Towards a Complete Characterization of Epistemic Reasoning: the Notion of Trust*

Francesco Fabiano[1][0000−0002−1161−0336]

Department of Mathematics, Computer Science and Physics,
University of Udine, Via delle Scienze 206, 33100 Udine, Italy
francesco.fabiano@uniud.it

Abstract. Designing autonomous agents, that interact with others to perform complex tasks, has always been one of the main objective of the Artificial Intelligence community.
For such systems to be employed in complex scenarios, where the information about others is key (*e.g.*, self-driving cars), it is necessary to define robust formalisms that allow each agent to act considering her beliefs on both: i) the state of the world; and ii) the other agents' perspective of it. The branch of AI that studies such formalisms is known in literature as Multi-Agent Epistemic Planning (MEP). The epistemic action-based language $m\mathcal{A}^\rho$, to the best of our knowledge, is the most comprehensive tool to model MEP domains but still lacks concepts that are necessary to reason on real-world scenarios.
In this paper we introduce the actions *(un)trustworthy announcement* and *(mis)trustworthy announcement* for $m\mathcal{A}^\rho$. These actions increase the language's expressiveness introducing the notion of trust, therefore allowing for a more profound representation of real-world scenarios. In particular, we will provide the characterization, along with some desired properties, of the aforementioned actions' transition functions. Finally, we will discuss the importance of formalizing the concept of trust in the MEP problem.

Keywords: Epistemic Action Languages · Planning · Multi-agent · Knowledge/Belief Representation.

1 Introduction

Recently, techniques derived from the fields of automated reasoning and knowledge representation have been heavily exploited in both our daily life and in the industry. The natural evolution of such applications, *i.e.*, systems that involve hundreds of agents each acting upon her beliefs to achieve her own goals (*e.g.*, self-driving cars), is going to be widely deployed in just few years. The branch of AI interested in studying and modeling such agent-based technologies is referred

to as *automated planning*. In particular, *multi-agent planning* [1, 6–8, 13] provides a powerful tool to model scenarios comprised of agents that interact with each other. To maximize the potentials of such autonomous systems each agent should be able to reason on both: i) her perspective of the "concrete" world; and ii) her beliefs of the other agents' perspective of the environment—that is, their viewpoint of the "concrete" world and of the others' perspective of it. The planning problem in this new setting is referred to as *multi-agent epistemic planning* in the literature.

Nevertheless, as said in [8] '*reasoning about knowledge and beliefs is not as direct as reasoning on the "physical" state of the world*'. Already existing epistemic action languages [2, 3, 8, 14] are able to model several families of problems and to study their information flows but cannot comprehensively reason on aspects like trust, dishonesty, deception, and incomplete knowledge. In order to exploit epistemic reasoning in complex real-world scenarios, *e.g.*, economy, security, justice and politics, it is then necessary to increase the expressiveness of the aforementioned languages.

In this paper we expand the language $m\mathcal{A}^{\rho}$ [8], to the best of our knowledge the most comprehensive epistemic language, with a formalization of the concept of *Trust*. We do so by introducing two different actions that formalize the information sharing when the idea of trust is involved:

i) *(un)trustworthy announcement* and;
ii) *(mis)trustworthy announcement*.

In particular, i) *(un)trustworthy announcement* formalizes the situation when the *untrusty* agents will not change their beliefs about the world no matter what the announcer says; and ii) *(mis)trustworthy announcement* captures the scenarios where the announcer, when not trusted, is believed to have a systematic faulty perception of the announced environment's properties. This leads the *untrusty* agents to believe the opposite of what has been announced.

The paper is organized as follows: Section 2 will present the field of epistemic reasoning. The background will be then concluded with Section 3 where we will introduce the epistemic action language $m\mathcal{A}^{\rho}$. In Section 4 we will present the semantics of the newly formalized actions along with some desired properties, formally demonstrated in the Supplementary Documentation (available at http://clp.dimi.uniud.it/sw/). Finally, in Section 5 we will discuss the impact of the new actions and some possible future developments.

Moreover, in the Supplementary Documentation we also provide the formalization of the *(un/mis)trustworthy announcement* actions for $m\mathcal{A}^{*}$ [2], the language on which $m\mathcal{A}^{\rho}$ is based on.

2 Epistemic reasoning

The research on autonomous reasoners has lead, among other things, to the formalization of the well-known planning problem [15] and to the introduction of several *modal logics* [5,16,17] used to describe different properties of the *world*.

Different logics allow diverse types of reasoning and bring with them different implications in terms of expressiveness and complexity.

In particular, *Dynamic Epistemic Logic* (DEL), the foundation of *multi-agent epistemic planning* (MEP), is used to reason not only on the state of the world but also on *information change*. As said in [17], 'information *is something relative to a subject who has a certain perspective on the world, called an* agent*, and that is meaningful as a whole, not just loose bits and pieces. This makes us call it* knowledge *and, to a lesser extent,* belief'. Concretely, DEL provides a formalization that allows to model and reason about the agents' perspective of the world and of the other agents' viewpoint (on both the world and the others' perspective). Therefore, DEL and MEP are tools that can be exploited when (possibly nested) knowledge/belief needs to be taken into consideration. Some examples of such domains can be ethical reasoning, economical or political strategies and juridic scenarios.

In what follows, we will provide a short description of the basic concepts that define DEL and MEP. As it is beyond the scope of this work to give an exhaustive introduction, the interested reader can refer to [10] for a complete characterization of such concepts.

Let $\mathcal{AG}$ be a finite set of agents s.t. $|\mathcal{AG}| = n$ with $n \geq 1$ and let $\mathcal{F}$ be a set of propositional variables, called *fluents*. Each *world* is described by a subset of elements of $\mathcal{F}$ (intuitively, those that are "true"). Moreover, in epistemic logic each agent $\mathsf{ag} \in \mathcal{AG}$ is associated to an epistemic modal operator $\mathbf{B}_{\mathsf{ag}}$ that represents the knowledge/belief of ag herself. Finally, epistemic *group operators* $\mathbf{E}_\alpha$ and $\mathbf{C}_\alpha$ are also introduced in epistemic logic. Intuitively, $\mathbf{E}_\alpha$ and $\mathbf{C}_\alpha$ represent the knowledge/belief of a group of agents α and the *common knowledge/belief* of α, respectively. To be more precise, as in [2], we have that:

Definition 1 (Fluent formula). *A* fluent formula *is a propositional formula built using fluents in $\mathcal{F}$ as propositional variables and the propositional operators $\wedge, \vee, \Rightarrow, \neg$. A* fluent atom *is a formula composed of just an element $\mathsf{f} \in \mathcal{F}$; a* fluent literal *is either a fluent atom $\mathsf{f} \in \mathcal{F}$ or its negation $\neg\mathsf{f}$.*

With a slight abuse of notation, we will refer to fluent literals simply as *fluents*.

Definition 2 (Belief formula). *A* belief formula *is defined as follows:*

- *A fluent formula is a belief formula;*
- *If φ is a belief formula and $\mathsf{ag} \in \mathcal{AG}$, then $\mathbf{B}_{\mathsf{ag}}\varphi$ is a belief formula;*
- *If φ_1, φ_2 and φ_3 are belief formulae, then $\neg\varphi_3$ and $\varphi_1 \,\mathsf{op}\, \varphi_2$ are belief formulae, where $\mathsf{op} \in \{\wedge, \vee, \Rightarrow\}$;*
- *If φ is a belief formula and $\emptyset \neq \alpha \subseteq \mathcal{AG}$ then $\mathbf{E}_\alpha\varphi$ and $\mathbf{C}_\alpha\varphi$ are belief formulae.*

Example 1. Let us consider the formula $\mathbf{B}_{\mathsf{ag}_1}\mathbf{B}_{\mathsf{ag}_2}\varphi$. This formula expresses that the agent ag_1 believes that the agent ag_2 believes that φ is true. The formula $\mathbf{B}_{\mathsf{ag}_1}\neg\varphi$ expresses that the agent ag_1 believes that φ is false.

Let us also introduce the notion of *multi-agent epistemic planning domain*. Intuitively, an epistemic planning domain contains all the necessary information to define a planning problem in a multi-agent epistemic scenario.

Definition 3 (Multi-agent epistemic planning domain). *We define a multi-agent epistemic domain as the tuple* $D = \langle \mathcal{F}, \mathcal{AG}, \mathcal{A}, \varphi_i, \varphi_g \rangle$ *where:*

- $\mathcal{F}$ *is the set of all the* fluents *of* D*;*
- $\mathcal{AG}$ *is the set of the* agents *of* D*;*
- $\mathcal{A}$ *represents the set of all the* actions *of* D*;*
- φ_i *is the belief formula that describes the* initial conditions *of the planning process; and*
- φ_g *is the belief formula that represents the* goal *condition.*

Moreover, from now on, with the term *action instance* we will indicate an element of the set $\mathcal{AI} = \mathcal{A} \times \mathcal{AG}$. Intuitively, an action instance $\mathsf{a}\langle\mathsf{ag}\rangle$ identifies the execution of the action a by the agent ag.

Given a domain D we will refer to its components through the *parenthesis* operator. For instance to access the elements $\mathcal{F}$ and $\mathcal{AG}$ of D we will use the more compact notation $D(\mathcal{F})$ and $D(\mathcal{AG})$, respectively.

Furthermore, we will indicate a state of an epistemic planning domain as *e-state*. Intuitively, an e-state contains all the information needed to encode both the concrete properties of the world and the knowledge/belief relations. The language $m\mathcal{A}^{\rho}$, derived by the language $m\mathcal{A}^{*}$ [2] (based on the "classical" *Kripke structures*), expresses the idea of e-state through *possibilities* [8, 11]. In the following section, we will provide a short introduction for $m\mathcal{A}^{\rho}$.

3 The Epistemic action language $m\mathcal{A}^{\rho}$

Let us briefly introduce the epistemic action language $m\mathcal{A}^{\rho}$ [8]. Let us note that the fundamental concepts of the language are inherited from $m\mathcal{A}^{*}$, the action language firstly introduced in [2] on which $m\mathcal{A}^{\rho}$ is based.

First, we need to define the three different types of action used by $m\mathcal{A}^{\rho}$ to model the e-states update:

- *World-altering* action (also called *ontic*): used to modify certain properties (*i.e.*, fluents) of the world;
- *Sensing* action: used by an agent to refine her beliefs about the world;
- *Announcement* action: used by an agent to affect the beliefs of other agents.

The action language also allows to specify, for each action instance $\mathsf{a}\langle\mathsf{ag}\rangle$, the observability relation of each agent. Namely, an agent x may be *fully observant* ($\mathsf{x} \in F$), *partially observant* ($\mathsf{x} \in P$), or *oblivious* ($\mathsf{x} \in O$) w.r.t. $\mathsf{a}\langle\mathsf{ag}\rangle$. If an agent is fully observant, then she is aware of both the execution of the action instance and its effects; she is partially observant if she is only aware of the action execution but not of the outcomes; she is oblivious if she is ignorant of the execution of the action. More precisely, given an action instance $\mathsf{a}\langle\mathsf{ag}\rangle$, a fluent literal f, a fluent formula ϕ and the belief formula φ, the syntax of $m\mathcal{A}^{\rho}$ is defined as follows:

- **executable** a **if** φ: captures the *executability conditions*;
- a **causes** f **if** φ: captures the effects of *ontic* actions;
- a **determines** f **if** φ: captures the effects of *sensing* actions;
- a **announces** ϕ **if** φ: captures the effects of *announcement* actions;
- ag **observes** a **if** φ: captures *fully observant* agents for an action; and
- ag **aware_of** a **if** φ: captures *partially observant* agents for a given action.

Notice that if we do not state otherwise, an agent will be considered oblivious. Finally, statements of the form **initially** φ and **goal** φ capture the initial and goal conditions, respectively.

The language $m\mathcal{A}^{\rho}$, introduced in [8,9], bases the e-states representation on the idea of *possibility*, firstly defined in [11]. Possibilities are *non-well-founded* objects and, therefore, exploit concepts such as *recursion* and *bisimulation*. In particular, the former is used to define the idea of e-state update while the latter is needed to capture the idea of e-state equality. Due to space constraints we will illustrate only the main ideas and intuitions behind the semantics of $m\mathcal{A}^{\rho}$ addressing the reader to [11] and [8] for a complete introduction to possibilities and $m\mathcal{A}^{\rho}$, respectively. Let us now introduce more formally the concept of possibility.

Definition 4 (Possibility [11]).

- *A* possibility u *is a function that assigns to each fluent* $\mathsf{f} \in \mathcal{F}$ *a truth value* $\mathsf{u}(\mathsf{f}) \in \{0,1\}$ *and to each agent* $\mathsf{ag} \in \mathcal{AG}$ *an information state* $\mathsf{u}(\mathsf{ag}) = \sigma$*;*
- *An information state* σ *is a set of possibilities.*

Intuitively, a possibility u allows to capture the concept of e-state by: i) encoding a possible world through the truth values $\mathsf{u}(\mathsf{f})$ $\forall \mathsf{f} \in \mathcal{F}$; and ii) capturing the beliefs of an agent $\mathsf{ag} \in \mathcal{AG}$ thanks to the assignment of information states $\mathsf{u}(\mathsf{ag})$. Since possibilities are non-well-founded objects, the concepts of *state* and *possible world* collapse. In fact, a possibility contains both the information of a possible world and the information about the agents' beliefs (represented by other possibilities).

Definition 5 (Entailment w.r.t. possibilities [9]). *Let the belief formulae* $\varphi, \varphi_1, \varphi_2$*, a fluent* f*, an agent* ag*, a (non-empty) group of agents* α*, and a possibility* u *be given.*

1. $\mathsf{u} \models \mathsf{f}$ *if* $\mathsf{u}(\mathsf{f}) = 1$*;*
2. $\mathsf{u} \models \neg\varphi$ *if* $\mathsf{u} \not\models \varphi$*;*
3. $\mathsf{u} \models \varphi_1 \vee \varphi_2$ *if* $\mathsf{u} \models \varphi_1$ *or* $\mathsf{u} \models \varphi_2$*;*
4. $\mathsf{u} \models \varphi_1 \wedge \varphi_2$ *if* $\mathsf{u} \models \varphi_1$ *and* $\mathsf{u} \models \varphi_2$*;*
5. $\mathsf{u} \models \boldsymbol{B}_{\mathsf{ag}}\varphi$ *if for each* $\mathsf{v} \in \mathsf{u}(\mathsf{ag})$ *it holds that* $\mathsf{v} \models \varphi$*;*
6. $\mathsf{u} \models \boldsymbol{E}_{\alpha}\varphi$ *if for all* $\mathsf{ag} \in \alpha$ *it holds that* $\mathsf{u} \models \boldsymbol{B}_{\mathsf{ag}}\varphi$ *;*
7. $\mathsf{u} \models \boldsymbol{C}_{\alpha}\varphi$ *if* $\mathsf{u} \models \boldsymbol{E}^{k}_{\alpha}\varphi$ *for every* $k \geq 0$*, where* $\boldsymbol{E}^{0}_{\alpha}\varphi = \varphi$ *and* $\boldsymbol{E}^{k+1}_{\alpha}\varphi = \boldsymbol{E}_{\alpha}(\boldsymbol{E}^{k}_{\alpha}\varphi)$*.*

For the sake of readability we will omit the complete specification of the ontic, sensing and annoumcement transition function. The interested reader is referred to [8].

4 (un/mis)Trustworthy announcement

Following we will provide a formal definition of the actions *(un)trustworthy announcement* and *(mis)trustworthy announcement* that capture two scenarios where the concept of trust influences the communication between agents. That is, an agent can or cannot trust what another agent is telling her and act consequently. We will provide a formal definition of e-state update for these actions for $m\mathcal{A}^{\rho}$. The expression '`ag` **t_announces/m_announces** `a` **if** φ' is the syntax to indicate that the agent `ag` is executing an *(un/mis)trustworthy announcement*.

In defining the actions we consider a static and globally visible version of 'trust' that can be formalized with a simple function $\mathcal{T} : \mathcal{AG} \times \mathcal{AG} \mapsto \{0, 1\}$. For the sake of readability we will consider only the case where $\mathcal{T}$ is a static and globally visible function. Let us notice that having $\mathcal{T}$ to be dynamic is easily achievable. In particular, we just need to define how $\mathcal{T}$ may vary, *e.g.*, making the function depending on some fluents value. For the sake of simplicity let us imagine $\mathcal{T}$ to be fixed and not dependent from the plan execution. On the other hand, making $\mathcal{T}$ not globally visible—*i.e.*, each agent knows her own version of the *trust* function—is not straightforward. The problem arise when two agents have different views of the same trust relation leading to the generation of *non-consistent beliefs*, an open problem in the MEP community. We leave the investigation of this scenario as future work.

To clarify the e-state update after the execution of the new actions we will also present a graphical representation of the transition function application.

The examples of execution will be based on a variation of the *Grapevine* domain [12]. Let us now present this new domain, referred to as *Trust_Grapevine*:

Domain 1 (Trust_Grapevine) *$n \geq 2$ agents are located in $k \geq 2$ rooms. Each agent knows $j \geq 0$ secrets. An agent can move freely to each other room, and she can share a "secret" with the agents that are in the room with her. Moreover the agents will be aware of the execution of announcements made in adjacent rooms without actually knowing the truth value of the announced fluent. Each agent can or cannot* trust *(or* mistrust*) what another agent shares.*

Let us notice that since the idea of trust is involved each agent, in order to learn a secret, needs to witness an announcement of agents that she trusts, making the newly presented domain slightly more intricate than the original Grapevine.

4.1 (un)Trustworthy announcement

We can now introduce the transition function of the action *(un)trustworthy announcement* for $m\mathcal{A}^{\rho}$. Intuitively, this action models an announcement where the listening agents can or cannot trust the announcer. That is: i) the trusty agents will update their belief consistently with what has been announced; and ii) the *untrusty*[1] ones will maintain their beliefs about the world and will update

[1] The agents that are fully observant w.r.t. announcement but that do not trust the announcer.

their perspective on the beliefs of the trusty agents. Let us recall that the sets $F_{\mathsf{a}}, P_{\mathsf{a}}, O_{\mathsf{a}}$ represent the set of fully observant, partially observant and oblivious agents w.r.t. to the execution of an action instance $\mathsf{a}\langle\mathsf{ag}\rangle$, respectively.

Let a domain D, its set of action instances $D(\mathcal{AI})$, and the set $\mathcal{S}$ of all the possibilities reachable from $D(\varphi_i)$ with a finite sequence of action instances be given. The transition function $\Phi : D(\mathcal{AI}) \times \mathcal{S} \to \mathcal{S} \cup \{\emptyset\}$ for the *(un)trustworthy announcement* relative to D is defined as follows.

Definition 6 ($m\mathcal{A}^{\rho}$ *(un)trustworthy announcement* transition function). *Allow us to use the compact notation* $\mathsf{u}(\mathcal{F}) = \{\mathsf{f} \mid \mathsf{f} \in D(\mathcal{F}) \wedge \mathsf{u} \models \mathsf{f}\} \cup \{\neg\mathsf{f} \mid \mathsf{f} \in D(\mathcal{F}) \wedge \mathsf{u} \not\models \mathsf{f}\}$ *for the sake of readability. Let an action instance* $\mathsf{a}\langle\mathsf{ag}\rangle \in D(\mathcal{AI})$ *where agent* $\mathsf{ag} \in D(\mathcal{AG})$ *announces the fluent formula* ϕ *and a possibility* $\mathsf{u} \in \mathcal{S}$ *be given.*
If a *is not executable in* u*, then* $\Phi(\mathsf{a},\mathsf{u}) = \emptyset$ *otherwise* $\Phi(\mathsf{a},\mathsf{u}) = \mathsf{u}'$*, where:*

$$e(\mathsf{a},\mathsf{u}) = \begin{cases} 0 & \text{if } \mathsf{u} \models \phi \\ 1 & \text{if } \mathsf{u} \models \neg\phi \end{cases}$$

$$\mathsf{u}'(\mathcal{F}) = \mathsf{u}(\mathcal{F})$$

$$\mathsf{u}'(\mathsf{ag_i}) = \begin{cases} \mathsf{u}(\mathsf{ag_i}) & \text{if } \mathsf{ag_i} \in O_{\mathsf{a}} \\ \bigcup\limits_{\mathsf{w}\in\mathsf{u}(\mathsf{ag_i})} \Upsilon(\mathsf{a},\mathsf{w}) \cup \Psi(\mathsf{a},\mathsf{w}) & \text{if } \mathsf{ag_i} \in P_{\mathsf{a}} \\ \bigcup\limits_{\mathsf{w}\in\mathsf{u}(\mathsf{ag_i})} \Upsilon(\mathsf{a},\mathsf{w}) & \text{if } \mathsf{ag_i} \in F_{\mathsf{a}} \wedge e(\mathsf{a},\mathsf{u}) = 1 \\ \bigcup\limits_{\mathsf{w}\in\mathsf{u}(\mathsf{ag_i})} \Psi(\mathsf{a},\mathsf{w}) & \text{if } \mathsf{ag_i} \in F_{\mathsf{a}} \wedge e(\mathsf{a},\mathsf{u}) = 0 \end{cases}$$

with $\Upsilon(\mathsf{a},\mathsf{w}) = \mathsf{w}'$ *such that*

$$\mathsf{w}'(\mathcal{F}) = \mathsf{w}(\mathcal{F})$$

$$\mathsf{w}'(\mathsf{ag_i}) = \begin{cases} \mathsf{w}(\mathsf{ag_i}) & \text{if } \mathsf{ag_i} \in O_{\mathsf{a}} \\ \bigcup\limits_{\mathsf{v}\in\mathsf{w}(\mathsf{ag_i})} \Phi(\mathsf{a},\mathsf{v}) & \text{if } \mathsf{ag_i} \in P_{\mathsf{a}} \\ \bigcup\limits_{\mathsf{v}\in\mathsf{w}(\mathsf{ag_i})} \Upsilon(\mathsf{a},\mathsf{v}) & \text{if } \mathsf{ag_i} \in F_{\mathsf{a}} \wedge \mathcal{T}(\mathsf{ag_i},\mathsf{ag}) = 0 \\ \bigcup\limits_{\mathsf{v}\in\mathsf{w}(\mathsf{ag_i}):e(\mathsf{a},\mathsf{v})=1} \Upsilon(\mathsf{a},\mathsf{v}) & \text{if } \mathsf{ag_i} \in F_{\mathsf{a}} \wedge \mathcal{T}(\mathsf{ag_i},\mathsf{ag}) = 1 \end{cases}$$

and $\Psi(\mathsf{a},\mathsf{w}) = \mathsf{w}'$ *such that*

$$\mathsf{w}'(\mathcal{F}) = \mathsf{w}(\mathcal{F})$$

$$\mathsf{w}'(\mathsf{ag_i}) = \begin{cases} \mathsf{w}(\mathsf{ag_i}) & \text{if } \mathsf{ag_i} \in O_{\mathsf{a}} \\ \bigcup\limits_{\mathsf{v}\in\mathsf{w}(\mathsf{ag_i})} \Phi(\mathsf{a},\mathsf{v}) & \text{if } \mathsf{ag_i} \in P_{\mathsf{a}} \\ \bigcup\limits_{\mathsf{v}\in\mathsf{w}(\mathsf{ag_i})} \Psi(\mathsf{a},\mathsf{v}) & \text{if } \mathsf{ag_i} \in F_{\mathsf{a}} \wedge \mathcal{T}(\mathsf{ag_i},\mathsf{ag}) = 0 \\ \bigcup\limits_{\mathsf{v}\in\mathsf{w}(\mathsf{ag_i}):e(\mathsf{a},\mathsf{v})=0} \Psi(\mathsf{a},\mathsf{v}) & \text{if } \mathsf{ag_i} \in F_{\mathsf{a}} \wedge \mathcal{T}(\mathsf{ag_i},\mathsf{ag}) = 1 \end{cases}$$

with $1 \leq \mathsf{i} \leq |D(\mathcal{AG})|$.

Intuitively this transition function allows, through the use of Υ and Ψ, to model the idea that the *untrusty* agents maintain their beliefs while knowing that the trusty ones updated their point of view of the "physical world" (and viceversa).

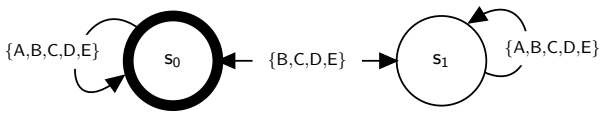
An example of execution As mentioned above, we will provide a graphical representation of the newly introduced transition function. Following [9], we will represent a possibility as a graph where the nodes correspond the possible worlds while the edges encode the beliefs of the agents. The thicker node represents the pointed possibility. To extract the point of view of the agents from a graph we need to follow the entailment rules (Definition 5) starting from the pointed possibility. Let us now briefly describe the example initial state (based on Domain 1). Since we are only interested in showing how to e-state update works we will omit the actions and goal description.

Example 2 (Five Agents Trust_Grapevine).

- The example has five agents: A, B, C, D and E;
- A, B, C are located in the same room (`room_1`) while D is in a room (`room_2`) adjacent to `room_1` and E is located in `room_3`, not adjacent to `room_1`;
- Agents B and D trust A while C and E do not;
- Agent A knows `secret_a`;
- The value of `secret_a` is `true`;
- Initially everyone knows the position of each agent and that only A knows the value of `secret_a`.

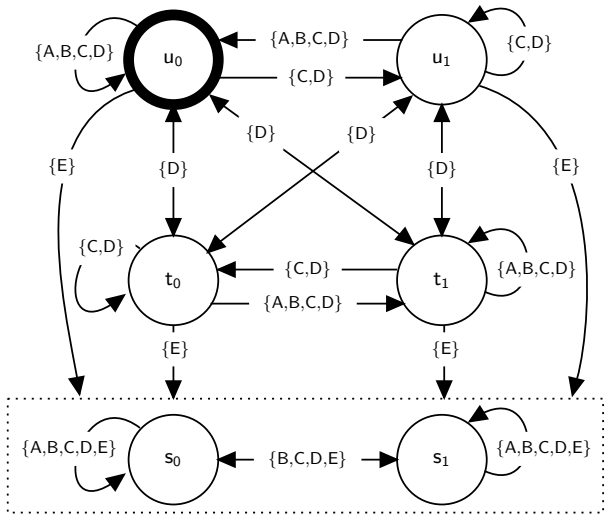
Let us now present a graphical representation of the above described initial state, in Figure 1.

In Figure 2 instead we represent the e-state generated after the execution of the *(un)trustworthy announcement* action instance announce_secret_a⟨A⟩ (`ann_a` for brevity). In `ann_a` A announces the value of `secret_a`. Let us note that from the position of the agents we know that A, B, C $\in F_{\texttt{ann\_a}}$, D $\in P_{\texttt{ann\_a}}$ and E $\in O_{\texttt{ann\_a}}$.



s_0	secret_a, at_1_A, t_1_B, at_1_C, at_2_D, at_3_E
s_1	at_1_A, t_1_B, at_1_C, at_2_D, at_3_E

Fig. 1. The initial e-state described in Example 2. The bottom Table presents the fluents interpretation of each possibility; for clarity only the positive fluents are reported.



$u_0/t_0/s_0$	secret_a, at_1_A, t_1_B, at_1_C, at_2_D, at_3_E
$u_1/t_1/s_1$	at_1_A, t_1_B, at_1_C, at_2_D, at_3_E

Fig. 2. The e-state obtained after executing the *(un)trustworthy announcement* action ann_a in the e-state represented in Figure 1.

4.2 (mis)Trustworthy announcement

In Defintion 6 we assumed that an agent $\mathsf{ag_i}$, that does not trust the announcer, will not change her beliefs about what has been announced. That is, an *untrusty* agent will not change her perspective on the "physical" state of the world. Let us notice that this type of trust captures the idea that, for the *untrusty* agents, the announcer is not reliable and the information she is providing is not worth taking into consideration as it can be not accurate.

Depending on the scenario it could be necessary to model a stronger concept of untrust. In particular, it could be required to design an *(un)trustworthy announcement* such that the *untrusty* agents will believe the contrary of what has been announced (while still believing that the announcer believes what she announced). We will call this type of action *(mis)trustworthy announcement*. The formalization of such variation of the action presented in Defintion 6 is as follows.

Definition 7 ($m\mathcal{A}^{\rho}$ *(mis)trustworthy announcement* transition function). *Let an action instance* $\mathsf{a}\langle\mathsf{ag}\rangle \in D(\mathcal{AI})$ *where agent* $\mathsf{ag} \in D(\mathcal{AG})$ *announces the fluent formula* ϕ *and a possibility* $\mathsf{u} \in \mathcal{S}$ *be given.*

If a *is not executable in* u*, then* $\Phi(\mathsf{a},\mathsf{u}) = \emptyset$ *otherwise* $\Phi(\mathsf{a},\mathsf{u}) = \mathsf{u}'$*, where:*

$$e(\mathsf{a},\mathsf{u}) = \begin{cases} 0 & \textit{if } \mathsf{u} \models \phi \\ 1 & \textit{if } \mathsf{u} \models \neg\phi \end{cases}$$

$$\mathsf{u}'(\mathcal{F}) = \mathsf{u}(\mathcal{F})$$

$$\mathsf{u}'(\mathsf{ag_i}) = \begin{cases} \mathsf{u}(\mathsf{ag_i}) & \textit{if } \mathsf{ag_i} \in O_\mathsf{a} \\ \bigcup\limits_{\mathsf{w}\in\mathsf{u}(\mathsf{ag_i})} \Upsilon(\mathsf{a},\mathsf{w}) \cup \Psi(\mathsf{a},\mathsf{w}) & \textit{if } \mathsf{ag_i} \in P_\mathsf{a} \\ \bigcup\limits_{\mathsf{w}\in\mathsf{u}(\mathsf{ag_i})} \Upsilon(\mathsf{a},\mathsf{w}) & \textit{if } \mathsf{ag_i} \in F_\mathsf{a} \wedge e(\mathsf{a},\mathsf{u}) = 1 \\ \bigcup\limits_{\mathsf{w}\in\mathsf{u}(\mathsf{ag_i})} \Psi(\mathsf{a},\mathsf{w}) & \textit{if } \mathsf{ag_i} \in F_\mathsf{a} \wedge e(\mathsf{a},\mathsf{u}) = 0 \end{cases}$$

with $\Upsilon(\mathsf{a},\mathsf{w}) = \mathsf{w}'$ *such that*

$$\mathsf{w}'(\mathcal{F}) = \mathsf{w}(\mathcal{F})$$

$$\mathsf{w}'(\mathsf{ag_i}) = \begin{cases} \mathsf{w}(\mathsf{ag_i}) & \textit{if } \mathsf{ag_i} \in O_\mathsf{a} \\ \bigcup\limits_{\mathsf{v}\in\mathsf{w}(\mathsf{ag_i})} \Phi(\mathsf{a},\mathsf{v}) & \textit{if } \mathsf{ag_i} \in P_\mathsf{a} \\ \bigcup\limits_{\mathsf{v}\in\mathsf{w}(\mathsf{ag_i}):e(\mathsf{a},\mathsf{v})=0} \Upsilon(\mathsf{a},\mathsf{v}) & \textit{if } \mathsf{ag_i} \in F_\mathsf{a} \wedge \mathcal{T}(\mathsf{ag_i},\mathsf{ag}) = 0 \\ \bigcup\limits_{\mathsf{v}\in\mathsf{w}(\mathsf{ag_i}):e(\mathsf{a},\mathsf{v})=1} \Upsilon(\mathsf{a},\mathsf{v}) & \textit{if } \mathsf{ag_i} \in F_\mathsf{a} \wedge \mathcal{T}(\mathsf{ag_i},\mathsf{ag}) = 1 \end{cases}$$

and $\Psi(\mathsf{a},\mathsf{w}) = \mathsf{w}'$ *such that*

$$\mathsf{w}'(\mathcal{F}) = \mathsf{w}(\mathcal{F})$$

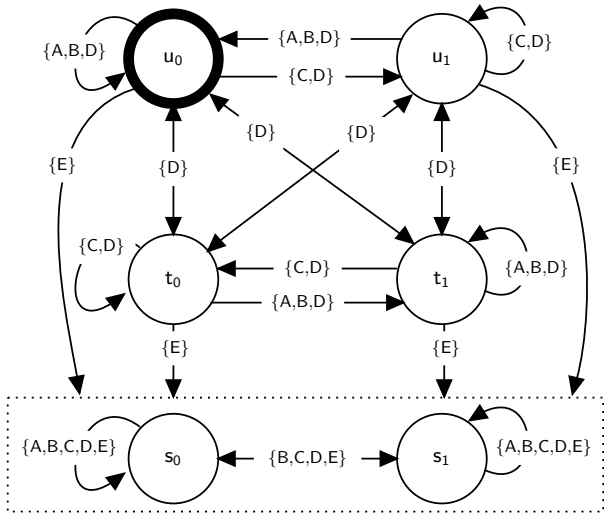
$$\mathsf{w}'(\mathsf{ag_i}) = \begin{cases} \mathsf{w}(\mathsf{ag_i}) & \textit{if } \mathsf{ag_i} \in O_\mathsf{a} \\ \bigcup\limits_{\mathsf{v}\in\mathsf{w}(\mathsf{ag_i})} \Phi(\mathsf{a},\mathsf{v}) & \textit{if } \mathsf{ag_i} \in P_\mathsf{a} \\ \bigcup\limits_{\mathsf{v}\in\mathsf{w}(\mathsf{ag_i}):e(\mathsf{a},\mathsf{v})=1} \Psi(\mathsf{a},\mathsf{v}) & \textit{if } \mathsf{ag_i} \in F_\mathsf{a} \wedge \mathcal{T}(\mathsf{ag_i},\mathsf{ag}) = 0 \\ \bigcup\limits_{\mathsf{v}\in\mathsf{w}(\mathsf{ag_i}):e(\mathsf{a},\mathsf{v})=0} \Psi(\mathsf{a},\mathsf{v}) & \textit{if } \mathsf{ag_i} \in F_\mathsf{a} \wedge \mathcal{T}(\mathsf{ag_i},\mathsf{ag}) = 1 \end{cases}$$

with $1 \leq \mathsf{i} \leq |D(\mathcal{AG})|$.

Let us note that the transition functions introduced in Definitions 6 and 7 only differ in the specification of Υ and Ψ for the *untrusty* fully observant agents. This difference is needed to represent the fact that in the case of *(un)trustworthy announcement* the *untrusty* agents maintain their beliefs while in the *(mis)trustworthy* one they will believe the opposite of what has been announced.

An example of execution As for the *(un)trustworthy announcement*, we will provide an example of *(mis)trustworthy announcement* execution. The initial

state is identical the one introduced in Example 2. The only difference is that now the action announce_secret_a⟨A⟩ (or ann_a for brevity) is a *(mis)trustworthy announcement* instead of a *(un)trustworthy announcement*. The initial state is, therefore, represented in Figure 1 while the e-state obtained after the execution of the *(mis)trustworthy announcement* is shown in Figure 3.



$u_0/t_0/s_0$	secret_a, at_1_A, t_1_B, at_1_C, at_2_D, at_3_E
$u_1/t_1/s_1$	at_1_A, t_1_B, at_1_C, at_2_D, at_3_E

Fig. 3. The e-state obtained after executing the *(mis)trustworthy announcement* action ann_a in the e-state represented in Figure 1.

4.3 Desired properties

In [8] are listed some useful properties that correctly capture certain intuitions concerning the effects of the various types of actions in $m\mathcal{A}^{\rho}$. Similarly, in what follows, we will provide some properties that the e-state update, after executing the *(un/mis)trustworthy announcement*, meets. Due to space constraint we will provide the formal demonstrations that these properties hold in the Supplementary Documentation[2]. As usual, we will indicate the sets of partially observant and oblivious agents (w.r.t. the action instance $\mathsf{a}\langle\mathsf{ag}\rangle$) with P_{a} and O_{a}, respectively. Moreover, we will indicate the set of trusty fully observant agents with F_{a} while will indicate the set of *untrusty* fully observant with U_{a}.

[2] Available at http://clp.dimi.uniud.it/sw/

Proposition 1 ((un)Trustworthy announcement properties). *Let* $\mathsf{a}\langle\mathsf{ag}\rangle$ *be an* (un)trustworthy announcement action instance *where* ag ***t_announces*** ϕ. *Let* e *be an e-state and let* e' *be its updated version (that is,* $\Phi(\mathsf{a},\mathsf{e}) = \mathsf{e}'$*), then in* $m\mathcal{A}^{\rho}$ *it holds that:*

1. $\mathsf{e}' \models \mathbf{C}_{F_\mathsf{a}}\phi$;
2. $\mathsf{e}' \models \mathbf{C}_{U_\mathsf{a}}(\mathbf{C}_{F_\mathsf{a}}\phi)$;
3. $\mathsf{e}' \models \mathbf{C}_{P_\mathsf{a}}(\mathbf{C}_{F_\mathsf{a}}\phi \vee \mathbf{C}_{F_\mathsf{a}}\neg\phi)$;
4. $\mathsf{e}' \models \mathbf{C}_{F_\mathsf{a}\cup U_\mathsf{a}}(\mathbf{C}_{P_\mathsf{a}}(\mathbf{C}_{F_\mathsf{a}}\phi \vee \mathbf{C}_{F_\mathsf{a}}\neg\phi))$;
5. *for every agent* $\mathsf{y} \in U_\mathsf{a}$, $\mathsf{e}' \models \mathbf{B}_\mathsf{y}\phi/\mathbf{B}_\mathsf{y}\neg\phi/(\neg\mathbf{B}_\mathsf{y}\phi \wedge \neg\mathbf{B}_\mathsf{y}\neg\phi)$ *iff* $\mathsf{e} \models \mathbf{B}_\mathsf{y}\phi/\mathbf{B}_\mathsf{y}\neg\phi/(\neg\mathbf{B}_\mathsf{y}\phi \wedge \neg\mathbf{B}_\mathsf{y}\neg\phi)$;
6. *for every agent* $\mathsf{y} \in O_\mathsf{a}$ *and a belief formula* φ, $\mathsf{e}' \models \mathbf{B}_\mathsf{y}\varphi$ *iff* $\mathsf{e} \models \mathbf{B}_\mathsf{y}\varphi$; *and*
7. *for every pair of agents* $\mathsf{x} \in F_\mathsf{a} \cup U_\mathsf{a} \cup P_\mathsf{a}$ *and* $\mathsf{y} \in O_\mathsf{a}$ *and a belief formula* φ, *if* $\mathsf{e} \models \mathbf{B}_\mathsf{x}\mathbf{B}_\mathsf{y}\varphi$ *then* $\mathsf{e}' \models \mathbf{B}_\mathsf{x}\mathbf{B}_\mathsf{y}\varphi$.

The properties presented in Proposition 1 try to capture some fundamental aspects of an *(un)trustworthy announcement* action. Intuitively:

1. Captures the idea that all the trusty fully observant agents should believe i) what has been announced; and ii) that all the other trusty fully observant agents believe what has been announced and so on *ad infinitum* (that is why we use the $\mathbf{C}$ operator).
2. Models the fact that all the *untrusty* agents believe that all the trusty ones have common belief on what has been announced.
3. Captures that the partially observants believe that the trusty fully observants have common knowledge on what has been announced while the partially observants themselves do not know the announced value.
4. States that the fully observant agents have common knowledge on the previous property.
5. Models the idea that all the *untrusty* agents do not modify their beliefs about the announced values.
6. Captures the fact that the oblivious agents do not change their beliefs.
7. States that the observant agents (trusty, *untrusty* and partial) believe that the oblivious agents did not change their beliefs.

As we did for the *(un)trustworthy announcement*, let us identify some properties also for the *(mis)trustworthy announcement* action.

Proposition 2 ((mis)Trustworthy announcement properties). *Let* $\mathsf{a}\langle\mathsf{ag}\rangle$ *be a* (mis)trustworthy announcement action instance *where* ag ***m_announces*** ϕ. *Let* e *be an e-state and let* e' *be its updated version (that is,* $\Phi(\mathsf{a},\mathsf{e}) = \mathsf{e}'$*), then in* $m\mathcal{A}^{\rho}$ *properties* $1, 2, 3, 4, 6$ *and* 7 *of Proposition 1 hold. In addition,*

a. $\mathsf{e}' \models \mathbf{C}_{U_\mathsf{a}}\neg\phi$;
b. $\mathsf{e}' \models \mathbf{C}_{F_\mathsf{a}}(\mathbf{C}_{U_\mathsf{a}}\neg\phi)$;
c. $\mathsf{e}' \models \mathbf{C}_{P_\mathsf{a}}(\mathbf{C}_{U_\mathsf{a}}\phi \vee \mathbf{C}_{U_\mathsf{a}}\neg\phi)$;

Proposition 2 describes the core ideas behind a *(mis)trustworthy announcement* action. While properties $1, 2, 3, 4, 6$ of Proposition 1 have already been described, the intuitive meaning of the remaining ones is as follows.

a. Captures the idea that all the *untrusty* fully observant agents should believe i) the contrary of what has been announced; and ii) that all the other *untrusty* fully observant agents believe the negation of what has been announced and so on *ad infinitum* (that is why we use the **C** operator).
b. Models the fact that all the trusty agents believe that all the *untrusty* ones have common belief on the negation of what as been announced.
c. Captures that the partially observants believe that the *untrusty* fully observant have common knowledge on what has been announced, while the partially observant themselves do not know the announced value.

5 Conclusions and Future Works

In this paper we introduced the notion of trust in the field of multi-agent epistemic planning. In particular, we provided a formalization for two actions, *i.e.*, *(un)trustworthy announcement* and *(mis)trustworthy announcement*, that model two different scenarios of information sharing when the concept of trust is involved. The former action captures the idea that whenever an agent does not trust another she considers the announcer as an unreliable source of information, and therefore does not change her beliefs about the world. The latter, on the other hand, describes the situation where the *untrusty* agents will believe the contrary of what has been announced while still believing that the announcer believes what she announced. Both of the newly presented actions have been formalized for, at the best of our knowledge, the most comprehensive epistemic action-based language: $m\mathcal{A}^{\rho}$. In particular, in Section 4 we presented the transition functions of the actions along with their desired properties (formally demonstrated in the Supplementary Documentation).

As already mentioned, the idea of trust is presented as static and globally visible. While making it dynamic would not increase the "difficulty" of the transition functions, allowing each agent to have her own point of view on the trust relations would require a redesign of the e-state updates. In particular, to formalize this type of trust, the idea of *non-consistent belief* is necessary. Since such concept is still an open issue in the MEP community, we leave the formalization of e-state update when trust depends on the agents' point of view as future work. Finally, another concept that arises when trust is taken into consideration is the idea of *lies*. Modeling this concept would require major modifications of Defintion 6 and, as for the dynamic version of trust, the idea of non-consistent belief. Capturing subtle concepts such as lies and misconception is not straightforward and will provide a contribution on its own. The difficulty of characterizing such ideas derives from the complexity of devising a transition function that correctly captures all the possible nested beliefs of the domain's agents. We, therefore, leave the investigation of lies as future work. A more immediate future work is the introduction of the new actions in EFP 2.0 [8] and PLATO [4], a C++ solver (based on $m\mathcal{A}^{*}$ and $m\mathcal{A}^{\rho}$) and an ASP solver (based on $m\mathcal{A}^{\rho}$) respectively. PLATO in particular, given its nature of logical reasoner, may provide a more suitable environment to implement and test the newly introduced actions.

6 Acknowledgments

The author wishes to thank Dovier Agostino, Pontelli Enrico and Burigana Alessandro for the illuminating discussions on epistemology and the anonymous Reviewers for their comments that allowed to improve the presentation.

This research is partially supported by the Università di Udine PRID ENCASE project, and by GNCS-INdAM 2017–2020 projects.

References

1. Allen, M., Zilberstein, S.: Complexity of decentralized control: Special cases. In: Advances in Neural Information Processing Systems. pp. 19–27 (2009)
2. Baral, C., Gelfond, G., Pontelli, E., Son, T.C.: An action language for multi-agent domains: Foundations. CoRR **abs/1511.01960** (2015), http://arxiv.org/abs/1511.01960
3. Bolander, T., Andersen, M.B.: Epistemic planning for single-and multi-agent systems. Journal of Applied Non-Classical Logics **21**(1), 9–34 (2011). https://doi.org/10.1016/0010-0277(83)90004-5
4. Burigana, A., Fabiano, F., Dovier, A., Pontelli, E.: Modelling multi-agent epistemic planning in asp. Theory and Practice of Logic Programming **20**(5), 593–608 (2020). https://doi.org/10.1017/S1471068420000289, https://doi.org/10.1017/S1471068420000289
5. Chagrov, A.: Modal Logic. Oxford University Press (1997)
6. De Weerdt, M., Clement, B.: Introduction to planning in multiagent systems. Multiagent and Grid Systems **5**(4), 345–355 (2009). https://doi.org/10.3233/MGS-2009-0133
7. Dovier, A., Formisano, A., Pontelli, E.: Autonomous agents coordination: Action languages meet CLP() and Linda. Theory and Practice of Logic Programming **13**(2), 149–173 (2013). https://doi.org/10.1016/S0004-3702(00)00031-X
8. Fabiano, F., Burigana, A., Dovier, A., Pontelli, E.: EFP 2.0: A multi-agent epistemic solver with multiple e-state representations. In: Proceedings of the Thirtieth International Conference on Automated Planning and Scheduling, Nancy, France, October 26-30, 2020. pp. 101–109. AAAI Press (2020), https://aaai.org/ojs/index.php/ICAPS/article/view/6650
9. Fabiano, F., Riouak, I., Dovier, A., Pontelli, E.: Non-well-founded set based multi-agent epistemic action language. In: Proceedings of the 34th Italian Conference on Computational Logic. CEUR Workshop Proceedings, vol. 2396, pp. 242–259. Trieste, Italy (June 19-21 2019), http://ceur-ws.org/Vol-2396/paper38.pdf
10. Fagin, R., Halpern, J.Y.: Reasoning about knowledge and probability. Journal of the ACM (JACM) **41**(2), 340–367 (1994). https://doi.org/10.1145/174652.174658
11. Gerbrandy, J., Groeneveld, W.: Reasoning about information change. Journal of Logic, Language and Information **6**(2), 147–169 (1997). https://doi.org/10.1023/A:1008222603071
12. Kominis, F., Geffner, H.: Beliefs in multiagent planning: From one agent to many. In: Proceedings of the International Conference on Automated Planning and Scheduling, ICAPS. pp. 147–155 (2015)
13. Lipovetzky, N., Geffner, H.: Best-first width search: Exploration and exploitation in classical planning. In: Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence. pp. 3590–3596. San Francisco, California, USA (February 4-9 2017), http://aaai.org/ocs/index.php/AAAI/AAAI17/paper/view/14862

14. Muise, C.J., Belle, V., Felli, P., McIlraith, S.A., Miller, T., Pearce, A.R., Sonenberg, L.: Planning over multi-agent epistemic states: A classical planning approach. In: Proc. of AAAI. pp. 3327–3334 (2015)
15. Russell, S., Norvig, P.: Artificial Intelligence: A Modern Approach. Prentice Hall Press, Upper Saddle River, NJ, USA, 3rd edn. (2009)
16. Smullyan, R.R.: First-order logic, vol. 43. Springer Science & Business Media (2012)
17. Van Ditmarsch, H., van Der Hoek, W., Kooi, B.: Dynamic epistemic logic, vol. 337. Springer Science & Business Media (2007)