Review of Analytic Philosophy
Vol. 4 No. 1 (2024)

Research article

A priori Belief Updates as a Method for Agent Self-recovery

Giorgio Cignarale and Roman Kuznets

Abstract:

Standard epistemic logic is concerned with describing agents' epistemic attitudes given the current set of alternatives the agents consider possible. While distributed systems can be (and often are) discussed without mentioning epistemics, it has been well established that epistemic phenomena lie at the heart of what agents, or processes, can and cannot do. Dynamic epistemic logic (DEL) aims to describe how epistemic attitudes of the agents/processes change based on the new information they receive, e.g., based on their observations of events and actions in a distributed system. In a broader philosophical view, this appeals to an *a posteriori* kind of reasoning, where agents update the set of alternatives considered possible based on their "experiences." Until recently, there was little incentive to formalize *a priori* reasoning, which plays a role in designing and maintaining distributed systems, e.g., in determining which states must be considered possible by agents in order to solve the distributed task at hand, and consequently in updating these states when unforeseen situations arise during runtime. With systems becoming more and more complex and large, the task of fixing design errors "on the fly" is shifted to individual agents, such as in the increasingly popular self-adaptive and self-organizing (SASO) systems. Rather than updating agents' *a posteriori beliefs*, this requires modifying their *a priori beliefs* about the system's global design and parameters. The goal of this paper is to provide a formalization of such *a priori* reasoning by using standard epistemic semantic tools, including Kripke models and DEL-style updates, and provide heuristics that would pave the way to streamlining this inherently nondeterministic and *ad hoc* process for SASO systems.

Keywords:
Distributed systems, Dynamic epistemic logic, *a priori* beliefs, Philosophy of computation, Self-adaptive and self-organizing systems

Received November 16, 2023; accepted August 30, 2024; online published October 9, 2024; https://doi.org/10.18494/SAM.RAP.2024.0021

1. Introduction

Epistemic logic (Hintikka, 1962) reasons about knowledge and/or beliefs of agents in a multiagent system. *Distributed systems* are a type of multiagent systems, with agents often referred to as *processes*, where these processes must coordinate their actions by communicating via either message passing or shared memory in order to accomplish some task (Lynch, 1996; Coulouris *et al*., 2011). Formal epistemic modeling (Fagin *et al*., 1995) has proved to be useful for characterizing the distributed system's evolution over time (Halpern & Moses, 1990), for deriving impossibility results (Moses & Tuttle, 1988), and for determining what processes can compute based on their local state in a given model (Ben-Zvi & Moses, 2014; Goren & Moses, 2020; Castañeda *et al*., 2022). The reason why epistemic analysis is always relevant in distributed scenarios was recently formalized as the Knowledge of Preconditions Principle by Moses (2016); it is so universal that it even applies to fault-tolerant distributed systems (Kuznets *et al*., 2019a,b; Fruzsa *et al*., 2021; Schlögl & Schmid, 2023). Dynamic epistemic logic (DEL) (Plaza, 1989; Gerbrandy & Groeneveld, 1997; Baltag *et al*., 1998; van Ditmarsch *et al*., 2007) provides tools for analyzing change in agents' epistemic attitudes in response to new information.

The epistemic analysis of distributed systems (Fagin *et al*., 1995) and of epistemic puzzles (van Ditmarsch & Kooi, 2015) routinely relies on agents' common knowledge of the model (Artemov, 2020). In effect, this is used to model agents' common *a priori* assumptions and enables agents to reason about (higher-order) reasoning of other agents. Note that this *a priori* knowledge differs from what agents learn through communication, independently of whether that communication is public, as in many epistemic puzzles, or private, as is more common in distributed systems. The information agents learn while playing out a puzzle or during a distributed system run is experience-based, *a posteriori* knowledge. Accordingly, dynamic epistemic logic implements knowledge change through model modifications that reorganize and shrink the already available possibilities, in contrast to the initial epistemic model creating the common space of these possibilities for the agents based on the puzzle description or distributed system specification.

Therefore, the system designer's task of creating a distributed system to given specifications can be viewed as creating common *a priori* knowledge for the agents. The role of *a priori* knowledge in the design cycle of distributed systems is analyzed by Cignarale *et al*. (2023). As argued there, mistakes in a system design would normally require the system designer to initiate the recovery process that amounts to the *a priori* knowledge update, better termed *a priori belief update*, due to the fallibility assumption inherent in the situation when system

behavior does not match the desired specifications. The fallibility here applies not only to the agents but also to the system designer who failed to account for some factors and/or behaviors. This picture of *a posteriori* experiences triggering *a priori* belief updates is largely based on a new philosophical approach to the *a priori* vs. *a posteriori* distinction proposed by Tahko (2008, 2011), where it is termed the *bootstrapping relation*. It is important to note that *a priori* knowledge/belief is characterized there as *modal*, i.e., relating to the set of possible states conceived by agents, and *fallible*, in the sense that that the actual world (or a faithful copy thereof) need not be among this set of possible states.

In the case of traditional distributed systems, aberrant *a posteriori* behavior prompts the system designer to trigger a new iteration of the design cycle: in a new design phase, she will re-adapt the *a priori* system assumptions so as to match the intended system behavior, redesign the affected parts of the implementation, and finally deploy and restart them, initializing the agents with updated *a priori* assumptions. Given the trend towards more and more complex and growing distributed systems, however, discovering and recovering from such design errors is increasingly becoming prohibitively costly: the ability to predict and/or monitor possible behaviors of such a system decreases exponentially, whereas the redesign costs increase dramatically.

This trend fueled the development of *self-adaptive and self-organizing systems* (*SASO systems*) (Berns & Ghosh, 2009; Tomforde *et al*., 2014) that have self-reflection and self-adaption capabilities. SASO systems allow processes to access and operate with their own representation of the system, which in turn enables them to update certain design assumptions on their own. In other words, in addition to *a posteriori* belief updates, which can be handled by traditional DEL methods, agents in SASO systems are expected to perform *a priori* belief updates with the goal of self-correcting their behavior, in response to situations not envisioned by the system designer.

This paper is devoted to the development of an epistemic formalization of self-recovery capabilities for agents by means of *a priori* belief updates, implemented in the form of DEL-inspired updates. We focus on self-recovery from an inconsistent state of beliefs and, following Plaza (1989), illustrate our methods using (variants of) standard epistemic puzzles such as the consecutive numbers and muddy children puzzle.

Let us first illustrate how faulty *a priori* assumptions can derail the progress, say, in the consecutive numbers puzzle and how human agents might still be able to find a solution by adjusting their *a priori* beliefs.

Example 1 (Consecutive numbers). Two agents a and b are privately told a natural number each. In addition, they are publicly told that the two numbers are consecutive (making it common knowledge). Suppose that a is told number 1 and b is told number 2. They are allowed to state whether they know the other's number or not, but not allowed to communicate their own number. Ordinarily, this instance of the puzzle is solved by the following dialog.

- a: I don't know your number.
- b: I don't know your number either.
- a: Now I know your number.
- b: Now I know yours too.

Here, the first statement by a is uninformative. The first statement by b makes it clear that b's number is not 0, which enables a to conclude that b's number must be 2. Since this determination would not have happened were a to hold number 3, now b can conclude that a's number is 1. The standard epistemic modeling of this example involves a Kripke model that each agent is supposed to build based on the rules of the puzzle in a way that makes this model commonly known to both agents. This commonality is based, in Lewis's telling, on the "suitable ancillary premises regarding [agents'] rationality, inductive standards, and background information" (Lewis, 1969, p. 53). While agents in epistemic puzzles are routinely considered to be perfect reasoners, which takes care of rationality and inductive standards, the question of background information is much less clear cut.

For our twist on the original formulation, imagine that, unknown to each other, a and b learned different definitions of natural numbers in school: a starts them from 1, while for b number 0 is also natural. What Lewis called belief in common background information and we call *a priori beliefs of the agent* is not shared by the agents preventing them from achieving common knowledge. Since natural numbers are routinely assumed (including in the formulation of the consecutive numbers puzzle) to be a well-defined object, each agent incorrectly believes that the other agent shares their definition of natural numbers. As is to be expected, the common knowledge of the model and of the situation at hand shatters, leading to the following possible conversation:

- a: I know your number.
- b: Wait, what? But that is impossible, unless... Ah, I see. Then I know your number too.

Here, agent a does not consider (1, 0) to be a legitimate pair, hence, (correctly) concludes that the numbers are (1, 2). Agent b, on the other hand, expects a to consider (1, 0) and, hence, does not understand a's reasoning. Indeed, according to b, if a had 1, he would have hesitated between (1, 0) and (1, 2), while if a had 3, he would have hesitated between (3, 2) and (3, 4). Thus, a's statement is incompatible with b's view of the world. In the proposed conversation, b does what is natural for a human reasoner: she questions her *a priori* assumptions, conceives that there is an alternative understanding of natural numbers as starting from 1, realizes that this is compatible with a's behavior, and updates her *a priori* beliefs (about a's *a priori* beliefs). Moreover, after this update, a's claim to knowing b's number is only compatible with a having 1: were a's number 3, the hesitation between (3, 2) and (3, 4) would have persisted. This update of b's *a priori* beliefs enables her to both explain the situation and arrive at the correct conclusion.

Standard (dynamic) epistemic reasoning, on the other hand, does not provide an adequate explanation. Epistemically, b's beliefs are supposed to become inconsistent. In fault-tolerant systems, this often translates to b considering herself and/or the other agent fully byzantine[1] and thus completely unreliable (Kuznets *et al*., 2019a), making the puzzle unsolvable.

Contributions. To the best of our knowledge, there is currently no epistemic modeling and analysis framework that explicitly considers *a priori* beliefs and their dynamics, and the issue of epistemic modeling of self-recovery has not been addressed in the literature. The goal of the present study is to model *a priori* belief updates epistemically, thus, providing agents with self-correcting capabilities. While the process remains highly nondeterministic in general, as there are multiple possible ways to resolve design mistakes, we provide some general guidelines and heuristics to guide possible future implementations of such self-recovery operations in SASO systems.

Paper organization. We provide some basic definitions of epistemic logic in Section 2. In Section 3 we discuss the guiding principles behind our epistemic approach to agents' *a priori* beliefs, highlighting their private nature (Section 3.1), their usefulness in conflict resolutions (Section 3.2), and their limitations in the higher-order case (Section 3.3). In Section 4, we introduce the novel *a priori* belief update mechanism for self-recovery (Section 4.1), and we

1 Fully byzantine agents can deviate arbitrarily from their original protocols as well as have false memories and erroneous perceptions. Moreover, their goals might not be known to correct agents. As such, they are attributed potentially inconsistent epistemic attitudes and cannot be trusted by correct agents.

show its fruitfulness in variants of popular epistemic puzzles (Section 4.2), including *a priori* belief updates triggered by public announcements (Section 4.3), simultaneous (and independent) *a priori* updates by several agents (Section 4.4), and *a priori* updates (Section 4.5) that do not achieve the desired goals. Some heuristics for the update synthesis problem are provided in Section 4.6. Section 5 lists some useful properties of *a priori* belief updates, and finally, conclusions are provided in Section 6.

2. Formal Preliminaries

Throughout the paper, we assume a fixed finite set $\mathcal{A} \neq \varnothing$ of *agents*. As is common, we employ Kripke semantics to reason about agents' epistemic states. Since we are interested in the dynamics of belief chance, we use PAL, the logic of public announcements as the simplest version of DEL.

Definition 2 (Language). The epistemic language with public announcements for agents from $\mathcal{A}$ is defined by

$$\varphi ::= p \mid \neg\varphi \mid (\varphi \wedge \varphi) \mid B_i\varphi \mid [\varphi]\varphi$$

where $p \in Prop$ is an *atomic proposition* (or simply *atom*) and $i \in \mathcal{A}$.

Definition 3 (Kripke models). A *Kripke model* $\mathcal{M} = \langle W, R, V \rangle$ is a triple comprising

- a set of *possible worlds* $W \neq \varnothing$;
- A function $R: \mathcal{A} \to 2^{W \times W}$ that assigns to each agent $i \in \mathcal{A}$ a binary relation $R(i) \subseteq W \times W$, called *accessibility relation*, which is usually denoted R_i instead of $R(i)$;
- a *valuation function* $V : Prop \to 2^W$ that assigns to each atom $p \in Prop$ a set $V(p) \subseteq W$ of worlds where p holds.

We use the notation $R_i(u) := \{v \in W \mid uR_iv\}$ for the set of all worlds that agent i considers possible in a world $u \in W$. A *pointed Kripke model* is a pair $(\mathcal{M}, v)$ where $\mathcal{M}$ is a Kripke model and $v \in W$ represents the *real* (or *actual*) *world*.

Definition 4 (Truth). *Truth of a formula φ in a world w of a Kripke model* $\mathcal{M} = \langle W, R, V \rangle$ is defined recursively: for atoms, $\mathcal{M}, w \vDash p$ iff $w \in V(p)$; boolean connectives behave classically; $\mathcal{M}, w \vDash B_i\varphi$ iff $\mathcal{M}, u \vDash \varphi$ for all $u \in R_i(w)$; finally, $\mathcal{M}, w \vDash [\varphi]\psi$ iff either it is not the case that $\mathcal{M}, w \vDash \varphi$ or it is the case that both $\mathcal{M}, w \vDash \varphi$ and $\mathcal{M} \mid \varphi, w \vDash \psi$, where $\mathcal{M} \mid \varphi := \langle W', R', V' \rangle$ is defined as

- $W' := \{u \in W \mid \mathcal{M}, u \vDash \varphi\}$ (note that $w \in W'$ whenever the clause with $\mathcal{M} \mid \varphi$ is used);
- $R'_i := R_i \cap (W' \times W')$ for each $i \in \mathcal{A}$;
- $V'(p) := V(p) \cap W'$ for each $p \in Prop$.

Strictly speaking, $\mathcal{M}, w \vDash \varphi$ and $\mathcal{M} \mid \varphi$ are defined by mutual recursion on φ, as is standard in DEL. Formula φ is *false at world w*, denoted $\mathcal{M}, w \nvDash \varphi$, iff it is not true at w.

Definition 5 (Binary relation types). A binary relation $R \subseteq W \times W$ is called

- *reflexive* iff wRw for all $w \in W$;
- *transitive* iff for all $w, v, u \in W$ we have wRu whenever wRv and vRu;
- *euclidean* iff for all $w, v, u \in W$ we have vRu whenever wRv and wRu;
- *symmetric* iff for all $w, v \in W$ we have vRw whenever wRv;
- an *equivalence relation* iff it is reflexive, transitive, and euclidean;
- a *partial equivalence relation* iff it is transitive and symmetric;
- an *introspective relation* iff it is transitive and euclidean.

Proposition 6. *An equivalence relation is also symmetric, hence, introspective and a partial equivalence relation. A partial equivalence relation is also euclidean, hence, introspective.*

Definition 7 (Model types). We call a Kripke model $\langle W, R, V \rangle$

- *epistemic* iff all R_a are equivalence relations;
- *introspective* iff all R_a are introspective;
- *quasi-epistemic* iff all R_a are partial equivalence relations.

Proposition 8. *An equivalence relation $R_a \subseteq W \times W$ partitions W into equivalence classes, or* a-clusters*, such that for each equivalence class $E \subseteq W$, we have uR_av for any $u, v \in E$ and neither uR_au' nor $u'R_au$ for any $u \in E$ and $u' \in W \setminus E$. A partial equivalence relation produces a similar partition but of a subset $W \setminus I$, where $I \subseteq W$ consists of isolated worlds, i.e., neither iR_aw nor wR_ai for any $w \in W$ and $i \in I$.*

Definition 9 (Composition and iteration of relations). For any binary relations $Q, Q' \subseteq W \times W$ on a set W, their *composition*

$$Q \circ Q' := \{(w, v) \in W \times W \mid (\exists u \in W)(wQu \text{ and } uQ'v\}.$$

Let Q^k for $k \geq 0$ be defined recursively by $Q^0 := \{(w, w) \mid w \in W\}$ and $Q^{k+1} := Q \circ Q^k$.

Definition 10 (Mutual and common accessibility). For a Kripke model $\langle W, R, V\rangle$ we define the

- *mutual accessibility relation* $R_{\mathcal{A}} := \bigcup_{a \in \mathcal{A}} R_a$ that corresponds to the mutual belief of all agents;
- *common accessibility relation* $R^*_{\mathcal{A}} := \bigcup_{k=1}^{\infty} R^k_{\mathcal{A}}$ that corresponds to the common belief of all agents.

We introduce the notion of agent i's submodel, which is similar to the notion of a rooted generated submodel (Chagrov & Zakharyaschev, 1997), except that the upward closure is with respect to chains of accessibility arrows that start from R_i.

Definition 11 (Agent's submodel). Let $(\mathcal{M}, v)$ with $\mathcal{M} = \langle W, R, V\rangle$ be a pointed model and $i \in \mathcal{A}$ be an agent with *consistent beliefs*, i.e., such that $R_i(v) \neq \varnothing$. The *submodel accessible by i at* $(\mathcal{M}, v)$, or *i's part/submodel of* $(\mathcal{M}, v)$ is the Kripke model $\mathcal{M}^i_v := \langle W', R', V'\rangle$ such that

- $W' := \left(R_i \circ R^*_{\mathcal{A}}\right)(v)$ is the set of all worlds that are common-belief accessible from any point in $R_i(v)$, including all worlds from $R_i(v)$;
- $R'_j := R_j \cap (W' \times W')$ for each $j \in \mathcal{A}$;
- $V'(p) := V(p) \cap W'$ for each $p \in Prop$.

For reasons of uniformity, we sometimes abuse the terminology and say that i's submodel is empty when $R_i(v) = \emptyset$.

It is common in epistemic modeling to homogeneously treat all worlds as equally possible. At the same time, to represent a specific scenario, the *actual world* is treated differently when considering pointed models. In epistemic scenarios, to ensure the factivity of knowledge, all agents are assumed to consider this actual world possible. However, in *fallibilistic* scenarios, this assumption must be relaxed to allow for false (or inconsistent) beliefs. We go one step further by exploiting the fact that factive beliefs can be achieved without making the actual world possible for any of the agents. Indeed, if an agent considers a duplicate of the actual world[2] possible, then the beliefs of this agent are factive. To implement our *a priori* belief updates, we abandon the homogeneity of Kripke models and divide worlds into functionally separate categories with the actual world being one such category.

While the actual world in a pointed model is objectively different from all other possible worlds, given that *a priori* belief updates are supposed to be fully private, it makes sense to extend this distinction to the subjective view of the agents: in the actual world, for each agent i, there is a subjective difference between worlds i actually considers possible, which we call *actually possible worlds*, and worlds i considers but has enough information to discount as actual possibilities, which we call *virtually possible worlds*. For instance, in the muddy children puzzle, children who see muddy faces of other children do not actually consider it possible that everyone is clean but do consider the possible world where everyone is clean. Such far-fetched possible worlds are, in fact, necessary to compute higher-order beliefs, i.e., imagine how other children think.

In our design of *a priori* belief updates, we functionally separate actually possible worlds from virtually possible worlds for the simple reason that information about them is extracted from different sources. While the actually possible worlds, which govern simple beliefs, can be adapted directly by the agent in question based on its own reasoning and observations, the beliefs of other agents, which are governed by virtually possible worlds, require collating the newly conceived possibilities with the pre-existing pattern of their beliefs. Thus, in deviation from how puzzles are typically modeled, we often partition Kripke models into several disjoint parts: the singleton actual world, the actually possible worlds for an agent i, and the virtually possible worlds for agent i. This separation enables us, among other things, to modify agent i's beliefs about agent j's beliefs without affecting the (actual) beliefs of agent j.

2 To be precise, a world bisimilar to the actual world.

This property ensures the minimal change of our *a priori* updates, a property commonly desirable in belief revision.

3. *A priori* Beliefs in Epistemic Logic

It is important to differentiate between *a priori* beliefs that can and cannot be represented syntactically. We call the former *explicit a priori beliefs* and define them as beliefs that can be represented by one epistemic formula in the object language. Thus, beliefs that can be represented by a finite set of formulas are explicit because the conjunction of the set provides an equivalent description. However, not all *a priori* beliefs are explicit. For instance, factivity of everyone's knowledge is usually described as either a frame property (reflexivity) or an infinite set of formulas ($B_a\varphi \to \varphi$ for all a and φ) but cannot be reduced to the truth of one formula only. We call such *a priori* beliefs *implicit*.

In so far as epistemic puzzles and distributed systems deal with *a priori* beliefs (usually without calling them that), it is done by providing the initial Kripke models that agents proceed to modify based on the information they receive. In other words, the implicit form of *a priori* beliefs is dominant, and converting it into an explicit representation may well be an unrealistic task even for simple epistemic puzzles (Artemov, 2022).

3.1 Privacy of *a priori* beliefs

In the typical modeling of epistemic puzzles, agents are assumed to have common factual *a priori* beliefs, as represented by the commonly known epistemic model (Artemov, 2020). This ensures the homogeneity of reasoning by all agents and creates no problems as long as agents have no need to modify their *a priori* beliefs. At the same time, for our setting, it is unreasonable to assume that the internal reasoning process leading one agent to modify its *a priori* beliefs would be noticeable by other agents, let alone be commonly known among them, even in cases where they start with commonly known priors.

Hence, we abandon the assumption of the common knowledge of the model and treat each agent as having its own private *a priori* beliefs represented by the submodel of this agent (cf. Definition 11). If submodels of several agents overlap, this is treated as a coincidence, especially in view of the fact that each agent, being only aware of its own submodel, would be oblivious to any such overlaps.

Note that if the initial model is a connected epistemic model, the submodel of each agent is the whole model. We interpret this situation as each agent believing to have the common

model that all agents share, but leave the possibility of one or several agents being wrong, typically as a result of a later *a priori* update.

To summarize, (i) agents' reasoning is represented by a pointed Kripke model $(\mathcal{M}, v)$, but instead of assuming common knowledge of $(\mathcal{M}, v)$, an agent i is only guaranteed to know i's submodel $\mathcal{M}_v^i$, which may or may not be different from submodels $\mathcal{M}_v^j$ visible by other agents j; and (ii) in addition to the standard public update mechanism for public announcements, we employ a *private* model update mechanism for *a priori* belief updates of an agent i that results in i's submodel becoming disjoint from other agents' submodels after i performs an *a priori* update.

3.2 *A priori* belief updates as conflict resolution

We still maintain agents' reasoning within a pointed Kripke model, i.e., each agent i is still logically omniscient w.r.t. the *a posteriori* information contained in i's submodel $\mathcal{M}_v^i$, of the given pointed Kripke model $(\mathcal{M}, v)$. The observations that agent i is making during the run may come into conflict with its *a priori* beliefs, which would manifest as $R_i(v) = \emptyset$, causing i's beliefs to become inconsistent. As long as such contradiction does not arise, i continues *a posteriori* reasoning in the standard epistemic manner or DEL manner for publicly announced information. In light of the Knowledge of Preconditions Principle (Moses, 2016), which states that $Bi\varphi$ is a precondition for an action whenever φ is, inconsistent beliefs make it impossible for the agent to act *correctly*. Indeed, even if the agent is supposed to choose between mutually exclusive actions A if φ holds or B otherwise, the inconsistent agent would have to perform both due to believing everything, including both φ and $\neg\varphi$. This provides a good incentive for the agent to reexamine its *a priori* beliefs and try updating them in such a way as to restore their consistency. Semantically, this is achieved by i creating a new submodel $\mathcal{M}_v^i$ (disjoint from the existing model) for itself in place of the current empty one. This is exactly the desired functionality for SASO systems: if an agent finds out during runtime that its *a priori* beliefs are inadequate, i.e., if an *a posteriori* epistemic update of an agent violates some of its *a priori* beliefs, the agent may initiate *a priori* reasoning, aiming at finding new *a priori* beliefs that comply with the *a posteriori* epistemic status. If an appropriate solution is found, an *a priori belief update* is privately invoked for installing a new private submodel within the current Kripke model.

Needless to say, there are many conceivable ways for resolving conflicts arising from inconsistencies in SASO systems, mainly because, in accordance with the Duhem–Quine thesis (Quine, 1951), it is not always possible to isolate the specific hypothesis (*a priori* belief

in our case) as the culprit for that inconsistency, even for a restricted set of explicit *a priori* beliefs. Thus, we will not try to provide deterministic algorithms for choosing an appropriate *a priori* update. At the same time, we do not leave this process completely *ad hoc*. A new submodel is constructed from several building blocks representing the agent's new guesses regarding the actually possible worlds, the virtually possible worlds, and their relationship, as described in Section 4.[3]

3.3 Higher-order *a priori* reasoning

The question of higher-order belief updates remains outside the scope of this paper. In other words, if an agent a detects some inconsistency in the beliefs of another agent b, we do not force a to try and guesstimate an *a priori* belief update by b, even if agent b is, in a's estimation, likely to perform it. Since *a priori* updates are likely to be nondeterministic, there is little reason to assume that a would be able to exactly match the thought processes of another agent b. In Lewis's terminology, once the belief in shared "inductive standards and background information" fails, so does the ground for a common understanding of the situation. In practice, this means that a would generally lose the ability to interpret b's actions or gain information from them.[4] In effect, from this moment on, a would treat b as a fully byzantine agent.

4. *A priori* Belief Updates

The aim of this section is to describe the semantic mechanism an agent a can use to perform *a priori* belief updates, in order to attempt self-recovery when a discovers that its *a posteriori* observations are in conflict with its *a priori* assumptions. Note that this means that a's part of the model is empty, making it impossible for a to use the current pointed model to recover a consistent epistemic state. Much like epistemic puzzles are described by semantic models, the *a priori* belief updates are also semantic: agent a tries to reimagine the epistemic situation as a *trial model* that is based on the agents' previous experiences, modifications of explicit *a priori* beliefs, and/or *ad hoc* guesses. However, a generally has no reason to ascribe

3 A method for Kripke model synthesis that uses actually and virtually possible worlds to satisfy specific explicit *a priori* assumptions is provided by Cignarale *et al*. (2024).

4 It might be possible to actually communicate an agent's *a priori* beliefs, but of course, such a communication action would correspond to an *a posteriori* update concerning *a priori* beliefs. Modeling this complex interaction is outside the scope of this paper

these internal attempts to the thinking of other agents. Hence, to model higher-order beliefs for all other agents, a uses some *backup model* that is typically derived from a's previous experiences and knowledge of *a priori* beliefs of other agents.[5]

4.1 General update mechanism

We first formulate this update mechanism to be as general as possible, requiring only the basic coherency restrictions on the trial and backup models. In Section 4.6, we will discuss some strategies autonomous agents may employ to generate these models. We do stress, however, that trial models are intended to be *ad hoc* guesses. The number of restrictions on the trial model should be inversely proportional to how wrong agent's beliefs are expected to be: the further away from reality the agent may have strayed, the fewer restrictions should be imposed on its imagination in the process of recovering a consistent state.

For instance, in the consecutive numbers puzzle (Example 1), the trial model should not include non-integer numbers or violate laws of arithmetic. On the other hand, if agents are expected to not be fully attentive to the formulation of the puzzle, a trial model may include pairs of numbers that are not consecutive to account for possible misunderstandings.

Definition 12 (*A priori* belief update). An *a priori belief update* for agent a is a tuple

$$\mathsf{U} = (\mathcal{M}^a, U^a, \mathcal{M}^{\neg a}, \mapsto) \tag{1}$$

where

- *trial model* $\mathcal{M}^a = \langle W^a, R^a, V^a \rangle$ is a quasi-epistemic Kripke model,
- $U^a \subseteq W^a$ is an a-cluster within $\mathcal{M}^a$ (note that $U^a \neq \varnothing$),
- *backup model* $\mathcal{M}^{\neg a} = \langle W^{\neg a}, R^{\neg a}, V^{\neg a} \rangle$ is a quasi-epistemic Kripke model,
- $S \mapsto W^{\neg a}$ is a *correspondence function* from some subset $S \subseteq W^a$ of the domain of the trial model in the domain of the backup model, i.e., a partial function from W^a to $W^{\neg a}$ that identifies some of the trial worlds with backup worlds

5 An important exception to this rule is the scenario where a suspects itself to be the only one to have erred. Then it would be natural for a to try to adapt itself to the alleged thinking of other agents, using that model for both trial and backup models.

such that for any atoms $p \in Prop$ and for any trial worlds $u, v \in W^a$ and backup worlds $u', v' \in W^{\neg a}$ the following *coherency conditions* are fulfilled.

- *Atomic coherency*: if $u \mapsto u'$, then $u \in V^a(p) \Leftrightarrow u' \in V^{\neg a}(p)$, i.e., only propositionally equivalent worlds can be identified.
- *Reasoning coherency*: for each agent $b \neq a$, if $u \mapsto u'$ and $v \mapsto v'$, then $u\, R_b^a\, v \Leftrightarrow u'\, R_b^{\neg a}\, v'$, i.e., the identification respects indistinguishabilities of all agents but a.
- *Simulation coherency*: for each agent $b \neq a$, if $u \mapsto u'$ and $u'\, R_b^{\neg a}\, v'$, then there exists $v \in W^a$ such that $u\, R_b^a\, v$ and $v \mapsto v'$, i.e., the trial model simulates the backup model for all agents but a.

Intuitively, U^a represents the local state of a, i.e., those worlds that a considers actually possible. Relations R_b^a for $b \neq a$ determine whether these new worlds constructed by a would have been distinguishable for other agents, were they aware of a's new trial vision of the world. However, since they are unaware of a's trial model $\mathcal{M}^a$, these indistinguishabilities need to be transferred to the backup model $\mathcal{M}^{\neg a}$, which represents a's virtually possible worlds, used by a to understand how the other agents imagine the epistemic situation. The coherency conditions on $\mapsto$ ensure compatibility between a's actually possible worlds and the virtually possible worlds considered by a. In particular, the trial model should simulate the backup model for agents other than a because a's understanding of their epistemic state cannot be worse than a's impression of their own understanding.

The result of applying an *a priori* belief update to a pointed Kripke model is described by the following definition.

Definition 13 (Result of *a priori* belief update). Let $\mathsf{U} = (\mathcal{M}^a, U^a, \mathcal{M}^{\neg a}, \mapsto)$ be an *a priori* belief update with $\mathcal{M}^a = \langle W^a, R^a, V^a \rangle$ and $\mathcal{M}^{\neg a} = \langle W^{\neg a}, R^{\neg a}, V^{\neg a} \rangle$, and let $(\mathcal{M}, v)$ be a pointed Kripke model with introspective model $\mathcal{M} = \langle W, R, V \rangle$ such that $R_a(v) = \varnothing$. The result of agent a applying U to $(\mathcal{M}, v)$ is a pointed Kripke model $(\mathcal{M} \circledcirc_a \mathsf{U}, v)$ where $\mathcal{M} \circledcirc_a \mathsf{U} := \langle W', R', V' \rangle$ such that[6]

- $W' := W \sqcup U^a \sqcup W^{\neg a}$;
- $R'_a := R_a \sqcup R_a^{\neg a} \sqcup (\{v\} \times U^a) \sqcup (U^a \times U^a)$;

6 In light of our strategy of partitioning the model into disjoint parts, we use the disjoint union operation $\sqcup$ to ensure that these parts do not overlap.

- $R'_b := R_b \sqcup R_b^{\neg a} \sqcup \{(u, v') \in U^a \times W^{\neg a} \mid (\exists v \in W^a) u R_b^a v \mapsto v'\}$ for each agent $b \neq a$;
- $V'(p) := V(p) \sqcup V^{\neg a}(p) \sqcup \left(V^a(p) \cap U^a\right)$ for any $p \in Prop$.

Remark 14. The requirement of simulation coherency is asymmetric. It is worth asking why it should not be a bidirectional bisimulation instead. Recall that both trial and backup models describe a's attempt to provide an alternative explanation for the apparent inconsistency of its beliefs. The difference is that the trial model $\mathcal{M}^a$ is a's suggestion of how things "actually are" whereas the backup model $\mathcal{M}^{\neg a}$ is a's attempt to imagine how this new picture of the world is viewed by all other agents based on some standard, default, or past views a expects of them. This puts a in a position of intellectual superiority akin to that of the system designer of a distributed system. Within this new point of view of a's creation, agent a *does* know better than the other agents, much like the system designer knows better than distributed agents.[7] Indeed, as mentioned before, it is the system designer who performs *a priori* belief updates in traditional distributed systems, which makes it reasonable for an agent to adapt a similar attitude for the same task. In a's view, $\mathcal{M}^a$ is an "improved" description of the world, which a hopes to be accurate, whereas $\mathcal{M}^{\neg a}$ may well be an erroneous description that a is aware of but rejects. If a thinks that other agents should consider certain possibilities, it is only rational for a to embed these possibilities into its own view of the world to ensure its accuracy by better describing the epistemic reasoning of other agents. At the same time, a can imagine others having blind spots and missing some possibilities a is now considering. A simulation in the opposite direction, from $\mathcal{M}^{\neg a}$ to $\mathcal{M}^a$ would have made such blind spots impossible and, hence, is not required by our definition. At the same time, if a's update is guided by the idea that other agents knew the true state of affairs all along and it is a who has been missing something, then the simulation in the opposite direction would make sense. Hence, while not requiring it, our definition does not preclude a bisimulation between $\mathcal{M}^a$ and $\mathcal{M}^{\neg a}$ either.

The properties of *a priori* belief updates are captured by the following theorem.

Theorem 15. *Let* $\mathsf{U} = (\mathcal{M}^a, U^a, \mathcal{M}^{\neg a}, \mapsto)$ *be an* a priori *belief update and* $(\mathcal{M}, v)$ *be a pointed Kripke model with introspective model* $\mathcal{M} = \langle W, R, V \rangle$ *such that* $R_a(v) = \varnothing$. *Then*

7 Note that this refers to the knowledge of the underlying model rather than belief in specific facts. In fact, if a considers more possibilities, a would believe fewer facts, presumably because contracting beliefs was necessary to avoid inconsistency.

1. $\mathcal{M} \circledcirc_a \mathsf{U}$ *is an introspective model;*
2. *for any purely propositional formula* φ

$$\mathcal{M}, v \vDash \varphi \Leftrightarrow \mathcal{M} \circledcirc_a \mathsf{U}, v \vDash \varphi.$$

3. *for any formula* φ *and any agent* $b \neq a$,

$$\mathcal{M}, v \vDash B_b\varphi \Leftrightarrow \mathcal{M} \circledcirc_a \mathsf{U}, v \vDash B_b\varphi;$$

4. $\mathcal{M}, v \vDash B_a\bot$, *but* $\mathcal{M} \circledcirc_a \mathsf{U}, v \nvDash B_a\bot$;

Proof. Let $\mathcal{M} \circledcirc_a \mathsf{U} = \langle W', R', V' \rangle$.

1. Transitivity and euclideanity of R'_a and R'_b for $b \neq a$ follow by construction due to all involved models being introspective. We demonstrate only several non trivial cases. (a) It cannot happen that $w\, R'_a\, v$ and $v\, R'_a\, u$ for some $u \in U^a$ because, by construction, $w\, R'_a\, v$ is equivalent to $w\, R_a\, v$, which would imply $v\, R_a\, v$ by euclideanity of R_a, whereas we assumed $R_a(v) = \varnothing$. (b) Let $b \neq a$. If $u\, R'_b\, v'$ because $u\, R_b^a\, v \mapsto v'$ and $v'\, R_b^{\neg a}\, z'$ for some $u \in U^a$, $v \in W^a$, and $v', z' \in W^{\neg a}$, then there exists $z \in W^a$ such that $v\, R_b^a\, z \mapsto z'$ by the simulation coherency condition. Hence, $u\, R_b^a\, z$ by transitivity of R_b^a, which implies $u\, R'_b\, z'$, as required. (c) Let $b \neq a$. If $u\, R'_b\, v'$ because $u\, R_b^a\, v \mapsto v'$ and $u\, R'_b\, z'$ because $u\, R_b^a\, z \mapsto z'$ for some $u \in U$, $v, z \in W^a$, and $v', z' \in W^{\neg a}$, then $v\, R_b^a\, z$ by euclideanity of R_b^a. Hence, $v'\, R_b^{\neg a}\, z'$ by the reasoning coherency condition, as required.
2. The statement easily follows from the fact that the propositional valuation at v does not change.
3. The statement easily follows from $R'_b\,(v) = R_b(v)$.
4. The statement easily follows from the fact that $R'_a\,(v) = U^a \neq \varnothing$. □

Remark 16. Note that Theorem 15.3 means that a performing a successful update of its *a priori* beliefs cannot resolve inconsistent beliefs of other agents (though a may erroneously believe to have resolved them). It is reasonable to expect each of the agents with inconsistent beliefs to perform such an operation. We will describe how to do it and why updates of different agents are completely independent from each other in Section 4.4.

4.2 Application to the muddy children puzzle

We now illustrate this *a priori* update mechanism by considering the muddy children puzzle (Fagin *et al*., 1995) with various additional *a priori* assumptions.

Example 17 (Standard muddy children puzzle). In the muddy children puzzle (MCP), n children are playing in the mud, and k of them get mud on their foreheads. Each can see whether there is mud on others but not on his/her own forehead. Father comes and announces, "At least one of you has a muddy forehead." Father then starts repeating the question, "If you know whether you are muddy or not, step forward." Under the assumption that children are perfect reasoners, pay complete attention, are truthful and the assumption that all this (and much more) is common knowledge among them, it is well known that the first $k - 1$ times Father asks, nobody steps forward. All k muddy children step forward the kth time Father asks, and the remaining $n - k$ clean children step forward the next $k + 1$ th time. For instance, for $n = 3$ and $k = 1$, i.e., with three children playing and one muddy child, the muddy child should immediately step forward, with the other two children stepping forward the second time.

We will use the standard epistemic modeling where the children are agents a, b, c, …. The muddiness of child a is represented by atom m_a that is true iff a is muddy. The fact that, before Father's first announcement, it is common knowledge that children do not know their own state but know that of others is formalized by requiring the validity of $\neg B_a m_a \wedge \neg B_a \neg m_a$ for all children a and of $B_a m_b \vee B_a \neg m_b$ for $a \neq b$. A Kripke model satisfying these and other requirements of the puzzle for the three agents is represented in Fig. 1(a). Father's role is that of an external announcer, making his epistemic state irrelevant to the solution of the puzzle. From a distributed perspective, his role is akin to that of the system designer. While typically modeled as a public announcement, his first announcement "at least one of you has a muddy forehead" can also be viewed as an *a priori* belief update of the system performed by the system designer.

Example 18 (MCP with false *a priori* assumption resolved by *a priori* belief update). Consider a variant of the muddy children puzzle where three children a, b, and c all commonly believe (*a priori*) that at least two of them are muddy,[8] i.e., they formed the common *a priori* belief

$$APB_2 = (m_a \wedge m_b) \vee (m_a \wedge m_c) \vee (m_b \wedge m_c).$$

(Note that APB_2 does not represent *all* common *a priori* beliefs of the children. Most of those are encoded in epistemic model $\mathcal{M}_0$.)

8 The origins of this common *a priori* belief are, in principle, immaterial for our example. A reader desiring a realistic explanation can easily find one, ranging from a cognitive bias, e.g., prejudice to guesstimating based on past games.

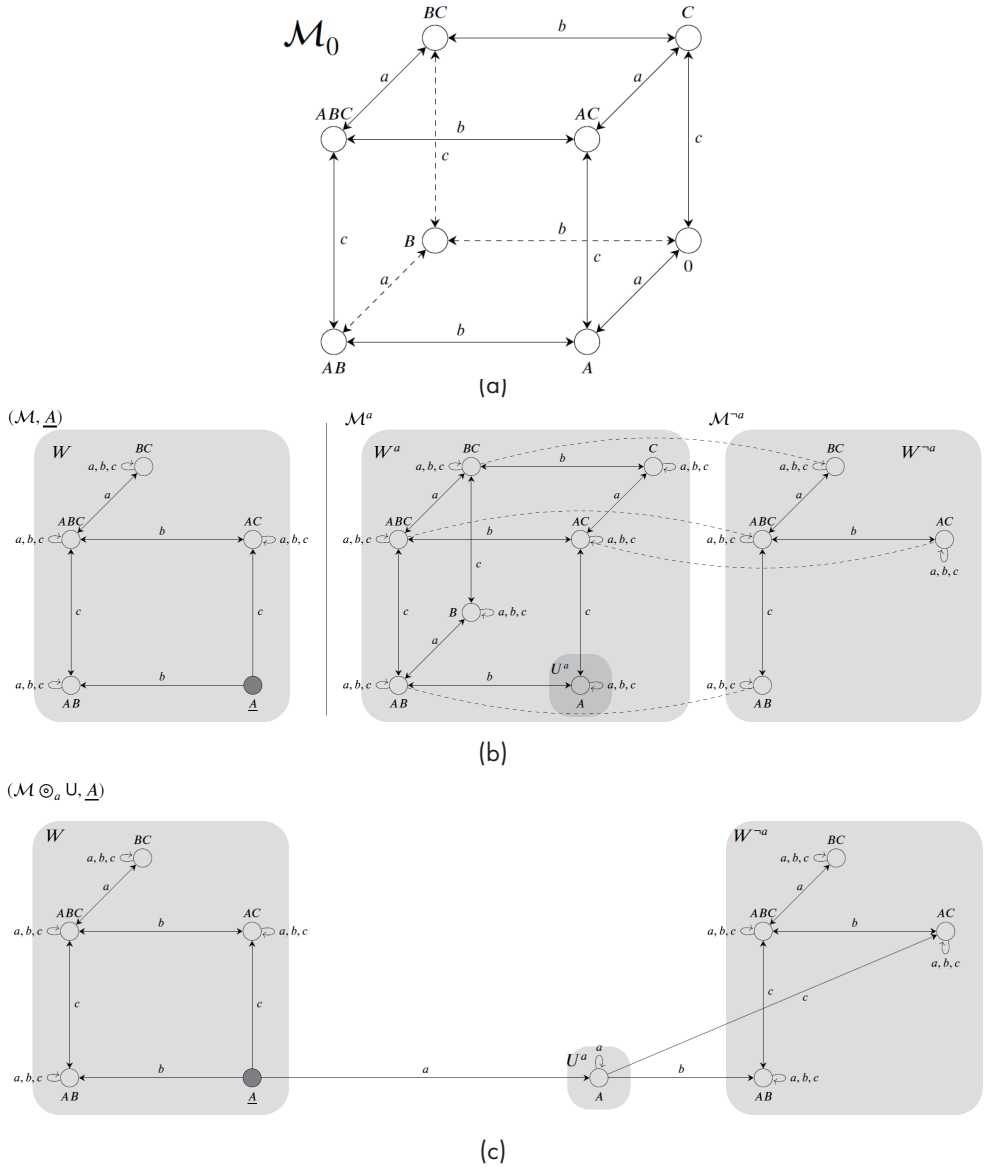


Fig. 1. (a) Epistemic model $\mathcal{M}_0$ for the muddy children puzzle with three agents of Example 17, before Father's announcements. D in the name of the state corresponds to m_d being true at the state, i.e., agent being muddy. No child is muddy in state 0. Bidirectional arrows (both dashed and solid) represent indistinguishability for agents, e.g., agent b cannot distinguish between states ABC and AC. Reflexive loops, present for all agents at every state, are omitted.
(b) Left: Initial pointed Kripke model $\mathcal{M}, \underline{A}$ restricted to APB_2.
Middle and Right: Elements of a priori belief update $\mathsf{U} = (\mathcal{M}^a, U^a, \mathcal{M}^{\neg a}, \mapsto)$ for agent a from Example 18. The correspondence relation $\mapsto$ is represented by dashed lines.
(c) Updated Kripke model $(\mathcal{M} \circledcirc_a \mathsf{U}, \underline{A})$ from Example 18. In particular, arrows from U^a to worlds in $W^{\neg a}$ are drawn in accordance with the correspondence relation $\mapsto$, and all worlds in U^a are reachable from $\underline{A}$ via a-arrows, as per Definition 13.

In terms of Kripke modeling, this can be viewed as a public announcement of APB_2, which results in model $\mathcal{M}_0$ shown in Fig. 1(a), which they commonly consider as the base model, shrinking to the Kripke model of four worlds only (AB, AC, BC, and ABC), making it unnecessary for Father to announce anything. Suppose that in this instance, however, only a is muddy, i.e., $m_a \wedge \neg m_b \wedge \neg m_c$. Thus, their common *a priori* belief APB_2 turns out to be false. This scenario is depicted as pointed model ($\mathcal{M}$, $\underline{A}$) in Fig. 1(b) (Left), which is obtained from the four-world model by adding a real world $\underline{A}$[9] and drawing arrows from it to the four worlds in accordance with what children would consider possible. Note that there are now one-directional arrows due to the fact that children have false beliefs as the result of false *a priori* assumptions.

While beliefs of all agents are false, only a can detect this because a's beliefs are inconsistent, unlike those of b and c:

$$\mathcal{M}, \underline{A} \vDash B_a\bot, \text{ while } \mathcal{M}, \underline{A} \nvDash B_b\bot \text{ and } \mathcal{M}, \underline{A} \nvDash B_c\bot.$$

Thus, a is the only child who is justified to update her *a priori* beliefs. Note that no public announcement was made, and that a's inconsistency is the result of a's reasoning about the epistemic situation.

The introduction of one-directional arrows changes the whole interpretation of the epistemic situation. Indeed, using the assumption of the common knowledge of the model, it could be tempting to suggest that agents b and c can detect a's inconsistency. It would not, however, match the underlying scenario. For instance, when b sees that a is muddy while c is not, we would expect b to conclude that b is the second muddy child and that a sees the mud on b's forehead (the model says as much by providing a unidirectional b-arrow from $\underline{A}$ to AB). To match this intuition, we are forced to abandon the assumption of common knowledge of the model. For pointed model ($\mathcal{M}$, $\underline{A}$), agent b is only aware of the world AB, the only possible one for b, and all worlds accessible from it by a sequence of arrows; agent c is only aware of the world AC, the only possible one for c, and all worlds accessible from it by a sequence of arrows, while a is not aware of any worlds, resulting in inconsistent beliefs. In particular, a has lost the ability to examine the reasoning of other agents. While it is natural for a to assume that other agents still operate within the last commonly considered model, i.e., the four-

9 It is named A because its propositional valuation is the same as that of A in $\mathcal{M}_0$, while the underline means it is the actual world.

world $\mathcal{M}^{\neg a}$, connecting this assumption to a's own reality requires this reality to be fleshed out by a, which is the purpose of $\mathcal{M}^{a}$ (and of its restriction U^a).

According to Definition 12, to perform an *a priori* belief update, agent a must come up with a trial model $\mathcal{M}^{a}$. Treating the process as ultimately a creative one, we eschew an explanation of how to do that (providing some heuristics in Section 4.6). Instead we consider several alternative trial models in this and following examples. The success of the *a priori* update depends largely on how good the choices made by a are, including crucially the choice of the trial model. If agent a manages to guess a trial model reflecting the actual world well (recall that the actual world is not considered possible by a), the *a priori* update is likely not only to resolve the inconsistency of a's beliefs, but also to result in new beliefs that are factive. This is the case we demonstrate in the current example. Other examples will demonstrate that a bad choice may not remove the inconsistency, or worse still, may create new consistent beliefs that are very far from reality.

In this example, agent a guesses the trial model $\mathcal{M}^{a}$ depicted in Fig. 1(b) (Middle), which fits well with what she observes where the singleton cluster $U^a = \{A\}$ of worlds from $\mathcal{M}^{a}$ represents the possibilities a considers actually possible based on her observations. As already mentioned, for the backup model, which a uses for computing what is considered by b and c, agent a takes the four-world model $\mathcal{M}^{\neg a}$ she herself used until recently. The elements of the *a priori* belief update $\mathsf{U} = (\mathcal{M}^{a}, U^a, \mathcal{M}^{\neg a}, \mapsto)$ are depicted in Fig. 1(b) (Right). In particular, since $\mathcal{M}^{a}$ extends $(\mathcal{M}, \underline{A})$ with three additional worlds where only one agent is muddy, agent a establishes the correspondence function $\mapsto$ to connect each world from $\mathcal{M}^{a}$ to a propositionally equivalent world of $\mathcal{M}^{\neg a}$ if such a world exists. It is easy to see that the coherency conditions for the *a priori* belief update are fulfilled. According to Def. 12, the result $(\mathcal{M} \circledcirc_a \mathsf{U}, \underline{A})$ of applying the update to the initial model is shown in Fig. 1(c). The two arrows from $A \in U^a$ for b and c to worlds in $W^{\neg a}$ are added because of the correspondence relation $\mapsto$. For instance, the b-arrow from A in U^a to AB in $\mathcal{M}^{\neg a}$ of the updated model is due to the b-arrow from A to AB in $\mathcal{M}^{a}$ and the correspondence from AB in $\mathcal{M}^{a}$ to AB in $\mathcal{M}^{\neg a}$. As a result,

$$\mathcal{M} \circledcirc_a \mathsf{U}, \underline{A} \vDash B_b(m_a \wedge m_b \wedge \neg m_c), \qquad \mathcal{M} \circledcirc_a \mathsf{U}, \underline{A} \vDash B_a B_b(m_a \wedge m_b \wedge \neg m_c),$$
$$\mathcal{M} \circledcirc_a \mathsf{U}, \underline{A} \vDash B_c(m_a \wedge \neg m_b \wedge m_c), \qquad \mathcal{M} \circledcirc_a \mathsf{U}, \underline{A} \vDash B_a B_c(m_a \wedge \neg m_b \wedge m_c),$$
$$\mathcal{M} \circledcirc_a \mathsf{U}, \underline{A} \vDash B_a(m_a \wedge \neg m_b \wedge \neg m_c), \qquad \mathcal{M} \circledcirc_a \mathsf{U}, \underline{A} \vDash B_b B_a(m_a \wedge m_b \wedge \neg m_c),$$
$$\mathcal{M} \circledcirc_a \mathsf{U}, \underline{A} \vDash B_c B_a(m_a \wedge \neg m_b \wedge m_c).$$

In other words, each agent believes itself to know the actual situation and to be muddy. All three of them will step forward upon Father's prompt. Moreover, the *a priori* update restored a's ability to understand the reasoning of others. Because a guessed $\mathcal{M}^{\neg a}$ correctly, she can correctly interpret b and c stepping forward. Agents b and c expect a to step forward. On the other hand, b cannot interpret c stepping forward because

$$\mathcal{M} \circledcirc_a \mathsf{U}, \underline{A} \vDash B_b(\neg B_c m_c \wedge \neg B_c \neg m_c)$$

and *vice versa*.

4.3 *A priori* belief update triggered by public announcements

One could say that Example 18 should have been modeled according to Fig. 1(c) from the very beginning because it fits better with the evidence observed by the agents than the model $\mathcal{M}^{\neg a}$ from Fig. 1(b) which they attempted to use. That would not address the question of how to turn an epistemic description of agents' observations and beliefs into complex epistemic scenarios, possibly with false common beliefs of a subset of agents. Our *a priori* belief updates provide a mechanism for building models for more complex epistemic situations based on existing models for simpler scenarios.

It is harder to find an alternative to *a priori* belief updates when the inadequacy of the *a priori* assumptions cannot be observed initially and is only uncovered through communication. As already mentioned, the idea of scrapping the original model and starting anew corresponds to the development cycle of ordinary distributed systems, where it is performed by a system designer. We aim to develop a mechanism for individual agents to do it, as befits SASO systems. Moreover, as we saw in Example 18, different agents are likely to discover the flaw in their assumptions at different times, which puts the idea of scrapping the *whole* model at odds with the Knowledge of Preconditions Principle (Moses, 2016) that requires an agent performing an action to know the reason for this action.

This was the situation in Example 1 where it was a's public announcement of knowing b's number that caused b to realize that some *a priori* assumptions must have been wrong. We now show how to perform *a priori* belief update in such a situation by representing b's reasoning in Example 1 in our formal framework.

Example 19 (Consecutive numbers with false a priori assumptions formalized). Let us describe our twist version of Example 1 as a Kripke model. As before, we label worlds according to

pairs of numbers held by the agents. Although this results in several worlds having the same label, we prefer to disambiguate when necessary rather than creating an overly complex nomenclature for worlds. As before, the actual world (1, 2) is distinguished by the underline. Unlike the MCP, an important *a priori* assumption is not common between agents a and b: a starts natural numbers from 1 while b also considers 0 to be natural. At the same time, none of them are initially aware of this disagreement. Hence, the pointed Kripke model ($\mathcal{N}$, (1, 2)) in Fig. 2(a) (Above) is introspective but not epistemic and consists of the actual world and two disjoint fragments, each only accessible to one of the two agents: the upper line of worlds represents what a (mistakenly) thinks is their common view of the situation, while the lower line of worlds represents what b (mistakenly) takes to be their common view. To distinguish similarly labeled worlds, we call the worlds from the upper (lower) line a-worlds (b-worlds). To simplify the notation without affecting the logical content of the puzzle, suppose a states b's number explicitly by saying "I know that you have number 2," which we represent by formula $B_a 2_b$. It is easy to see that $\mathcal{N}, w \not\models B_a 2_b$ for any b-world w; of the a-worlds, only world (1, 2) satisfies $B_a 2_b$, and so does the actual world. Accordingly, the result of a public update with $B_a 2_b$, as depicted in Fig. 2(a) (Below), contains only two worlds and $\mathcal{N} \mid B_a 2_b, (1, 2) \models B_b \bot$, prompting the "Wait, what?" comment in our informal rendition and triggering an *a priori* update by b.

To match the informal reasoning from Example 1, consider the nature of this public announcement: it was about a's beliefs rather than propositional facts. Hence, it is rational for b to return to the original model she considered before the public announcement: model $\mathcal{M}^b$ from Fig. 2(b) is nothing but b-worlds from Fig. 2(a) (Above) and U^b is the b-cluster of worlds (1, 2) and (3, 2) that b considered possible pre-announcement. What b should modify is how a sees the situation. For that b correctly imagines the abbreviated natural numbers, resulting in $\mathcal{M}^{\neg b}$ being the same as a-worlds from Fig. 2(a) (Above). Once again, the correspondence is determined by propositional truth of atoms n_a and m_b. It is easy to see that all coherency conditions are satisfied.

The result of b applying *a priori* update U to ($\mathcal{N} \mid B_a 2_b$, (1, 2)) is shown in Fig. 2(c) (Left). However, this is not the model explaining why b now knows a's number. Indeed, this cannot be the final stage of b's reevaluation since b's part of this model does not reflect the public announcement a made. Agent b still needs to apply the standard public announcement update to the only part of the model it is aware of, which is comprised of worlds from U^b and $W^{\neg b}$. This leaves b's part of the model with two worlds only, (1, 2) from U^b and (1, 2) from $W^{\neg b}$, as shown in Fig. 2(c) (Right). It can be easily seen that $((\mathcal{N} \mid B_a 2_b) \circledcirc_b \mathsf{U}) \mid_b B_a 2_b, (1, 2) \models B_b 1_a$, explaining that b has now figured out a's number too.

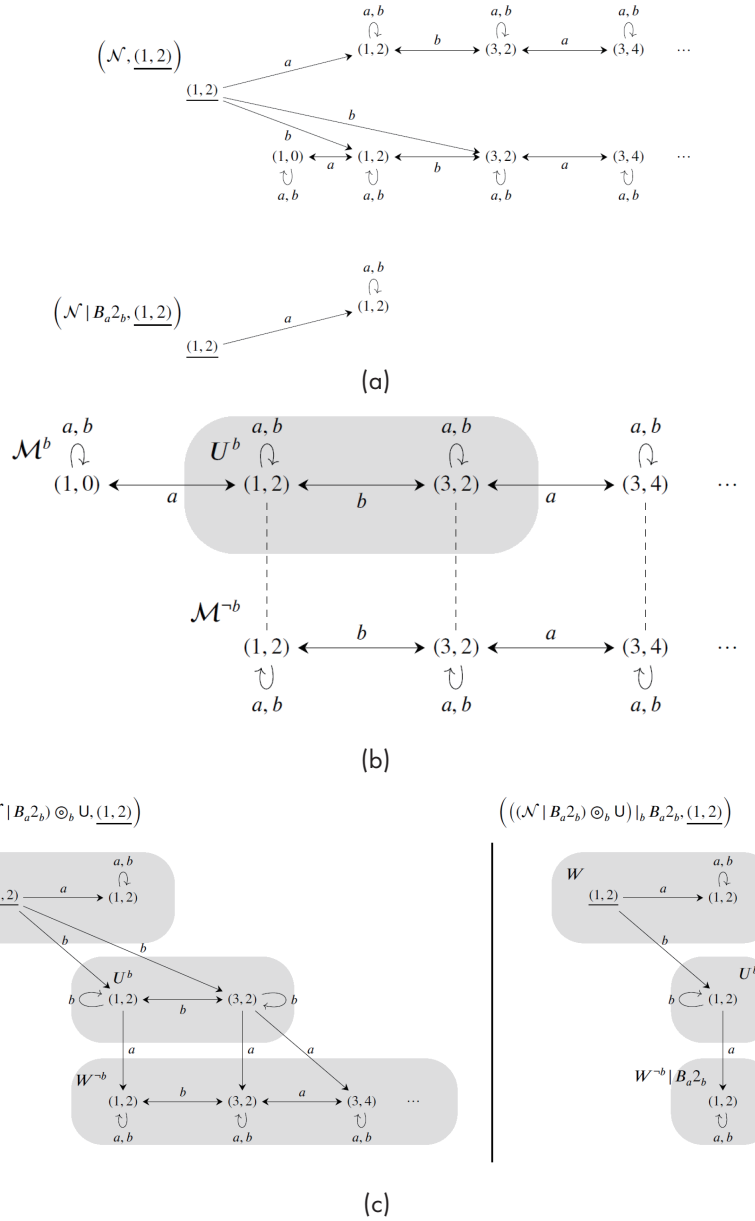


Fig. 2. (a) Above: Introspective pointed Kripke model $(\mathcal{N}, \underline{(1,2)})$ representing the initial state of both agents in the consecutive numbers puzzle from Example 19, where b considers 0 to be a natural number while a does not.
Below: $(\mathcal{N} \mid B_a 2_b, \underline{(1,2)})$ after a's public announcement that a knows b's number is 2.
(b) Elements of the *a priori* belief update $\mathsf{U} = (\mathcal{M}^b, U^b, \mathcal{M}^{\neg b}, \mapsto)$ for b in Example 19. The correspondence relation is represented by dashed lines. The grey rectangle is b's equivalence class U^b.
(c) Left: Result of b applying an *a priori* belief update in the inconsistent epistemic state $(\mathcal{N} \mid B_a 2_b, \underline{(1,2)})$ in Fig. 2(a).
Right: Result of b (re)applying a's public announcement of $B_a 2_b$ to b's part of the model.

Such a "private" public announcement update$|_b$ may not be standard, but it is warranted in this setting. Recall that b here assumes the role of the system designer and treats her part of the model as the whole model and the only model pre-announcement. It is natural then for b to update this partial model as if it is the whole model. This operation should not affect a's part of the model since b is not aware of it.

If, in the *a priori* updated model after the public announcement there is at least one world in U^a that makes the announced formula true, then the self-recovery operation is indeed successful, as stated by the following corollary:

Corollary 20. *Let* $\mathsf{U} = (\mathcal{M}^a, U^a, \mathcal{M}^{\neg a}, \mapsto)$ *be an* a priori *belief update triggered by the public announcement of* φ *and* $(\mathcal{M}, v)$ *be a pointed Kripke model with introspective model* $\mathcal{M} = \langle W, R, V \rangle$ *pre-announcement. If* $(\mathcal{M} \mid \varphi) \circledcirc_a \mathsf{U}, u \vDash \varphi$ *for some world* $u \in U^a$*, then* a*'s beliefs are consistent after the* a priori *update, i.e.,* $((\mathcal{M} \mid \varphi) \circledcirc_a \mathsf{U}) |_a \varphi, v \nvDash B_a \bot$.

Proof. The statement follows from Theorem 15.4. □

Note that because of Moorean sentences, there is generally no guarantee that a believes in φ after the public announcement.

4.4 Simultaneous *a priori* belief updates by several agents

So far we considered only scenarios where beliefs of exactly one agent become inconsistent, resulting in an *a priori* belief update for this agent. It is, of course, entirely possible that several agents become inconsistent simultaneously. Given that an *a priori* belief update is performed by an agent privately and has no effect on the rest of the model, it is straightforward to apply several such updates in parallel.

Example 21 (Simultaneous *a priori* belief updates triggered by a public announcement). Let us continue the scenario from Example 18 where we left off, i.e., after a has restored her consistency, Father asked the children if they know whether they are muddy, and all three children stepped forward, which is equivalent to the public announcement of "all step forward":

$$ASF = (B_a m_a \vee B_a \neg m_a) \wedge (B_b m_b \vee B_b \neg m_b) \wedge (B_c m_c \vee B_c \neg m_c). \tag{2}$$

The resulting model $\mathcal{M}' := (\mathcal{M} \circledcirc_a \mathsf{U}) \mid ASF$ is shown in Fig. 3(a) and has only two worlds remaining from Fig. 1(c), the actual world $\underline{A}$ and world A from U^a. Thus, beliefs of b and c have become inconsistent (there are no outgoing b- or c-arrows from the actual world $\underline{A}$) and a knows that this is the case (there are no outgoing b- or c-arrows from A, the only world a considers possible). At the same time, a still maintains consistency of beliefs and thinks of herself as muddy:

$$\mathcal{M}', \underline{A} \vDash B_b\bot \wedge B_c\bot \wedge B_aB_b\bot \wedge B_aB_c\bot \wedge \neg B_a\bot \wedge B_am_a.$$

At this point, b and c each independently performs its own *a priori* belief update triggered by the announcement ASF. Although a can expect them to do so, one could argue that, due to the nondeterministic *ad hoc* nature of *a priori* belief updates, a does not think she can guess how b and c would choose to update their beliefs, nor is such a guess necessary for a to be able to respond to Father. Accordingly, a remains satisfied in her belief that she is muddy and that her companions are confused.

Suppose that each of b and c chooses to do the same thing that a has done earlier, i.e., to use $\mathcal{M}^a$ from Fig. 1(b) (Middle) as both $\mathcal{M}^b$ and $\mathcal{M}^c$ and to use $\mathcal{M}^{\neg a}$ from Fig. 1(b) (Right) as both $\mathcal{M}^{\neg b}$ and $\mathcal{M}^{\neg c}$. Correspondence $\mapsto$ is also the same for both b and c. However, their local states $U^b = \{A, AB\}$ and $U^c = \{A, AC\}$ differ due to their differing points of view.

We present the state after b and c's simultaneous independent *a priori* update in Fig. 3(b). Then, each of them applies ASF to its private part of the model. No worlds in $W^{\neg b}/W^{\neg c}$ satisfy ASF, so all are pruned. World AB of U^b is rejected because there c does not know its state:

$$((\mathcal{M} \circledcirc_a \mathsf{U}) \mid ASF) \circledcirc_{b,c} \mathsf{U}', AB \nvDash B_cm_c \vee B_c\neg m_c;$$

similarly, AC of U^c does not survive the announcement ASF since there b is not sure of its state. This leaves only worlds A from U^b and U^c (and a's whole part of the model, which is not affected by the *a priori* thinking of b and c), resulting in Fig. 3(c). Note that now each agent believes that the beliefs of other two agents are inconsistent while they themselves correctly believe that a is the only muddy child. In effect, all children solved the problem correctly but lost trust in each other's reasoning, using

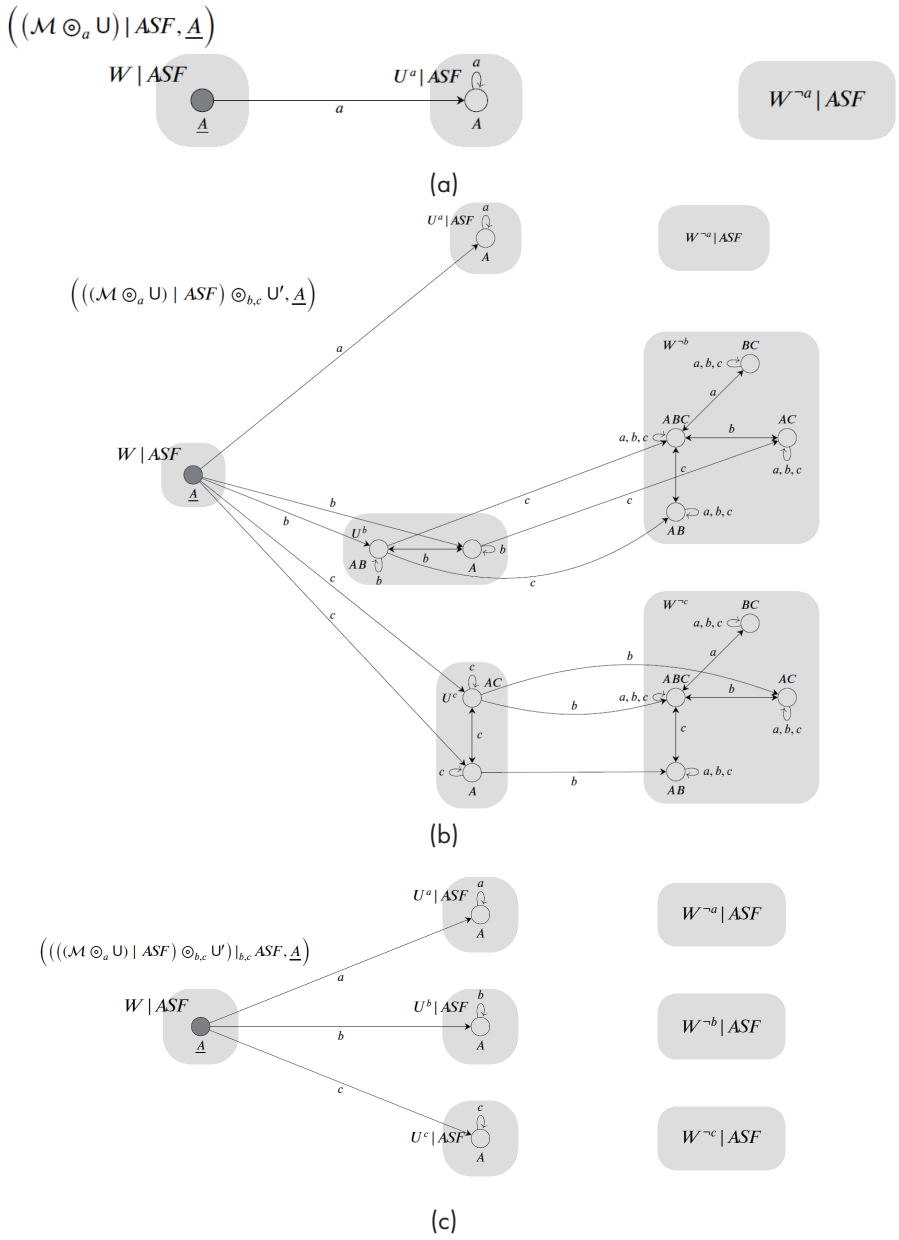


Fig. 3. (a) Pointed Kripke model representing the result of the public announcement of ASF on the *a priori* updated model $(\mathcal{M} \circledcirc_a \mathsf{U}, \underline{A})$ of Example 18, as described in Example 21.
(b) Pointed Kripke model showing the intermediate state of Example 21 after b and c had independently performed *a priori* belief updates in parallel but have not yet (re)applied the public announcement ASF.
(c) Pointed Kripke model illustrating the final stage of Example 21 after a priori belief updates of b and c and their application of the public announcement ASF to their newly created parts of the model.

$$\mathcal{M}'' := (((\mathcal{M} \circledcirc_a \mathsf{U}) \mid ASF) \circledcirc_{b,c} \mathsf{U}') \mid_{b,c} ASF$$

for the model, $E\varphi := B_a\varphi \wedge B_b\varphi \wedge B_c\varphi$ for mutual knowledge, and $\hat{E}\varphi := \neg E\neg\varphi$ for its dual:

$$\mathcal{M}'', \underline{A} \vDash E(m_a \wedge \neg m_b \wedge \neg m_c) \wedge \neg\hat{E}\bot \wedge B_a(B_b\bot \wedge B_c\bot) \wedge B_b(B_a\bot \wedge B_c\bot) \wedge B_c(B_a\bot \wedge B_b\bot).$$

Note that update U' in Fig. 3(b) represents two separate *a priori* belief updates: one by b (rectangles U^b and $W^{\neg b}$) and another by c (rectangles U^c and $W^{\neg c}$). Hence, here we take U' to be a partial function from the set $\mathcal{A}$ of agents to *a priori* belief update tuples, so that $\mathsf{U}'(b)$ is the *a priori* update performed by b while $\mathsf{U}'(c)$ is the one for c. This notation is similar to using R for accessibility relations of all agents. The domain of U' is then listed in the subscript to $\circledcirc$ as in $((\mathcal{M} \circledcirc_a \mathsf{U}) \mid ASF) \circledcirc_{b,c} \mathsf{U}'$. This does not cause any formal problems because, being private, individual *a priori* belief updates do not interfere with each other.

4.5 *A priori* belief updates need not yield (correct) solutions

In all examples considered so far, all agents performing the updates have succeeded in recovering a consistent epistemic state (though sometimes at the expense of expecting consistency from other agents). By no means do we claim that this is always the case. The *ad hoc* chosen *a priori* belief update U need not lead to the resolution of the agent's conundrum. Whether it does is rather a matter of luck. For instance, consider the setup of Example 19. Previously, we had agent b guess the exact model of the number line used by a, but b's guess can also be wrong. For instance, b might think that a mistakenly considers all integers rather than non negative integers only, making $\mathcal{M}^{\neg b}$ from Fig. 2(c) extend infinitely in both directions, retaining world (1, 0), and not resulting in a consistent state for b after $B_a 2_b$ is taken into account. If the agent is persistent, it should be expected that such an unsuccessful update is rejected and a new attempt is made to restore consistency using a different *a priori* belief update.

It is also possible that a wrong *a priori* guess does not result in an inconsistency but leads to wrong conclusions, undetected by any of the agents.

Example 22 (Simultaneous a priori belief updates creating consistent false beliefs). Consider a variant of Example 18 where all agents have the same initial explicit *a priori* common belief APB_2, but all are clean in actuality. In this case, all agents detect inconsistency from the start,

and each privately and independently performs an *a priori* belief update. Suppose that, using the same reasoning as agent a in Example 18 and agents b and c in Example 21, agents use a simultaneous update U'' such that

$$\mathsf{U}_a'' := (\mathcal{M}^a, \{A\}, \mathcal{M}^{\neg a}, \mapsto), \mathsf{U}_b'' := (\mathcal{M}^b, \{B\}, \mathcal{M}^{\neg b}, \mapsto), \mathsf{U}_c'' := (\mathcal{M}^c, \{C\}, \mathcal{M}^{\neg c}, \mapsto),$$

where the only difference to the previous case is in their local states $V^a = \{A\}$, $V^b = \{B\}$, and $V^c = \{C\}$ that are chosen based on what they observe. The resulting model is depicted in Fig. 4(a). It is easy to see that

$$\mathcal{M} \circledcirc_{a,b,c} \mathsf{U}'', \underline{0} \vDash B_a(m_a \wedge B_b m_b \wedge B_c m_c) \wedge B_b(m_b \wedge B_a m_a \wedge B_c m_c) \wedge B_c(m_c \wedge B_a m_a \wedge B_b m_b).$$

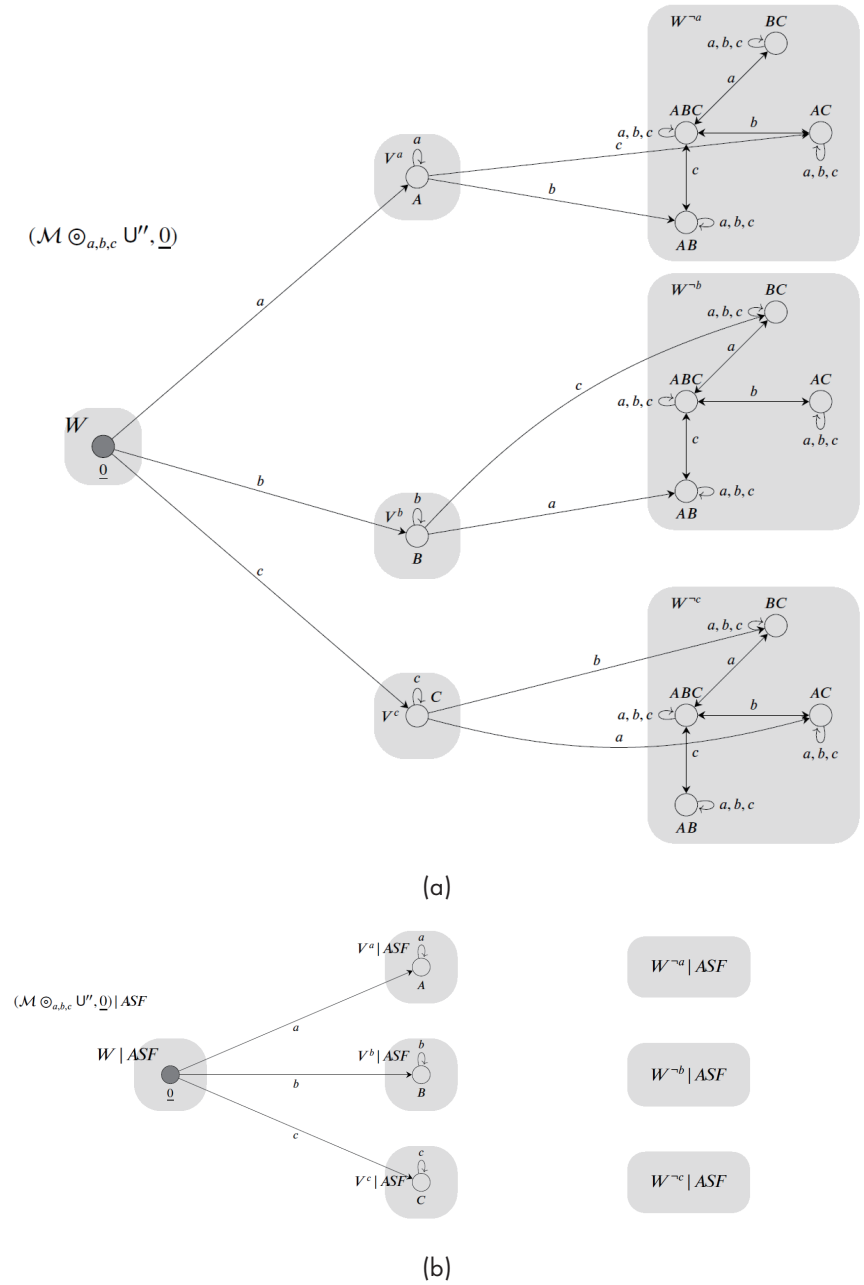
In other words, after this *a priori* belief update each child now erroneously believes that (I) it is the only muddy child and (II) other children "erroneously" believe themselves to be also muddy.[10] Hence, all children step forward, triggering public update ASF resulting in the model shown in Fig. 4(b). This behavior conforms to everyone's expectations, so all children preserve consistency of beliefs. However, each child thinks that this behavior should have puzzled the other two, so that after stepping forward they each think the others ––-have inconsistent beliefs:

$$(\mathcal{M} \circledcirc_{a,b,c} \mathsf{U}'') \mid ASF, \underline{0} \vDash B_a(B_b\bot \wedge B_c\bot) \wedge B_b(B_a\bot \wedge B_c\bot) \wedge B_c(B_a\bot \wedge B_b\bot).$$

4.6 Some heuristics for *a priori* belief updates

In developing a method for agents to "think outside the box," we did not want to put any boundaries or restrictions on the kinds of trial and backup models used in *a priori* belief updates, did not want to "box them in" as it were. At the same time, there exist rather regular methods of generating *a priori* belief updates, some of which we have already used. Let us outline some of these methods, which can easily be implemented in a form of an exhaustive trial-and-error search through finitely many possibilities.

10 The word *erroneously* is here in air quotes because it has a flavor of a double Gettier example (Gettier, 1963). Firstly, a believes that b considers itself to be muddy, b does in fact consider itself to be muddy, but not for the reason a expects, as manifested by the difference between b's part of the model representing b's beliefs and $W^{\neg a}$ representing a's rendition of b's beliefs. Secondly, a thinks that b is wrong, i.e., that b is clean, a is in fact correct, but not for the reason a expects, as manifested by the difference between the real world $\underline{0}$ and world A of V^a



(a)

(b)

Fig. 4. (a) Pointed Kripke model representing the intermediate state of Example 22 after all agents had independently performed ***a priori*** belief updates in parallel but before any public announcements have been made.
(b) Pointed Kripke model representing the final state of Example 22 after all agents had independently performed a priori belief updates in parallel and have applied the public announcement ASF.

Heuristics 1: Predefined master models If an agent is aware of several incompatible alternative models of reality but is not sure which one of them better suits its observations and/or is not sure which points of view other agents entertain, then *a priori* belief updates can be generated by assigning various combinations of these models to the agent itself and to other agents. This method was used in Example 19 for two different models of natural numbers.

Heuristics 2: Varying explicit a priori assumptions Even when all agents agree on the same model, e.g., a master model representing the general rules of the puzzle, they may differ in additional *a priori* assumptions each of them makes. Here the master model represents (common) implicit *a priori* assumptions while additional assumptions are private and explicit in that they are represented by formulas imposed on the master model. For instance, APB_2 is such an explicit assumption in Example 18 imposed on the implicit assumptions modeled by $\mathcal{M}_0$ from Example 17. In many of our MCP-related examples, agent a continues using $\mathcal{M}_0 \mid APB_2$ as $\mathcal{M}^{\neg a}$ to describe the reasoning of others while a itself switches to $\mathcal{M}_0 \mid APB_1$ as $\mathcal{M}^a$ for her own reasoning, where $APB_1 := m_a \vee m_b \vee m_c$ states that at least one child is muddy. If such an *a priori* belief update fails to restore her consistency, she can then switch to the full $\mathcal{M}_0$ or consider a more complex *a priori* restriction.

Heuristics 2a: Choosing explicit a priori assumptions based on prior failures Using the master model, if the initial model or one of the previous unsuccessful attempts to reach a consistent state were based on some explicit *a priori* beliefs APB, then it is reasonable to attempt new explicit beliefs of the form $APB' \wedge \neg APB$ in constructing $\mathcal{M}^a$ because the possibility of APB holding has already been rejected.

Heuristics 3: Learning from inconsistencies arising from public announcements If an agent a reaches an inconsistency as a result of the factual public announcement of φ, then φ is valuable information for guiding the *a priori* reasoning process; in fact, to recover consistency, φ should be true in at least one world of the equivalence class U^a of the trial model, so that when the public announcement is re-applied, the inconsistency would not arise again.

Heuristics 4: Changing the underlying logic It is also possible to relax the restrictions imposed on the trial and/or backup models. For instance, a may try to explain the beliefs of b and c by the lack of positive and/or negative introspection on their part, which would necessitate the relaxation of the requirements of transitivity and/or euclideanity on the backup model.

5. Properties of *A priori* Belief Updates

This would be an appropriate place to list the axioms of *a priori* belief updates, perhaps, akin to axioms of PAL for public announcements from Plaza (1989). Unfortunately, this does not seem to be any easier to do than providing a finite syntactic description of common epistemic puzzles. Even restricting our attention to finite models only, the question amounts to asking whether any finite combination of maximal consistent sets for the logic can be represented by a formula (or finitely many formulas). The problem is that there are uncountably many maximal consistent sets (for an infinite set of atomic propositions), hence, uncountably many finite combinations thereof, while only countably many formulas (cf. Artemov (2022)). It seems difficult, if not hopeless, to describe syntactically the result of an update when no syntactic description of the update exists.

Thus, in this section, we try to "model check" rather than axiomatize the results of *a priori* belief updates.

Theorem 23. *Let* $\mathsf{U} = (\mathcal{M}^a, U^a, \mathcal{M}^{\neg a}, \mapsto)$ *be a single-agent* a priori *belief update with trial model* $\mathcal{M}^a = \langle W^a, R^a, V^a\rangle$ *and backup model* $\mathcal{M}^{\neg a} = \langle W^{\neg a}, R^{\neg a}, V^{\neg a}\rangle$ *and let* $(\mathcal{M}, v)$ *be a pointed Kripke model with* $\mathcal{M} = \langle W, R, V\rangle$ *and* $R_a(v) = \varnothing$. *The following properties hold after agent* a*'s* a priori *belief update.*

1. *For any formula* ψ *that does not involve any modalities* B_b *for agents* $b \neq a$, *including for all purely propositional formulas, the following three statements are equivalent:*
   (*a*) $\mathcal{M} \circledcirc_a \mathsf{U}, v \vDash B_a\psi$;
   (*b*) $\mathcal{M}^a, u \vDash \psi$ *for all* $u \in U^a$;
   (*c*) $\mathcal{M}^a, u \vDash B_a\psi$ *for at least one* $u \in U^a$.
   In other words, agent a*'s factual beliefs and* a*'s beliefs about its own beliefs are fully determined by worlds from* U^a *of* $\mathcal{M}^a$.
2. $\mathcal{M}^{\neg a}, w \vDash \varphi$ *iff* $\mathcal{M} \circledcirc_a \mathsf{U}, w \vDash \varphi$ *for any* $w \in W^{\neg a}$ *and any formula* φ. *In other words, agent* a*'s higher-order beliefs about other agents' beliefs are fully determined by* $\mathcal{M}^{\neg a}$.

Proof.

1. The first statement easily follows from the fact that the identity relation on U^a is an a-bisimulation between $\mathcal{M}^a$ and $\mathcal{M} \circledcirc_a \mathsf{U}$ (where a-bisimulation means that forth and back conditions are restricted to R_a transitions).

2. This follows from the fact that the identity relation on $W^{\neg a}$ is a full bisimulation between Kripke models $\mathcal{M}^{\neg a}$ and $\mathcal{M} \circledcirc_a \mathsf{U}$. □

Remark 24. For an *a priori* update $\mathsf{U} = (\mathcal{M}^a, U^a, \mathcal{M}^{\neg a}, \mapsto)$, one might think that global properties of models $\mathcal{M}^a$ and $\mathcal{M}^{\neg a}$ would be transferred to a's part of the model after the *a priori* update U, so that after the update, a believes all global properties used in the update construction. This does hold for propositional validities, as follows from the preceding theorem. In fact, the truth of ψ in U^a of $\mathcal{M}^a$ is already sufficient to ensure a's post-update belief $B_a\psi$ if ψ is purely propositional. However, this transfer of global properties fails for epistemic formulas involving other agents. Consider, for instance, the *a priori* update U from Fig. 1(b) for Example 18. For formula ASF from (2), it is clear that $\mathcal{M}^a \vDash \neg ASF$ and $\mathcal{M}^{\neg a} \vDash \neg ASF$. After all, we know that, in the standard MCP, children do not step forward all at once unless all are muddy. However, in model $\mathcal{M} \circledcirc_a \mathsf{U}$ from Fig. 1(c), we have $\mathcal{M} \circledcirc_a \mathsf{U}, A \vDash ASF$, resulting in $\mathcal{M} \circledcirc_a \mathsf{U}, \underline{A} \vDash B_a ASF$. Thus, global assumptions about other agents' beliefs need not survive in the face of the mismatch between two different points of view: the one a reserves for itself vs. the one a assigns to others. This negative result also extends to implicit assumptions such as the factivity of beliefs, since in the same example, both $\mathcal{M}^a$ and $\mathcal{M}^{\neg a}$ represented agents with factive beliefs, but the model after the *a priori* update lacks reflexivity for b and c, predictably resulting in their beliefs not being factive.

Theorem 25. *When agent a's* a priori *update* U *triggered by a public statement φ of agent b results in agent a believing that agent b has inconsistent beliefs, (re)applying b's public statement after the* a priori *update affects neither a's factual beliefs, nor a's beliefs about the beliefs of a and b. In other words, if $\mathcal{M} \mid B_b\varphi, v \vDash B_a\bot$ and $(\mathcal{M} \mid B_b\varphi) \circledcirc_a \mathsf{U}, v \vDash B_a B_b \bot$, then for any formula ψ that does not involve any modalities B_c for agents $c \notin \{a, b\}$,*

$$(\mathcal{M} \mid B_b\varphi) \circledcirc_a \mathsf{U}, v \vDash B_a\psi \quad \Leftrightarrow \quad ((\mathcal{M} \mid B_b\varphi) \circledcirc_a \mathsf{U}) \mid_a B_b\varphi, v \vDash B_a\psi.$$

Proof. Indeed, $(\mathcal{M} \mid B_b\varphi) \circledcirc_a \mathsf{U}, v \vDash B_a B_b \bot$ means that there are no b-outgoing arrows from any world of U^a in $(\mathcal{M} \mid B_b\varphi) \circledcirc_a \mathsf{U}$. While the public update may affect worlds in $W^{\neg a}$ and, thereby, what a thinks of the beliefs of other agents (not a or b), the U^a part of the model before the public announcement update is ab-bisimilar to itself after the public announcement. □

6. Conclusions

To the best of our knowledge, this paper provides the first epistemic formalization of *a priori* belief updates, which are crucial for modeling of self-adaptive systems, where agents should possess self-correcting abilities.

We provide multiple examples demonstrating how such self-correction can be achieved in various versions of standard epistemic puzzles, e.g., consecutive numbers and muddy children, in cases where agents find themselves in an inconsistent epistemic state. We also provide examples of self-correction not resolving the inconsistency or resolving it in a way that results in several or even all agents arriving at false conclusions. In one case, the false conclusions include the inconsistency of other agents' beliefs, which prevents any future corrections no matter which communications occur. We prove the properties of *a priori* belief updates semantically, explain the difficulties of developing a logic of *a priori* updates, and provide counterexamples to several preservation properties, underscoring the arbitrary nature of *a priori* updates.

Related work. We believe that some aspects of the presented *a priori* belief updates could be implemented in plausibility models via plausibility change (van Ditmarsch, 2005; Baltag and Smets, 2008; van Benthem and Smets, 2015). However, this would require embedding all *a priori* updates an agent might ever consider into the initial model. Not only would this lead to a significant increase of the initial model size, but it would also mean that all possible self-correcting subroutines, as well as their order of application, are pre-programmed, turning them from autonomous *a priori* actions into deterministic *a posteriori* private updates. This approach may have merits for traditional distributed systems but is less suitable for SASO systems.

Future work. We aim to extend the update mechanism to more general action models (van Ditmarsch *et al*., 2007) and develop logics for restricted types of *a priori* belief updates, which may be applicable to the belief update synthesis problem. We also plan to address the question of communicating new *a priori* beliefs (i.e., modeling *a posteriori* updates about *a priori* beliefs), which would enable other agents to guide the recovery (or dually, state corruption) of the agent in question.

It would also be interesting to explore how these methods would look in the formalism of simplicial complexes (van Ditmarsch *et al*., 2021; Goubault *et al*., 2021), which are categorically dual to Kripke models. For instance, the operation of choosing a new local state, which

amounts to identifying a suitable equivalence class in a Kripke model, would, in a simplicial model, correspond to choosing a new singleton vertex. This appears to be a more basic operation which suggests that the process of *a priori* belief updates may look more natural when presented simplicially.

Acknowledgments

We are grateful to Hans van Ditmarsch for his discerning comments on an earlier version of this paper, as well as to Stephan Felber, Kristina Fruzsa, Rojo Randrianomentsoa, Hugo Rincón Galeana, Ulrich Schmid, Thomas Schlögl, and Tuomas Tahko for multiple illuminating and inspiring discussions. We thank the anonymous reviewers for their helpful comments. This research was funded in whole or in part by the Austrian Science Fund (FWF) projects ByzDEL [https://doi.org/10.55776/P33600], LoDEx [https://doi.org/10.55776/I6372], and ZK 35 — High-dimensional statistical learning:New methods to advance economic and sustainability policy[https://zk35.org/].

References

Artemov, S. (2020). Observable models. *Logical Foundations of Computer Science: In- ternational Symposium, LFCS 2020, Deerfield Beach, FL, USA, January 4–7, 2020, Proceedings,* Volume 11972 of *Lecture Notes in Computer Science*, 12–26. Springer. https://doi.org/10.1007/978-3-030-36755-8_2

Artemov, S. (2022). Towards syntactic epistemic logic. *Fundameta Informaticae*, 186 (1–4), 45–62. https://fi.episciences.org/10792/

Baltag, A., Moss, L. S., and Solecki, S. (1998). The logic of public announcements, common knowledge, and private suspicions. *Theoretical Aspects of Rationality and Knowledge: Proceedings of the Seventh Conference* (*TARK 1998*), 43–56. Morgan Kaufmann. http://tark.org/proceedings/tark_jul22_98/p43-baltag.pdf

Baltag, A. and Smets, S. (2008). A qualitative theory of dynamic interactive belief revision. *Logic and the Foundations of Game and Decision Theory* (*LOFT 7*), Volume 3 of Texts in *Logic and Games*, 11–58. Amsterdam University Press. https://www.jstor.org/stable/j.ctt46mz4h.4

van Benthem, J. and Smets, S. (2015). Dynamic logics of belief change. *Handbook of Epistemic Logic*, 313–393. College Publications.

Ben-Zvi, I. and Moses, Y. (2014). Beyond Lamport's *Happened-before*: On time bounds and the ordering of events in distributed systems. *Journal of the ACM*, 61(2), 13:1–13:26. https://doi.org/10.1145/2542181

Berns, A. and Ghosh, S. (2009). Dissecting self-* properties. *SASO 2009: Third IEEE International Conference on Self-Adaptive and Self-Organizing Systems,* 10–19. IEEE. https://doi.org/10.1109/SASO.2009.25

Castañeda, A., Gonczarowski, Y. A., and Moses, Y. (2022). Unbeatable consensus. *Distributed Computing*, 35(2), 123–143. https://doi.org/10.1007/s00446-021-00417-3

Chagrov, A. and Zakharyaschev, M. (1997). *Modal Logic*, Volume 35 of *Oxford Logic Guides*. Clarendon Press. https://doi.org/10.1093/oso/9780198537793.001.0001

Cignarale, G., Kuznets, R., and Schlögl, T. (2024). Minimizing agents' state corruption resulting from leak-free epistemic communication modeling. *Foundations of Information and Knowledge Systems: 13th International Symposium, FoIKS 2024, Sheffield, UK, April 8–11, 2024, Proceedings,* Volume 14589 of *Lecture Notes in Computer Science*, 165–181. Springer. https://doi.org/10.1007/978-3-031-56940-1_9

Cignarale, G., Schmid, U., Tahko, T., and Kuznets, R. (2023). The role of a priori belief in the design and analysis of fault-tolerant distributed systems. *Minds and Machines*, 33(2), 293–319. https://doi.org/10.1007/s11023-023-09631-3

Coulouris, G., Dollimore, J., Kindberg, T., and Blair, G. (2011). *Distributed Systems: Concepts and Design* (5th ed.). Addison–Wesley.

van Ditmarsch, H. (2005). Prolegomena to dynamic logic for belief revision. *Synthese*, 147(2), 229–275. https://doi.org/10.1007/s11229-005-1349-7

van Ditmarsch, H., Goubault, É., Lazić, M., Ledent, J., and Rajsbaum, S. (2021). A dynamic epistemic logic analysis of equality negation and other epistemic covering tasks. *Journal of Logical and Algebraic Methods in Programming*, 121, 100662. https://doi.org/10.1016/j.jlamp.2021.100662

van Ditmarsch, H., van der Hoek, W., and Kooi, B. (2007). *Dynamic Epistemic Logic*, Volume 337 of Synthese Library. Springer. https://doi.org/10.1007/978-1-4020-5839-4

van Ditmarsch, H. and Kooi, B. (2015). *One Hundred Prisoners and a Light Bulb*. Springer. https://doi.org/10.1007/978-3-319-16694-0

Fagin, R., Halpern, J. Y., Moses, Y., and Vardi, M. Y. (1995). *Reasoning About Knowledge*. MIT Press. https://doi.org/10.7551/mitpress/5803.001.0001

Fruzsa, K., Kuznets, R., and Schmid, U. (2021). *Fire! Proceedings of the Eighteenth Conference on Theoretical Aspects of Rationality and Knowledge*, *Beijing*, *China*, *June 25–27*, *2021*, Volume 335 of Electronic Proceedings in Theoretical Computer Science, 139–153. Open Publishing Association. https://doi.org/10.4204/EPTCS.335.13

Gerbrandy, J. and Groeneveld, W. (1997). Reasoning about information change. *Journal of Logic, Language, and Information*, 6(2), 147–169. https://doi.org/10.1023/A:1008222603071

Gettier, E. L. (1963). Is justified true belief knowledge? *Analysis*, 23(6), 121–123. https://doi.org/10.1093/analys/23.6.121

Goren, G. and Moses, Y. (2020). Silence. *Journal of the ACM*, 67(1), 3:1–3:26. https://doi.org/10.1145/3377883

Goubault, É., Ledent, J., and Rajsbaum, S. (2021). A simplicial complex model for dynamic epistemic logic to study distributed task computability. *Information and Computation*, 278, 104597. https://doi.org/10.1016/j.ic.2020.104597

Halpern, J. Y. and Moses, Y. (1990). Knowledge and common knowledge in a distributed environment. *Journal of the ACM*, 37(3), 549–587. https://doi.org/10.1145/79147.79161

Hintikka, J. (1962). *Knowledge and Belief: An Introduction to the Logic of the Two Notions*. Cornell University Press.

Kuznets, R., Prosperi, L., Schmid, U., and Fruzsa, K. (2019a). Causality and epistemic reasoning in byzantine multi-agent systems. *Proceedings of the Seventeenth Conference on Theoretical Aspects of Rationality and Knowledge*, *Toulouse*, *France*, *17–19 July 2019,* Volume 297 of *Electronic Proceedings in Theoretical Computer Science*, 293–312. Open Publishing Association. https://doi.org/10.4204/EPTCS.297.19

Kuznets, R., Prosperi, L., Schmid, U., and Fruzsa, K. (2019b). Epistemic reasoning with byzantine-faulty agents. *Frontiers of Combining Systems: 12th International Sympo- sium, FroCoS 2019*, *London*, *UK*, *September 4–6*, *2019*, *Proceedings*, Volume 11715 of *Lecture Notes in Artificial Intelligence*, 259–276. Springer. https://doi.org/10.1007/978-3-030-29007-8_15

Lewis, D. K. (1969). *Convention: A Philosophical Study*. Harvard University Press.

Lynch, N. A. (1996). *Distributed Algorithms*. Morgan Kaufmann.

Moses, Y. (2016). Relating knowledge and coordinated action: The knowledge of precondi- tions principle. *Proceedings Fifteenth Conference on Theoretical Aspects of Rationality and Knowledge*, *Carnegie Mellon University*, *Pittsburgh*, *USA*, *June 4–6*, *2015*, Volume 215 of *Electronic Proceedings in Theoretical Computer Science*, 231–245. Open Pub lishing Association. https://doi.org/10.4204/EPTCS.215.17

Moses, Y. and Tuttle, M. R. (1988). Programming simultaneous actions using common knowledge. *Algorithmica*, 3(1–4), 121–169. https://doi.org/10.1007/BF01762112

Plaza, J. A. (1989). Logics of public communications. *Proceedings of the Fourth International Symposium on Methodologies for Intelligent Systems: Poster Session Program*, 201–216. Oak Ridge National Laboratory.

Quine, W. V. (1951). Two dogmas of empiricism. *Philosophical Review*, 60(1), 20–43. https://doi.org/10.2307/2181906

Schlögl, T. and Schmid, U. (2023). A sufficient condition for gaining belief in byzantine fault- tolerant distributed systems. *Proceedings of the Nineteenth conference on Theoretical Aspects of Rationality and Knowledge*, *Oxford*, *United Kingdom*, *28–30th June 2023*, Volume 379 of *Electronic Proceedings in Theoretical Computer Science*, 487–497. Open Publishing Association. https://doi.org/10.4204/EPTCS.379.37

Tahko, T. E. (2008). A new definition of a priori knowledge: In search of a modal basis.Metaphysica, 9(1), 57–68. https://doi.org/10.1007/s12133-007-0022-7

Tahko, T. E. (2011). A priori and a posteriori: A bootstrapping relationship. *Metaphysica*, 12(2), 151–164. https://doi.org/10.1007/s12133-011-0083-5

Tomforde, S., Hähner, J., von Mammen, S., Gruhl, C., Sick, B., and Geihs, K. (2014). "Know thyself" — Computational self-reflection in intelligent technical systems. *SaSow 2014: 2014 IEEE Eighth International Conference on Self-Adaptive and Self-Organizing Systems Workshops*, 150–159. IEEE. https://doi.org/10.1109/SASOW.2014.25

About the Author

Giorgio Cignarale is a Ph.D. student, member of the LogiCS doctoral college at TU Wien in Vienna, Austria, under the supervision of Roman Kuznets and Ulrich Schmid.

✉ giorgio.cignarale@tuwien.ac.at

Roman Kuznets received a Ph.D. in Computer Science from the Graduate Center of the City University of New York, USA. He worked at the University of Bern in Switzerland and at TU Wien in Vienna, Austria, where, among others, he led the project Reasoning about Knowledge in Byzantine Distributed Systems (ByzDEL). He now works at the Institute of Computer Science of the Czech Academy of Sciences.

✉ roman@logic.at