springer, 2012.

[49] P. van Emde Boas, J. Groenendijk, and M. Stokhof. The Conway paradox: Its solution in an epistemic framework. In *Truth, Interpretation and Information: Selected Papers from the Third Amsterdam Colloquium*, pages 159–182. Foris Publications, Dordrecht, 1984.