Towards the Role of Theory of Mind in Explanation

Maayan Shvo   Toryn Q. Klassen   Sheila A. McIlraith

Department of Computer Science
University of Toronto
Toronto, Canada

Vector Institute
Toronto, Canada

Schwartz Reisman Institute for Technology and Society
Toronto, Canada

XLoKR 2020

VECTOR
INSTITUTE

CIFAR

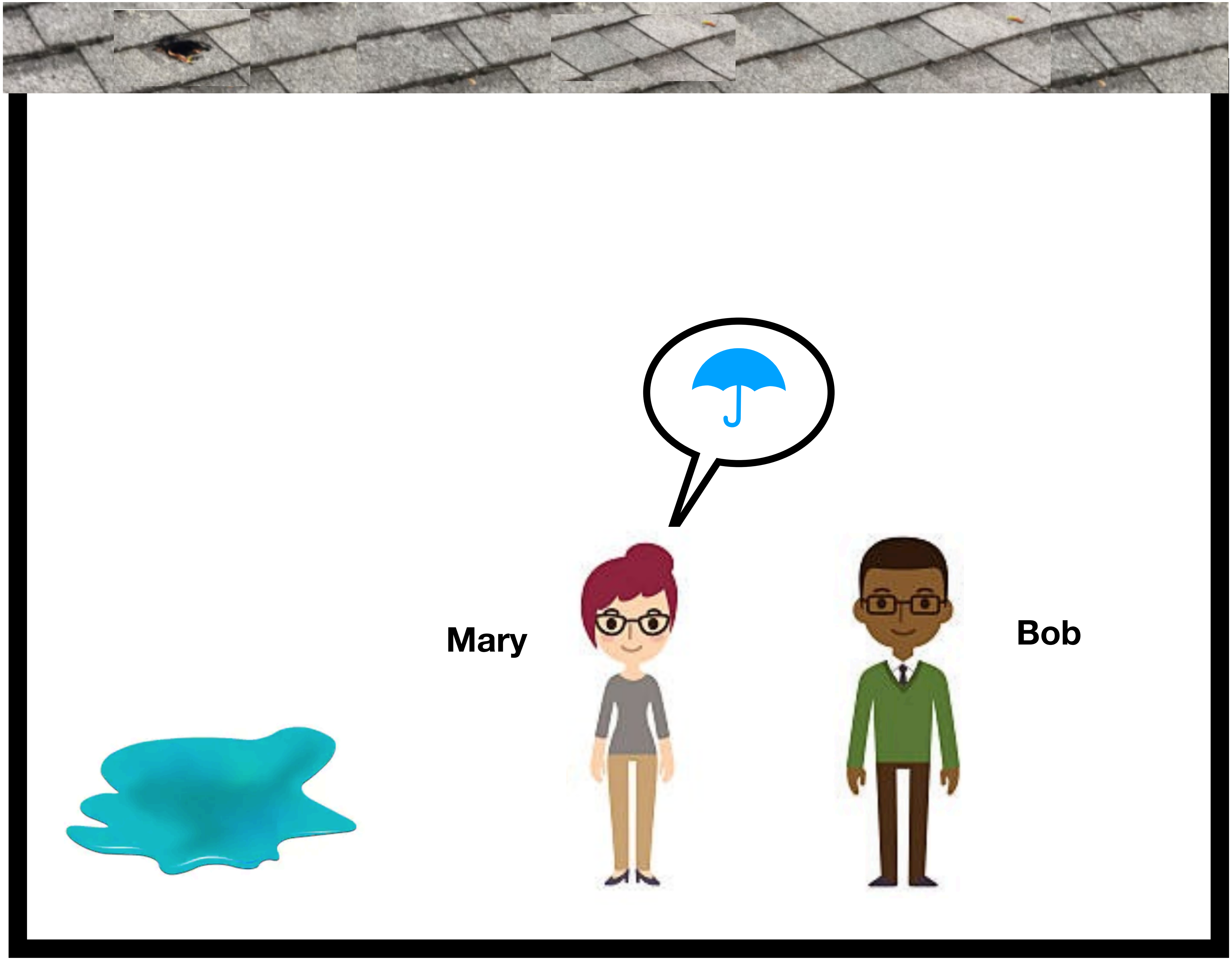
Canada CIFAR AI Chairs Program

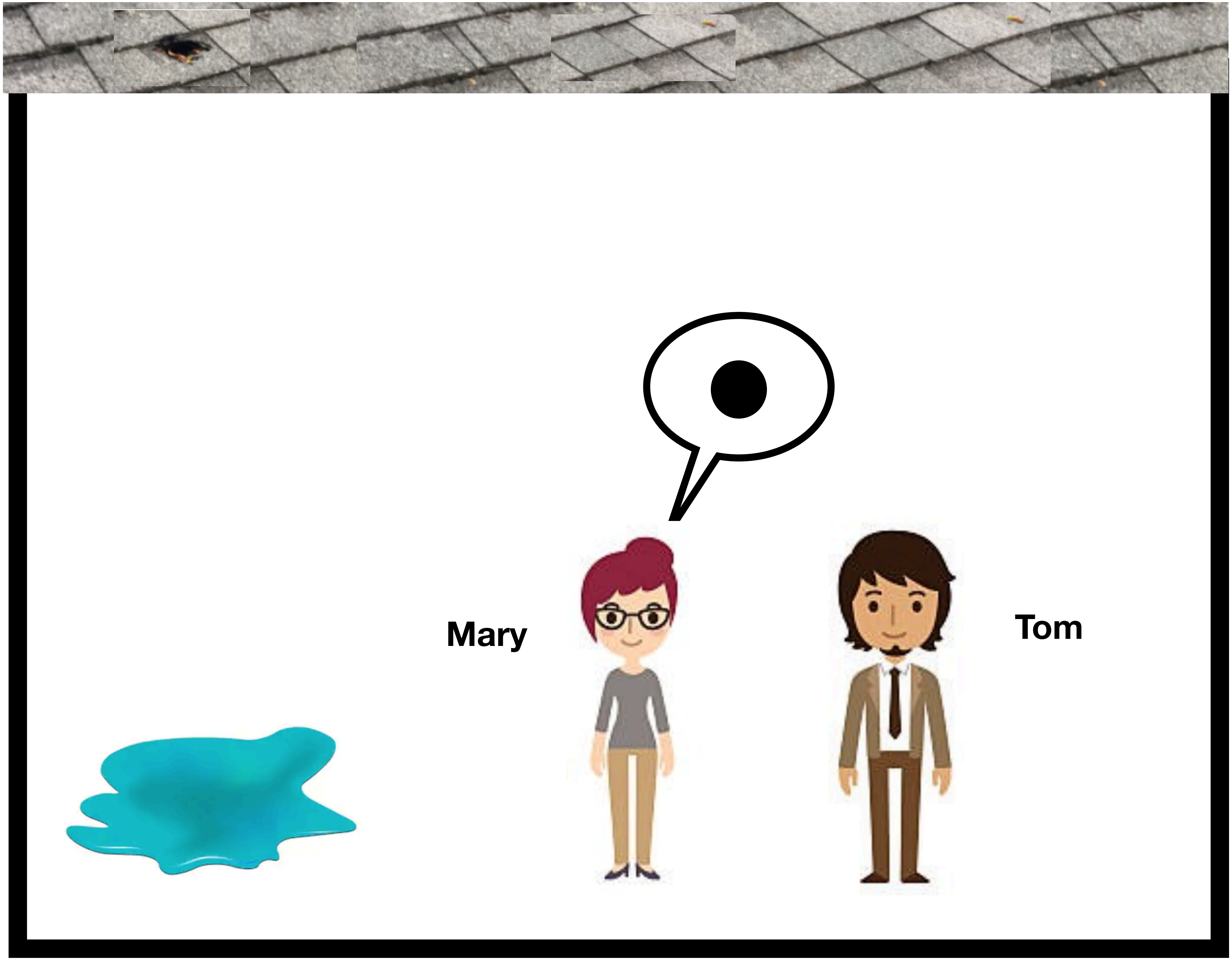
Theory of Mind

The ability to attribute mental states (e.g., beliefs, goals) to oneself, and to others.

Running Example

Mary, Bob and Tom are housemates sharing a house. While Tom was away on a business trip, Mary and Bob noticed a hole in the roof of their house and called a handyman to fix it. Before the handyman could come, however, it rained during the night and the floor got wet. Bob, who sleeps in a windowless room, did not notice the rain. Tom, who just got back from his trip that day, noticed the rain but did not know about the hole in the roof. Mary saw Tom return to the house at night and so knew that Tom knew that it had rained. In the morning, when trying to explain the wet floor to Bob, Mary tells him that it had rained during the night and when explaining to Tom she tells him that she and Bob had discovered a hole in the roof (adding that the handyman will arrive the next day).


Mary
Bob



Mary
Tom

Theory of Mind in Explanation

(Weiner, 1980)
(Gärdenfors, 1988)
(Cawsey, 1991)
(Slugoski et al., 1993)
(Halpern and Pearl, 2005)
(Chakraborti et al., 2017)
(Chandrasekaran et al., 2017)
(Westberg et al., 2019)
(Miller et al., 2019)

Theory of Mind in Explanation - Desiderata

- Multiple explainers and explainees

Theory of Mind in Explanation - Desiderata

- Multiple explainers and explainees

- Multiple agent types with different internal belief representations

Theory of Mind in Explanation - Desiderata

- Multiple explainers and explainees

- Multiple agent types with different internal belief representations

- Must allow for both the explainer and explainee to hold false beliefs

Theory of Mind in Explanation - Desiderata

- Multiple explainers and explainees

- Multiple agent types with different internal belief representations

- Must allow for both the explainer and explainee to hold false beliefs

- Explainer must be able to tailor explanations to the explainee’s beliefs

Theory of Mind in Explanation - Desiderata

- Multiple explainers and explainees
- Multiple agent types with different internal belief representations
- Must allow for both the explainer and explainee to hold false beliefs
- Explainer must be able to tailor explanations to the explainee's beliefs
- Explainer must reason about how the explainee assimilates explanations

Theory of Mind in Explanation - Building Blocks

Epistemic States

(Gärdenfors, 1988)
(Levesque, 1989)
(Boutilier and Becher, 1995)
(Halpern and Pearl, 2005)

Theory of Mind in Explanation - Building Blocks

Epistemic States

(Gärdenfors, 1988)
(Levesque, 1989)
(Boutilier and Becher, 1995)
(Halpern and Pearl, 2005)

Belief Revision

(Boutilier and Becher, 1995)
(Nepomuceno-Fernández et al., 2017)

Theory of Mind in Explanation - Building Blocks

Epistemic States

(Gärdenfors, 1988)
(Levesque, 1989)
(Boutilier and Becher, 1995)
(Halpern and Pearl, 2005)

Belief Revision

(Boutilier and Becher, 1995)
(Nepomuceno-Fernández et al., 2017)

☑Multiple explainers and explainees

☑Multiple agent types with different internal belief representations

☑Must allow for both the explainer and explainee to hold false beliefs

☑Explainer must be able to tailor explanations to the explainee's beliefs

☑Explainer must reason about how the explainee assimilates explanations

Our Belief-level Account of Explanation

$$\overrightarrow{e} = e_1, \ldots, e_n$$

e_i **is the epistemic state of agent i**

Our Belief-level Account of Explanation

$$\overrightarrow{e} = e_1, \ldots, e_n$$

e_i is the epistemic state of agent i

$$\overrightarrow{e} \vDash B_i \phi$$

Agent i believes phi to be true

Our Belief-level Account of Explanation

$$\overrightarrow{e} = e_1, \ldots, e_n$$

e_i is the epistemic state of agent i

$$\overrightarrow{e} \models B_i\phi$$

Agent i believes phi to be true

$$\overrightarrow{e} \models [\alpha]_i(B_i\beta \wedge \neg B_i\bot)$$

After agent i revises its beliefs with alpha, agent i will believe beta and not have inconsistent beliefs

Our Belief-level Account of Explanation

$$\overrightarrow{e} = e_1, \ldots, e_n$$

e_i is the epistemic state of agent i

$$\overrightarrow{e} \models B_i\phi$$

Agent i believes phi to be true

$$\overrightarrow{e} \models [\alpha]_i(B_i\beta \land \neg B_i\bot)$$

After agent i revises its beliefs with alpha, agent i will believe beta and not have inconsistent beliefs

$$Expl(i, \alpha, \beta) \triangleq [\alpha]_i(B_i\beta \land \neg B_i\bot)$$

Our Belief-level Account of Explanation

$$\overrightarrow{e} = e_1, \ldots, e_n$$

e_i is the epistemic state of agent i

$$\overrightarrow{e} \models B_i\phi$$

Agent i believes phi to be true

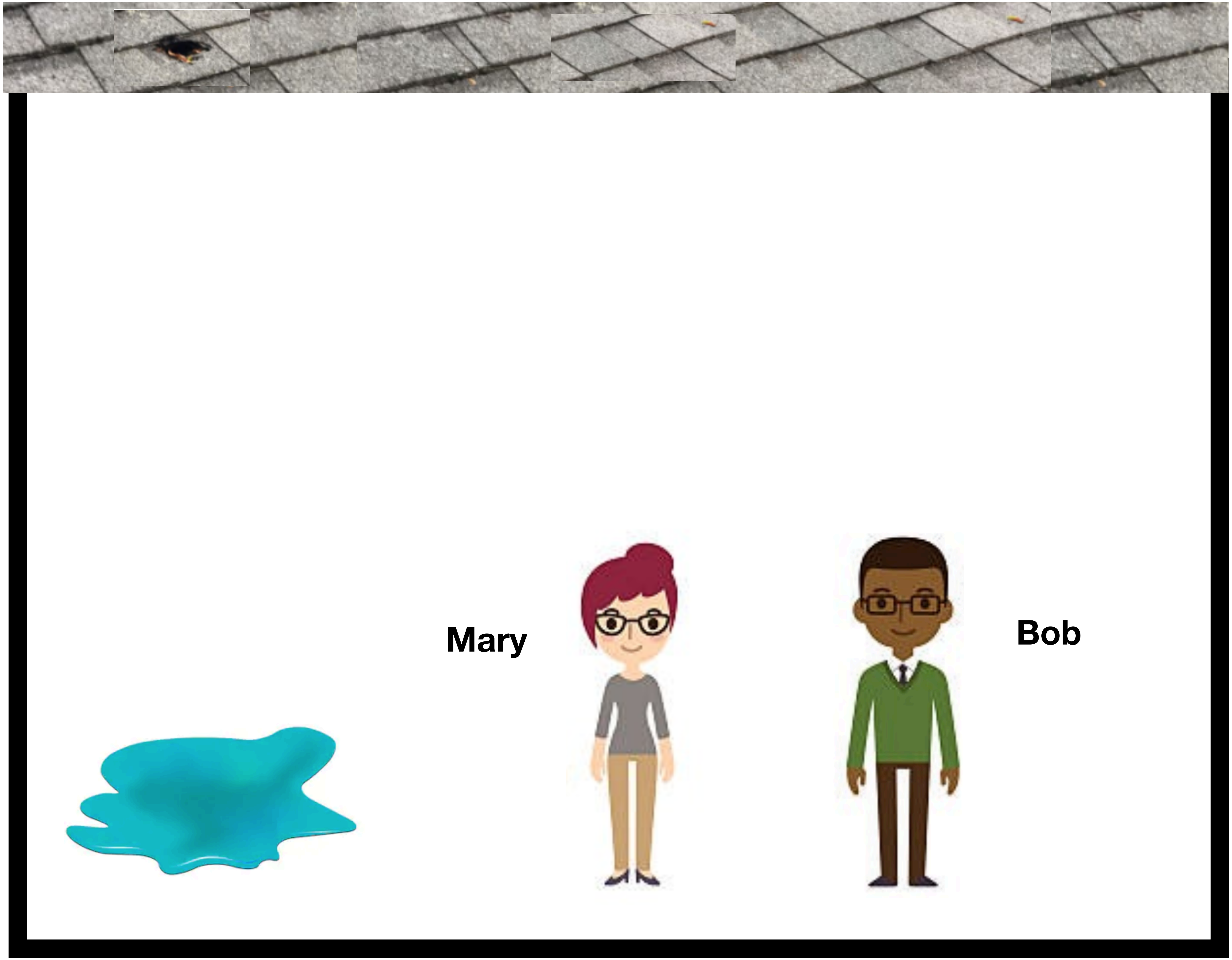
$$\overrightarrow{e} \models [\alpha]_i(B_i\beta \wedge \neg B_i\bot)$$

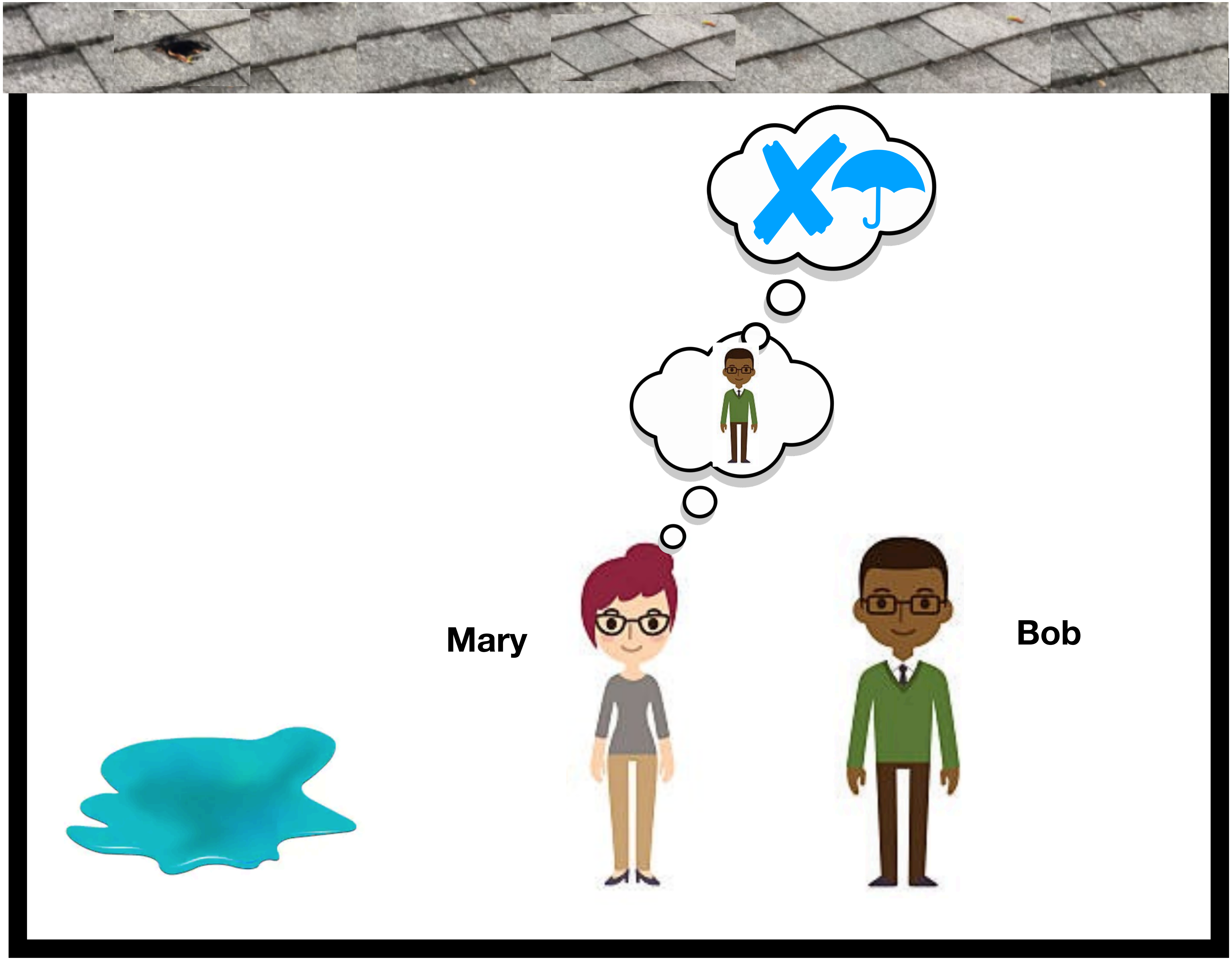
After agent i revises its beliefs with alpha, agent i will believe beta and not have inconsistent beliefs

$$Expl(i, \alpha, \beta) \triangleq [\alpha]_i(B_i\beta \wedge \neg B_i\bot)$$

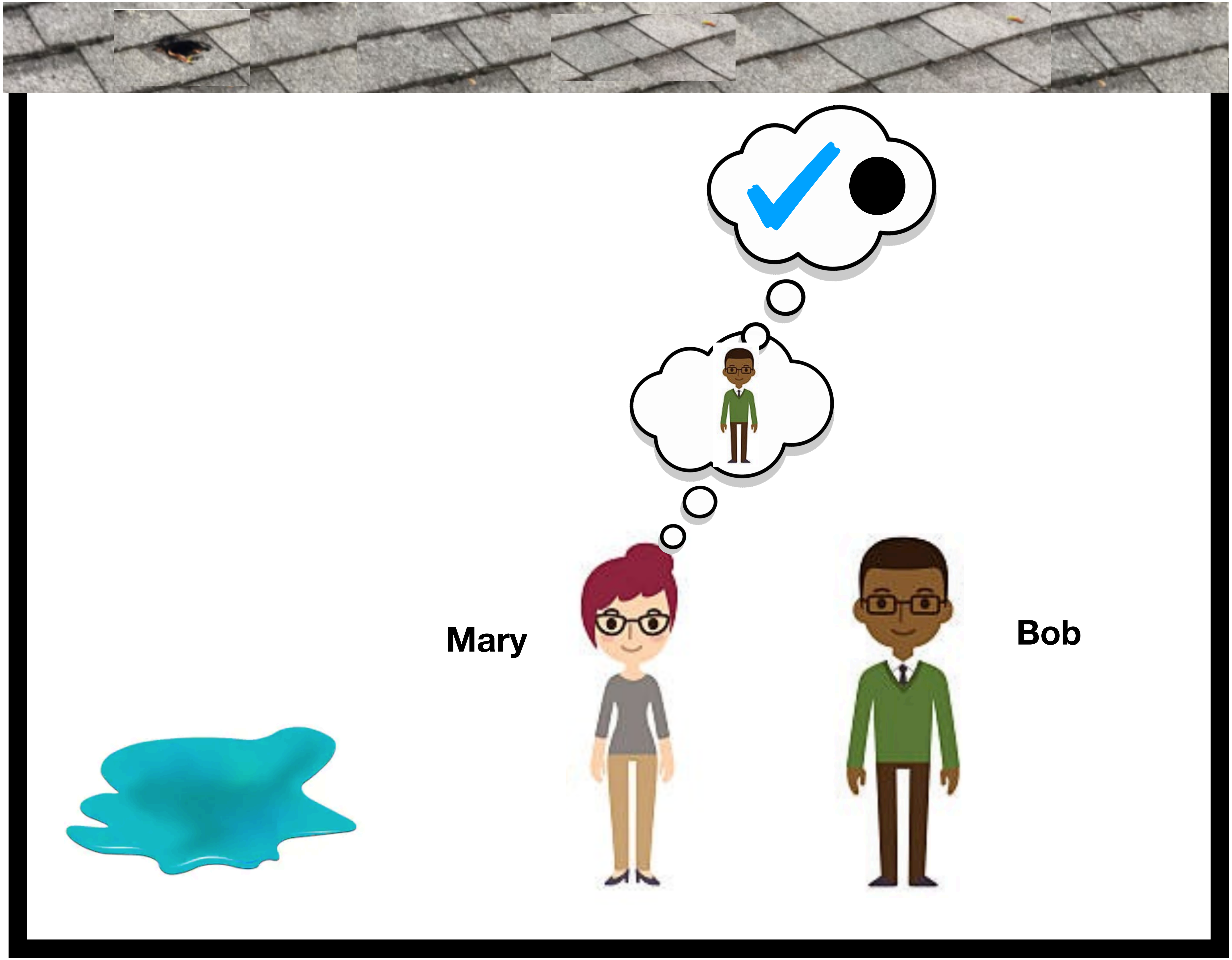
$$\overrightarrow{e} \models B_j Expl(i, \alpha, \beta)$$

Agent j believes that alpha is an explanation for beta for agent i

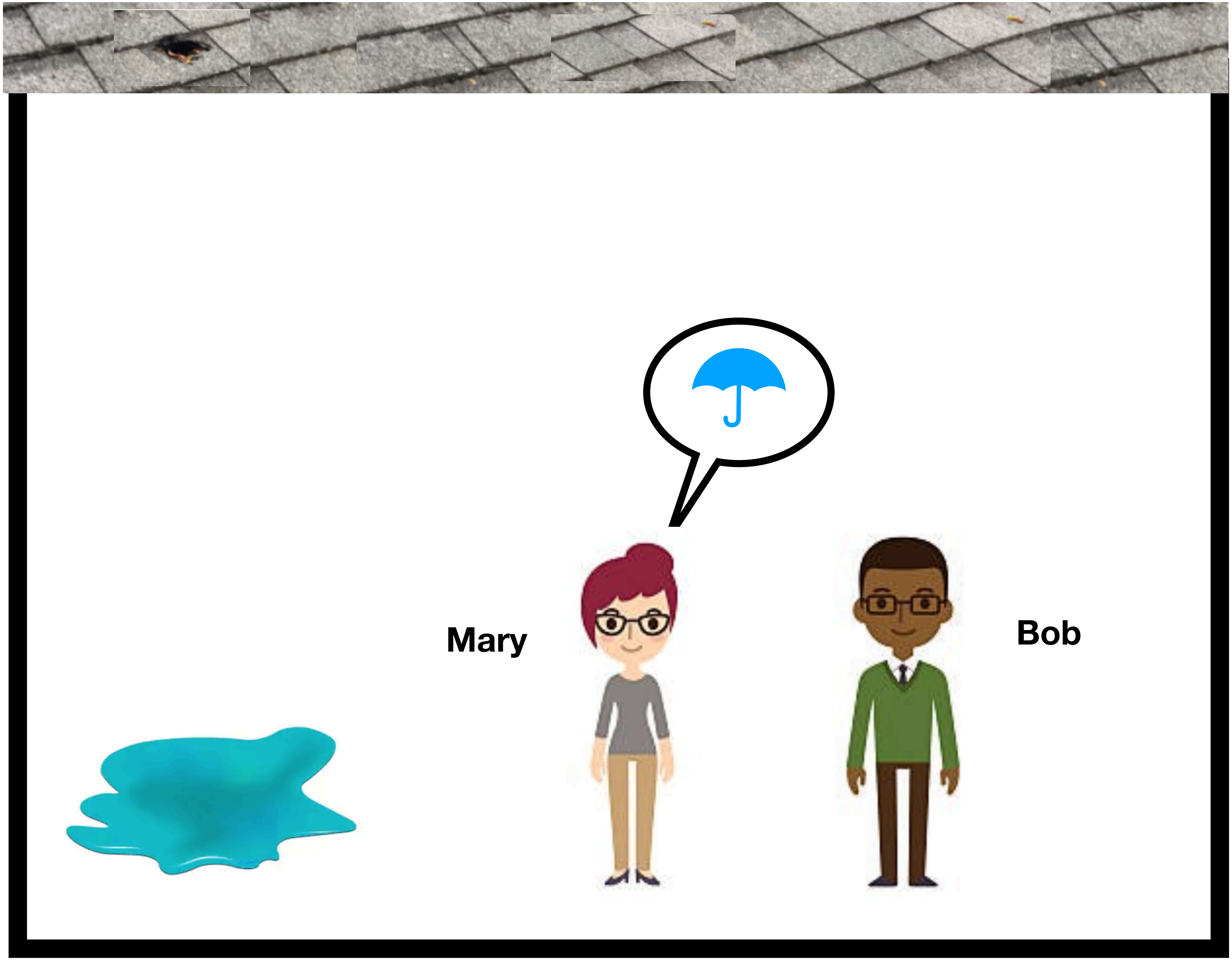

Mary
Bob



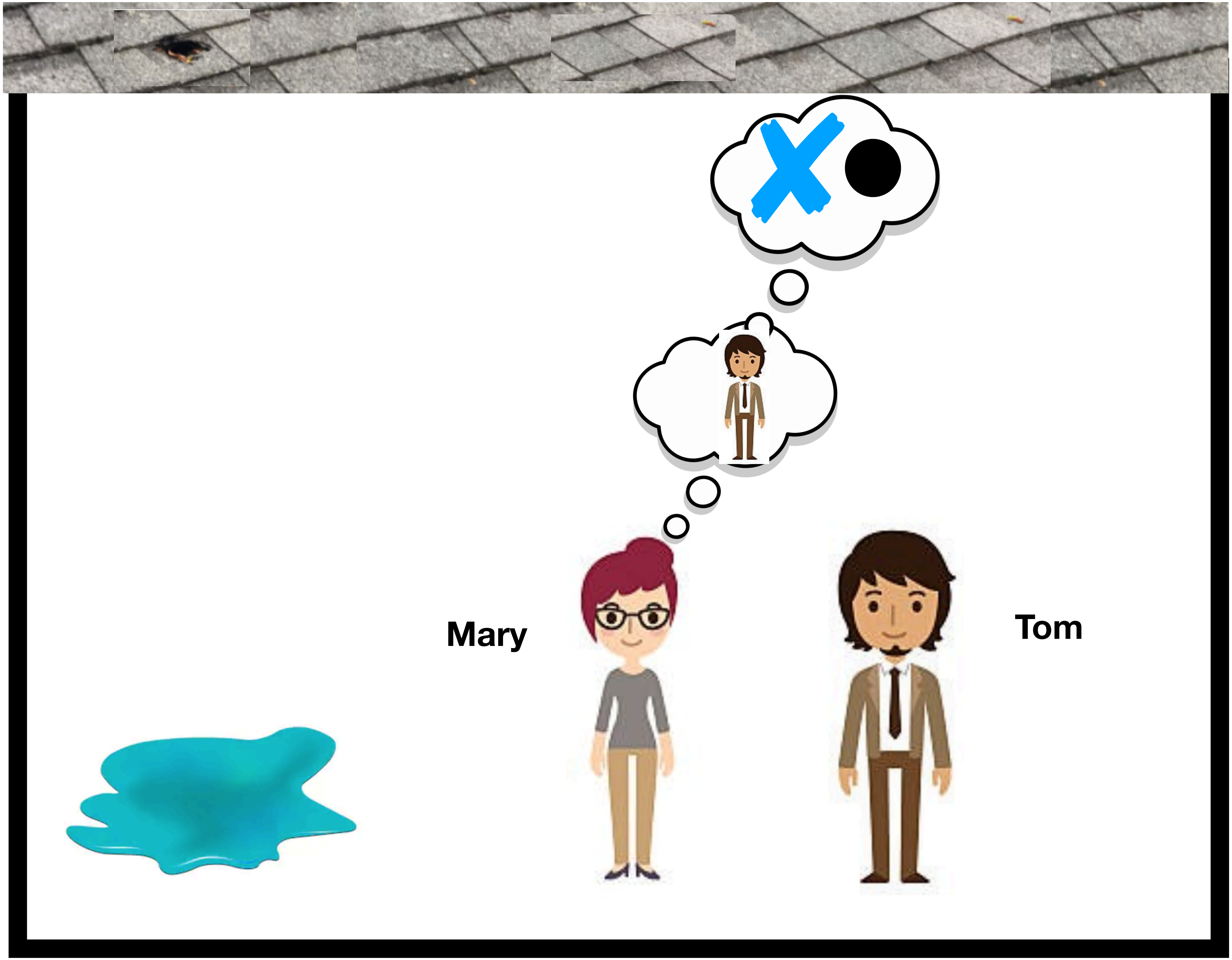
$$\overrightarrow{e} \models B_{Mary}B_{Bob}\neg rain$$



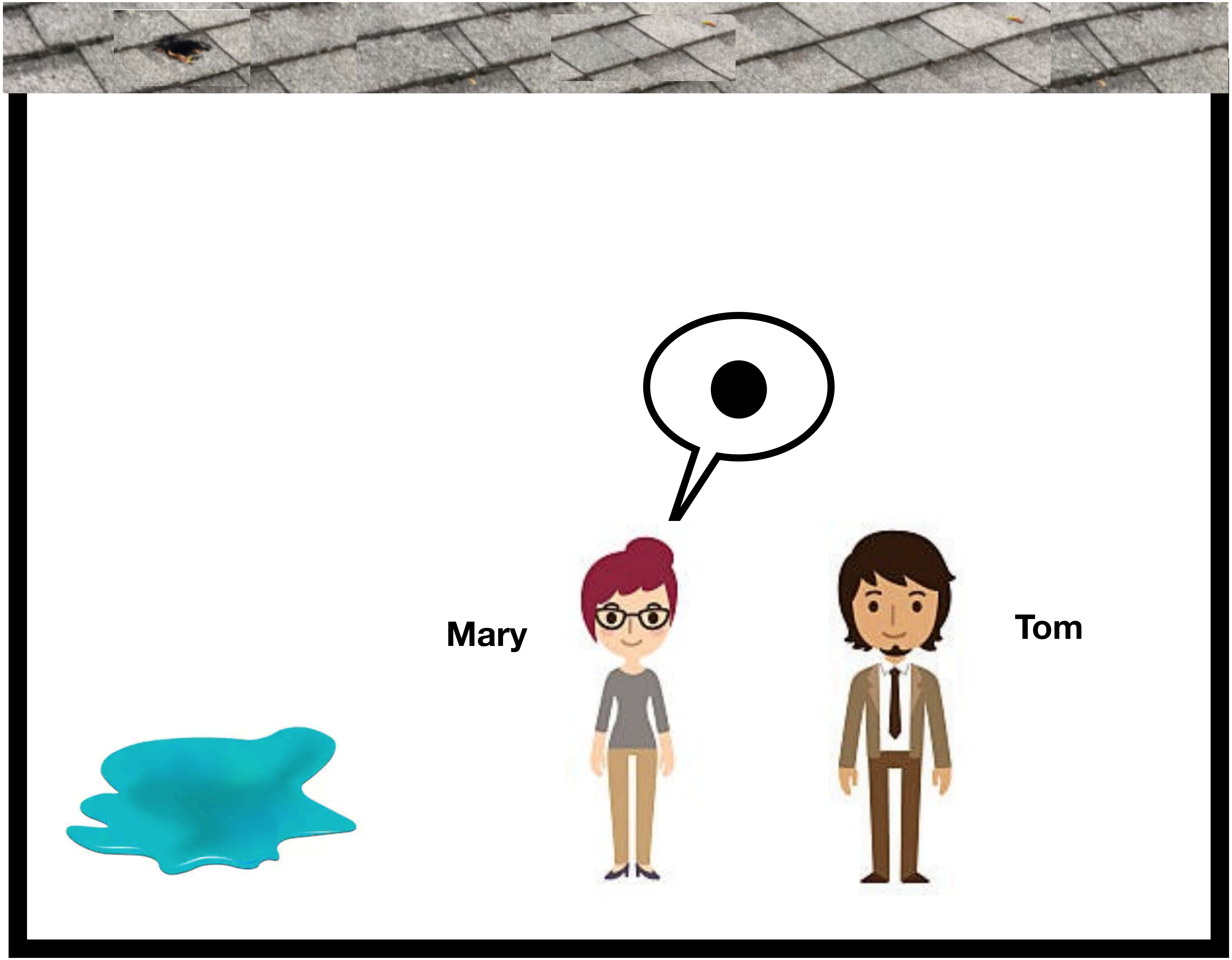
$$\overrightarrow{e} \models B_{Mary}B_{Bob}holeInRoof$$



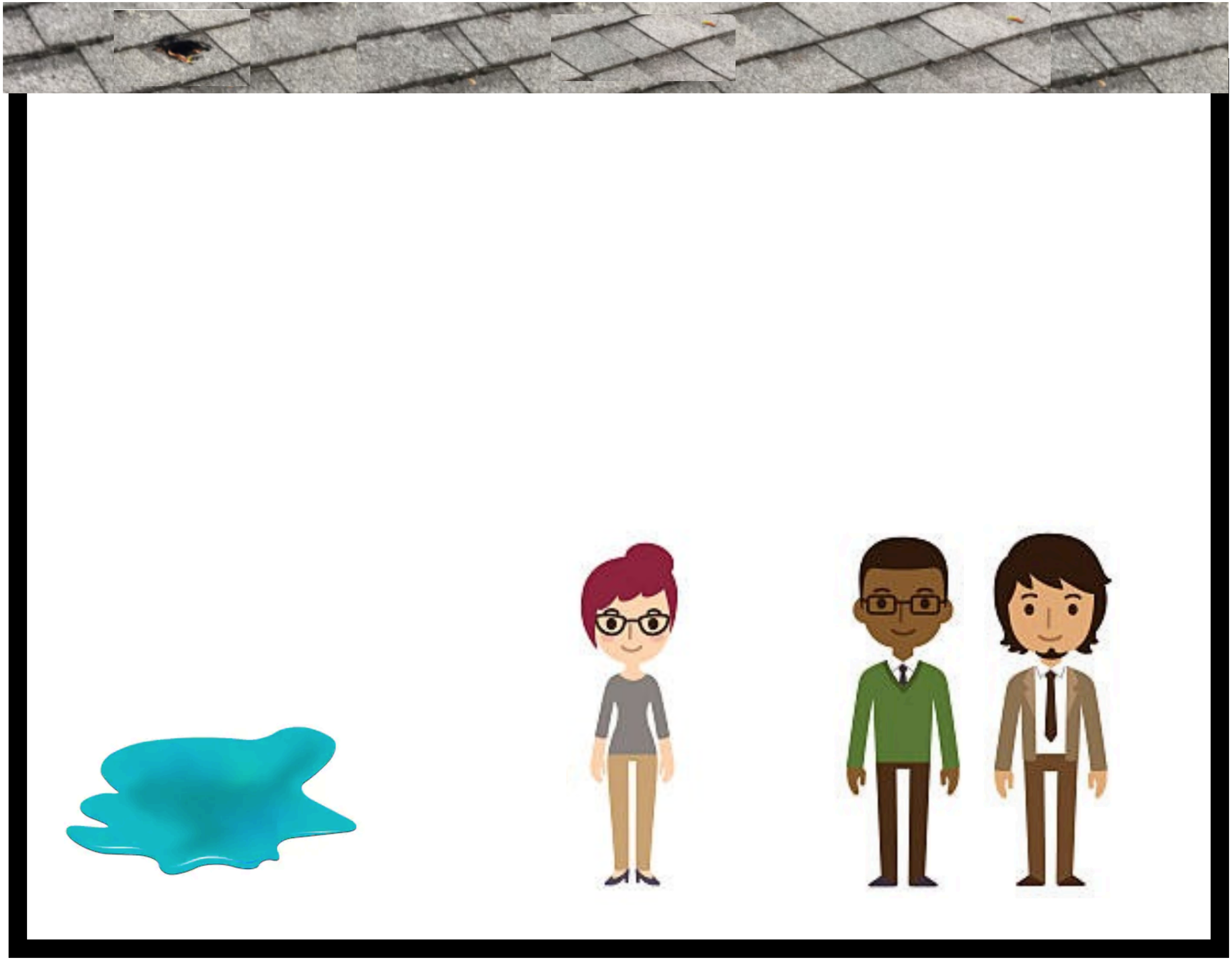
$$\vec{e} \models B_{Mary}Expl(Bob, rain, wetFloor)$$

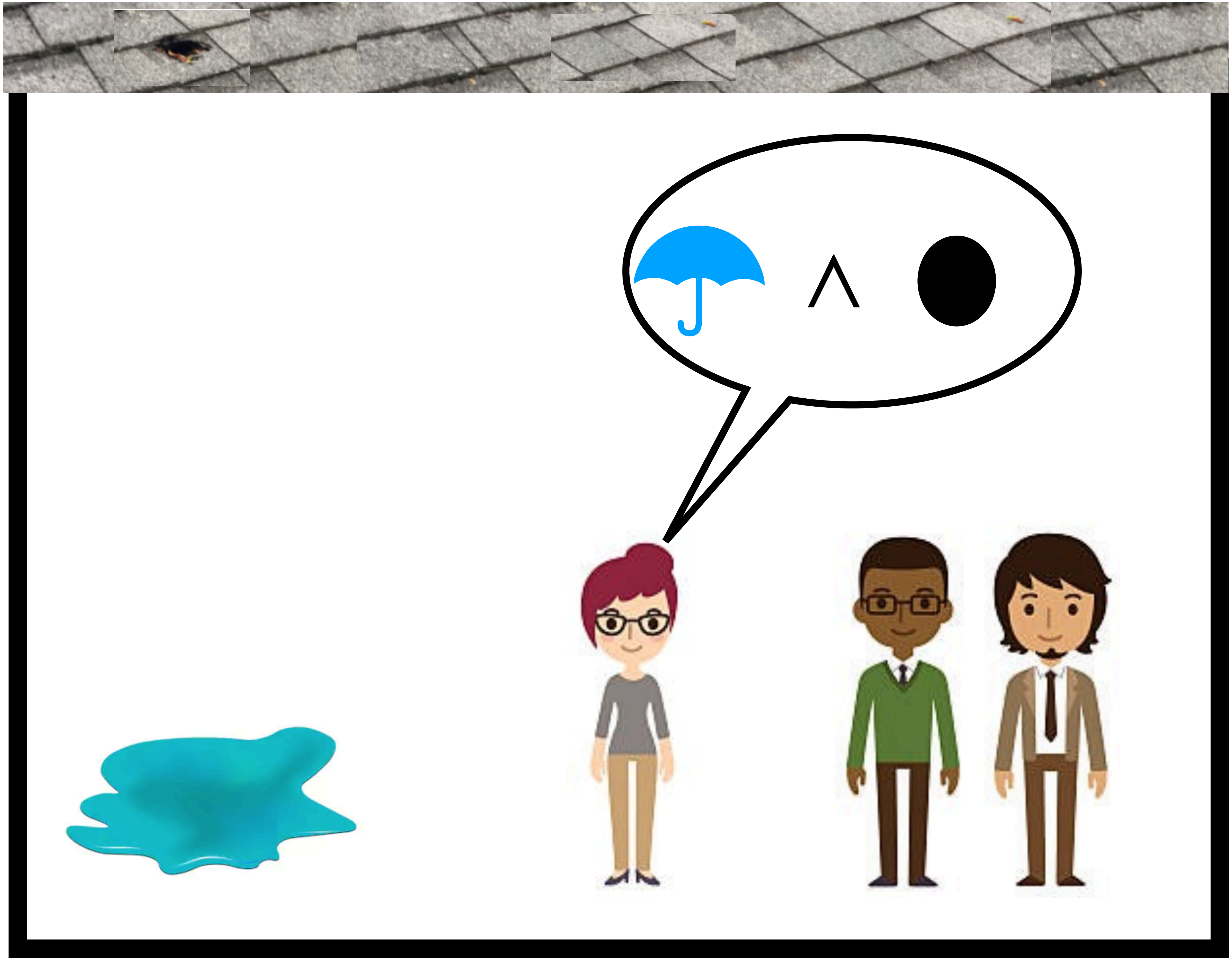


$$\vec{e} \models B_{Mary} B_{Tom} \neg holeInRoof$$

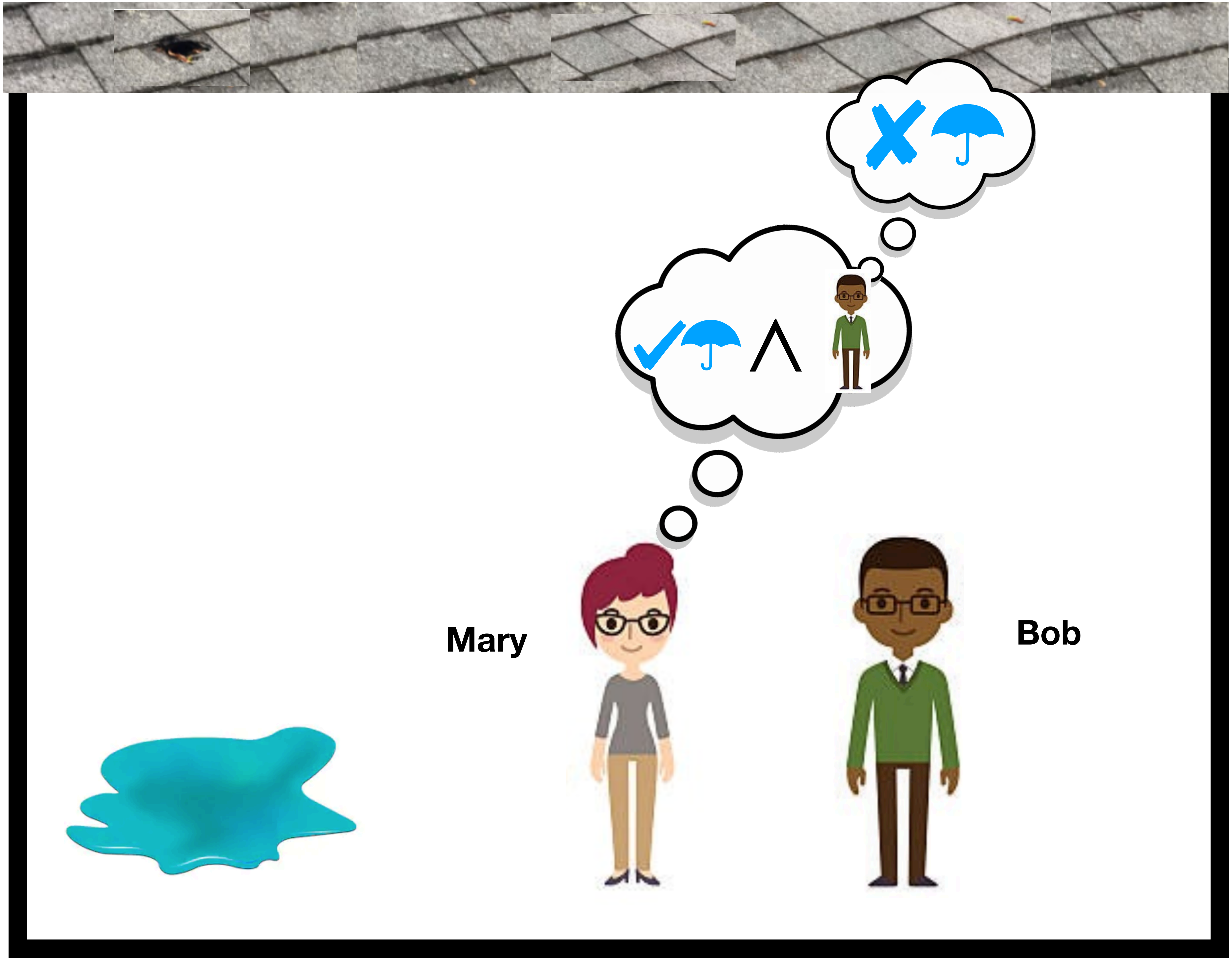


$$\overrightarrow{e} \models B_{Mary}Expl(Tom, holeInRoof, wetFloor)$$



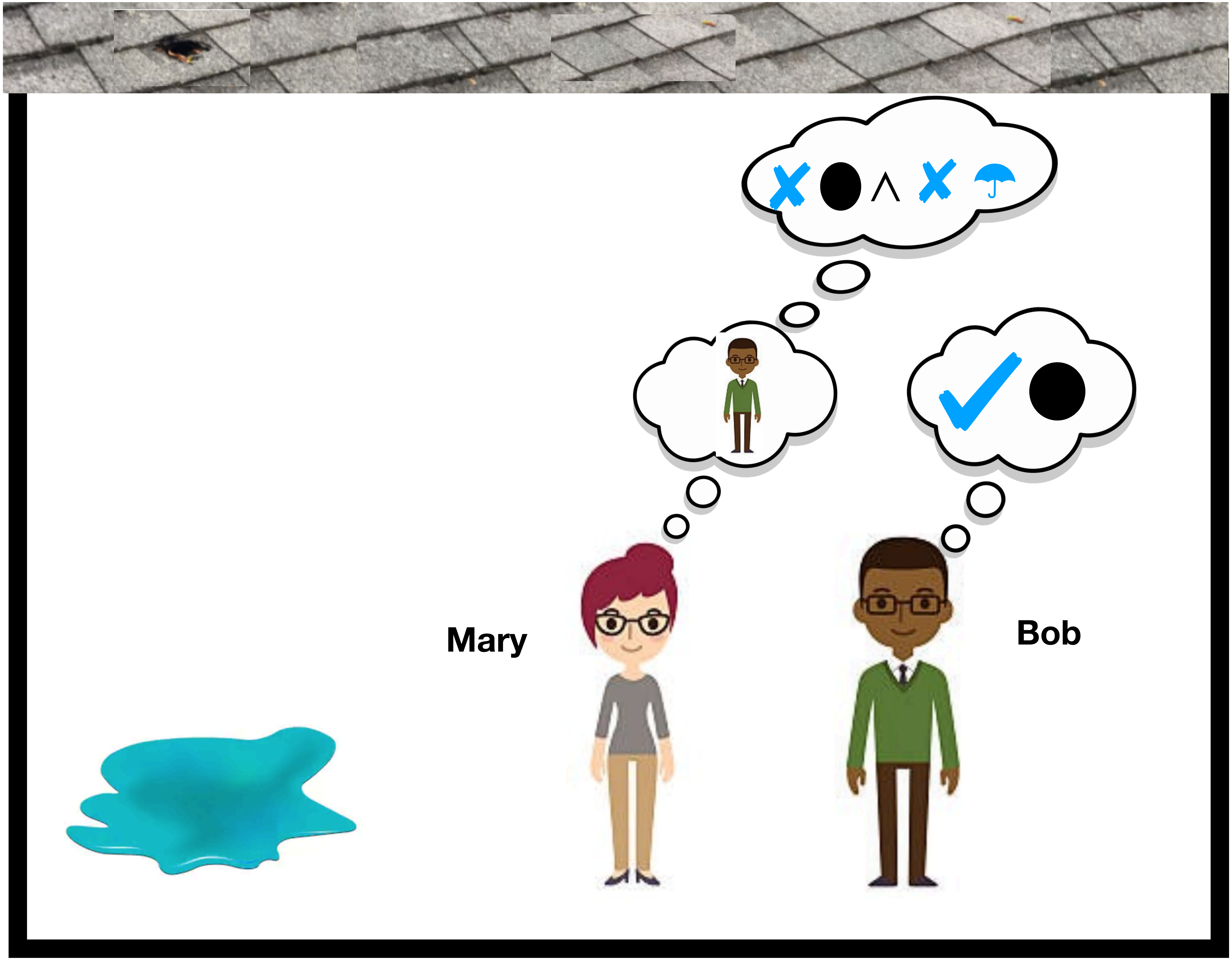


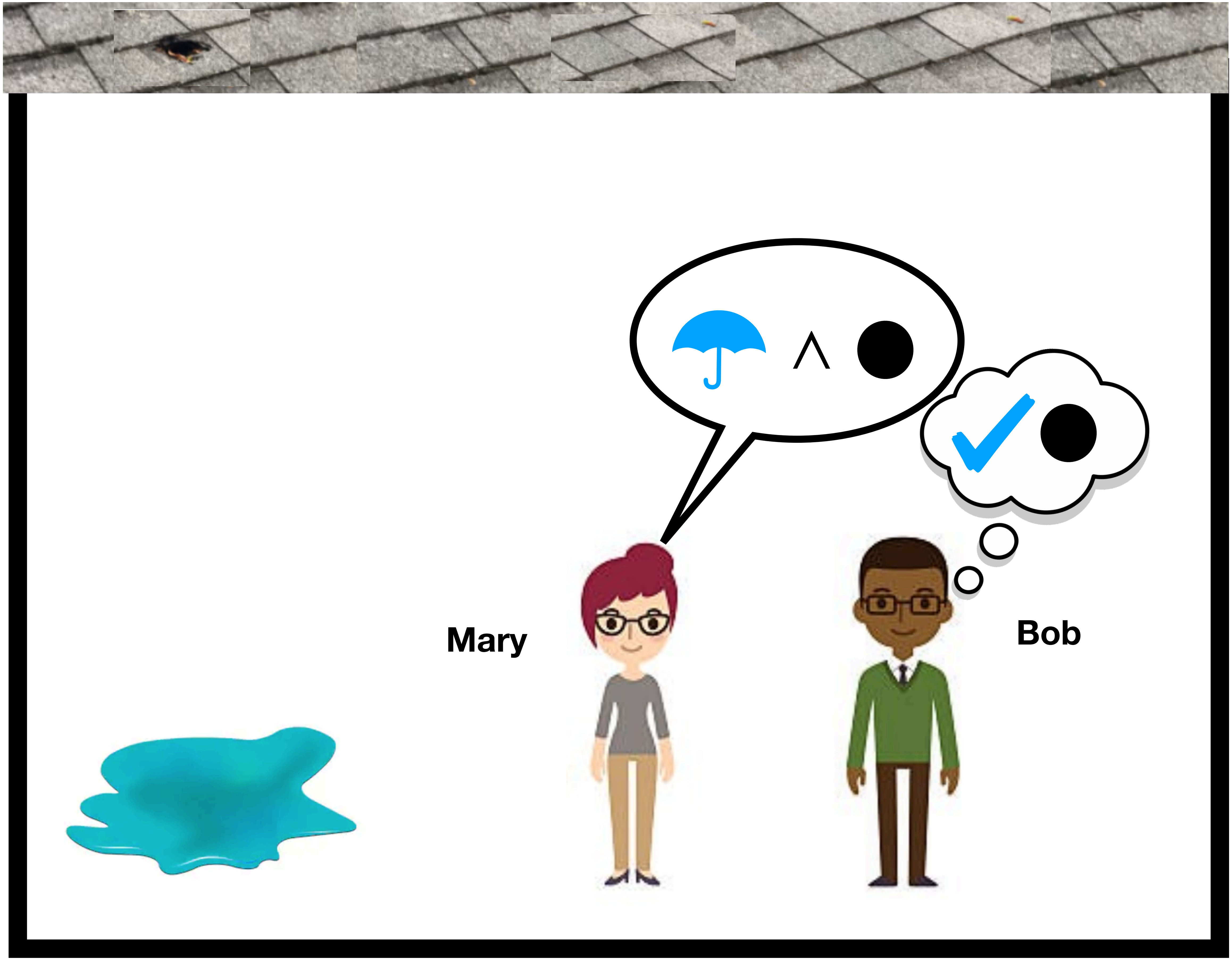
Explainer-Explainee Discrepancies

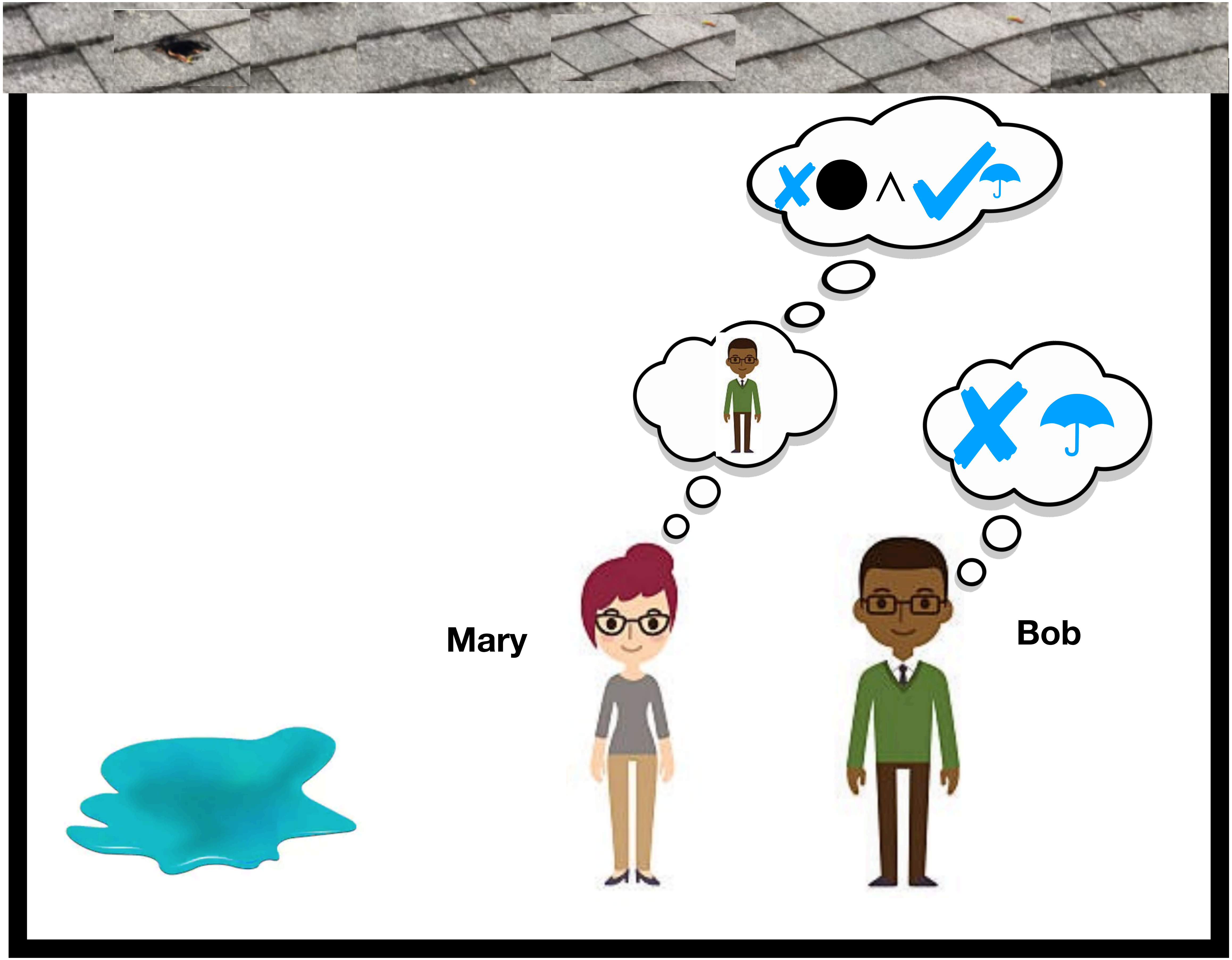


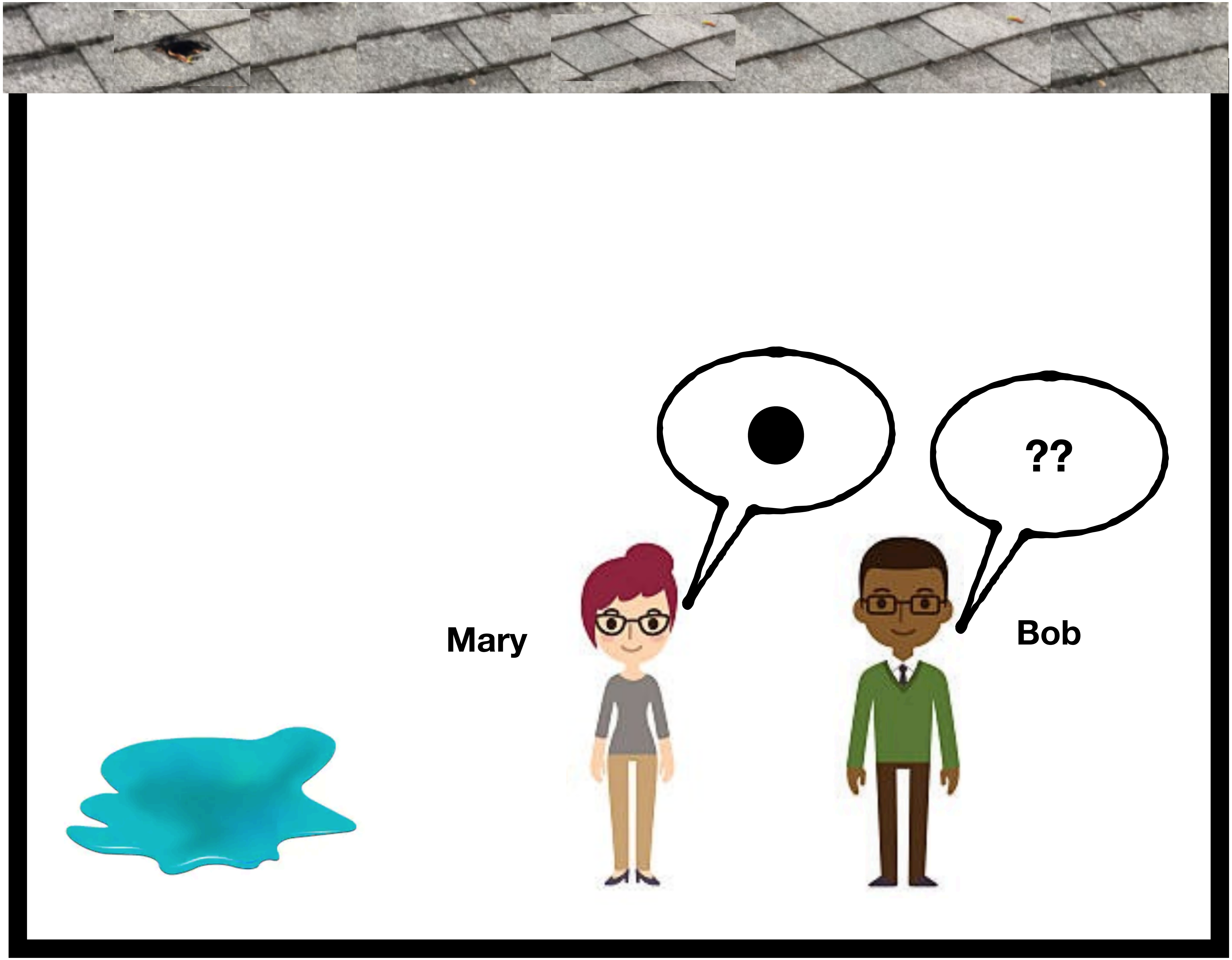
$$B_{Mary} rain \wedge B_{Mary} B_{Bob} \neg rain$$

The (In)Adequacy of the Explainer's Beliefs


✗● ∧ ✗☂
✓●
Mary
Bob


☂ ∧ ●
✓●
Mary
Bob


✗● ∧ ✓☂
✗☂
Mary
Bob


Mary
Bob
??

Summary (and Why You Should Read the Paper)

- We propose a belief-level account of explanation
- We appeal to generic epistemic states
- We appeal to a generic revision operator

☑Multiple explainers and explainees

☑Multiple agent types with different internal belief representations

☑Must allow for both the explainer and explainee to hold false beliefs

☑Explainer must be able to tailor explanations to the explainee’s beliefs

☑Explainer must reason about how the explainee assimilates explanations

Summary (and Why You Should Read the Paper)

- Explainer-Explainee Discrepancies
- The (In)Adequacy of the Explainer's Beliefs

Towards the Role of Theory of Mind in Explanation

Maayan Shvo **Toryn Klassen** **Sheila McIlraith**

maayanshvo@cs.toronto.edu