Mining and Validating Belief-based Agent Explanations

Ahmad Alelaimat, Aditya Ghose and Hoa Khanh Dam

Decision Systems Lab

School of Computing and Information Technology

University of Wollongong

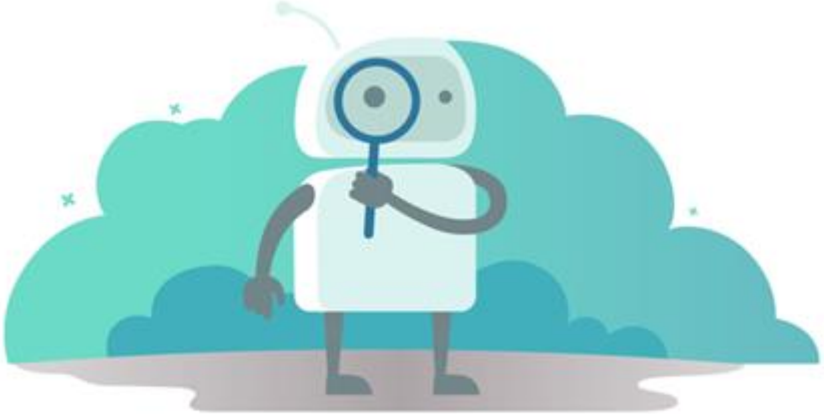
Decision System Lab
University of Wollongong

UNIVERSITY
OF WOLLONGONG
AUSTRALIA

Outline

- Explainable BDI Agents.
- Motivating Applications.
  - Startup competitor analysis.
  - Explaining plan selection
- Current Limitations.
- Mining and Validating Belief-based Explanations.
  - Updating Belief-based Explanations.
  - Mining Belief-based Explanations.
  - Validating the Explanation Process
- Conclusion.

Explainable Agents



- **What does it mean to have an explainable agent?**

  Not only a software entity that justifies its decisions but also communicates and delivers a meaningful explanations [1].

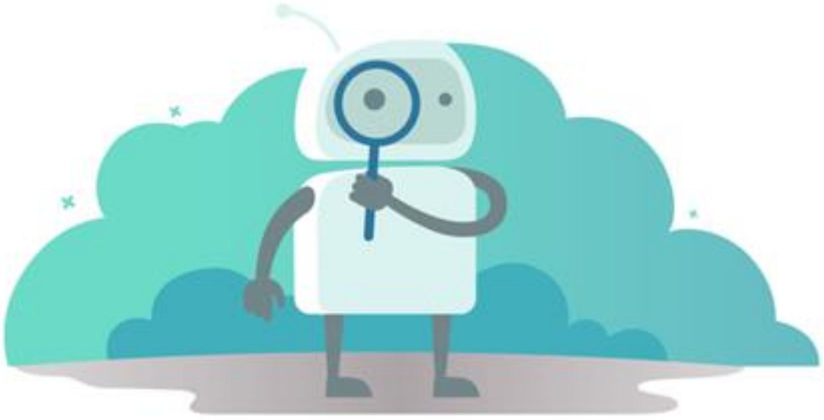
- **Why explainable agents?**

  Trust, collaboration, education, etc. [1].

- **What does it mean to have an explainable BDI agent?**

  A research question.

Explainable BDI Agents



- **Faithful explanations**
  - **A detailed explanation should reflect the agent system's processing.**
  - **Sacrifice how useful and accessible the explanation is to certain audiences [2].**
- **Unfaithful explanations.**
  - **People provide short explanations when asked to explain agent behavior [3].**
  - **Two common explanation styles: (1) a goal-based explanation and (2) a belief-based explanation.**
- **When we need belief-based explanations?**
  - **A research question.**

Motivating Application I

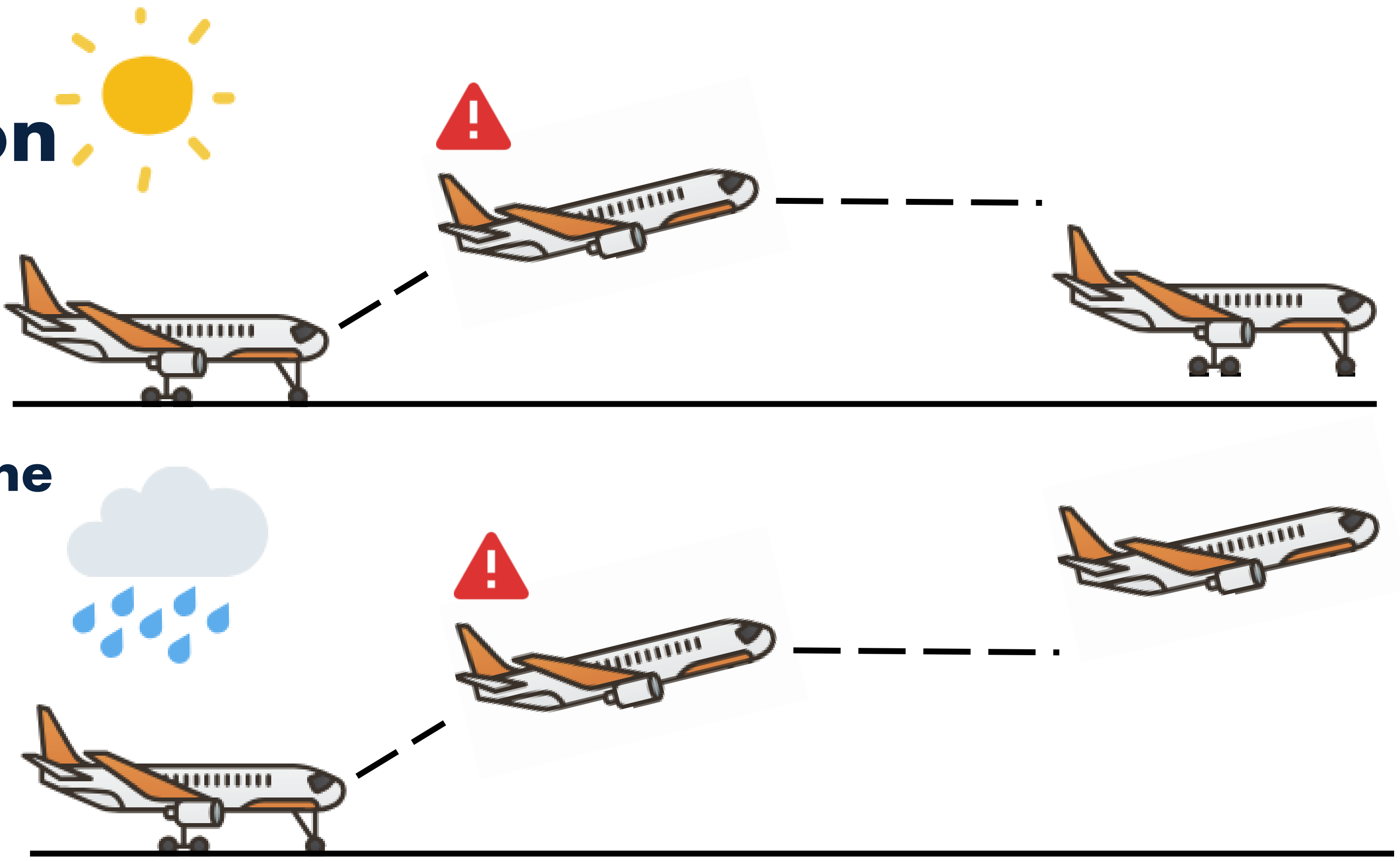
Startup competitor analysis

What must have been known for the target competitor to perform a particular task over another?

Motivating Application II

Explaining plan selection

Why must have been known for the pilot to select a particular plan among other applicable options?



Summarising the Weaknesses

- Much of the previous explanation generation approaches can theoretically do so, but assuming:

  1. Availability of explanation generation modules,
  2. Reliable observations, and
  3. Deterministic execution of plans.

- Explanation generation in competitive settings.

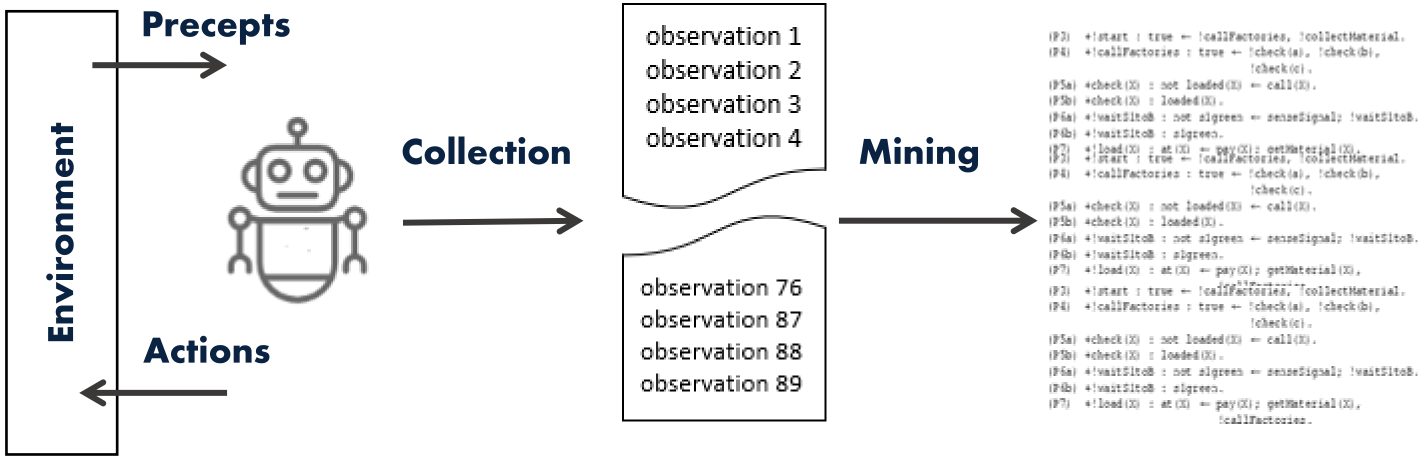
The Overall Approach

Given as inputs:

1. Audit Logs,
2. Plan library, and
3. update operator.

Compute:

- the belief-based explanations of every action referred to in the audit log.



```
(P3)  +!start : true ← !callFactories, !collectMaterial.
(P4)  +!callFactories : true ← !check(a), !check(b),
                               !check(c).
(P5a) +check(X) : not loaded(X) ← call(X).
(P5b) +check(X) : loaded(X).
(P6a) +!waitSltoB : not slgreen ← senseSignal; !waitSltoB.
(P6b) +!waitSltoB : slgreen.
(P7)  +!load(X) : at(X) ← pay(X); getMaterial(X),
(P3)  +!start : true ← !callFactories, !collectMaterial.
(P4)  +!callFactories : true ← !check(a), !check(b),
                               !check(c).
(P5a) +check(X) : not loaded(X) ← call(X).
(P5b) +check(X) : loaded(X).
(P6a) +!waitSltoB : not slgreen ← senseSignal; !waitSltoB.
(P6b) +!waitSltoB : slgreen.
(P7)  +!load(X) : at(X) ← pay(X); getMaterial(X),
(P3)  +!start : true ← !callFactories, !collectMaterial.
(P4)  +!callFactories : true ← !check(a), !check(b),
                               !check(c).
(P5a) +check(X) : not loaded(X) ← call(X).
(P5b) +check(X) : loaded(X).
(P6a) +!waitSltoB : not slgreen ← senseSignal; !waitSltoB.
(P6b) +!waitSltoB : slgreen.
(P7)  +!load(X) : at(X) ← pay(X); getMaterial(X),
                          !callFactories.
```

Agent Execution

Audit Logs

Explanations

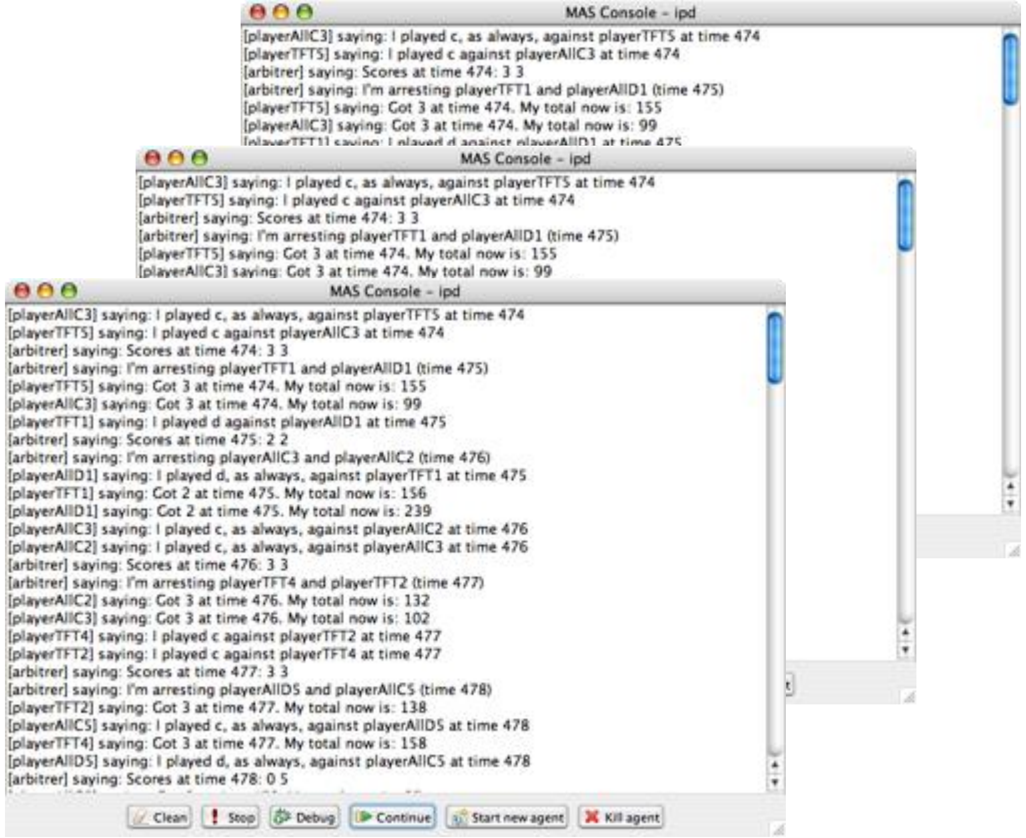
Audit Logs

- **We are interested in two modes of audit logging:**

1. **behaviour logs, and**
2. **belief logs**

- **Collecting such data can be done using audit logging tools such as:**

1. **Mind Inspector in Jason platform.**
2. **DTT in JACK platform.**



MAS Console – ipd

```
[playerAllC3] saying: I played c, as always, against playerTFT5 at time 474
[playerTFT5] saying: I played c against playerAllC3 at time 474
[arbitrer] saying: Scores at time 474: 3 3
[arbitrer] saying: I'm arresting playerTFT1 and playerAllD1 (time 475)
[playerTFT5] saying: Got 3 at time 474. My total now is: 155
[playerAllC3] saying: Got 3 at time 474. My total now is: 99
[playerTFT1] saying: I played d against playerAllD1 at time 475
[arbitrer] saying: Scores at time 475: 2 2
[arbitrer] saying: I'm arresting playerAllC3 and playerAllC2 (time 476)
[playerAllD1] saying: I played d, as always, against playerTFT1 at time 475
[playerTFT1] saying: Got 2 at time 475. My total now is: 156
[playerAllD1] saying: Got 2 at time 475. My total now is: 239
[playerAllC3] saying: I played c, as always, against playerAllC2 at time 476
[playerAllC2] saying: I played c, as always, against playerAllC3 at time 476
[arbitrer] saying: Scores at time 476: 3 3
[arbitrer] saying: I'm arresting playerTFT4 and playerTFT2 (time 477)
[playerAllC2] saying: Got 3 at time 476. My total now is: 132
[playerAllC3] saying: Got 3 at time 476. My total now is: 102
[playerTFT4] saying: I played c against playerTFT2 at time 477
[playerTFT2] saying: I played c against playerTFT4 at time 477
[arbitrer] saying: Scores at time 477: 3 3
[arbitrer] saying: I'm arresting playerAllD5 and playerAllC5 (time 478)
[playerTFT2] saying: Got 3 at time 477. My total now is: 138
[playerAllC5] saying: I played c, as always, against playerAllD5 at time 478
[playerTFT4] saying: Got 3 at time 477. My total now is: 158
[playerAllD5] saying: I played d, as always, against playerAllC5 at time 478
[arbitrer] saying: Scores at time 478: 0 5
```

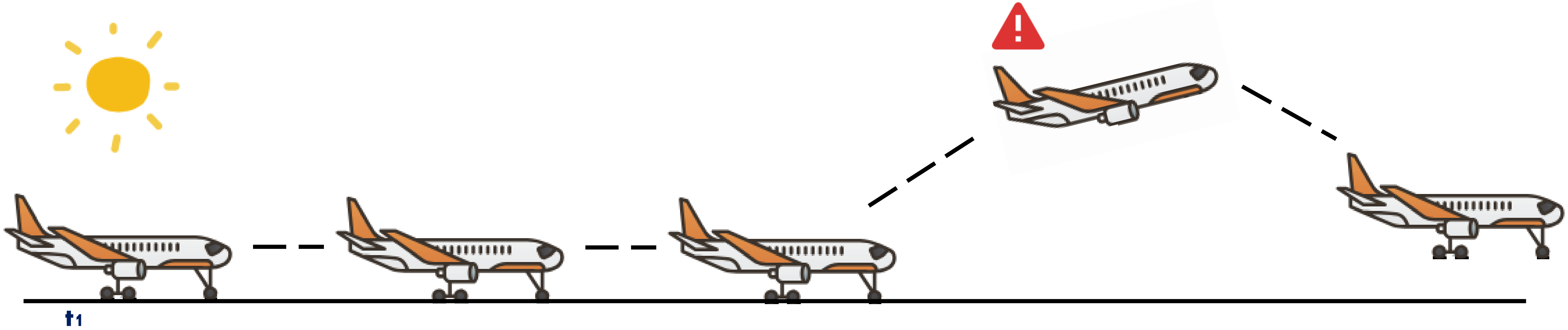
Updating Belief-based Explanations

- Updating Belief-based Explanations: At each action step in a plan execution, we accumulate the enabling beliefs of the preceding steps.

- Why we update belief-based explanations?
  - It could be used to contextualise explanations.
  - It could also be used to validate the mined explanations.

- Updated belief-based explanations are non-deterministic.

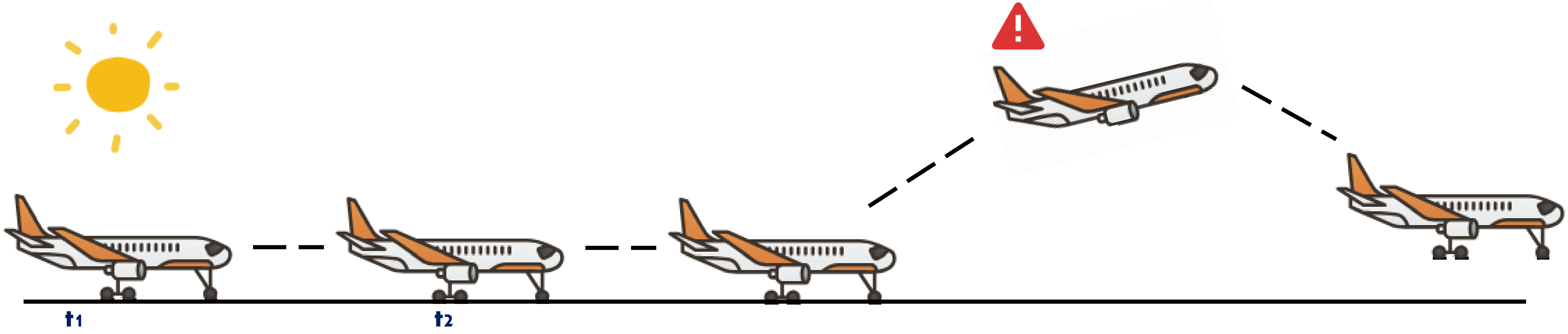
Updating Belief-based Explanations



Current beliefs It is sunny today
Cross wind

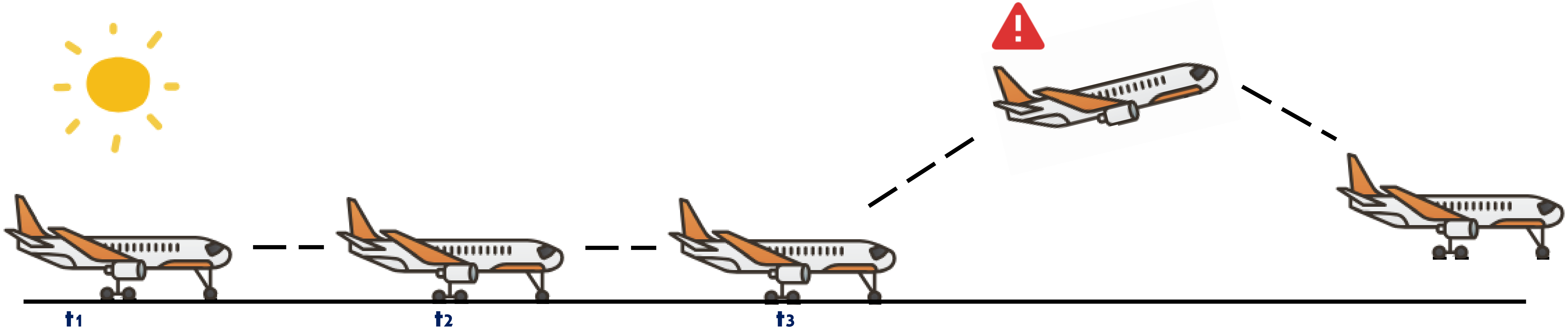
Updated beliefs It is sunny today
Cross wind

Updating Belief-based Explanations



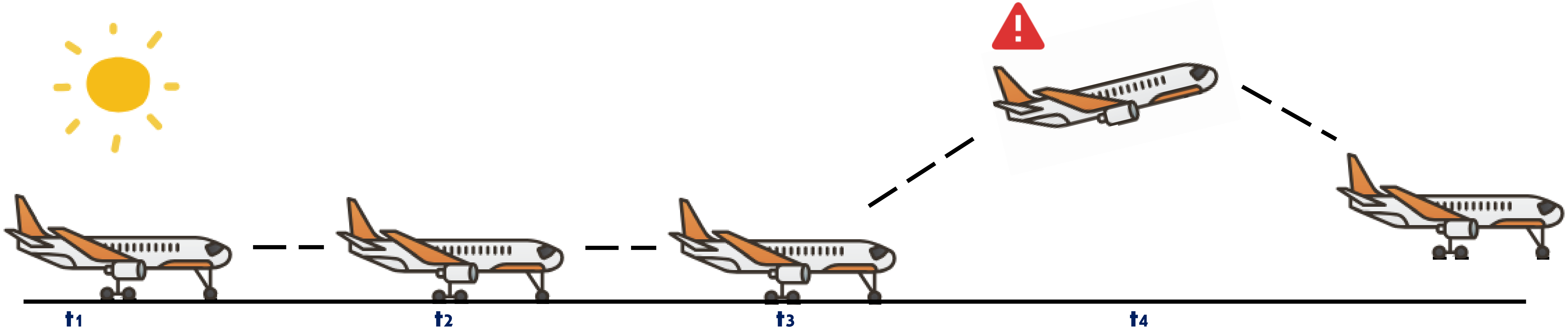
Current beliefs	**It is sunny today** **Cross wind**	**v1 = 129** **v2 = 145**
Updated beliefs	**It is sunny today** **Cross wind**	**It is sunny today** **Cross wind** **v1 = 129** **v2 = 145**

Updating Belief-based Explanations



Current beliefs	It is sunny today Cross wind	$v_1 = 129$ $v_2 = 145$	speed $= 135$
Updated beliefs	It is sunny today Cross wind	It is sunny today Cross wind $v_1 = 129$ $v_2 = 145$	It is sunny today Cross wind $v_1 = 129$ $v_2 = 145$ speed $= 135$

Updating Belief-based Explanations

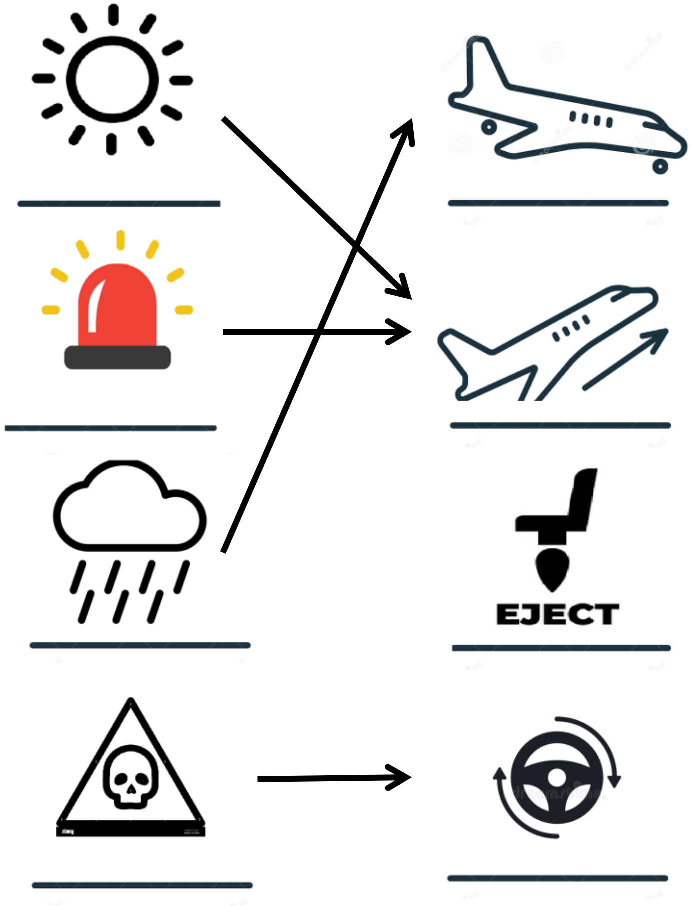


Current beliefs	**It is sunny today** **Cross wind**	v_1 = 129 v_2 = 145	**speed** = 135	**EFTO**
Updated beliefs	**It is sunny today** **Cross wind**	**It is sunny today** **Cross wind** v_1 = 129 v_2 = 145	**It is sunny today** **Cross wind** v_1 = 129 v_2 = 145 **speed** = 135	**It is sunny today** **Cross wind** v_1 = 129 v_2 = 145 **speed** = 150 **EFTO**

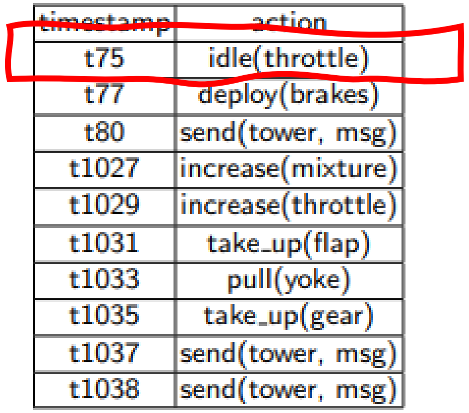
Mining Belief-based Explanations

- We are interested in discovering all the beliefs that are observed always, or most of the time, directly before the execution of each action referred to in the behavior log.

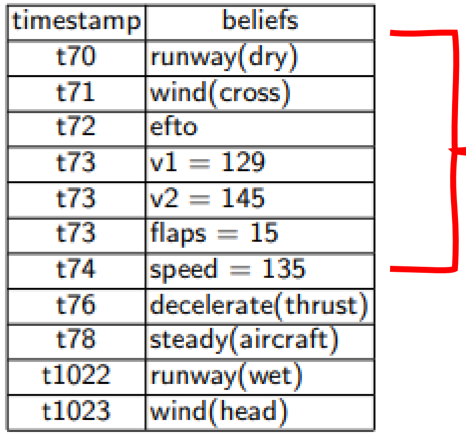
- Association rule learning can be an effective means for discovering regularities between beliefs and actions.



Mining Belief-based Explanations



timestamp	action
t75	idle(throttle)
t77	deploy(brakes)
t80	send(tower, msg)
t1027	increase(mixture)
t1029	increase(throttle)
t1031	take_up(flap)
t1033	pull(yoke)
t1035	take_up(gear)
t1037	send(tower, msg)
t1038	send(tower, msg)



timestamp	beliefs
t70	runway(dry)
t71	wind(cross)
t72	efto
t73	v1 = 129
t73	v2 = 145
t73	flaps = 15
t74	speed = 135
t76	decelerate(thrust)
t78	steady(aircraft)
t1022	runway(wet)
t1023	wind(head)

Explanation

The guiding intuition here is that:

1. **beliefs observed in the belief log immediately before executing an action can be the enabling beliefs of that action, and**

2. **persistent beliefs observed a long time before the execution of an action are typically not the enabling beliefs of that action but may be of that action plus some others.**

Validating the Explanation Process

To validate the mined explanations, it is useful to establish:

- **Soundness: a sound belief-based explanation is one that is mined correctly.**

- **Completeness: a complete belief-based explanation requires that all the enabling beliefs of a given action are mined.**

Conclusion

- **Faithful vs. Unfaithful explanations.**
- **Not all agents are explainable by design.**
- **Belief-based explanations for competitor analysis.**
- **Updating, mining and validating belief-based explanations.**
- **Next step: Goal-based explanations mining.**

Question?

References

[1] Anjomshoae, S., Najjar, A., Calvaresi, D. and Främling, K., 2019. Explainable agents and robots: Results from a systematic literature review. In 18th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2019), Montreal, Canada, May 13–17, 2019 (pp. 1078-1088). International Foundation for Autonomous Agents and Multiagent Systems.

[2] Phillips, P.J., Hahn, C.A., Fontana, P.C., Broniatowski, D.A. and Przybocki, M.A., 2020. Four principles of explainable artificial intelligence. Gaithersburg, Maryland, p.18.

[3] Broekens, J., Harbers, M., Hindriks, K., Van Den Bosch, K., Jonker, C. and Meyer, J.J., 2010. Do you get it? User-evaluated explainable BDI agents. In Multiagent System Technologies: 8th German Conference, MATES 2010, Leipzig, Germany, September 27-29, 2010. Proceedings 8 (pp. 28-39). Springer Berlin Heidelberg.