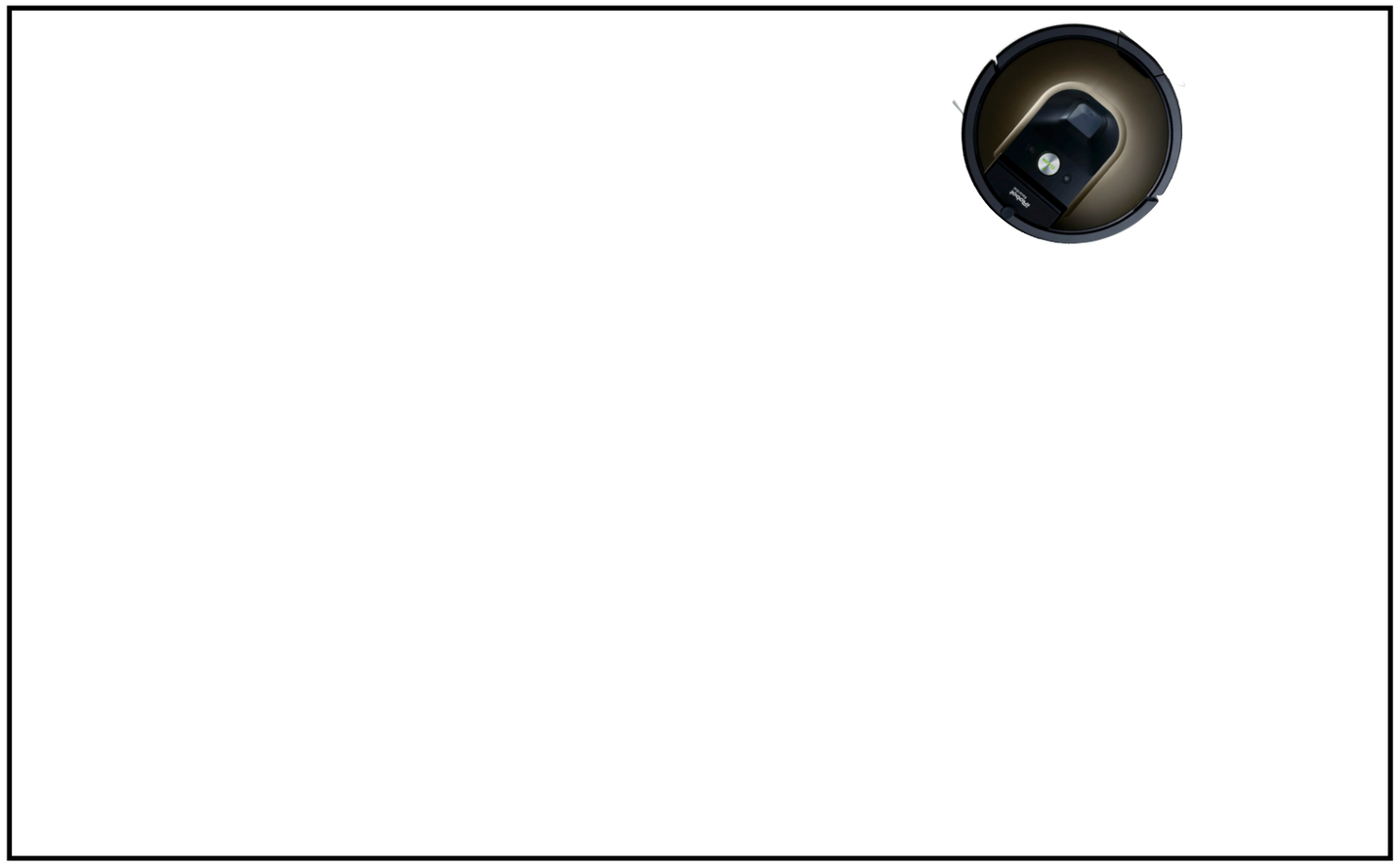
Explainable Planning and Decision Making

William Yeoh

Computer Science and Engineering
Washington University in St. Louis

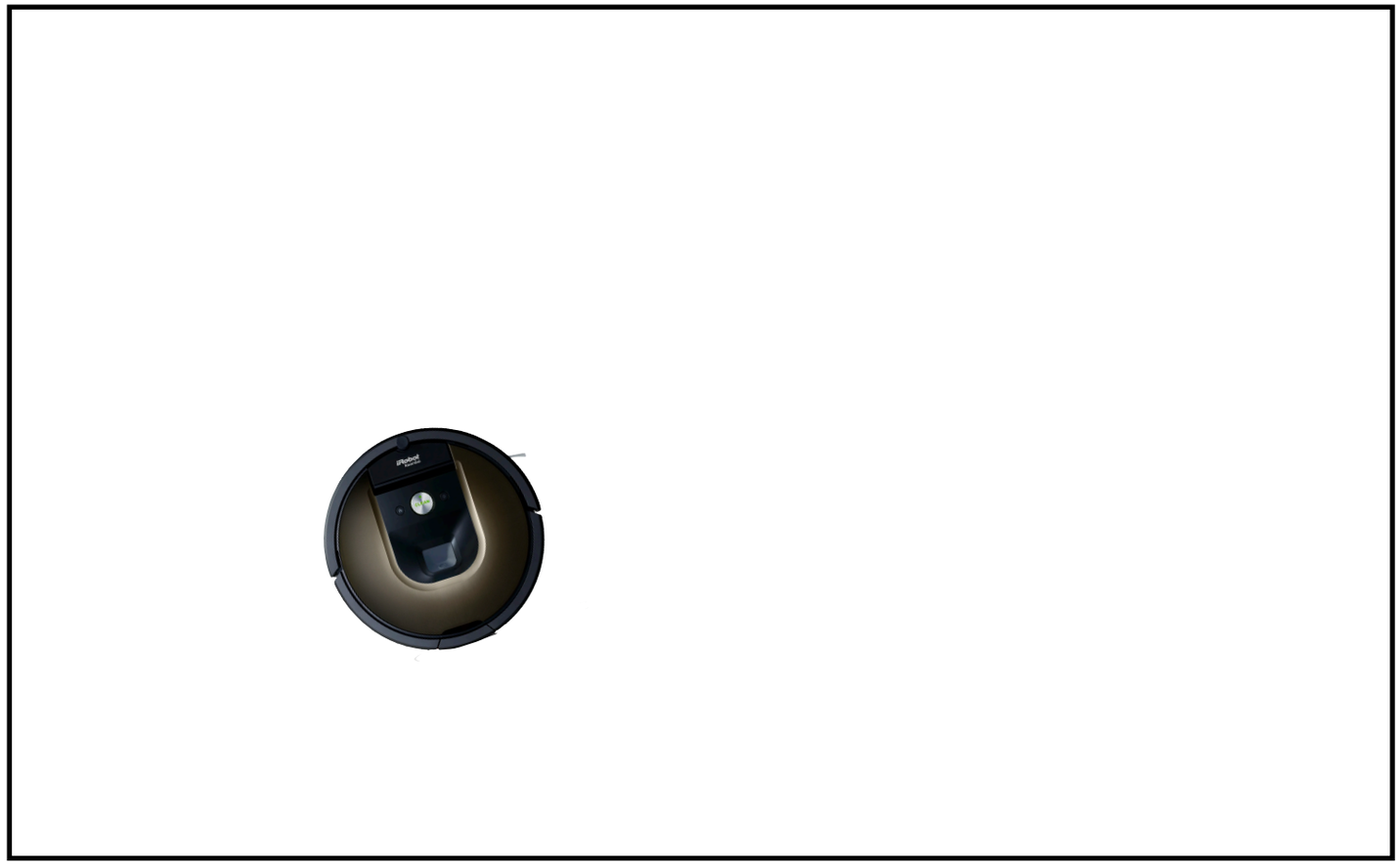
Explainable Planning



Why is the Roomba robot
behaving the way it is behaving?

- Why is it not moving parallel to the walls?
- Why is it moving back and forth?
- Why is it moving in circles?
- …

Explainable Planning



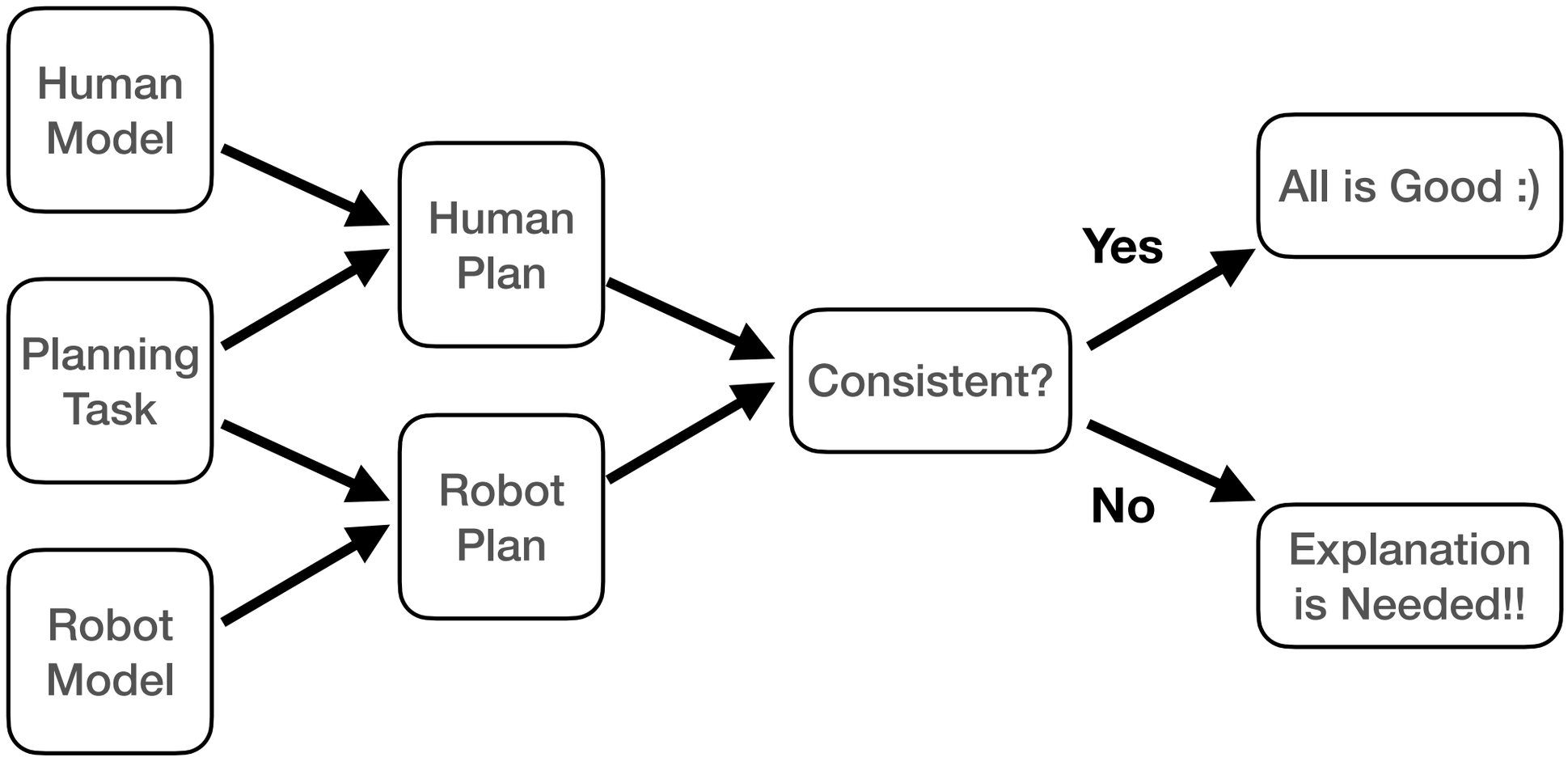
Why is the Roomba robot
behaving the way it is behaving?

- Why is it not moving parallel to the walls?
- Why is it moving back and forth?
- Why is it moving in circles?
- …

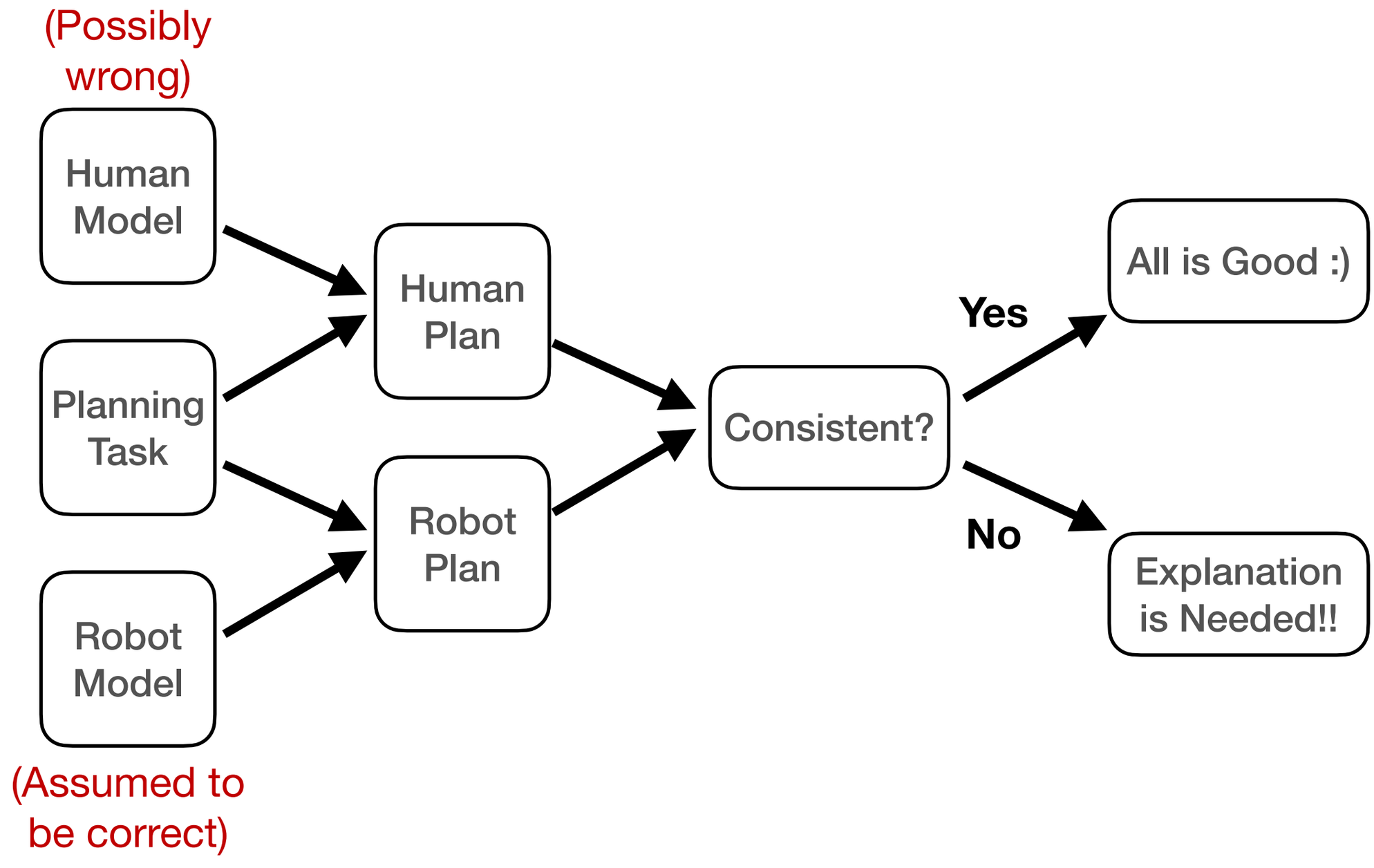
And the reason is …

- Because I have no idea how the walls are aligned
- Because I detected that this area is dirty even after I went over it
- Because my right wheel is stuck
- …

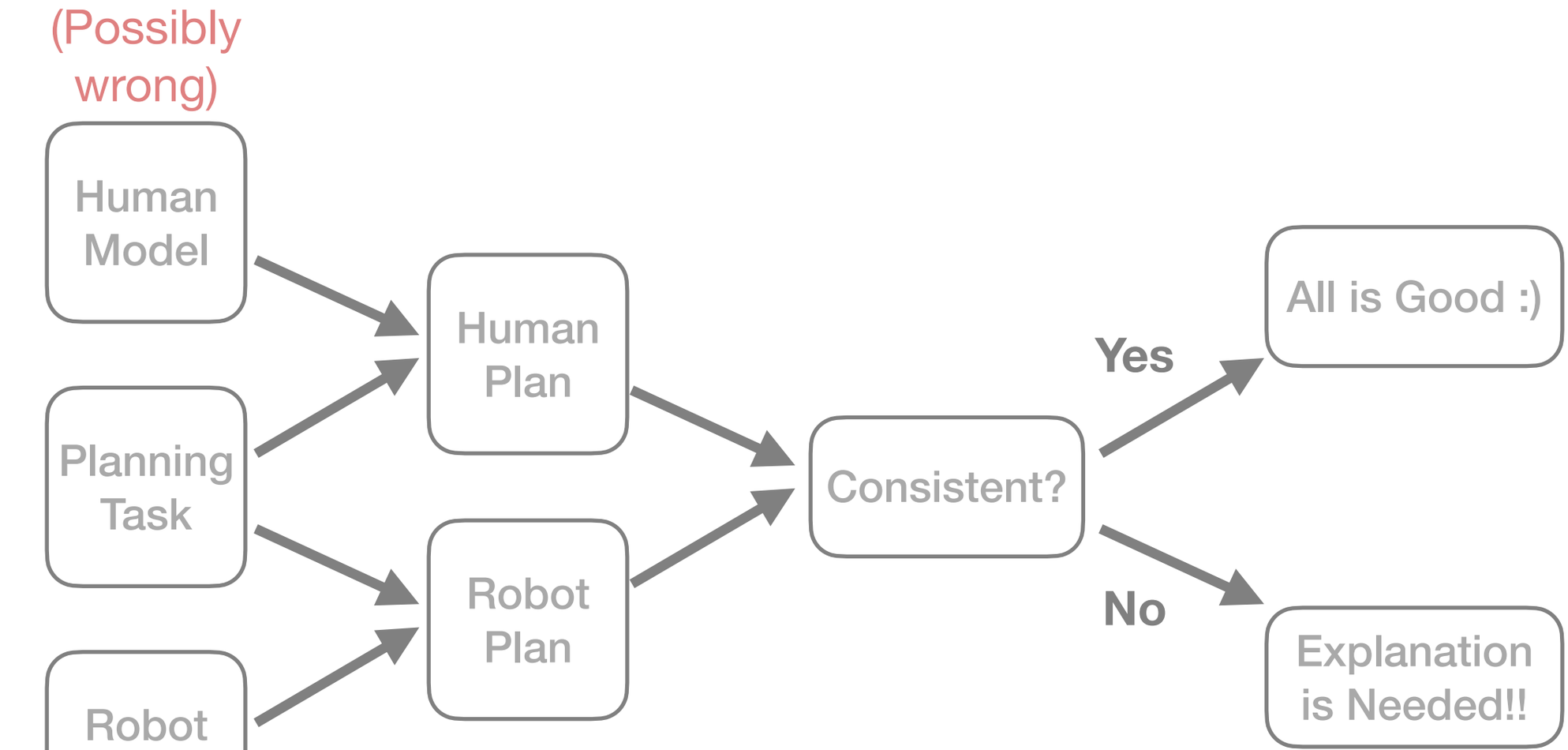
Explainable Planning



Explainable Planning



Explainable Planning



High-Level Idea: Reconcile the differences in the two models such that the human plan from the updated human model is consistent with the robot plan

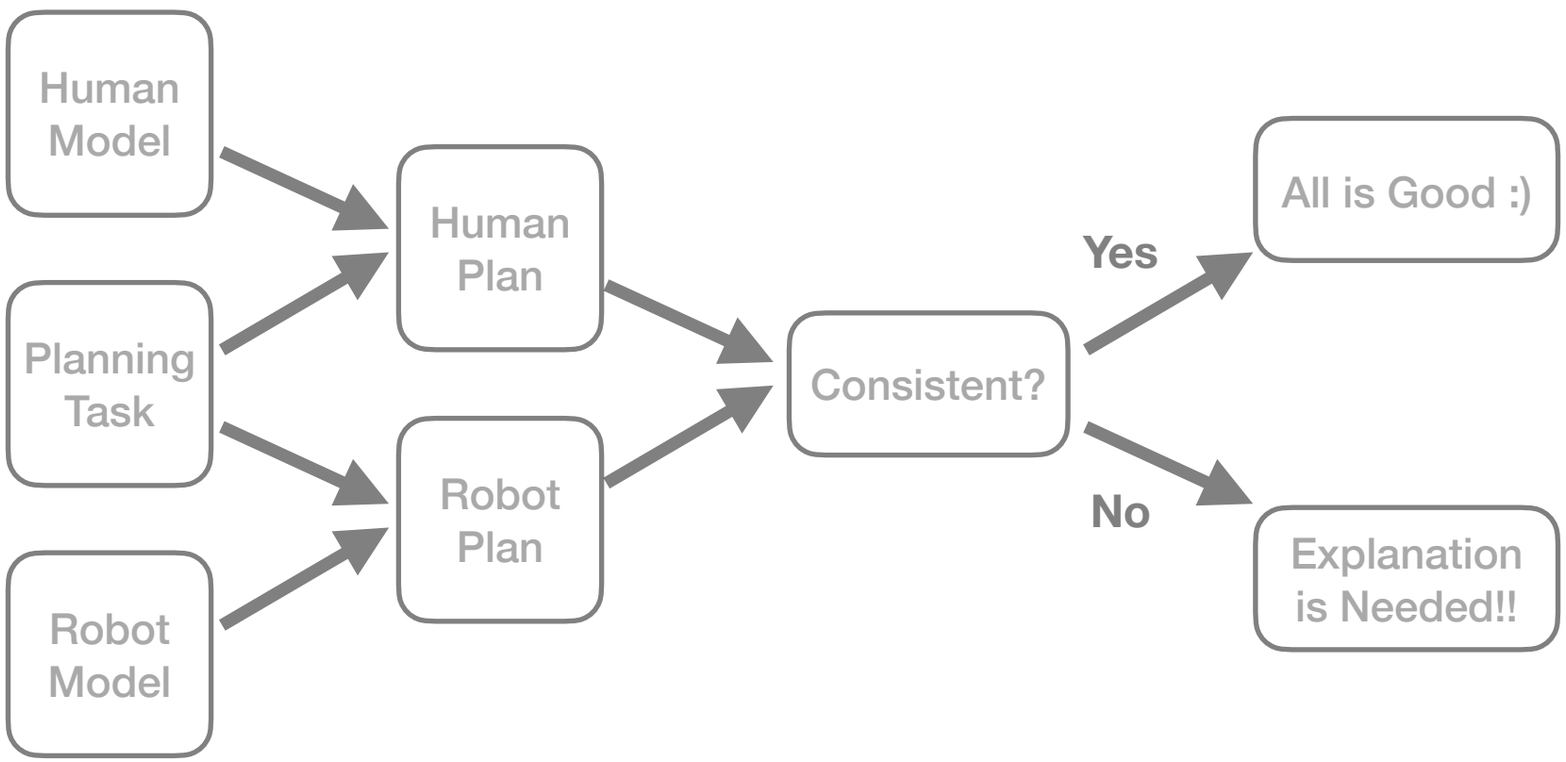
YODAlab

Model Reconciliation Problem

- **Model Reconciliation Problem:**
  - Given two knowledge bases KB_a and KB_h of the agent (robot) providing an explanation and the human receiving the explanation, respectively, and a query q such that KB_a entails q, the goal is to find an explanation (ϵ^+, ϵ^-), where $\epsilon^+ \subseteq KB_a$ and $\epsilon^- \subseteq KB_h$, such that $KB_h \cup \epsilon^+ \backslash \epsilon^-$ entails q.

Model Reconciliation Problem

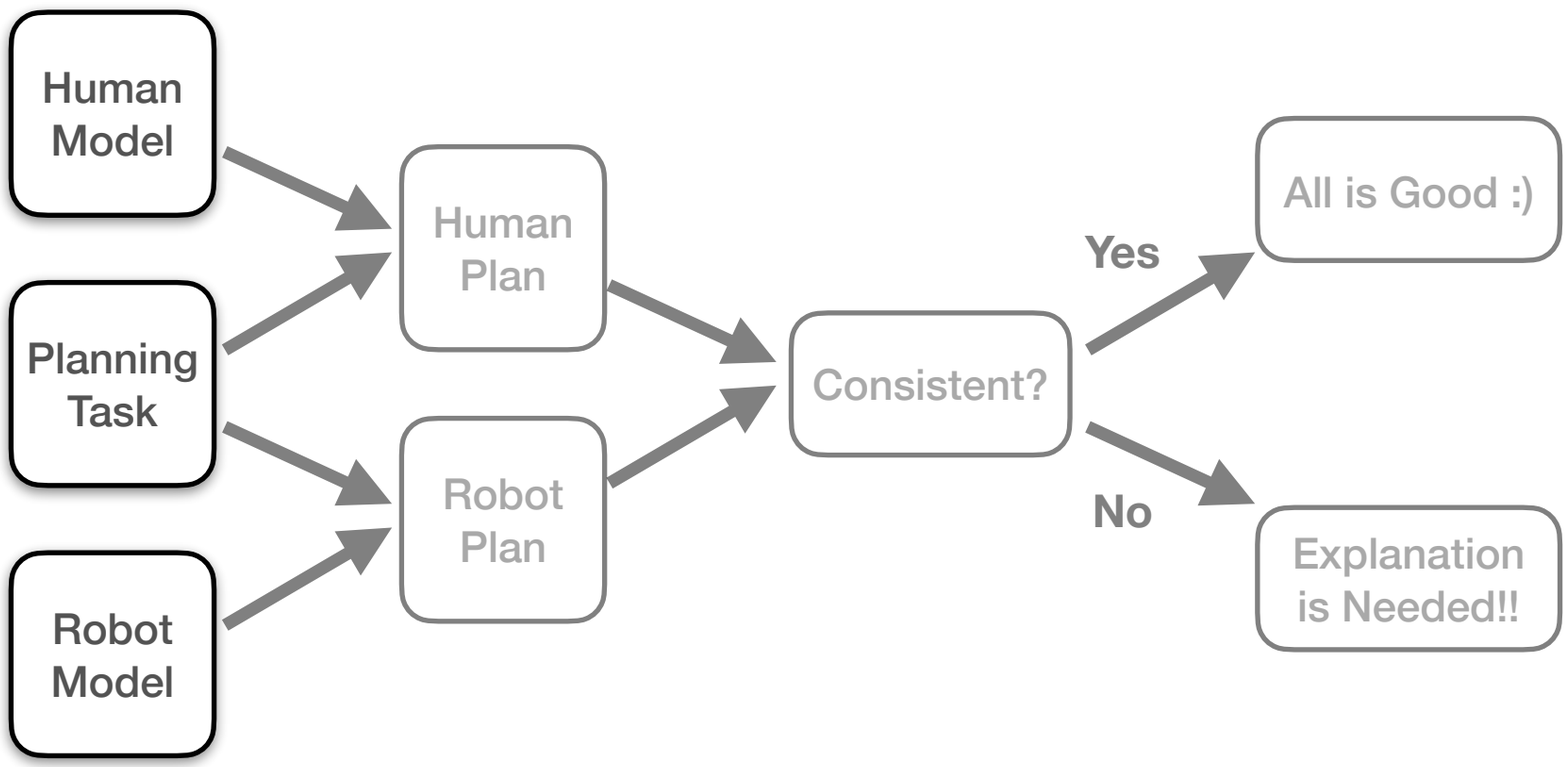
- **Model Reconciliation Problem:**
  - Given two knowledge bases KB_a and KB_h of the agent (robot) providing an explanation and the human receiving the explanation, respectively, and a query q such that KB_a entails q, the goal is to find an explanation (ϵ^+, ϵ^-), where $\epsilon^+ \subseteq KB_a$ and $\epsilon^- \subseteq KB_h$, such that $KB_h \cup \epsilon^+ \backslash \epsilon^-$ entails q.



YODAlab

Model Reconciliation Problem

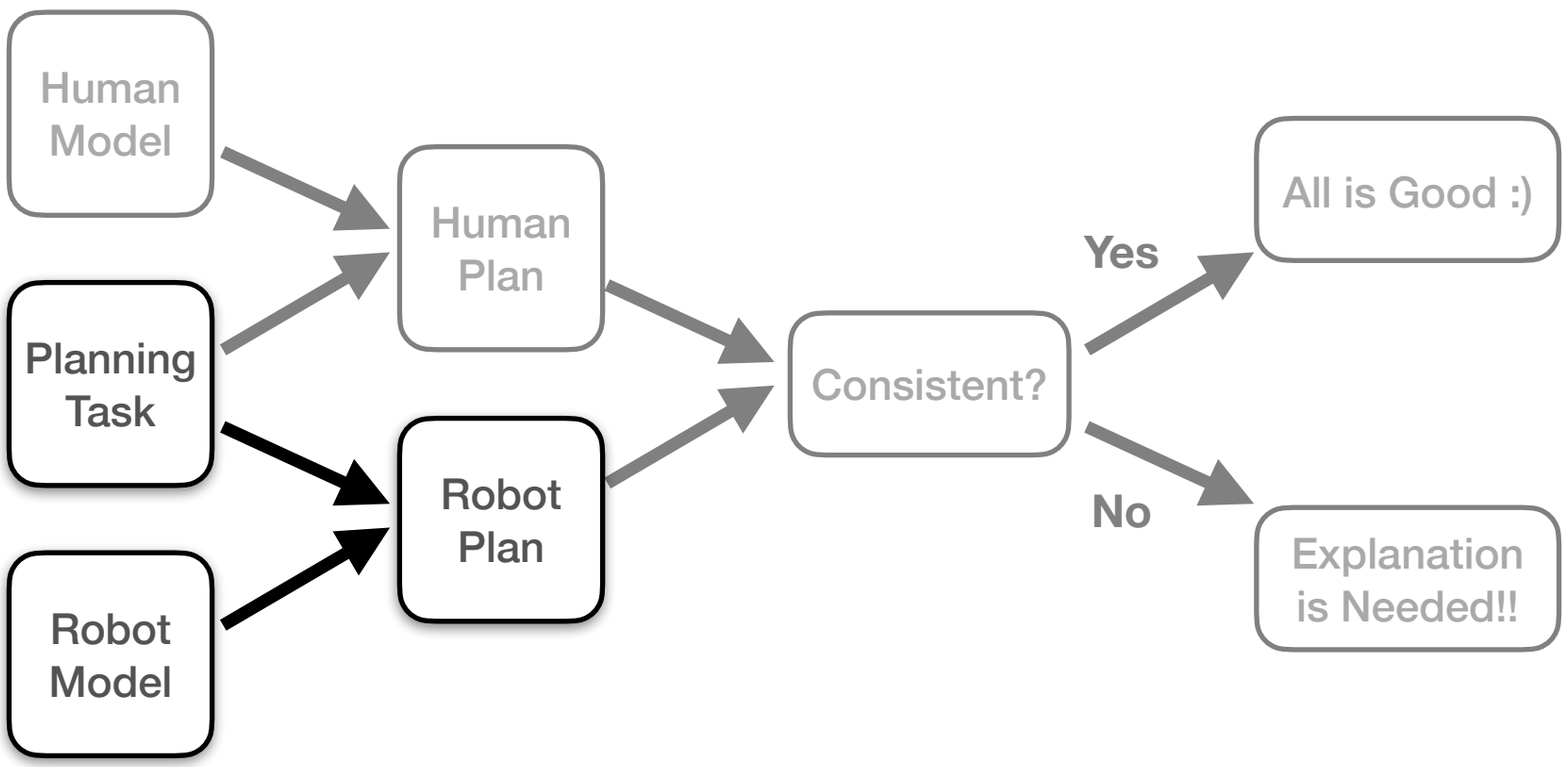
- **Model Reconciliation Problem:**
  - Given two knowledge bases KB_a and KB_h of the agent (robot) providing an explanation and the human receiving the explanation, respectively, and a query q such that KB_a entails q, the goal is to find an explanation (ϵ^+, ϵ^-), where $\epsilon^+ \subseteq KB_a$ and $\epsilon^- \subseteq KB_h$, such that $KB_h \cup \epsilon^+ \backslash \epsilon^-$ entails q.



YODAlab

Model Reconciliation Problem

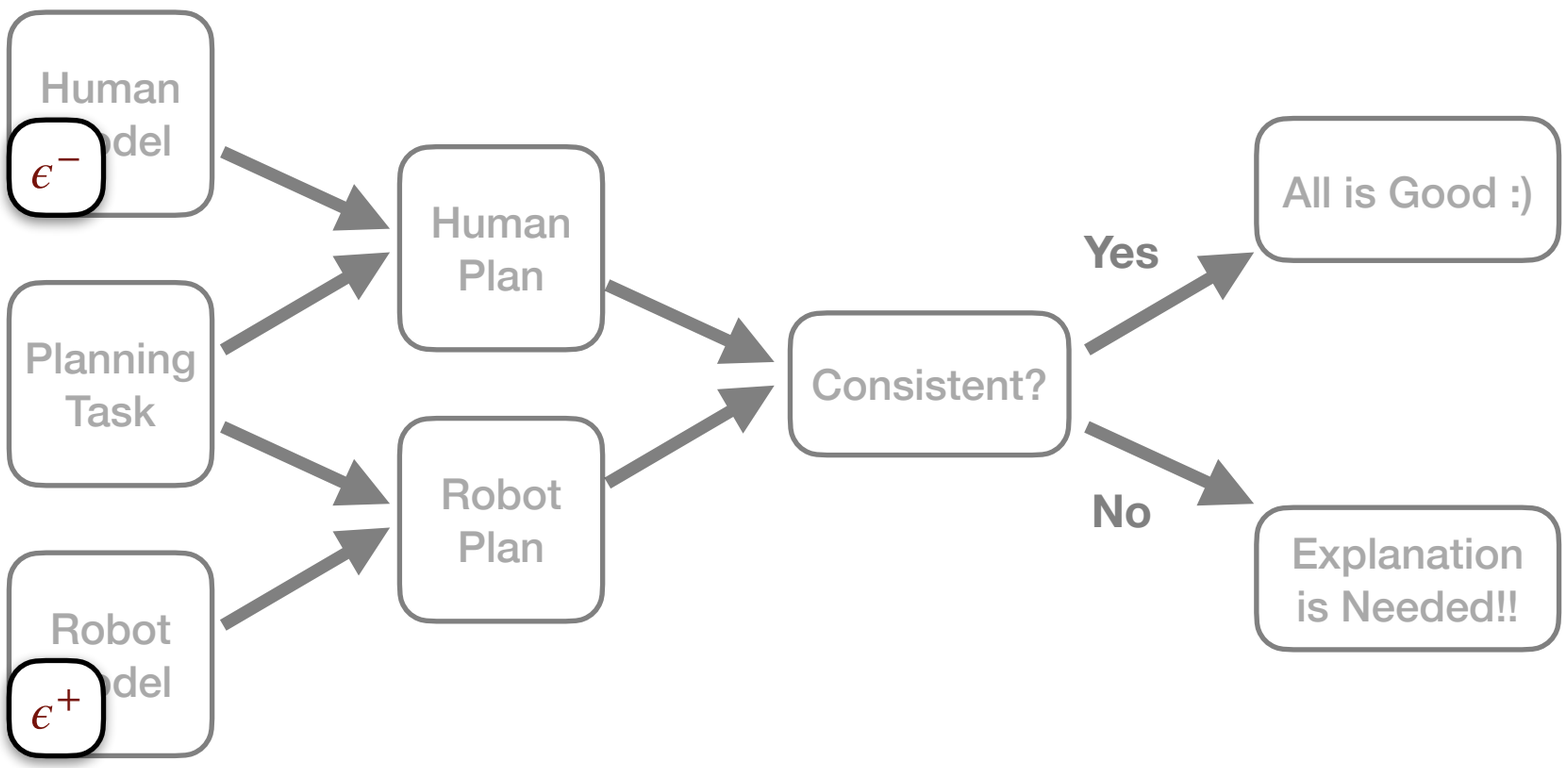
- **Model Reconciliation Problem:**
  - Given two knowledge bases KB_a and KB_h of the agent (robot) providing an explanation and the human receiving the explanation, respectively, and a query q such that KB_a entails q, the goal is to find an explanation (ϵ^+, ϵ^-), where $\epsilon^+ \subseteq KB_a$ and $\epsilon^- \subseteq KB_h$, such that $KB_h \cup \epsilon^+ \backslash \epsilon^-$ entails q.



YODAlab

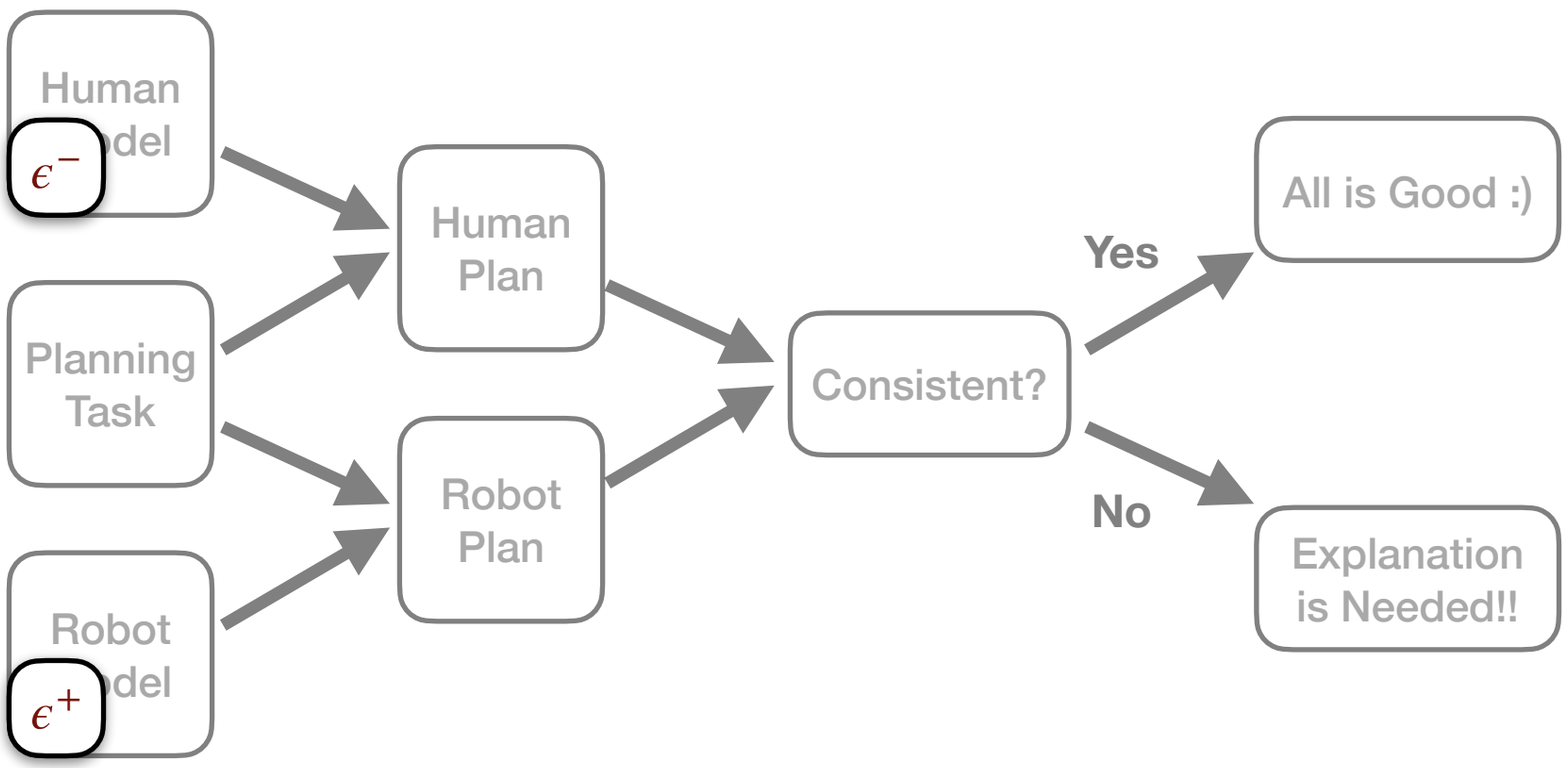
Model Reconciliation Problem

- **Model Reconciliation Problem:**
  - Given two knowledge bases KB_a and KB_h of the agent (robot) providing an explanation and the human receiving the explanation, respectively, and a query q such that KB_a entails q, the goal is to find an explanation (ϵ^+, ϵ^-), where $\epsilon^+ \subseteq KB_a$ and $\epsilon^- \subseteq KB_h$, such that $KB_h \cup \epsilon^+ \backslash \epsilon^-$ entails q.



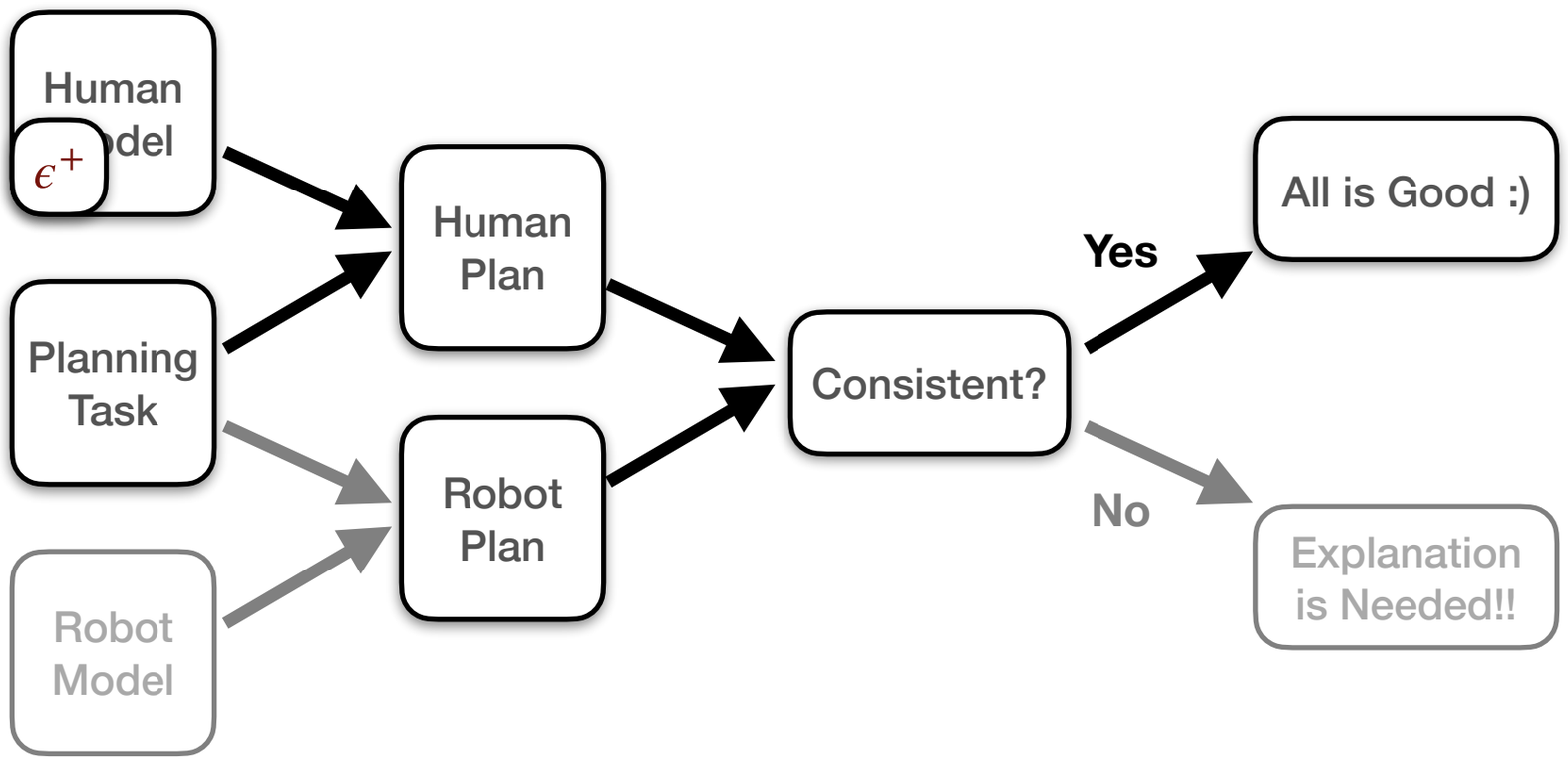
Model Reconciliation Problem

- **Model Reconciliation Problem:**
  - Given two knowledge bases KB_a and KB_h of the agent (robot) providing an explanation and the human receiving the explanation, respectively, and a query q such that KB_a entails q, the goal is to find an explanation (ϵ^+, ϵ^-), where $\epsilon^+ \subseteq KB_a$ and $\epsilon^- \subseteq KB_h$, such that $KB_h \cup \epsilon^+ \backslash \epsilon^-$ entails q.



Model Reconciliation Problem

- **Model Reconciliation Problem:**
  - Given two knowledge bases KB_a and KB_h of the agent (robot) providing an explanation and the human receiving the explanation, respectively, and a query q such that KB_a entails q, the goal is to find an explanation (ϵ^+, ϵ^-), where $\epsilon^+ \subseteq KB_a$ and $\epsilon^- \subseteq KB_h$, such that $KB_h \cup \epsilon^+ \backslash \epsilon^-$ entails q.



YODAlab

Model Reconciliation Problem

- **Model Reconciliation Problem:**
  - Given two knowledge bases KB_a and KB_h of the agent (robot) providing an explanation and the human receiving the explanation, respectively, and a query q such that KB_a entails q, the goal is to find an explanation (ϵ^+, ϵ^-), where $\epsilon^+ \subseteq KB_a$ and $\epsilon^- \subseteq KB_h$, such that $KB_h \cup \epsilon^+ \backslash \epsilon^-$ entails q.

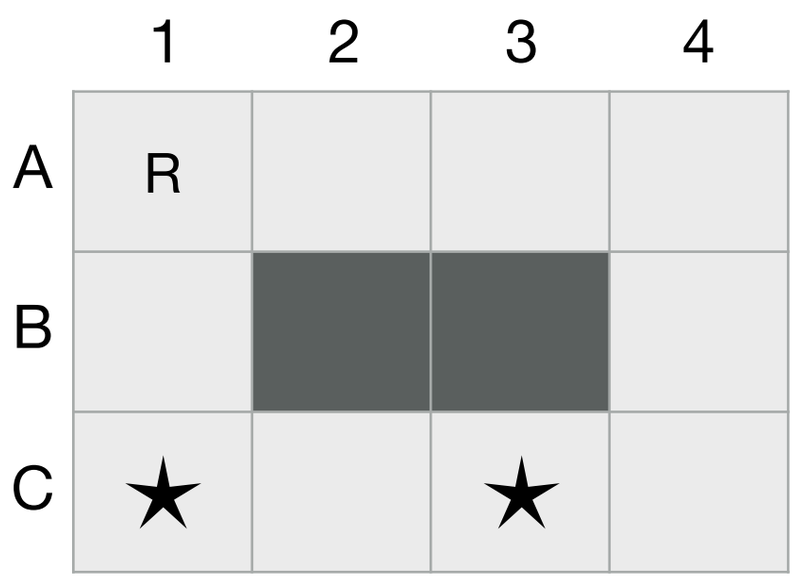
- Research questions:
  - How to effectively find an explanation (ϵ^+, ϵ^-)?
  - How to measure the goodness of explanations?
  - How to convey explanations to the human?
  - What if we don't know the model of the human?
  - What if the model of the agent (robot) is wrong?
  - …

Model Reconciliation Problem

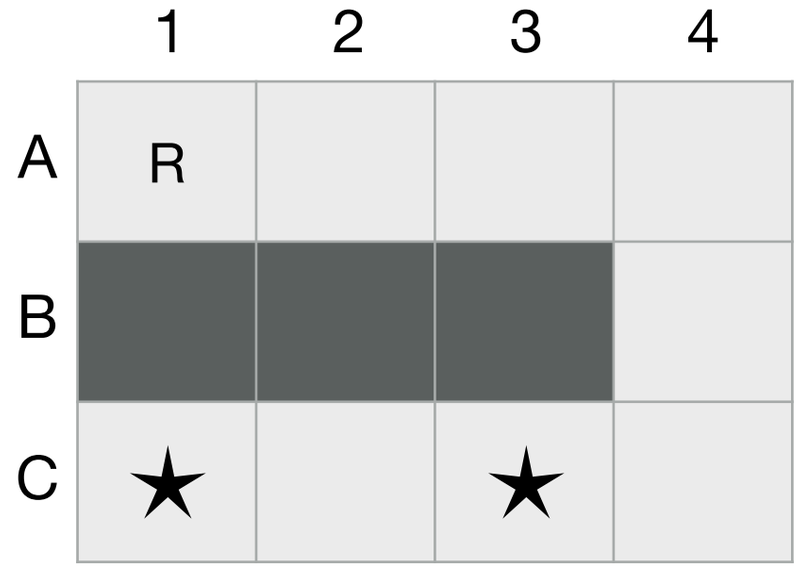
- **Model Reconciliation Problem:**
  - Given two knowledge bases KB_a and KB_h of the agent (robot) providing an explanation and the human receiving the explanation, respectively, and a query q such that KB_a entails q, the goal is to find an explanation (ϵ^+, ϵ^-), where $\epsilon^+ \subseteq KB_a$ and $\epsilon^- \subseteq KB_h$, such that $KB_h \cup \epsilon^+ \backslash \epsilon^-$ entails q.

- Research questions:
  - How to effectively find an explanation (ϵ^+, ϵ^-)?
  - How to measure the goodness of explanations?
  - How to convey explanations to the human?
  - What if we don't know the model of the human?
  - What if the model of the agent (robot) is wrong?
  - …

Illustrative Toy Example

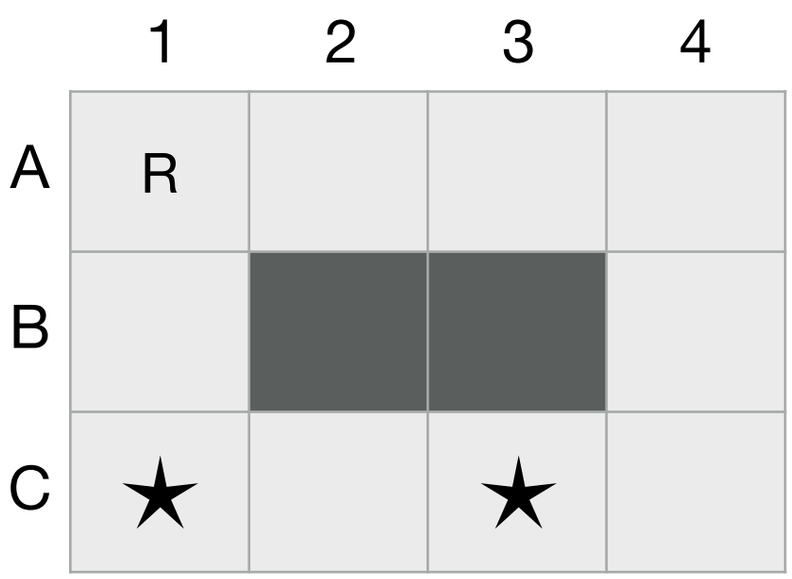


Human's Model

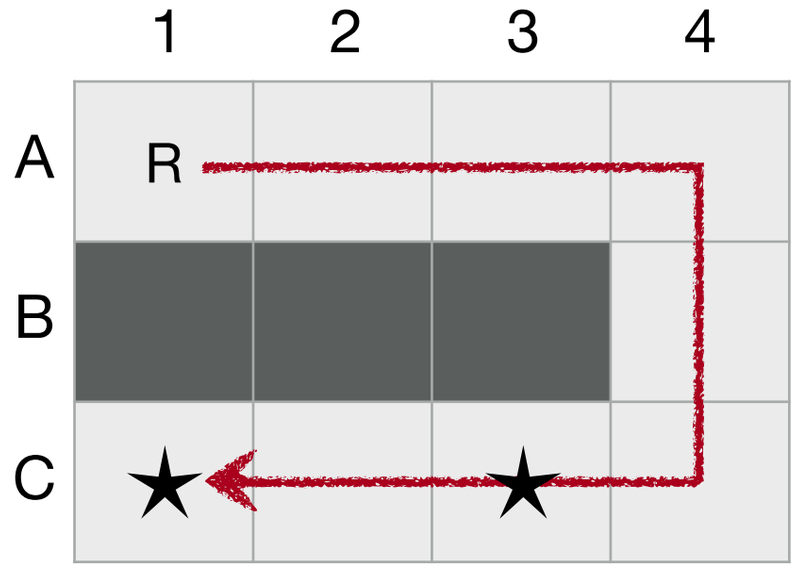


Robot's Model

Illustrative Toy Example

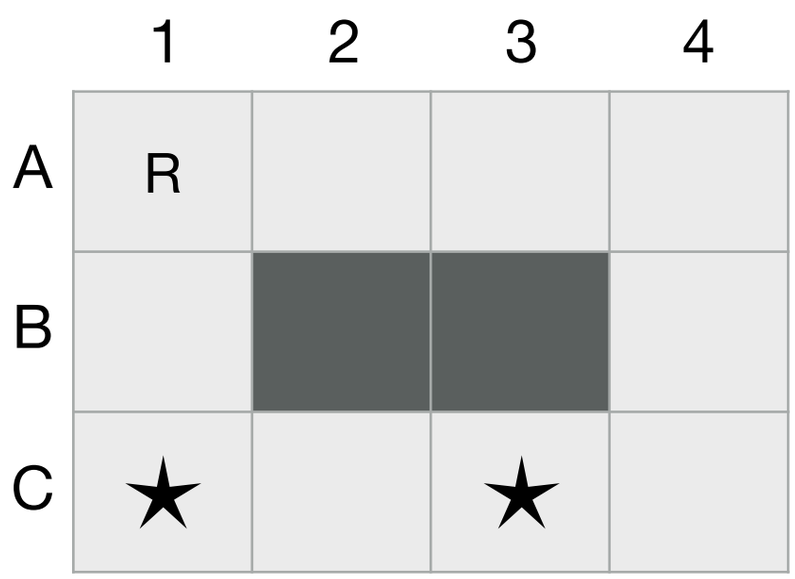


Human's Model

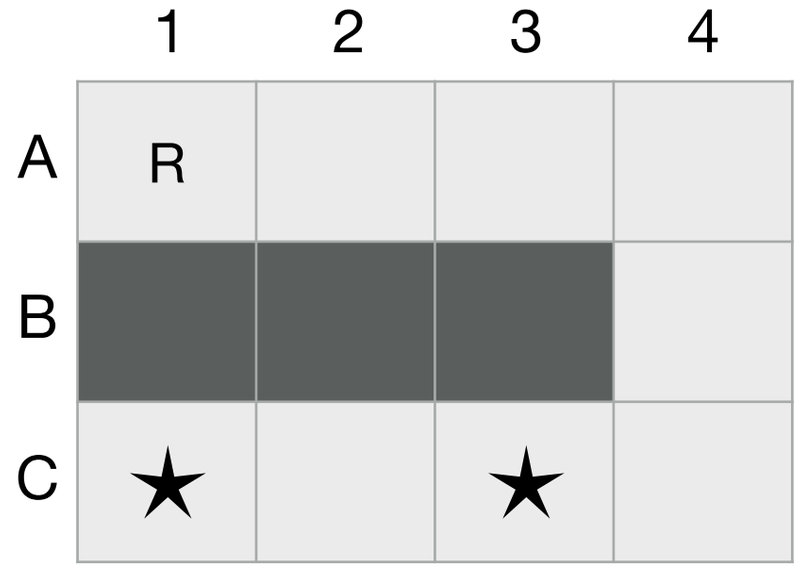


Robot's Model

Modeling Planning as SAT



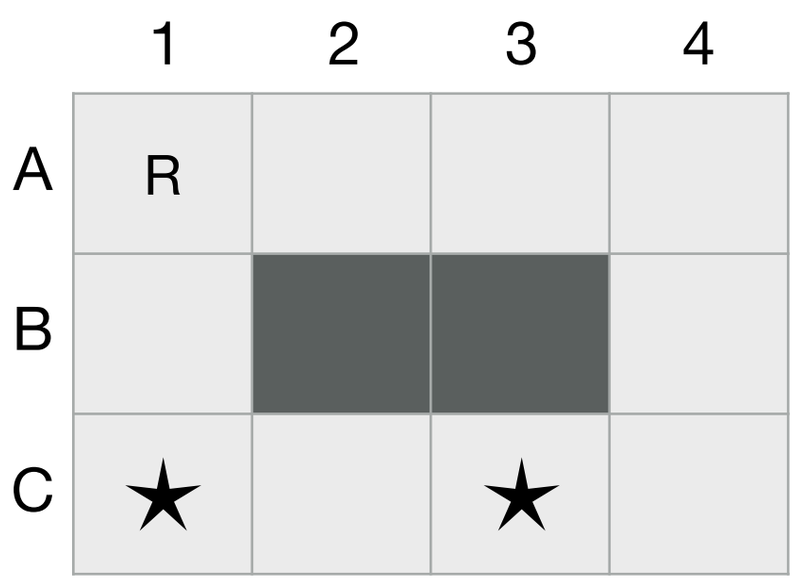
Human's Model



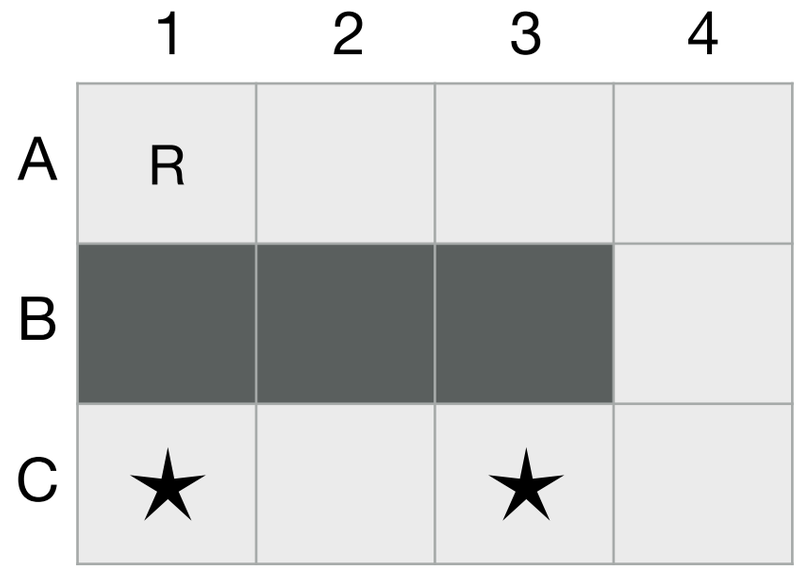
Robot's Model

- **Modeling Planning as a Boolean Satisfiability (SAT) Problem**
  - Starting robot location: $A1_0, \neg A2_0, \ldots, \neg C4_0$

Modeling Planning as SAT



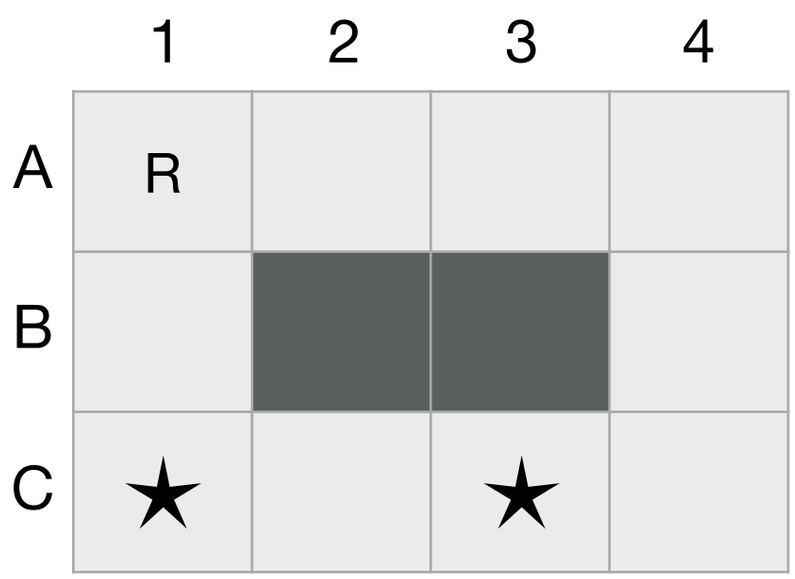
Human's Model



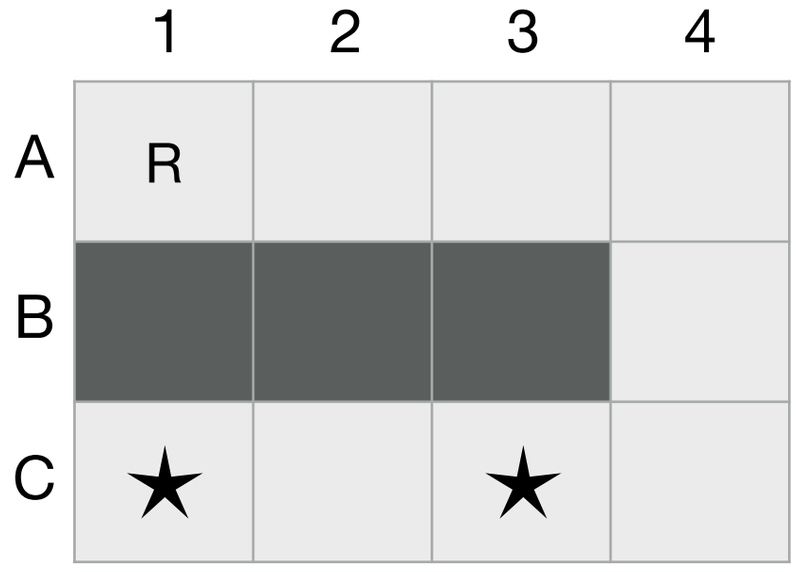
Robot's Model

- **Modeling Planning as a Boolean Satisfiability (SAT) Problem**
  - Starting robot location: $A1_0, \neg A2_0, \ldots, \neg C4_0$
  - Blocked locations: $\neg Blocked\_A1, \neg Blocked\_B1, \ldots$

Modeling Planning as SAT



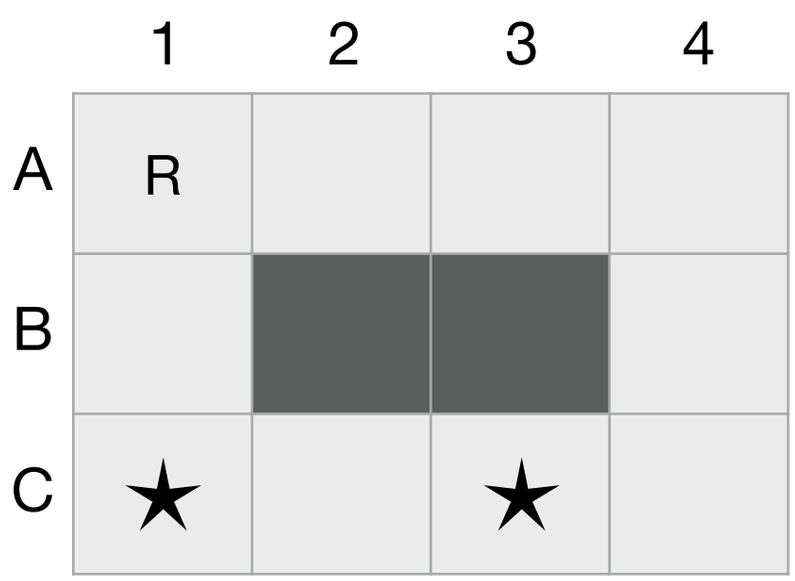
Human's Model



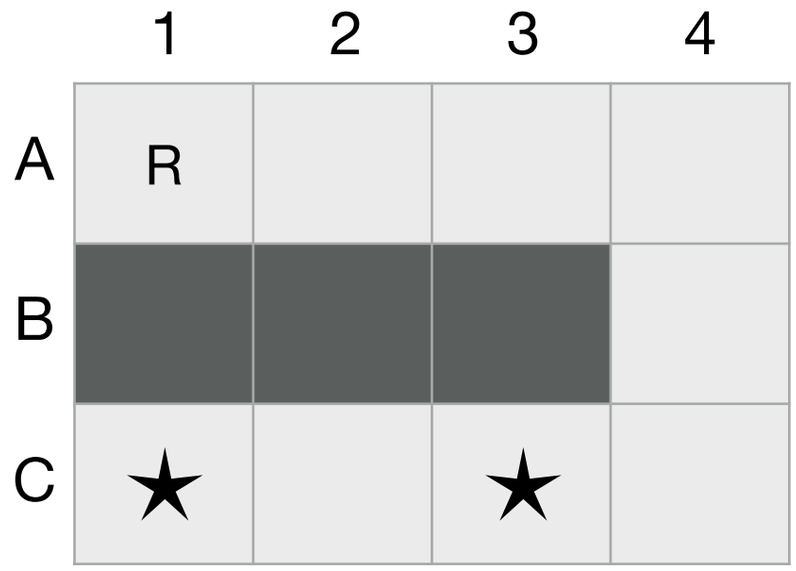
Robot's Model

- **Modeling Planning as a Boolean Satisfiability (SAT) Problem**
  - Starting robot location: $A1_0, \neg A2_0, \ldots, \neg C4_0$
  - Blocked locations: $\neg Blocked\_A1, \neg Blocked\_B1, \ldots$
  - Goal states: $C1_t \wedge C3_{t'}$ for some time steps t and t'

Modeling Planning as SAT



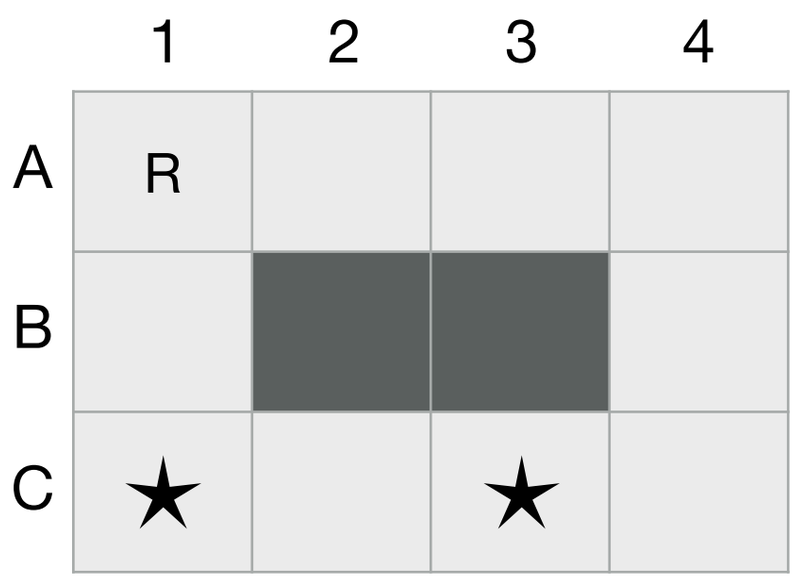
Human's Model



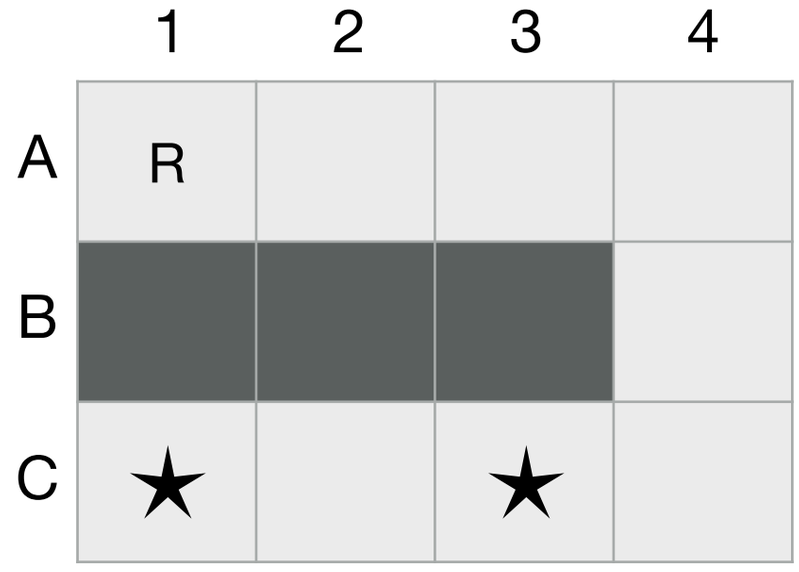
Robot's Model

- **Modeling Planning as a Boolean Satisfiability (SAT) Problem**
  - Starting robot location: $A1_0, \neg A2_0, \ldots, \neg C4_0$
  - Blocked locations: $\neg Blocked\_A1, \neg Blocked\_B1, \ldots$
  - Goal states: $C1_t \wedge C3_{t'}$ for some time steps t and t'
  - Actions of robots:
    - $\downarrow_t \wedge A1_t \wedge \neg Blocked\_B1 \rightarrow B1_{t+1}$
    - $\downarrow_t \wedge A1_t \wedge Blocked\_B1 \rightarrow A1_{t+1}$
    - …

Modeling Planning as SAT



Human's Model

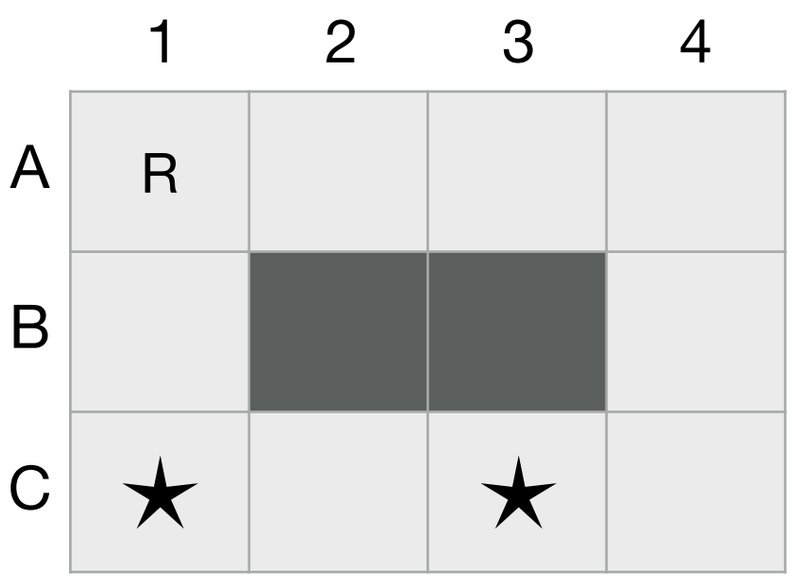


Robot's Model

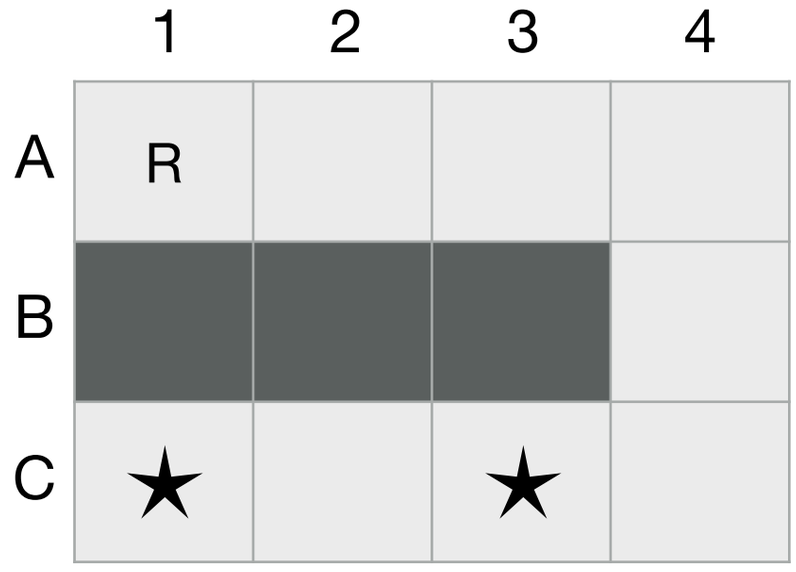
- **Modeling Planning as a Boolean Satisfiability (SAT) Problem**
  - Starting robot location: $A1_0, \neg A2_0, \ldots, \neg C4_0$
  - Blocked locations: $\neg Blocked\_A1, \neg Blocked\_B1, \ldots$
  - Goal states: $C1_t \wedge C3_{t'}$ for some time steps t and t'
  - Actions of robots:
    - $\downarrow_t \wedge A1_t \wedge \neg Blocked\_B1 \rightarrow B1_{t+1}$
    - $\downarrow_t \wedge A1_t \wedge Blocked\_B1 \rightarrow A1_{t+1}$
    - …

KB_h

Modeling Planning as SAT



Human's Model

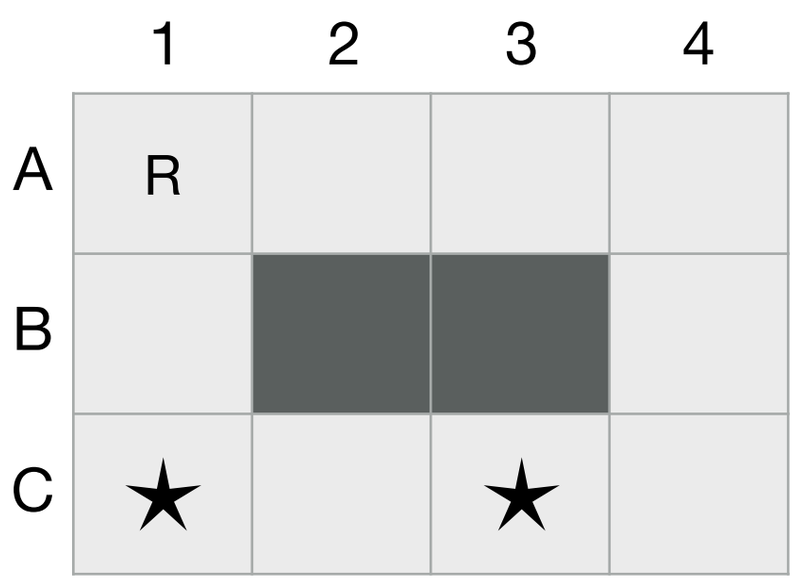


Robot's Model

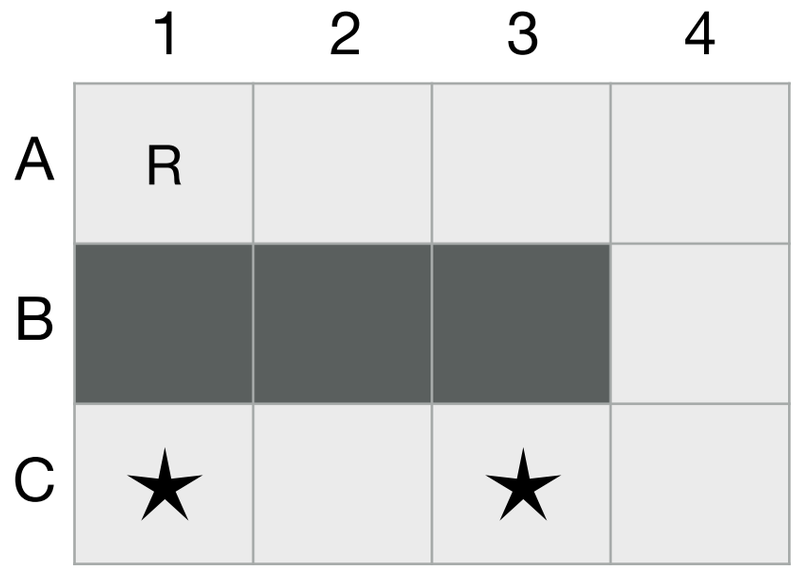
- **Modeling Planning as a Boolean Satisfiability (SAT) Problem**
  - Starting robot location: $A1_0, \neg A2_0, \ldots, \neg C4_0$
  - Blocked locations: $\neg Blocked\_A1, Blocked\_B1, \ldots$
  - Goal states: $C1_t \wedge C3_{t'}$ for some time steps t and t'
  - Actions of robots:
    - $\downarrow_t \wedge A1_t \wedge \neg Blocked\_B1 \rightarrow B1_{t+1}$
    - $\downarrow_t \wedge A1_t \wedge Blocked\_B1 \rightarrow A1_{t+1}$
    - …

KB_a

Modeling Planning as SAT



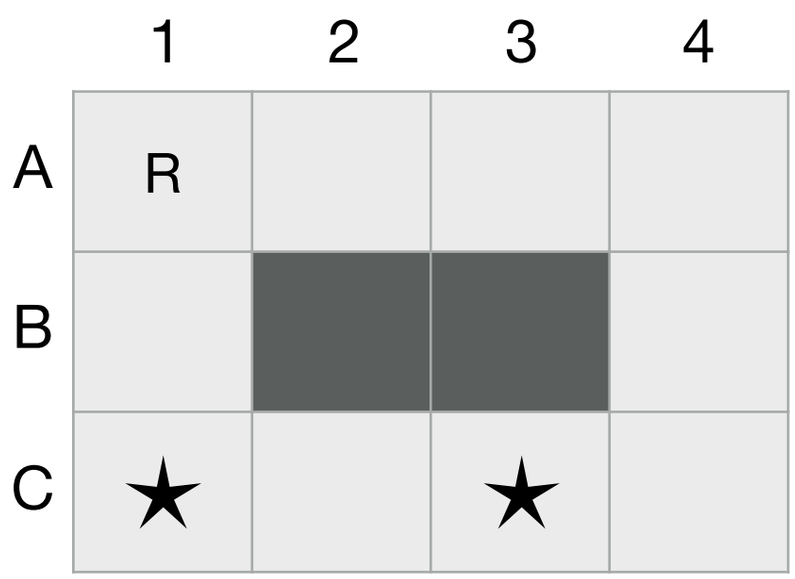
Human's Model



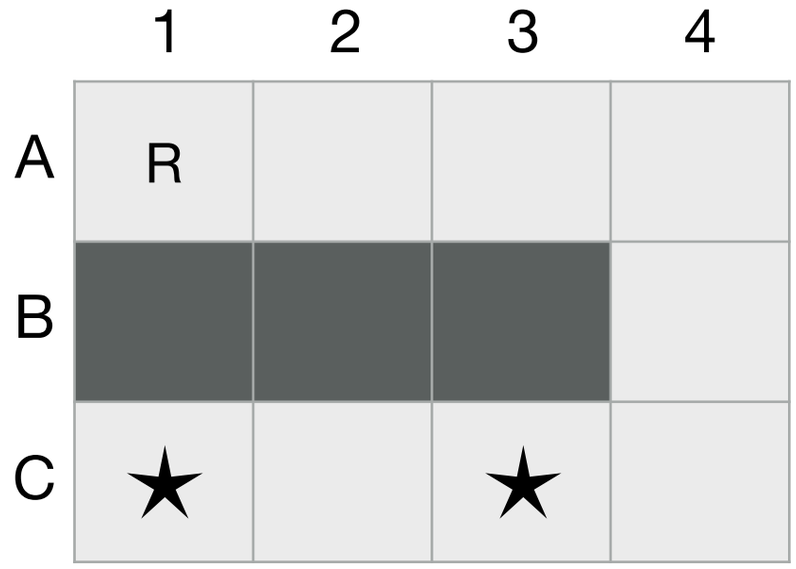
Robot's Model

- **Modeling Planning as a Boolean Satisfiability (SAT) Problem**
  - Goal: Find truth value assignments to sequence of actions such that the goal states are reached
  - A lot of efficient SAT solvers exists to solve this problem

Modeling Planning as SAT



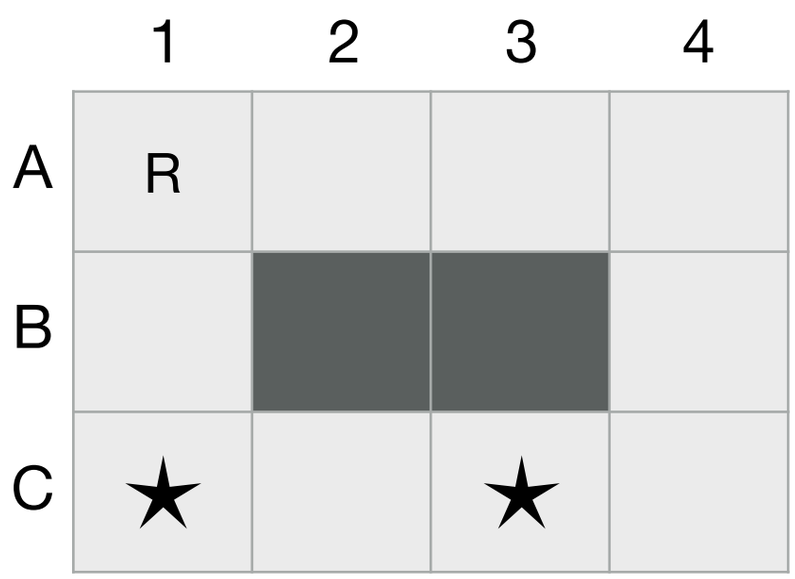
Human's Model



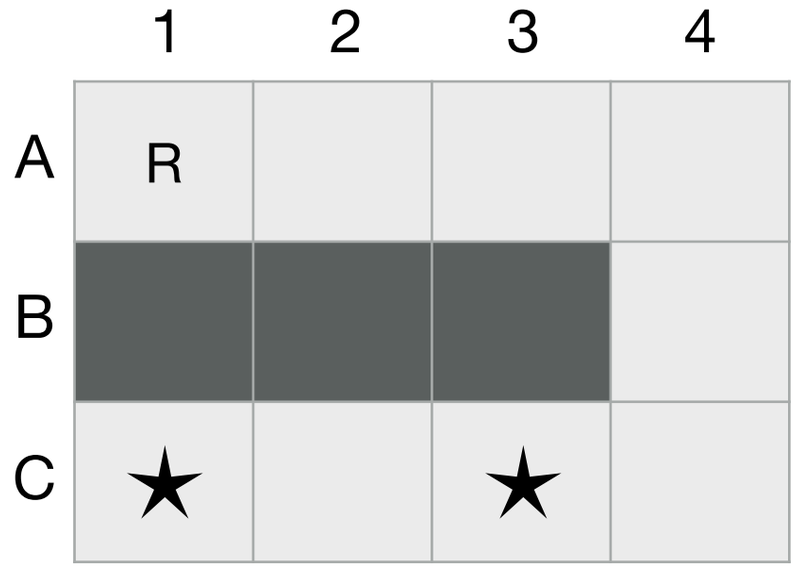
Robot's Model

- **Modeling Planning as a Boolean Satisfiability (SAT) Problem**
  - Goal: Find truth value assignments to sequence of actions such that the goal states are reached
    - For human: $\downarrow_0, \downarrow_1, \rightarrow_2, \rightarrow_3$ results in $C1_2, C3_4$
    - For robot: $\rightarrow_0, \rightarrow_1, \rightarrow_2, \downarrow_3, \downarrow_4, \leftarrow_5, \leftarrow_6, \leftarrow_7$ results in $C1_8, C3_6$
  - Human's question:
    - Why $\rightarrow_0, \rightarrow_1, \rightarrow_2, \downarrow_3, \downarrow_4, \leftarrow_5, \leftarrow_6, \leftarrow_7$ and not $\downarrow_0, \downarrow_1, \rightarrow_2, \rightarrow_3$?

Modeling Planning as SAT



Human's Model



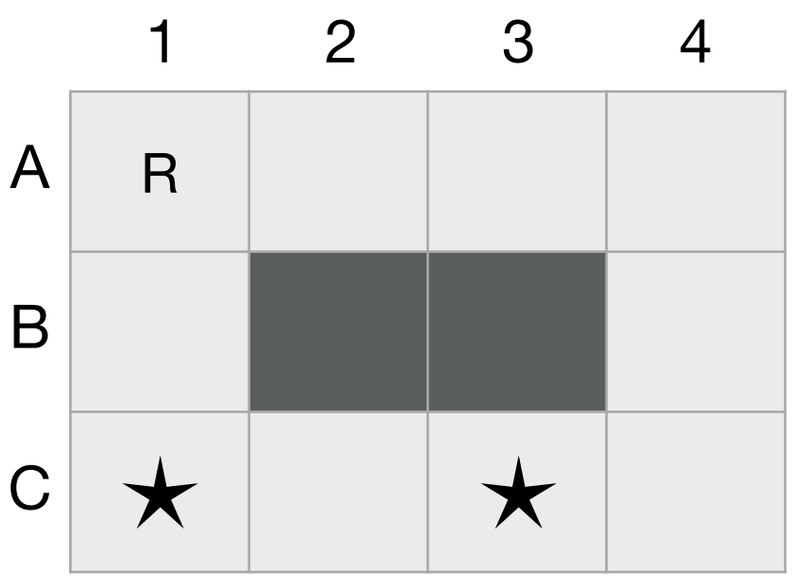
Robot's Model

- **Modeling Planning as a Boolean Satisfiability (SAT) Problem**
  - Goal: Find truth value assignments to sequence of actions such that the goal states are reached
    - For human: $\downarrow_0, \downarrow_1, \rightarrow_2, \rightarrow_3$ results in $C1_2, C3_4$
    - For robot: $\rightarrow_0, \rightarrow_1, \rightarrow_2, \downarrow_3, \downarrow_4, \leftarrow_5, \leftarrow_6, \leftarrow_7$ results in $C1_8, C3_6$
  - Human's question:
    - Why $\rightarrow_0, \rightarrow_1, \rightarrow_2, \downarrow_3, \downarrow_4, \leftarrow_5, \leftarrow_6, \leftarrow_7$ and not $\downarrow_0, \downarrow_1, \rightarrow_2, \rightarrow_3$?

Query q

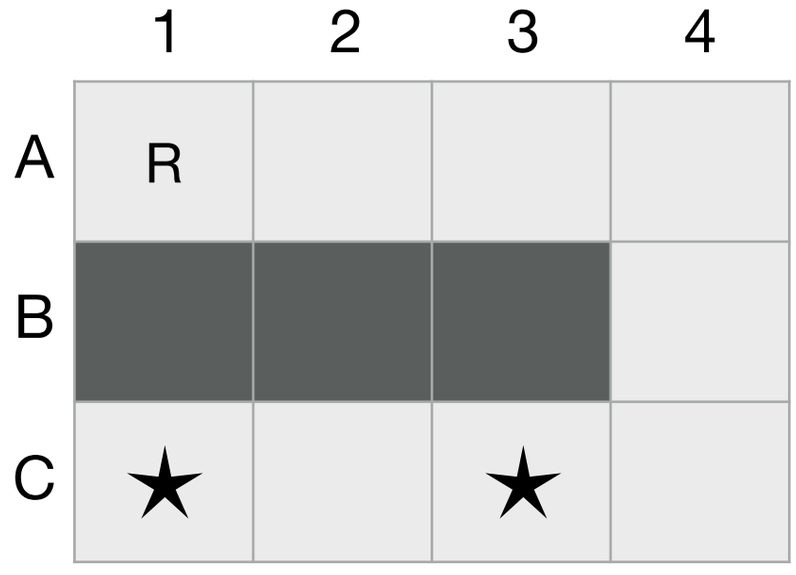
YODAlab

Model Reconciliation with SAT



Human's Model

- Blocked locations: $\neg Blocked\_A1, \neg Blocked\_B1, \ldots$
- Goal states: $C1_t \wedge C3_{t'}$ for some time steps t and t'
- ...

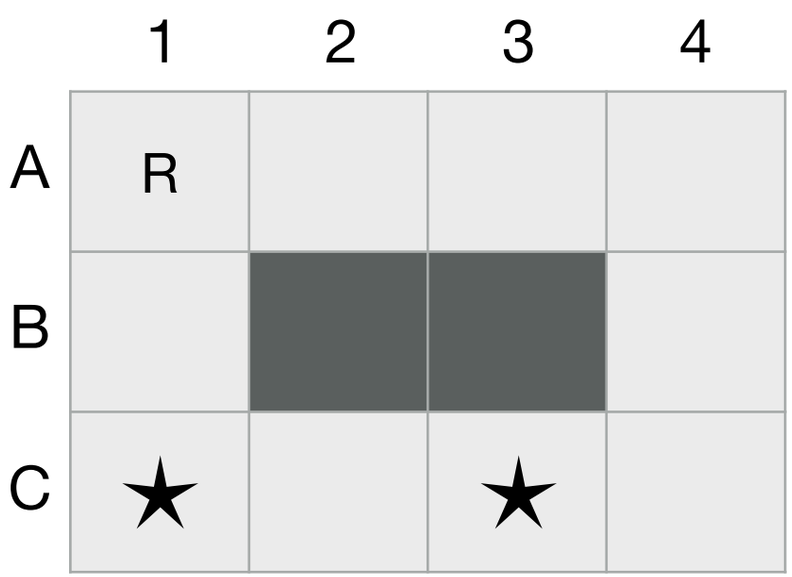


Robot's Model

- Blocked locations: $\neg Blocked\_A1, Blocked\_B1, \ldots$
- Goal states: $C1_t \wedge C3_{t'}$ for some time steps t and t'
- ...

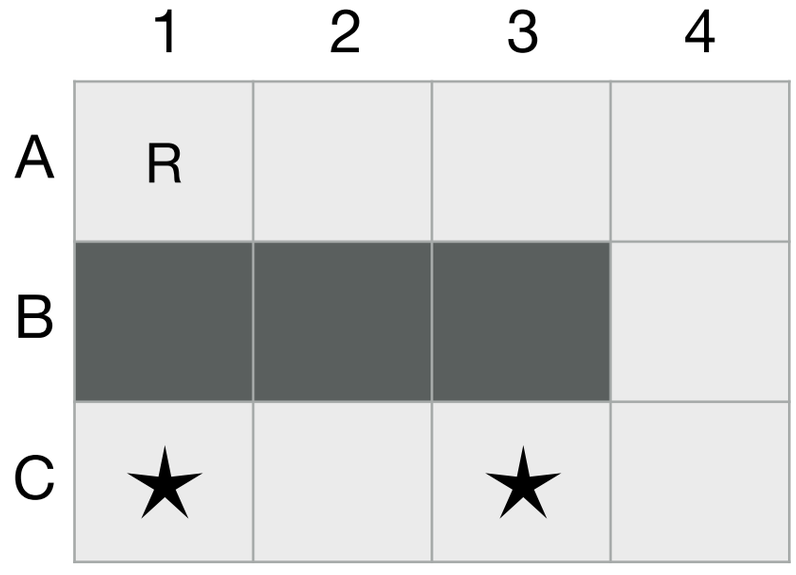
- **Model Reconciliation Problem:**
  - Find $\epsilon^- \subseteq KB_h$ to remove from human's model and $\epsilon^+ \subseteq KB_a$ from agent's model to add to human's model such that the query is entailed

Model Reconciliation with SAT



Human's Model

- Blocked locations: $\neg Blocked\_A1, \neg Blocked\_B1, \ldots$
- Goal states: $C1_t \wedge C3_{t'}$ for some time steps t and t'
- ...



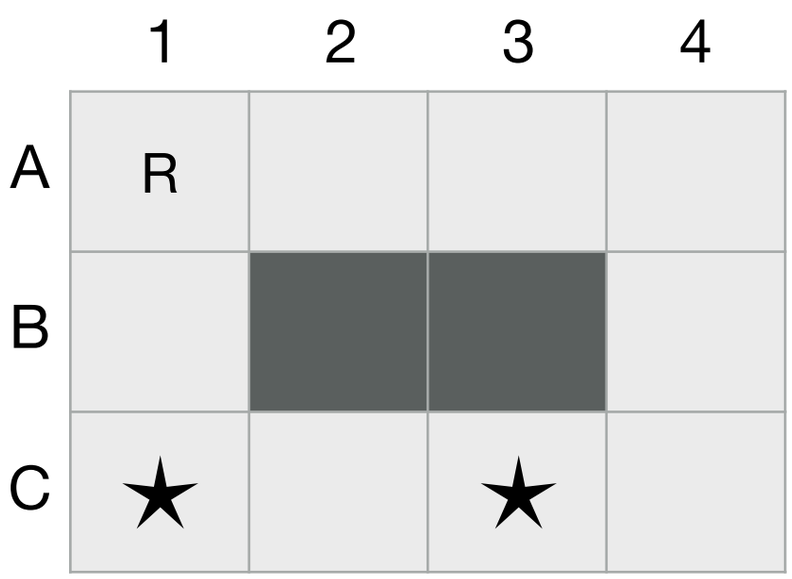
Robot's Model

- Blocked locations: $\neg Blocked\_A1, Blocked\_B1, \ldots$
- Goal states: $C1_t \wedge C3_{t'}$ for some time steps t and t'
- ...

- **Model Reconciliation Problem:**
  - Find $\epsilon^- \subseteq KB_h$ to remove from human's model and $\epsilon^+ \subseteq KB_a$ from agent's model to add to human's model such that the query is entailed
  - $\epsilon^- = \neg Blocked\_B1$ and $\epsilon^+ = Blocked\_B1$

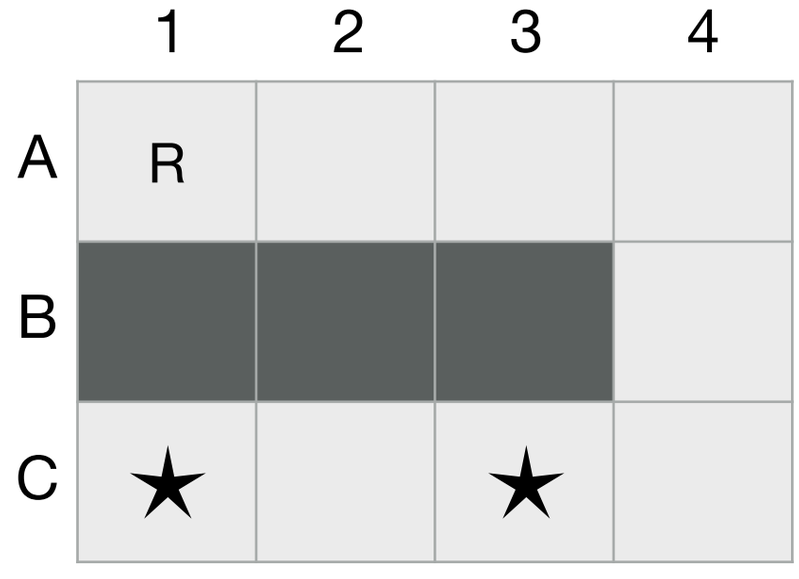
YODAlab

Model Reconciliation with SAT



Human's Model

- Blocked locations: $\neg Blocked\_A1, \neg Blocked\_B1, \ldots$
- Goal states: $C1_t \wedge C3_{t'}$ for some time steps t and t'
- ...



Robot's Model

- Blocked locations: $\neg Blocked\_A1, Blocked\_B1, \ldots$
- Goal states: $C1_t \wedge C3_{t'}$ for some time steps t and t'
- ...

- **Model Reconciliation Problem:**
  - Find $\epsilon^- \subseteq KB_h$ to remove from human's model and $\epsilon^+ \subseteq KB_a$ from agent's model to add to human's model such that the query is entailed
  - Solve using hitting sets [AAAI-21], logic programming [JELIA-21], etc.

Hitting Sets

- **Minimal Unsatisfiable Set (MUS):**
  - Given an inconsistent KB, a subset $U \subseteq KB$ is an MUS if U is unsatisfiable and $\forall U' \subset U$, U' is satisfiable

- **Minimal Correction Set (MCS):**
  - Given an inconsistent KB, a subset $C \subseteq KB$ is an MCS if $KB \backslash C$ is satisfiable and $\forall C' \subset C$, $KB \backslash C'$ is unsatisfiable

Hitting Sets

- **Minimal Unsatisfiable Set (MUS):**
  - Given an inconsistent KB, a subset $U \subseteq KB$ is an MUS if U is unsatisfiable and $\forall U' \subset U$, U' is satisfiable

- **Minimal Correction Set (MCS):**
  - Given an inconsistent KB, a subset $C \subseteq KB$ is an MCS if $KB \backslash C$ is satisfiable and $\forall C' \subset C$, $KB \backslash C'$ is unsatisfiable

- High-level ideas of our approach:
  - Utilize MUS to get ϵ^+ because MUS($KB_a \cup \neg q$) $\backslash \neg q$ is a support for query q in KB_a
  - Utilize MCS to get ϵ^- so that the updated KB_h is consistent
  - In reality: Exploits duality between MUS and MCS through hitting sets to speed up computation as well as allowing for approximations

Model Reconciliation Problem

- **Model Reconciliation Problem:**
  - Given two knowledge bases KB_a and KB_h of the agent (robot) providing an explanation and the human receiving the explanation, respectively, and a query q such that KB_a entails q, the goal is to find an explanation (ϵ^+, ϵ^-), where $\epsilon^+ \subseteq KB_a$ and $\epsilon^- \subseteq KB_h$, such that $KB_h \cup \epsilon^+ \backslash \epsilon^-$ entails q.

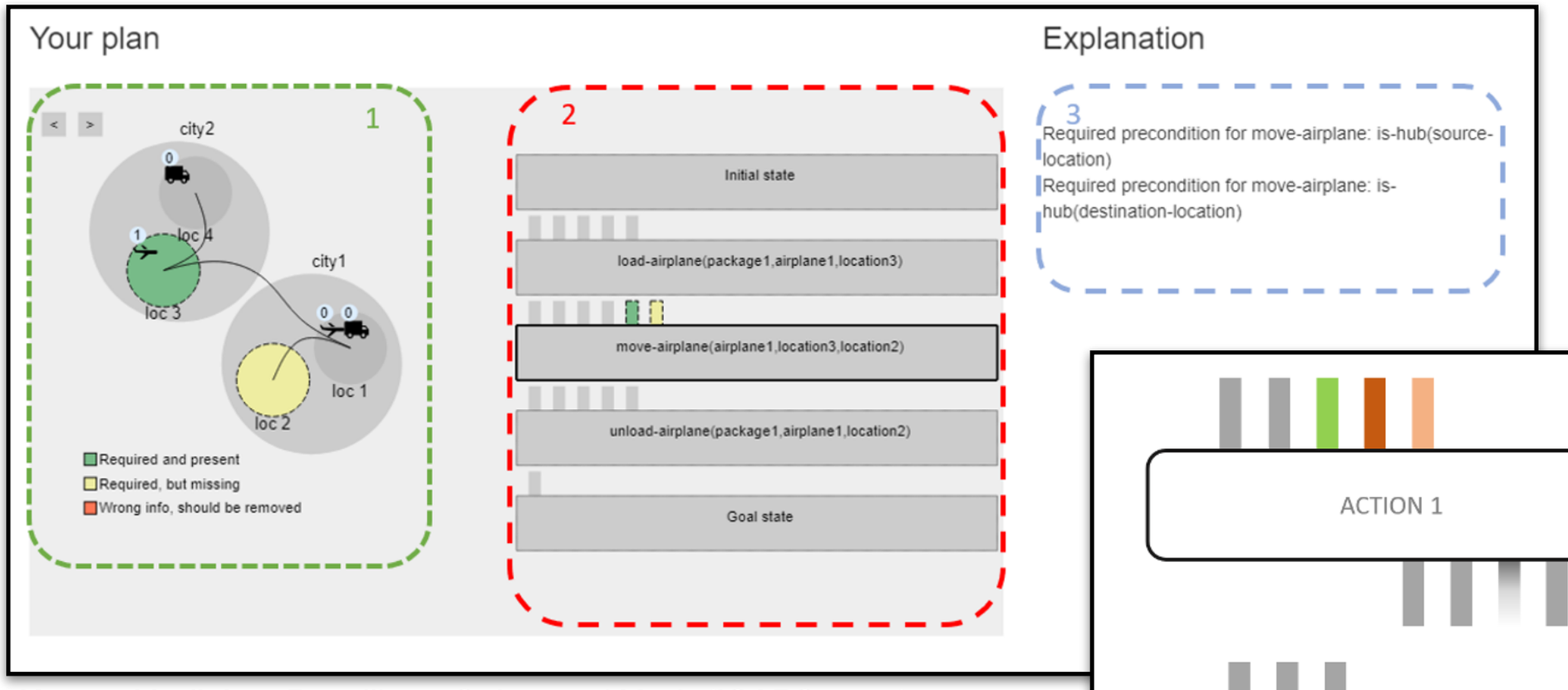
- Research questions:
  - How to effectively find an explanation (ϵ^+, ϵ^-)?
  - How to measure the goodness of explanations?
  - How to convey explanations to the human?
  - What if we don't know the model of the human?
  - What if the model of the agent (robot) is wrong?
  - …

Ongoing and Future Work

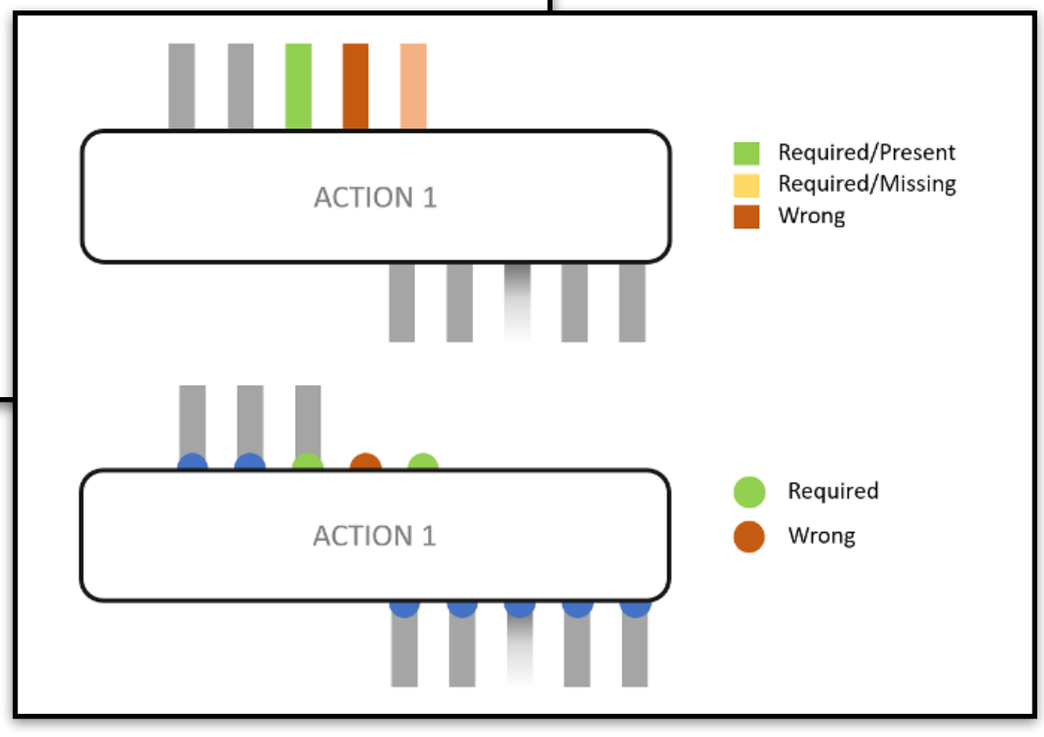
- **Measuring the goodness of an explanation:**
  - How easy it is to convey the explanation
  - How complex is the explanation (e.g., number of literals in the explanation)
  - How long is the explanation (e.g., number of rules that need to be removed/added)
  - How likely is the explanation accepted (e.g., the smaller the number of rules that we need to remove from the human's model, the better)
  - …

Ongoing and Future Work

- **Conveying explanations:** Using visualizations



Kumar, Vasileiou, Bancilhon, Ottley, and Yeoh (ICAPS 2022)



Bryce, Bonasso, Adil, Bell, and Kortenkamp (UISP 2017)

Ongoing and Future Work

- **Conveying explanations:** Argumentation-based dialogues
  - Goal: Human wants to understand why robot thinks c is true

KB_h: $e, e \rightarrow \neg c$; $i, i \rightarrow \neg f$

KB_a: $a, b, a \wedge b \rightarrow c$; $f, f \rightarrow h$; $h, h \rightarrow \neg e$

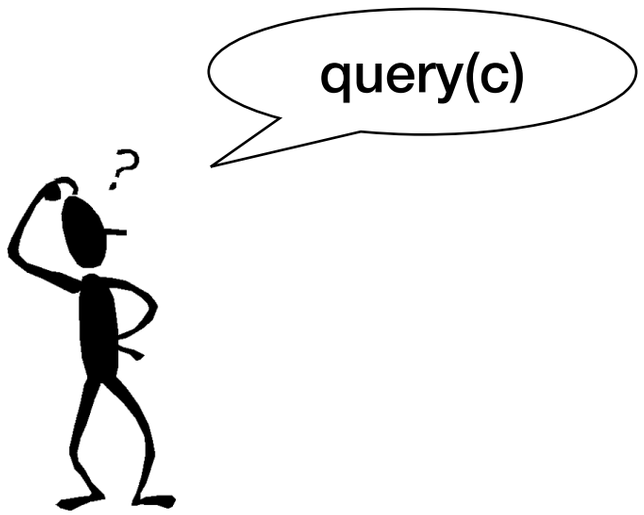
Vasileiou, Kumar, Yeoh, Son, and Toni (arXiv 2023)

Ongoing and Future Work

- **Conveying explanations:** Argumentation-based dialogues
  - Goal: Human wants to understand why robot thinks c is true

KB_h: $e, e \rightarrow \neg c$; $i, i \rightarrow \neg f$

KB_a: $a, b, a \wedge b \rightarrow c$; $f, f \rightarrow h$; $h, h \rightarrow \neg e$



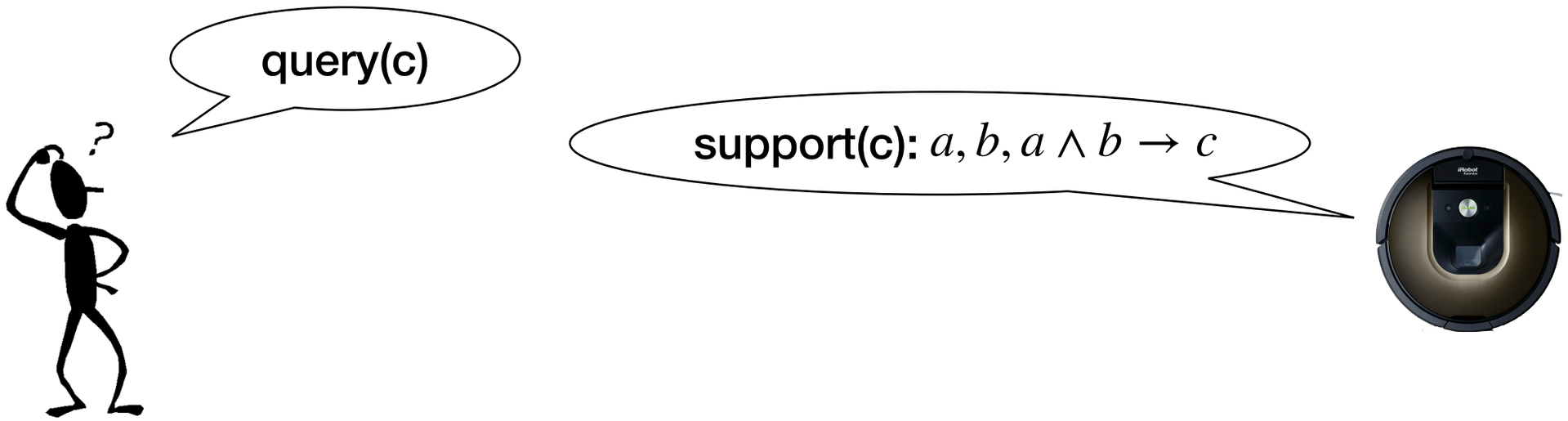
Vasileiou, Kumar, Yeoh, Son, and Toni (arXiv 2023)

Ongoing and Future Work

- **Conveying explanations:** Argumentation-based dialogues
  - Goal: Human wants to understand why robot thinks c is true

KB_h: $e, e \rightarrow \neg c$; $i, i \rightarrow \neg f$

KB_a: $a, b, a \wedge b \rightarrow c$; $f, f \rightarrow h$; $h, h \rightarrow \neg e$



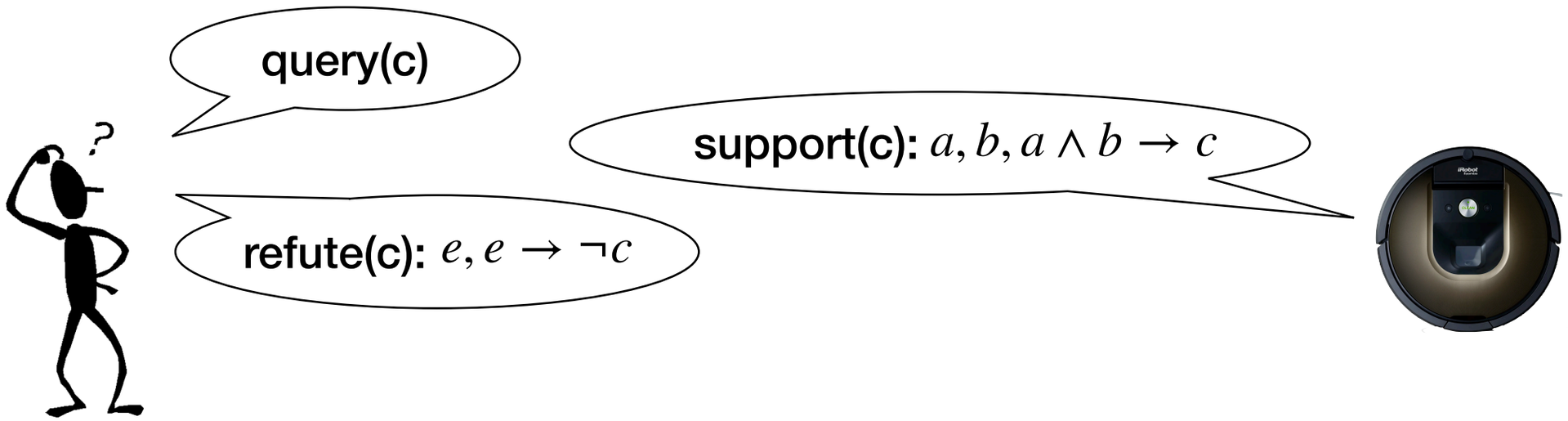
Vasileiou, Kumar, Yeoh, Son, and Toni (arXiv 2023)

Ongoing and Future Work

- **Conveying explanations:** Argumentation-based dialogues
  - Goal: Human wants to understand why robot thinks c is true

KB_h: $e, e \rightarrow \neg c$; $i, i \rightarrow \neg f$

KB_a: $a, b, a \wedge b \rightarrow c$; $f, f \rightarrow h$; $h, h \rightarrow \neg e$



Vasileiou, Kumar, Yeoh, Son, and Toni (arXiv 2023)

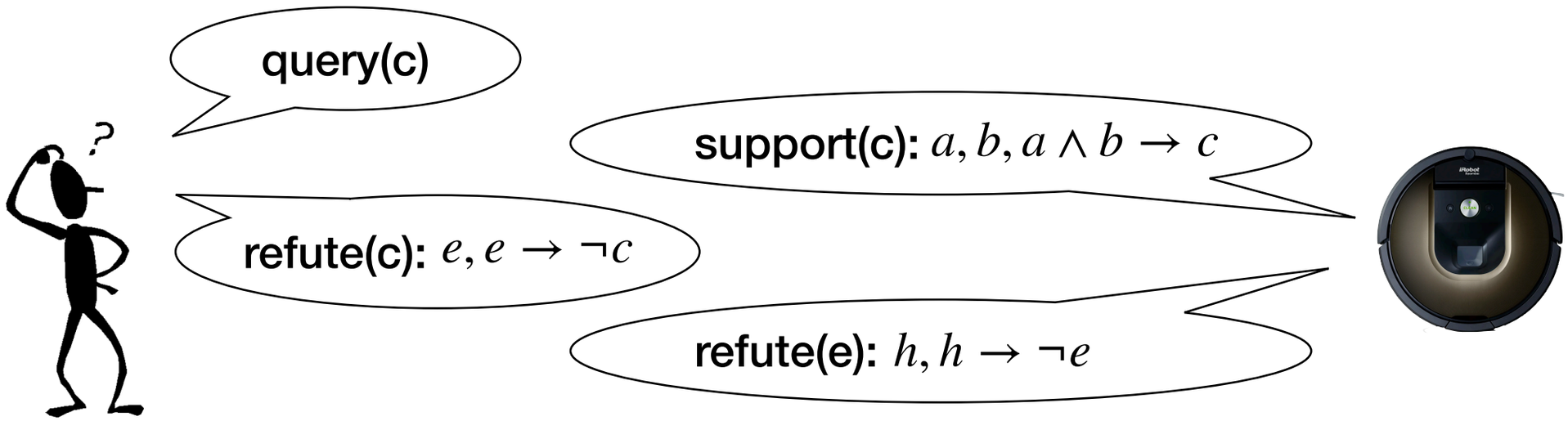
YODAlab

Ongoing and Future Work

- **Conveying explanations:** Argumentation-based dialogues
  - Goal: Human wants to understand why robot thinks c is true

KB_h: $e, e \rightarrow \neg c$; $i, i \rightarrow \neg f$

KB_a: $a, b, a \wedge b \rightarrow c$; $f, f \rightarrow h$; $h, h \rightarrow \neg e$



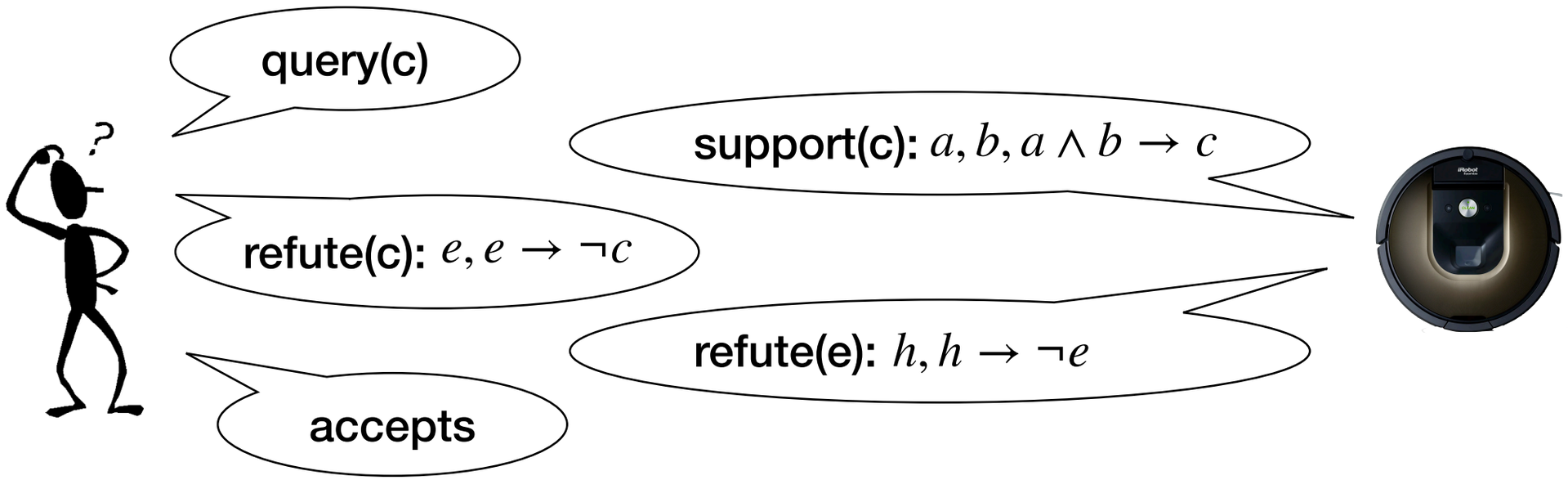
Vasileiou, Kumar, Yeoh, Son, and Toni (arXiv 2023)

Ongoing and Future Work

- **Conveying explanations:** Argumentation-based dialogues
  - Goal: Human wants to understand why robot thinks c is true

KB_h: $e, e \rightarrow \neg c$; $i, i \rightarrow \neg f$

KB_a: $a, b, a \wedge b \rightarrow c$; $f, f \rightarrow h$; $h, h \rightarrow \neg e$



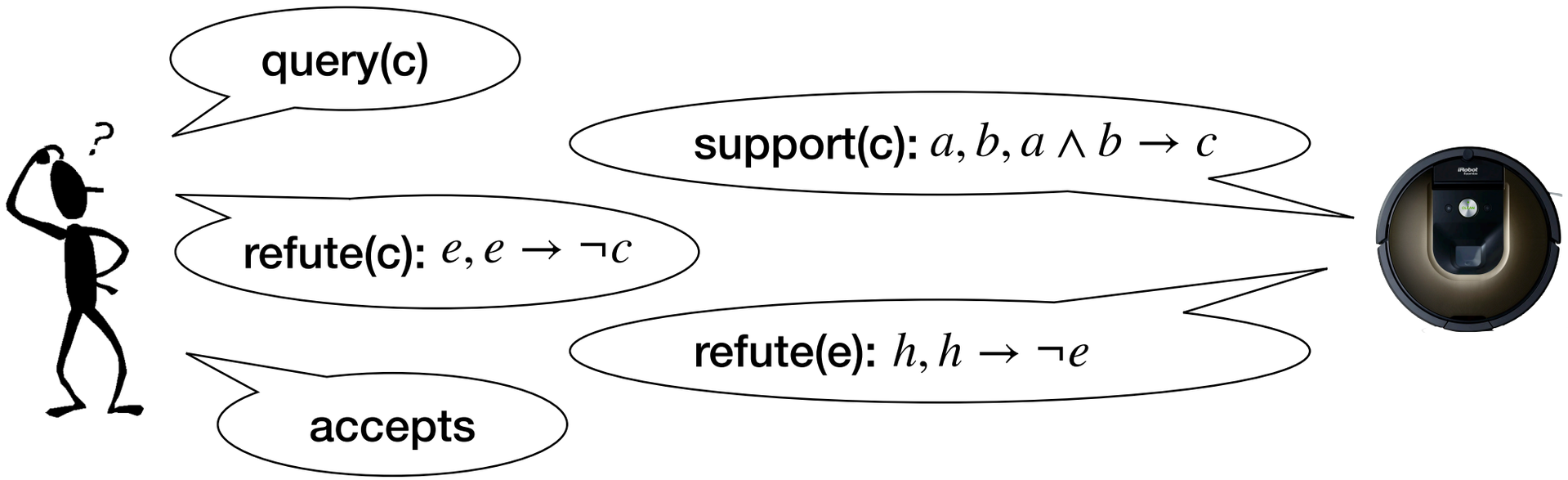
Vasileiou, Kumar, Yeoh, Son, and Toni (arXiv 2023)

Ongoing and Future Work

- **Conveying explanations:** Argumentation-based dialogues
  - Goal: Human wants to understand why robot thinks c is true

KB_h: $a, e \rightarrow \neg c$; $h, h \rightarrow \neg e$; $i, i \rightarrow \neg f$; $a, b, a \wedge b \rightarrow c$

KB_a: $a, b, a \wedge b \rightarrow c$; $f, f \rightarrow h$; $h, h \rightarrow \neg e$



Vasileiou, Kumar, Yeoh, Son, and Toni (arXiv 2023)

Ongoing and Future Work

- **Conveying explanations:** Argumentation-based dialogues
  - Goal: Human wants to understand why robot thinks c is true

KB_h $e, e \rightarrow \neg c$ $h, h \rightarrow \neg e$ $i, i \rightarrow \neg f$ $a, b, a \wedge b \rightarrow c$

$a, b, a \wedge b \rightarrow c$ $f, f \rightarrow h$ $h, h \rightarrow \neg e$ KB_a

- Further research questions:
  - How to learn and approximate humans' model based on interaction?
  - How to model how humans update their models? Does it depend on trust? If so, how to measure trust?
  - How to generalize to handle uncertainty and reason in the presence of uncertainty?
  - How to leverage LLMs for natural language conversations?

Ongoing and Future Work

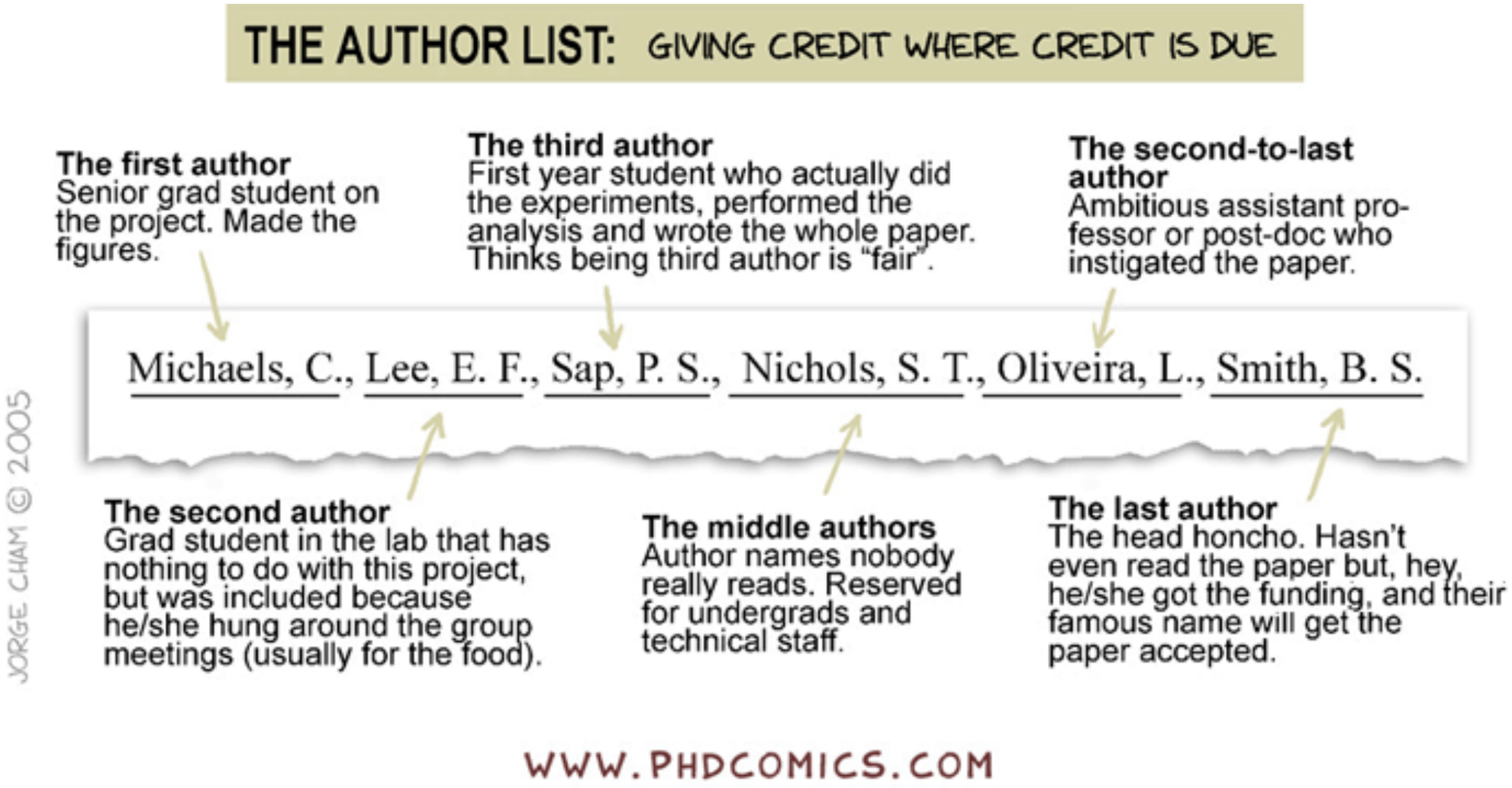
- **Conveying explanations:** Argumentation-based dialogues
  - Goal: Human wants to understand why robot thinks c is true

Medium-term Goal: LLM-like Explainable Agent
Able to provide convincing explanations that are grounded in reality in a natural conversational manner

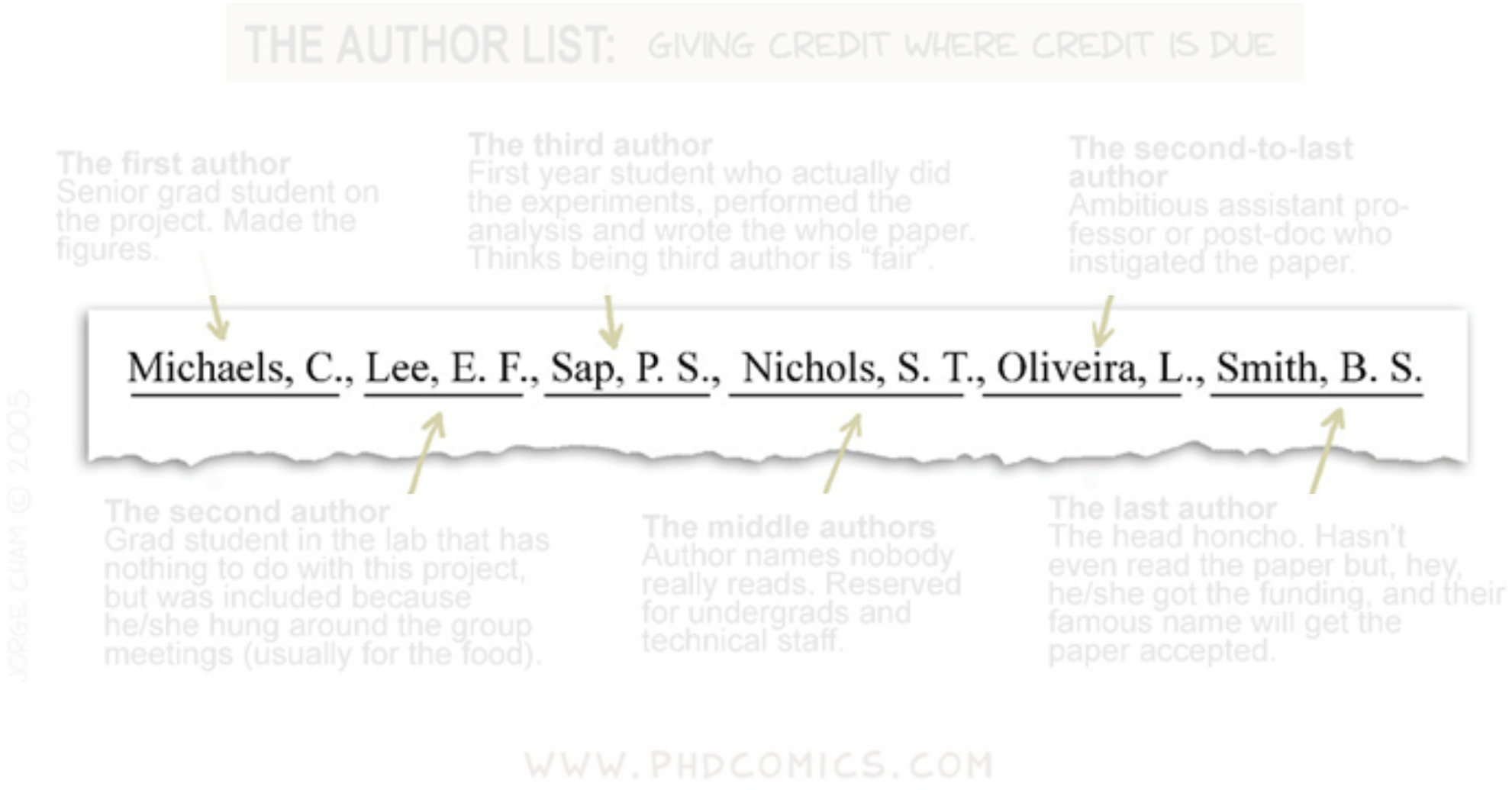
- Further research questions:
  - [illegible]?
  - [illegible]n
  - [illegible] of uncertainty?
  - How to leverage LLMs for natural language conversations?

Ultimate Goal: Collaborative Human-Agent Teams
Enable agents to work with humans to solve problems where neither agents nor humans can solve individually!!

Acknowledgments

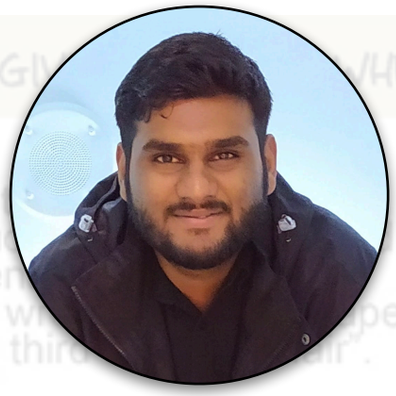


Acknowledgments



Acknowledgments

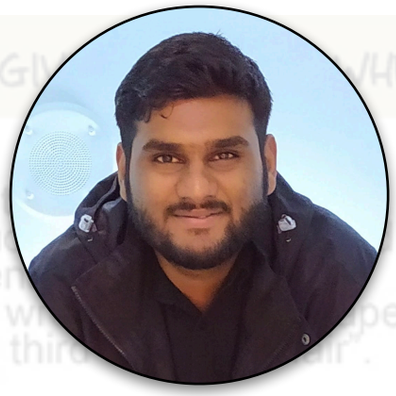
Student Collaborators



Michaels, C., Lee, E. F., Sap, P. S., Nichols, S. T., Oliveira, L., Smith, B. S.

Acknowledgments

Student Collaborators


Michaels, C., Lee, E. F., Sap, P. S., Nichols, S. T., Oliveira, L., Smith, B. S.

Faculty Collaborators