



AAAI 2025 Tutorial T04
Time: 2025-02-25 8:30-12:30

Location: 118A Pennsylvania Convention Center

Foundation Models Meet Embodied Agents



Manling Li
Northwestern



Yunzhu Li
Columbia



Jiayuan Mao
MIT



Wenlong Huang
Stanford



Northwestern
University



COLUMBIA



Stanford
University



AAAI 2025 Tutorial T04
Time: 2025-02-25 8:30-12:30

Location: 118A Pennsylvania Convention Center

Part I: Motivation and Overview

Manling Li, Assistant Professor at Northwestern University

AAAI Tutorial: Foundation Models Meet Embodied Agents



Northwestern
University



COLUMBIA



Stanford
University

What is a generalist agent?

What is a generalist agent?



Having a robot that can do many tasks, across many environments.

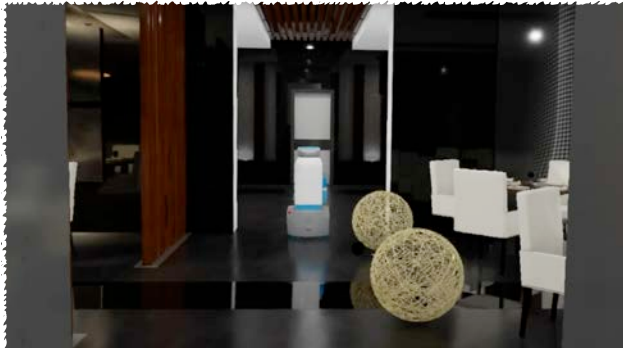
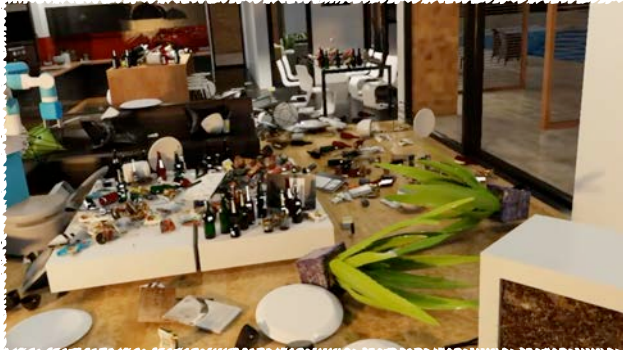
BEHAVIOR-1K

A 3D rendered scene of a messy living room. In the foreground, a white coffee table is cluttered with various items including bottles, plates, and food. The floor is covered with scattered debris like papers, bottles, and a large green plant. In the background, there are white chairs and a table. On the left side, a blue and white robot arm is visible, reaching towards the scene. The overall scene is brightly lit, suggesting a window or large light source.

simulating and benchmarking robot tasks that **matter** to humans

<https://behavior.stanford.edu/>

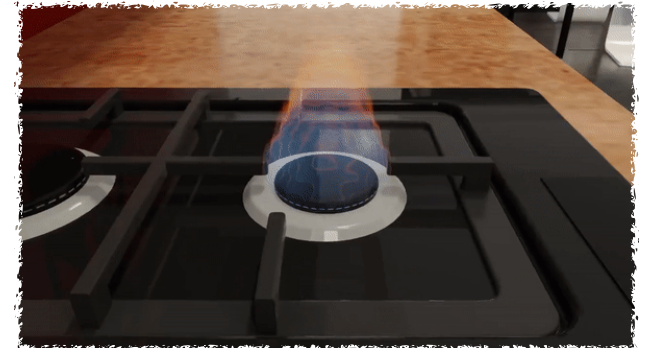
Observation



State: 3D assets & states



Transition Model



tasks that **matter**

What would you like a robot to help you with?



Cleaning the floor?

Images Generated by DALL-E 3

tasks that **matter**

What would you like a robot to help you with?



Folding Laundry?

Images Generated by DALL-E 3

<https://behavior.stanford.edu/>

tasks that **matter**

What would you like a robot to help you with?

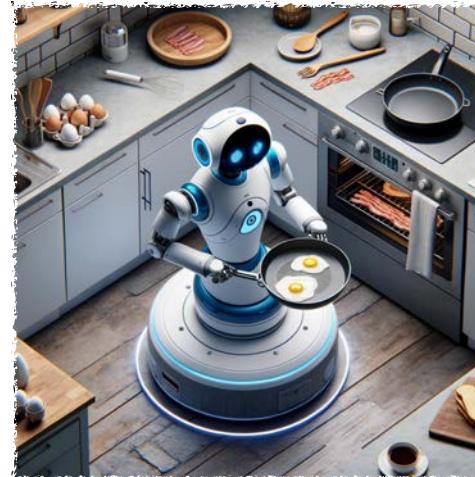


Cooking Breakfast?

Images Generated by DALL-E 3

tasks that **matter**

What would you like a robot to help you with?

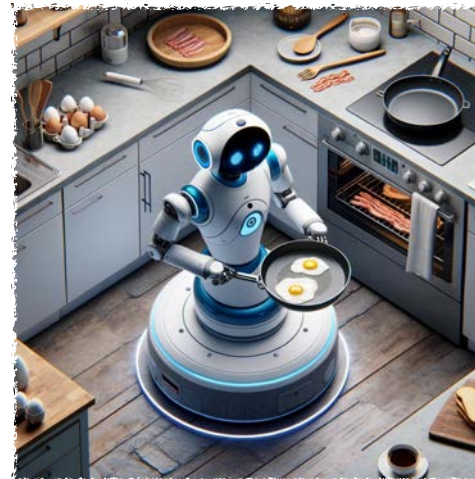
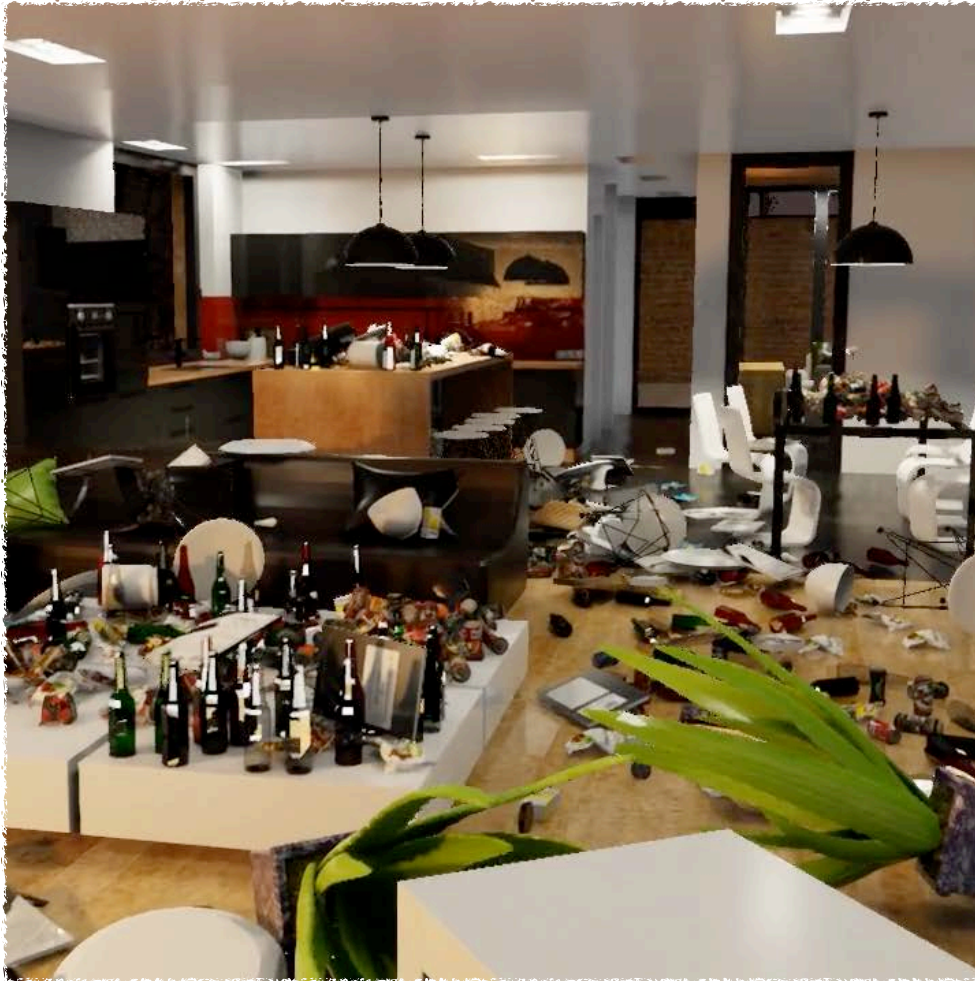


Opening gifts?

Images Generated by DALL-E 3

tasks that **matter**

What would you like a robot to help you with?



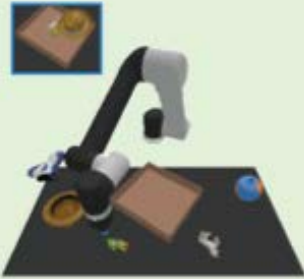
Cleaning after a wild party?

Images Generated by DALL-E 3

Robotic Manipulation



Human: pack the yoshi figure in the brown box



LEO:

$$P = [0.1, -0.2, 0]$$

$$R = [0, 0, 0, 1]$$

Task Planning



Human: Organize and tidy up the bedroom.

LEO:

1. *Clean the floor by sweeping to remove ...*
2. *Make the bed ...*
3. *Fold and organize ...*

Embodied Navigation



Human: Find the pillow

LEO:

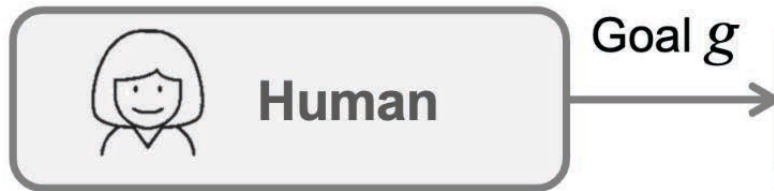
1. *Move forward*
2. *Turn right*
3. *Move forward*
4. ...



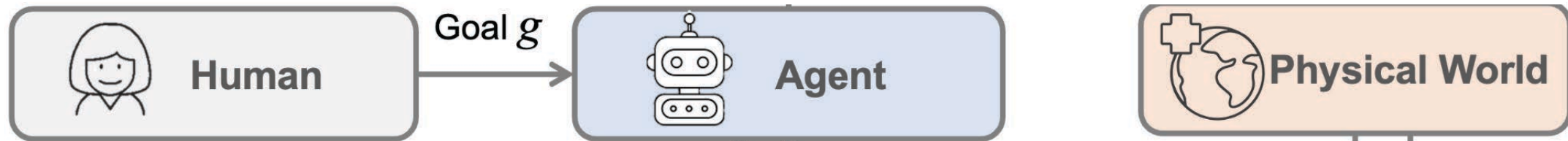
What is “embodied decision making”?

Can we leverage MDP as a guiding principle to categorize “foundation models”?

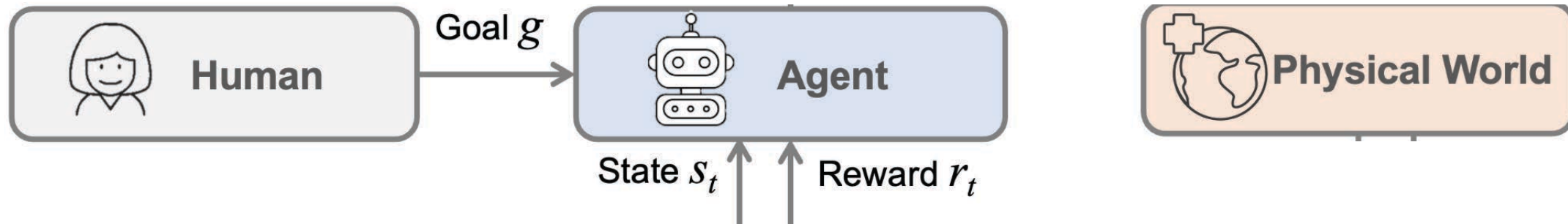
Let us go back to MDPs (Markov Decision Processes)



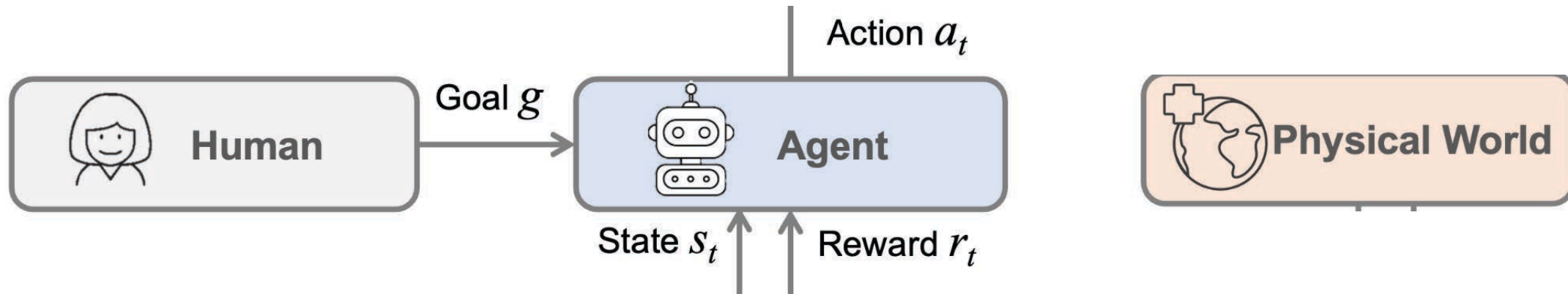
Let us go back to MDPs (Markov Decision Processes)



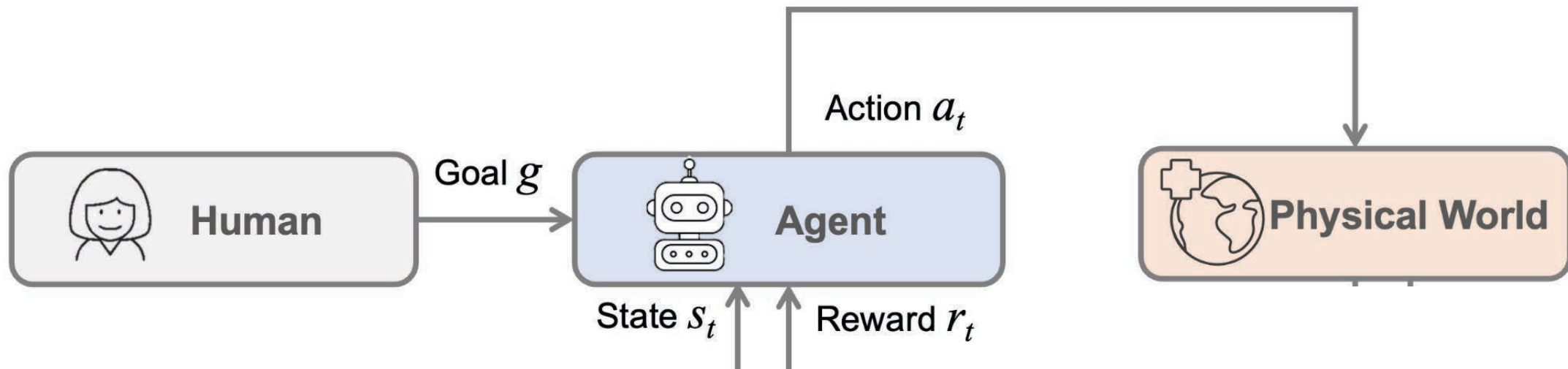
Let us go back to MDPs (Markov Decision Processes)



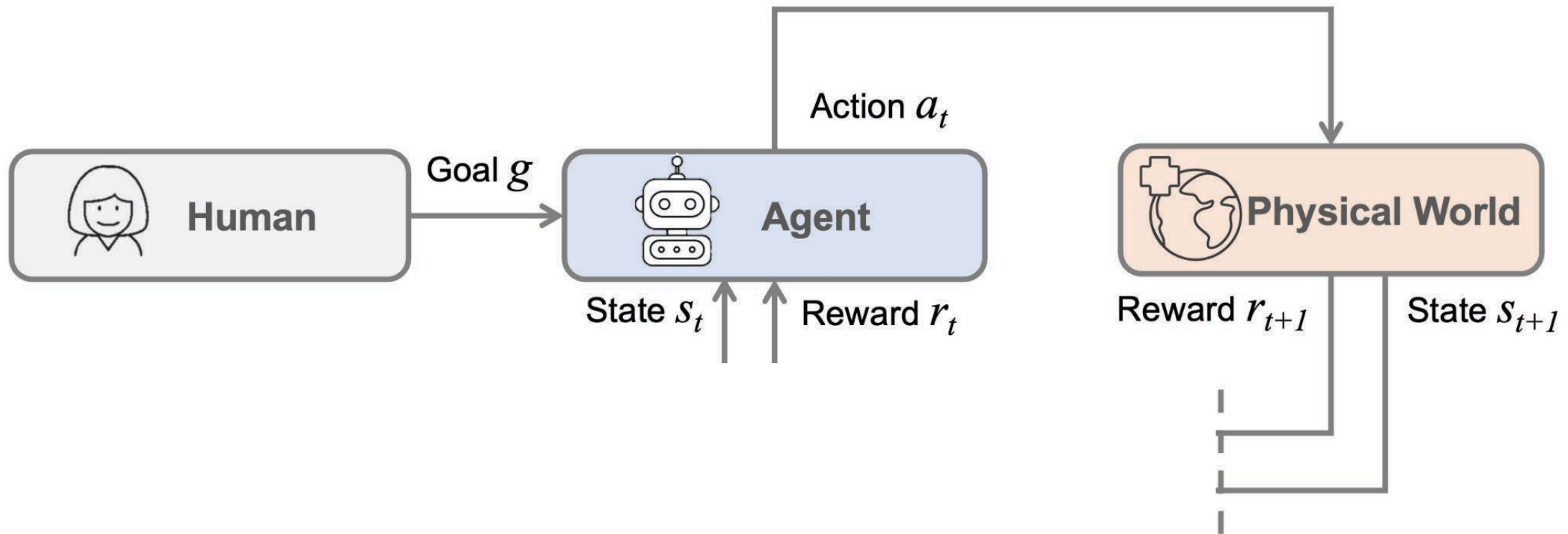
Let us go back to MDPs (Markov Decision Processes)



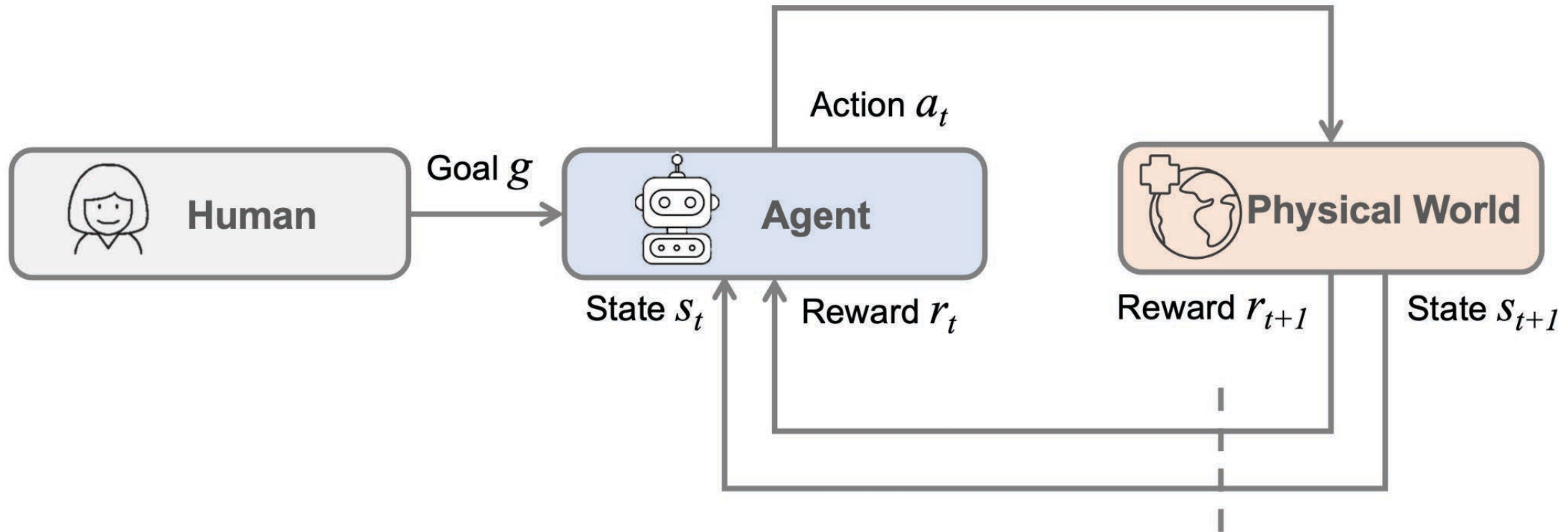
Let us go back to MDPs (Markov Decision Processes)



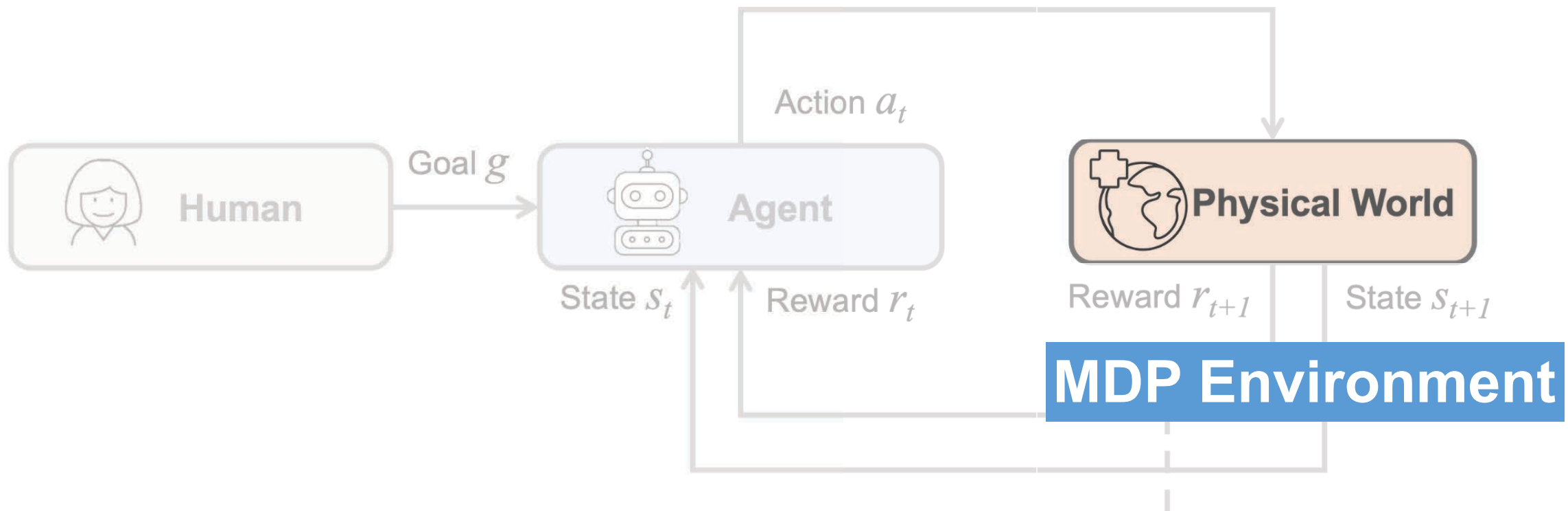
Let us go back to MDPs (Markov Decision Processes)



Let us go back to MDPs (Markov Decision Processes)



Let us go back to MDPs (Markov Decision Processes)



Open-ended Environments

Craft Glass Bridge



Build Oak House



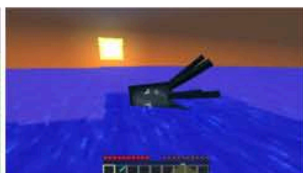
Make Ice Igloo



Combat Zombie



Fish Squid



Farm Sugar Cane



Find Ocean Monument



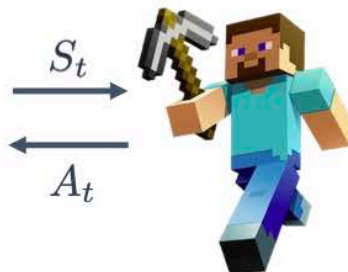
Explore Desert Temple



Treasure Hunt in End City



Generalist Agent



Internet-scale Knowledge Base

YouTube

Wiki

Features	Description	Screenshot	[hide]
<p>Water, Sand, Clay, Sugar Cane, Seagrass, Salmon, Squid, Drowned</p>	<p>Temperature: 0.5. Rainfall: 0.5. A biome that consists of water blocks in an elongated, curving shape similar to a real river. Rivers are a reliable source of clay. They are good for fishing, but drowned can spawn at night.</p>	<p>River</p>	

Reddit

me always bringing blocks to complete the staircase

190

r/Minecraft · Posted by u/Anime-ghostGirl 6 days ago

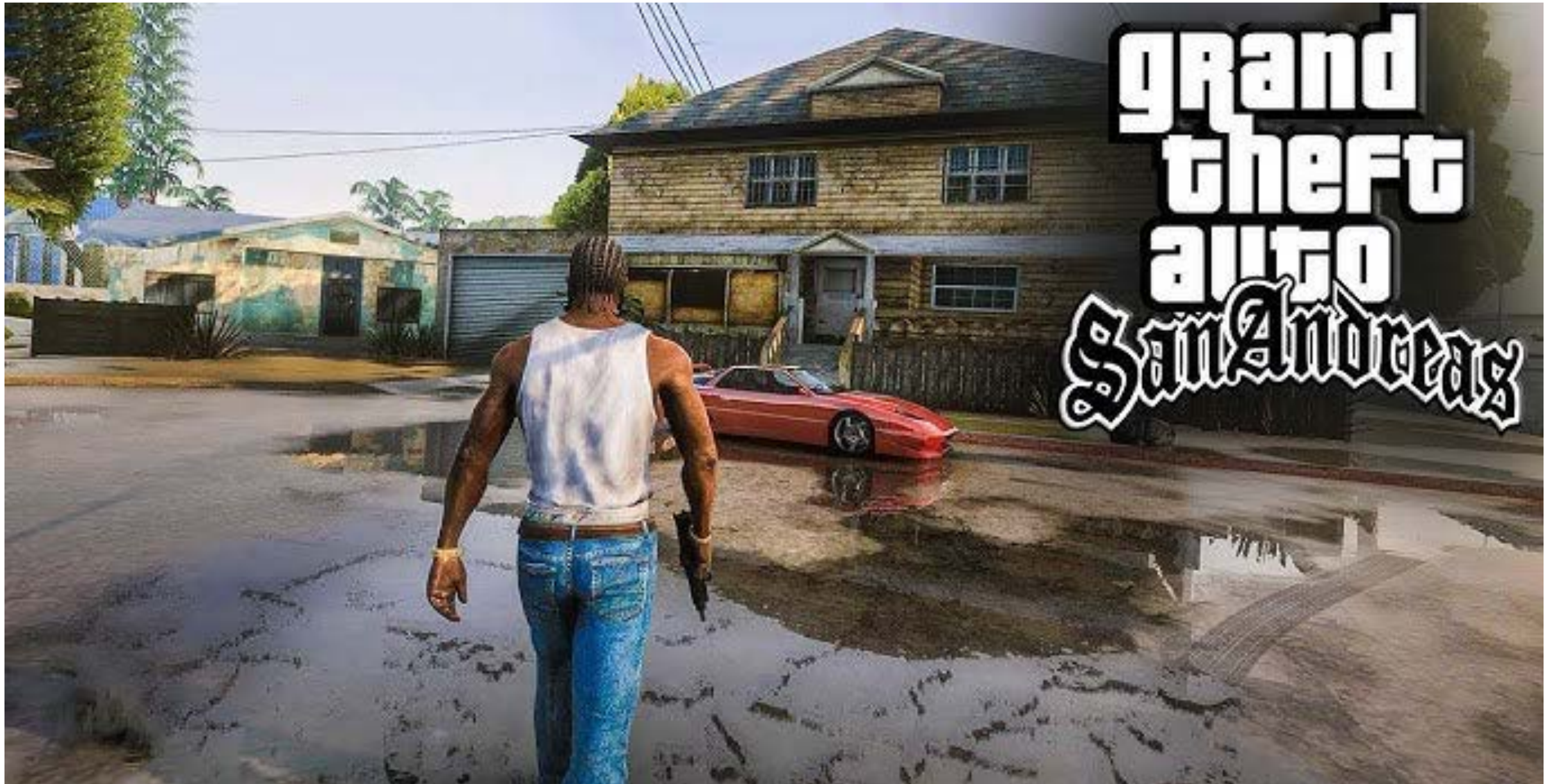
I present to you me struggling to get up stairs in the end city

i dig a staircase in the wall ^^

Or just use enderpearl.

Water is useful in a lot of situations. Early game, and late game

MineDojo



Heading OCR

Question: Tell me the heading text of this screenshot of webpage.
Answer: Discover, Appreciate, & Understand the Animal World!

Captioning

Question: What is the meta description of this website?
Answer: The world's largest & most trusted collection of animal facts, pictures and more!

WebQA

Question: What additional platform is mentioned for following the website's content?
Answer: YouTube Channel

VisualWebBench

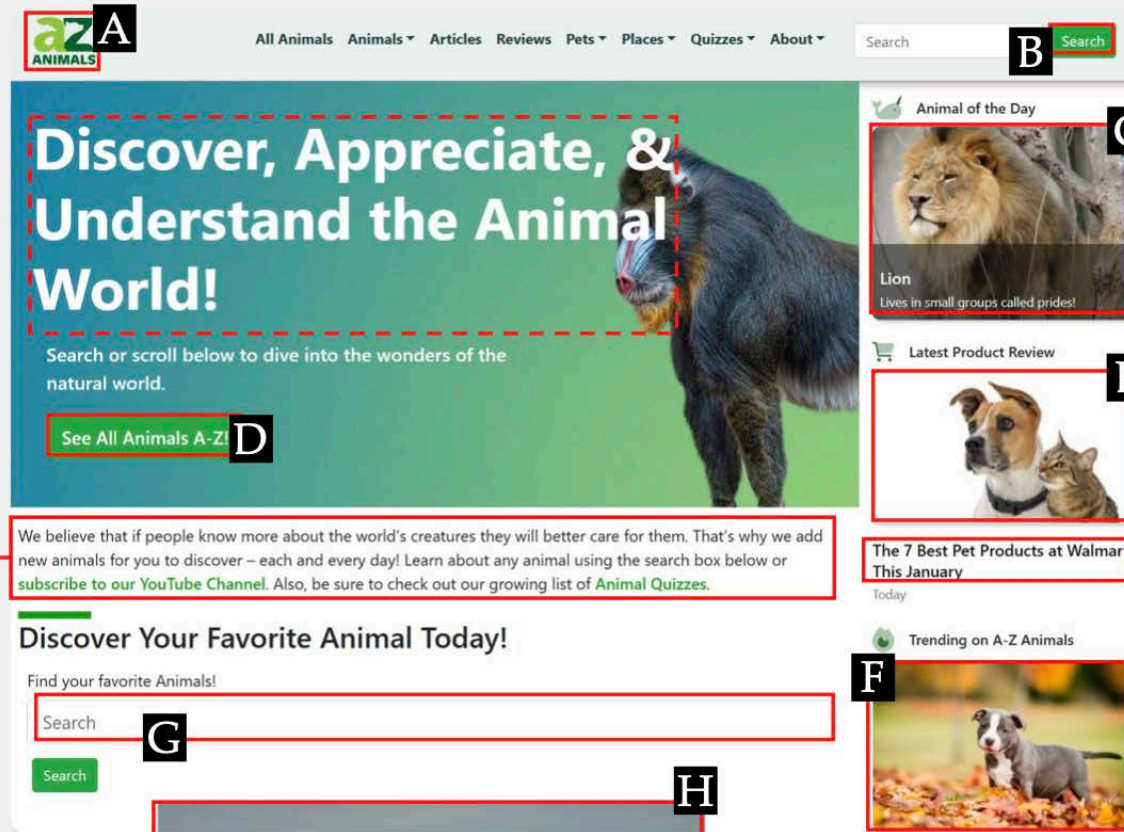
- Website-wise Task
- Element-wise Task
- Action-wise Task

Element OCR

Question: Tell me the text content in the red bounding box
Answer: We believe that if people know about the world's creatures they will better care for them. That's why we add new animals for you to discover ...

Element Grounding

Question: I have labeled bright IDs for some HTML elements in this website screenshot. Tell me which one is the element corresponding to the description: button with text "See All Animals A-Z!"
Answer: D



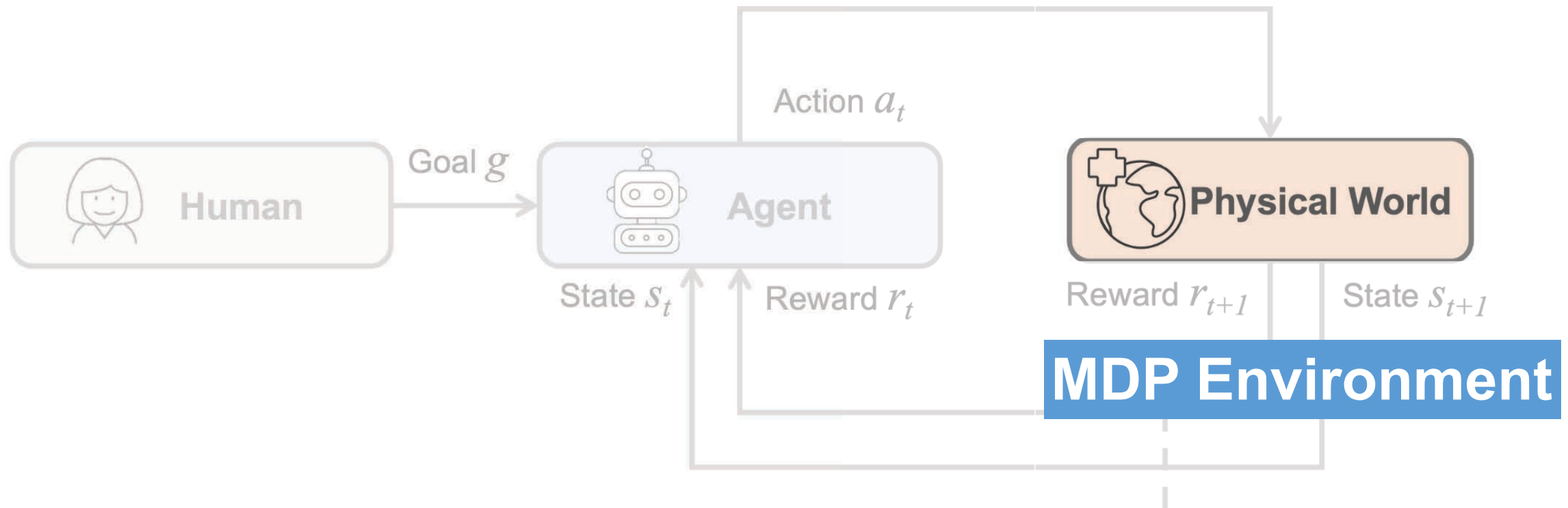
Action Grounding

Question: I have labeled bright IDs for some HTML elements in this website screenshot. Tell me which one I should click to complete the instruction: learn about the animal of the day
Answer: C

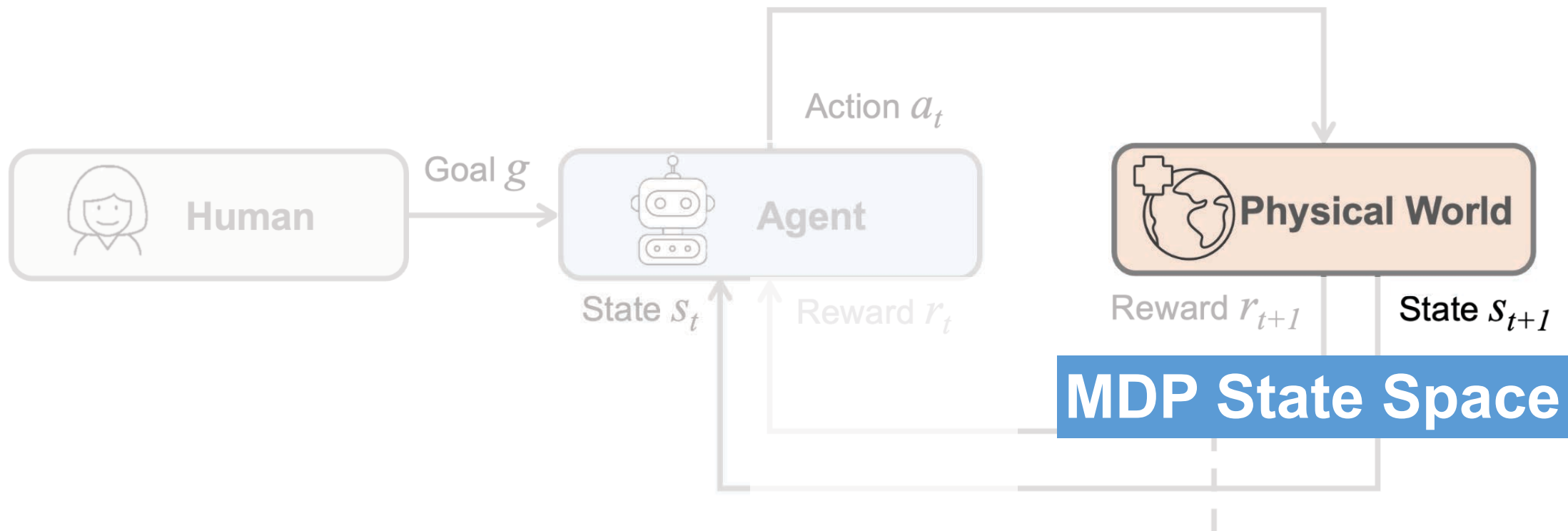
Action Prediction

Question: After clicking the element in the bounding box, which one is the best description of the new webpage?
 (A) Animal news, facts, ...
 (B) All animals A-Z List
 (C) The 7 best pet ...
 (D) Search any animals!
Answer: C

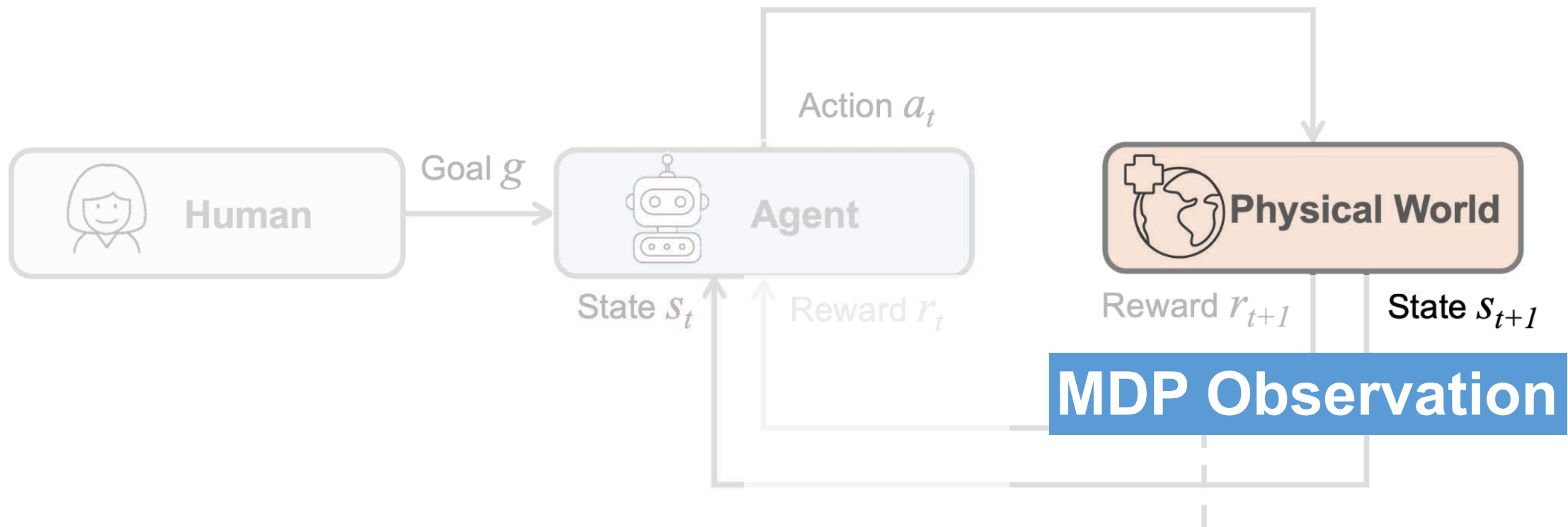
Let us go back to MDPs (Markov Decision Processes)



Let us go back to MDPs (Markov Decision Processes)



Let us go back to MDPs (Markov Decision Processes)



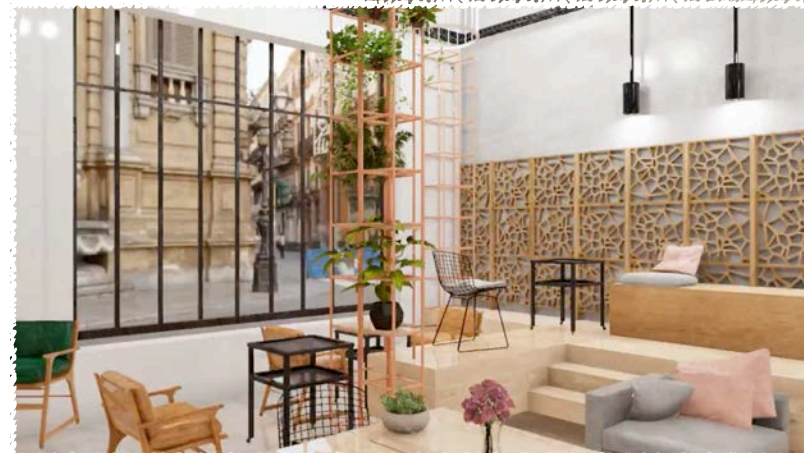


Environment : Observation (Rendered 2D Images)

Enabled by large dataset of realistic interactive **scenes** and objects



50 Scenes





Environment : Observation → State (3D Assets & States)

50 Scenes

Enabled by large dataset of realistic interactive scenes and **objects**



10000 Objects



Semantic



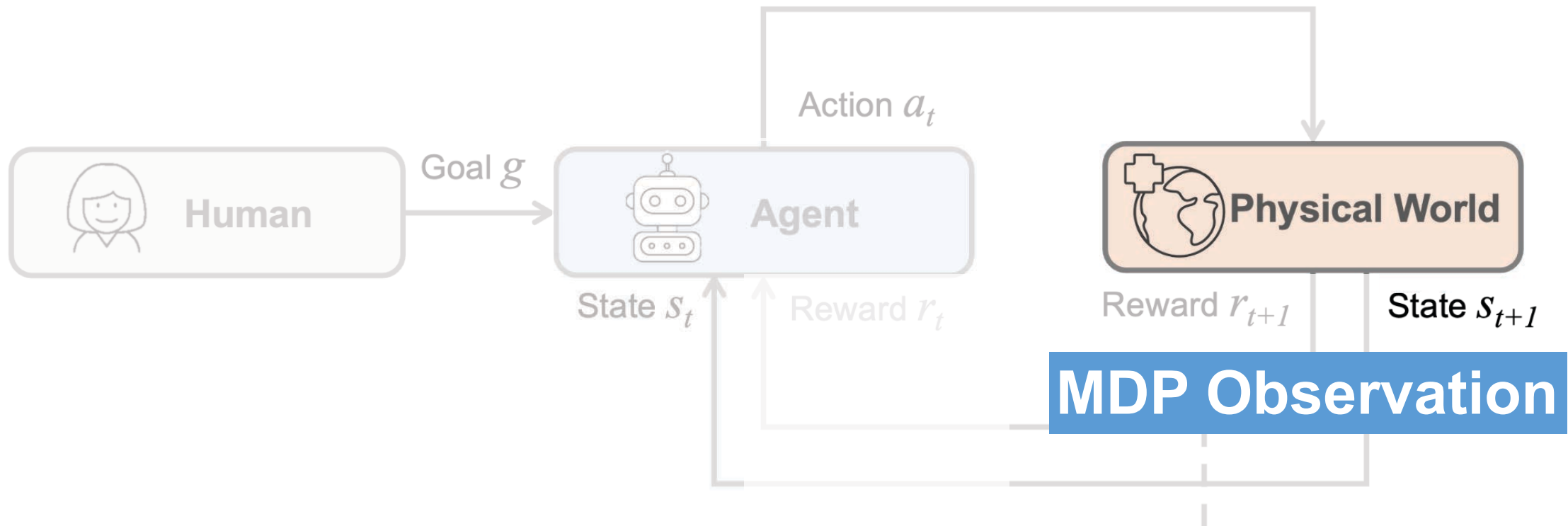
Properties: cookable, sliceable,
freezable, burnable, deformable
...
Cooking temperature: 58°C
...

Physical

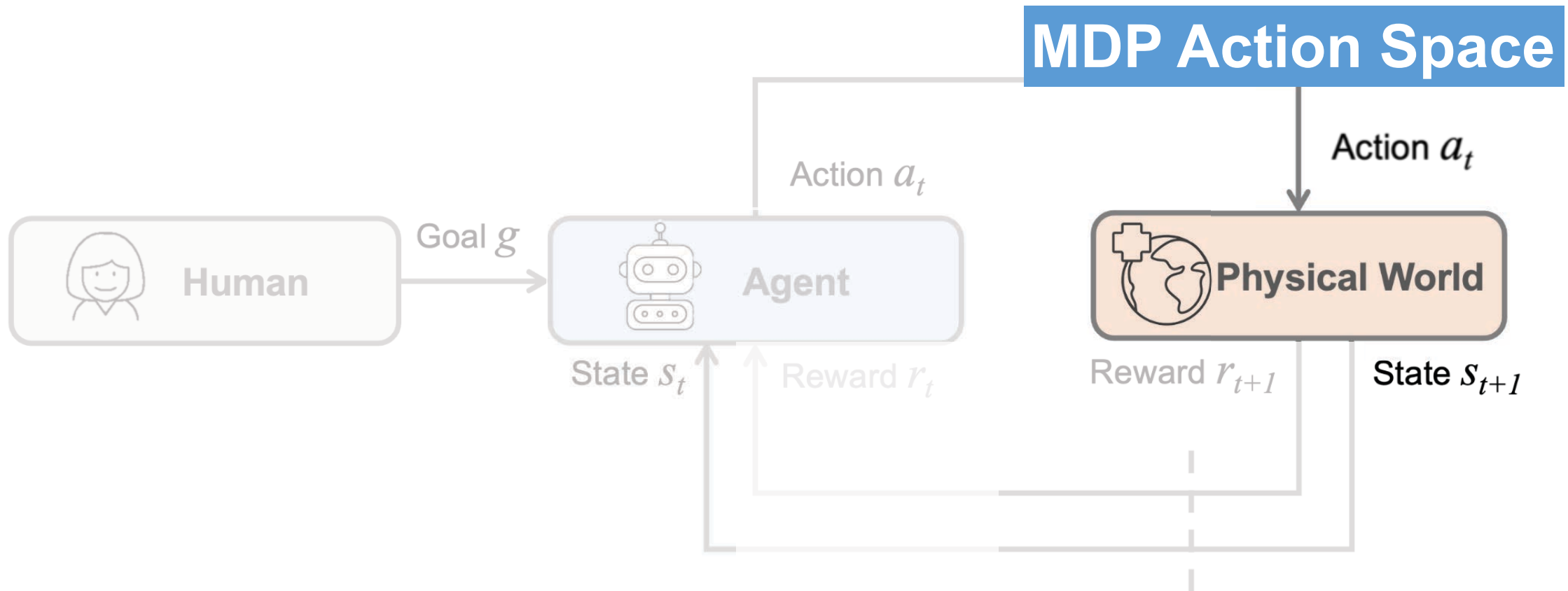


Articulation annotation
(joint type, origin, axis, limit)
Mass, friction, CoM, ...
Canonical size and orientation

Let us go back to MDPs (Markov Decision Processes)



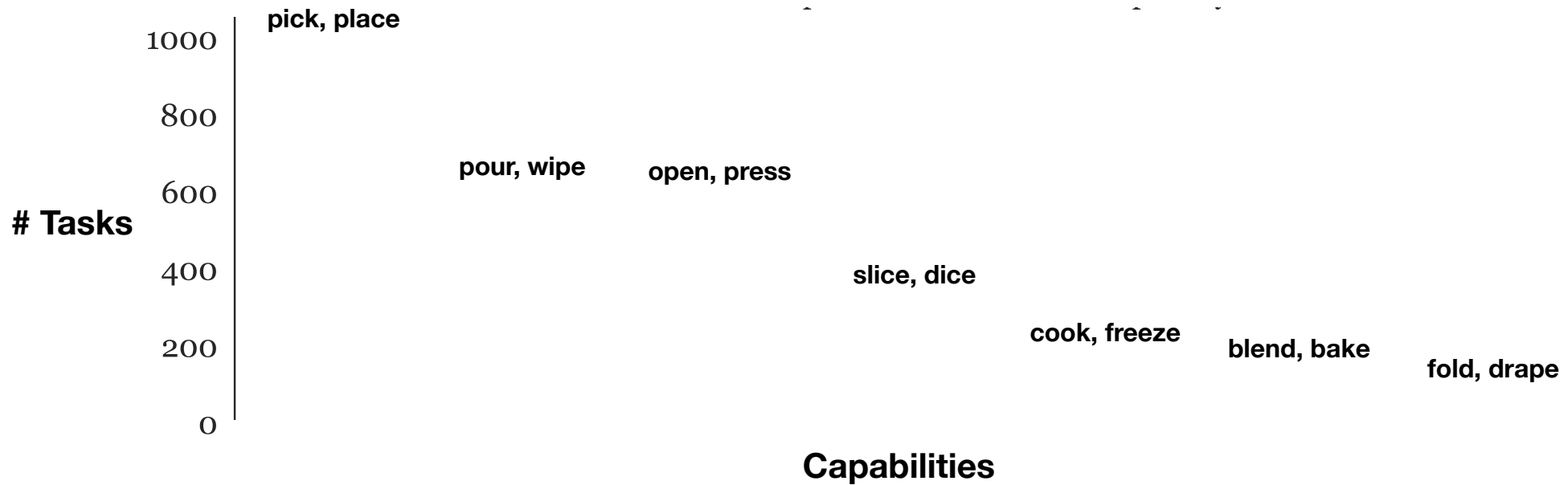
Let us go back to MDPs (Markov Decision Processes)

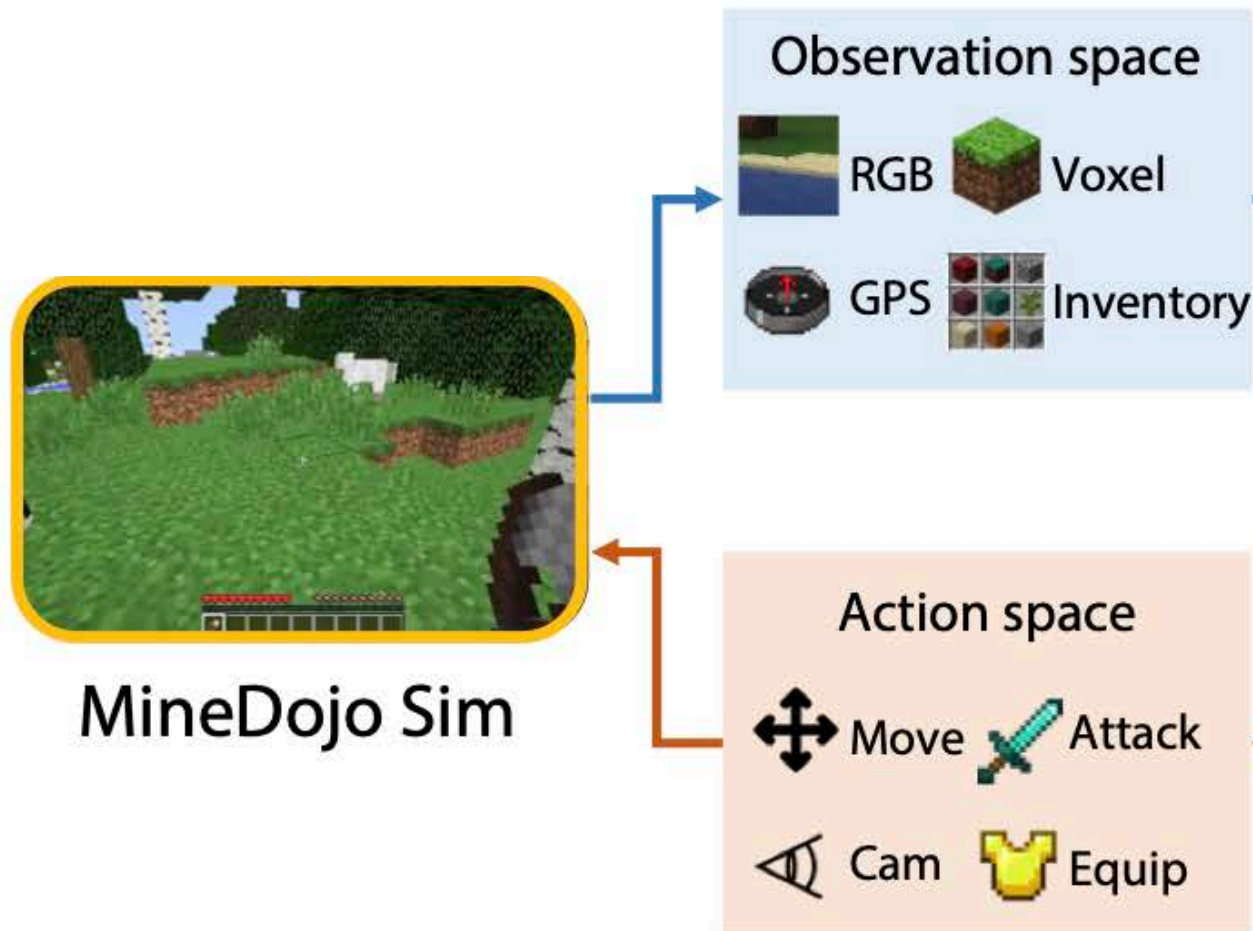




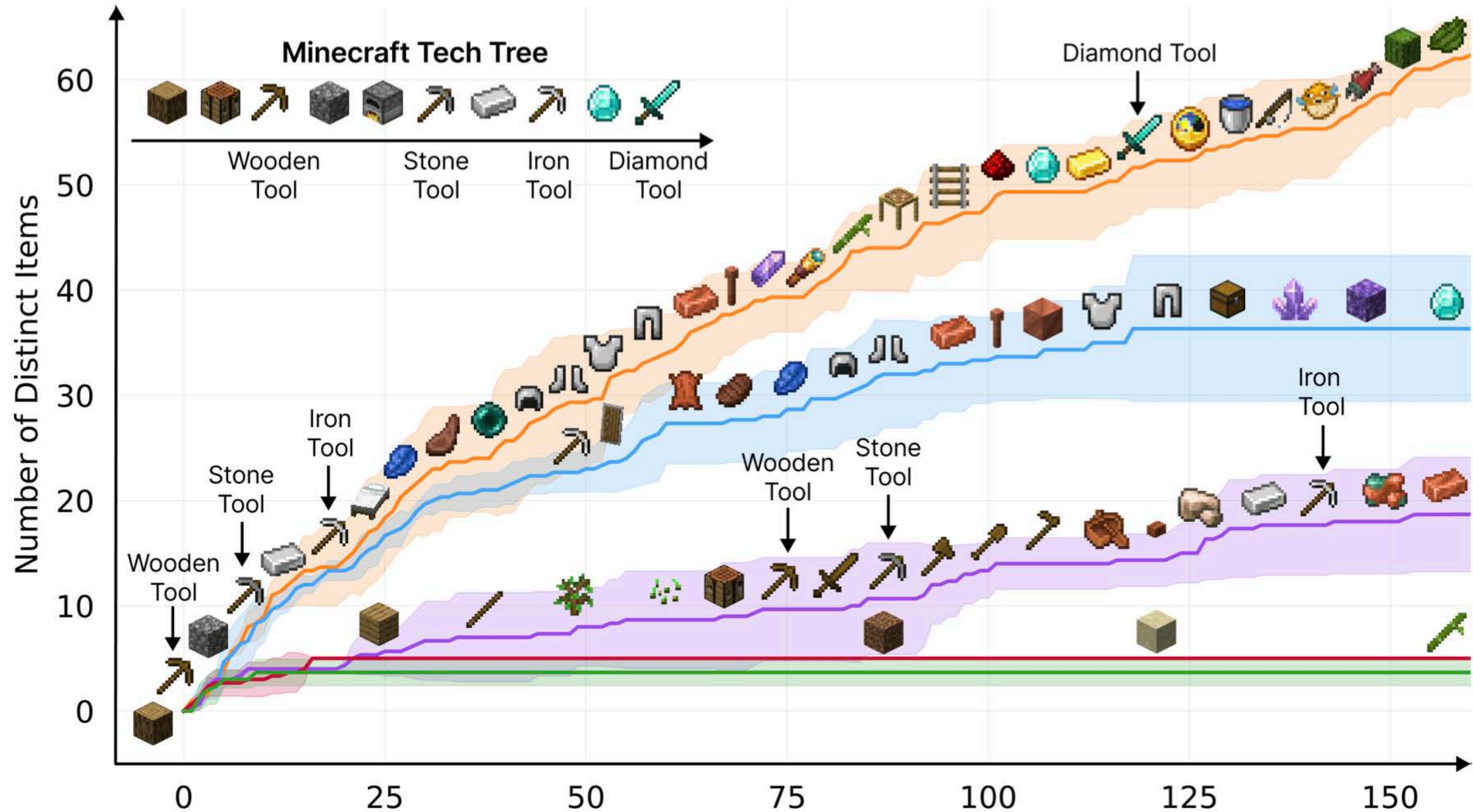
Action : can robots learn to **solve** these tasks?

→ What **capabilities** are needed?

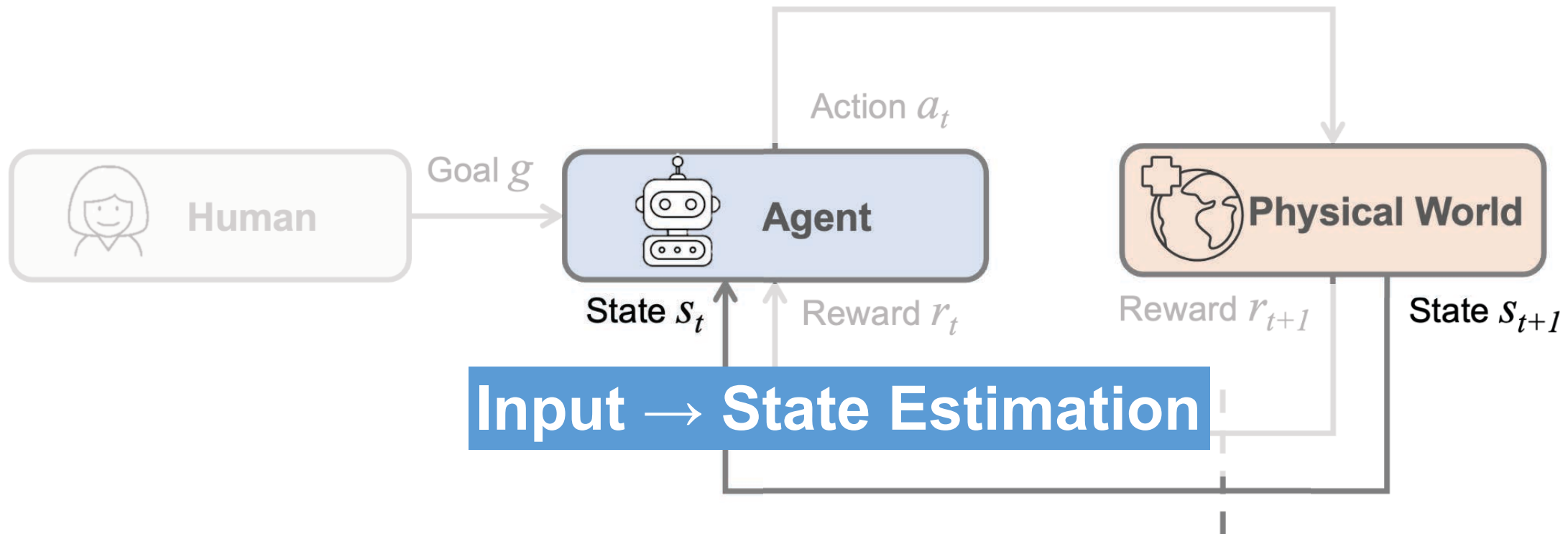




MDP Action Space: Skills



Let us go back to MDPs (Markov Decision Processes)



Perception / State Estimation

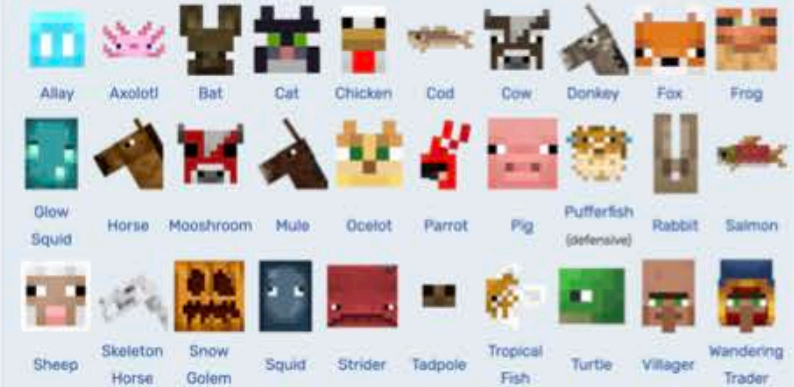
$$o \rightarrow s$$

Observation (2D rendered scenes)



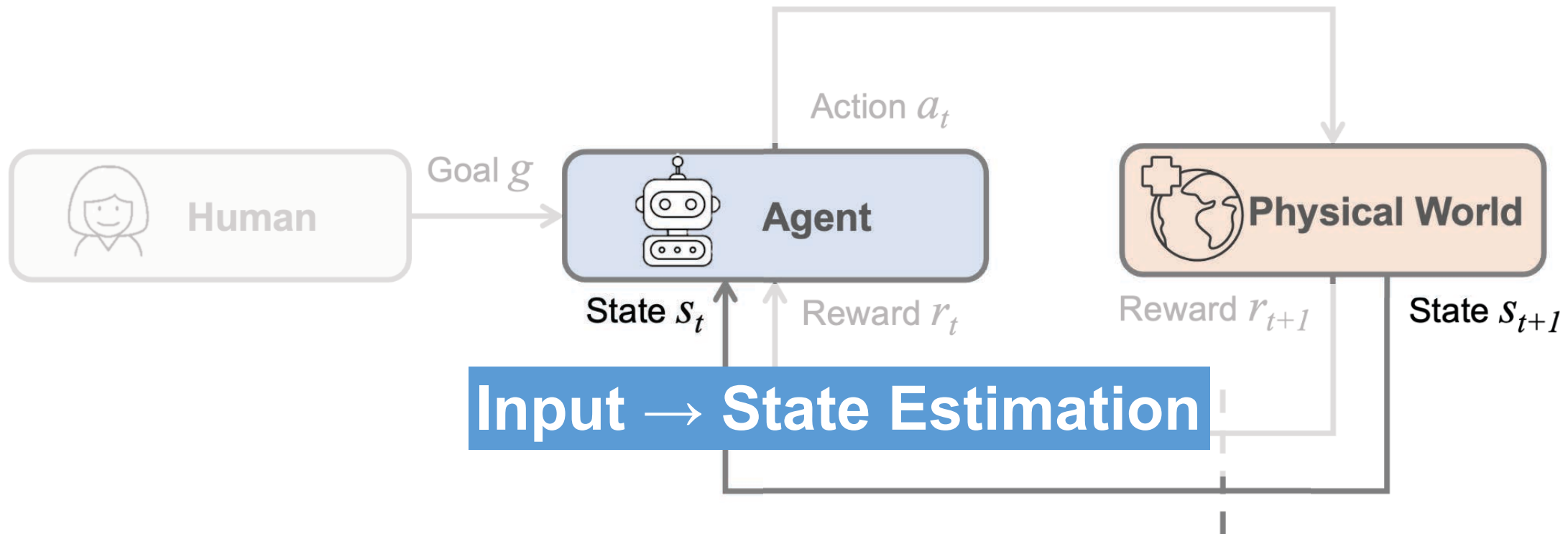
State (3D assets)

Passive mobs

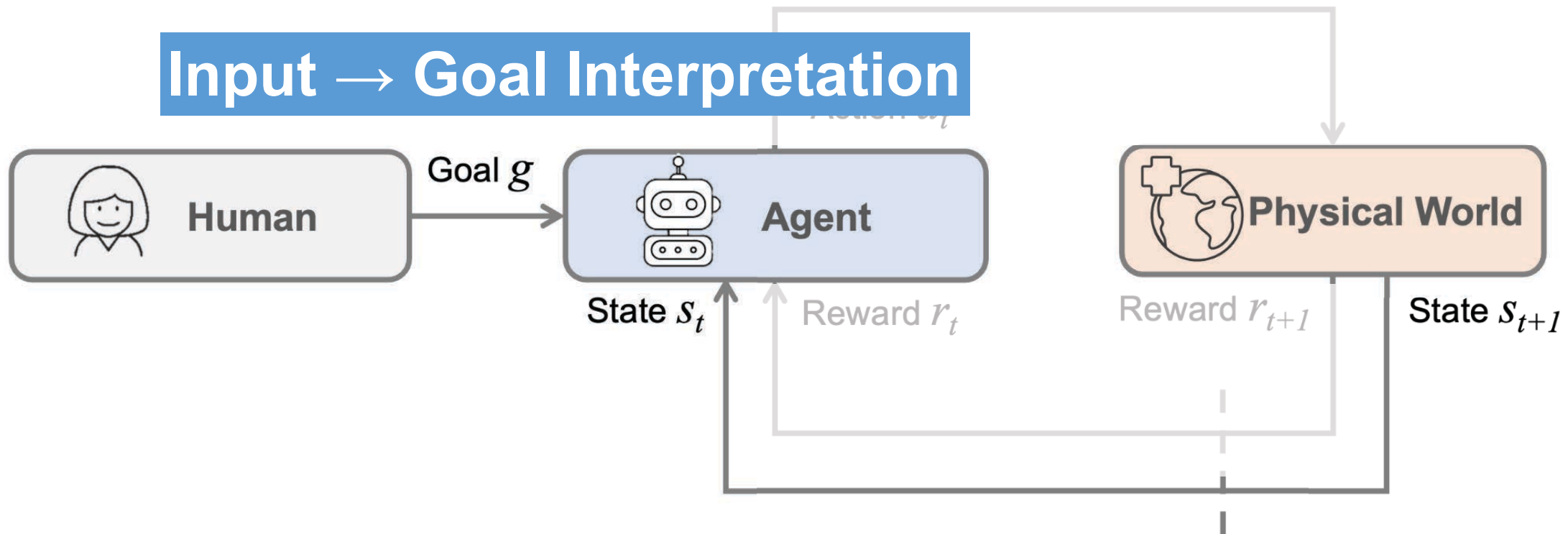


Name	Ingredients	Crafting recipe
Cake	Milk Bucket + Sugar + Egg + Wheat	
Golden Apple	Gold Ingot + Apple	

Let us go back to MDPs (Markov Decision Processes)



Input → Goal Interpretation



Goal Interpretation

g



Set up the table.

...

on_top_of (turkey, table)

...





Set up the table.

...

on_top_of (turkey, table)

...



Use the plates.

...

on_top_of (plate, table)

on_top_of (turkey, plate)

...





Goal : defines a task?

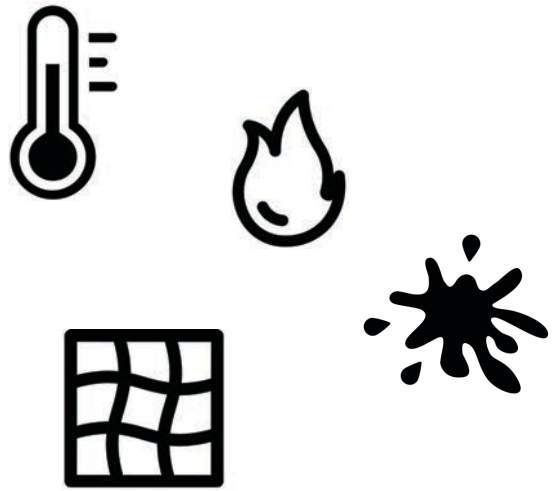
wikiHow
to do anything
YouTube

“Pack lunch”

What objects?



What properties?

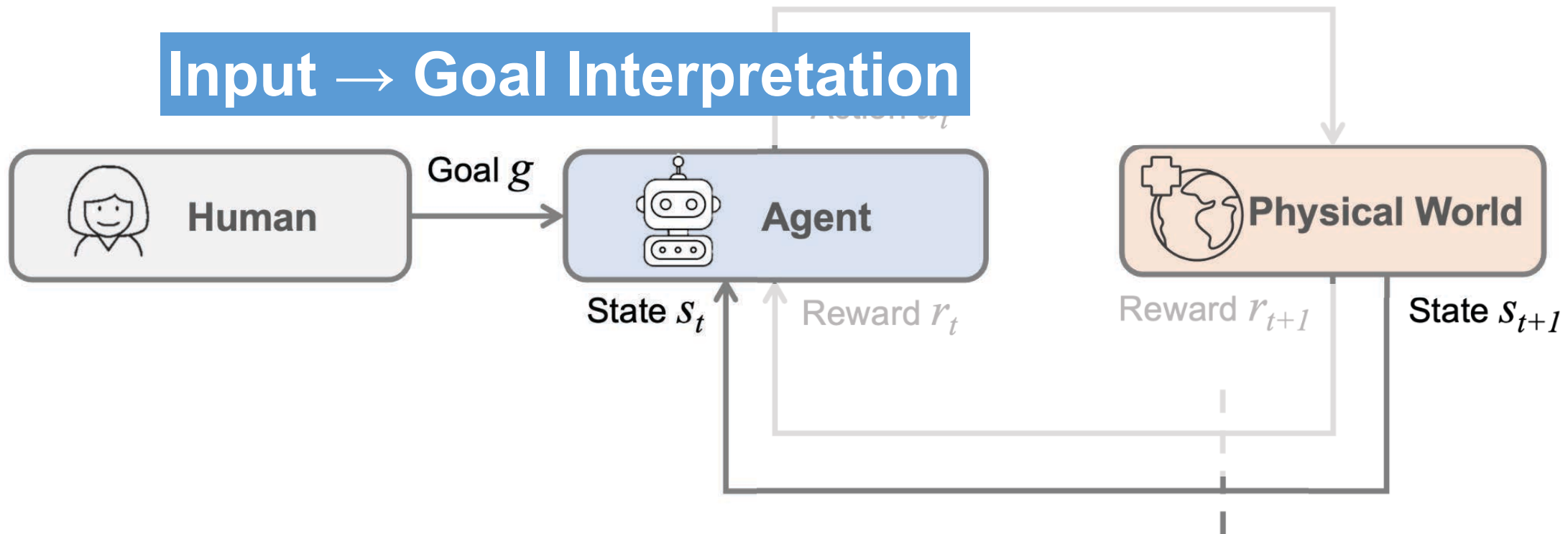


Start & Goal?

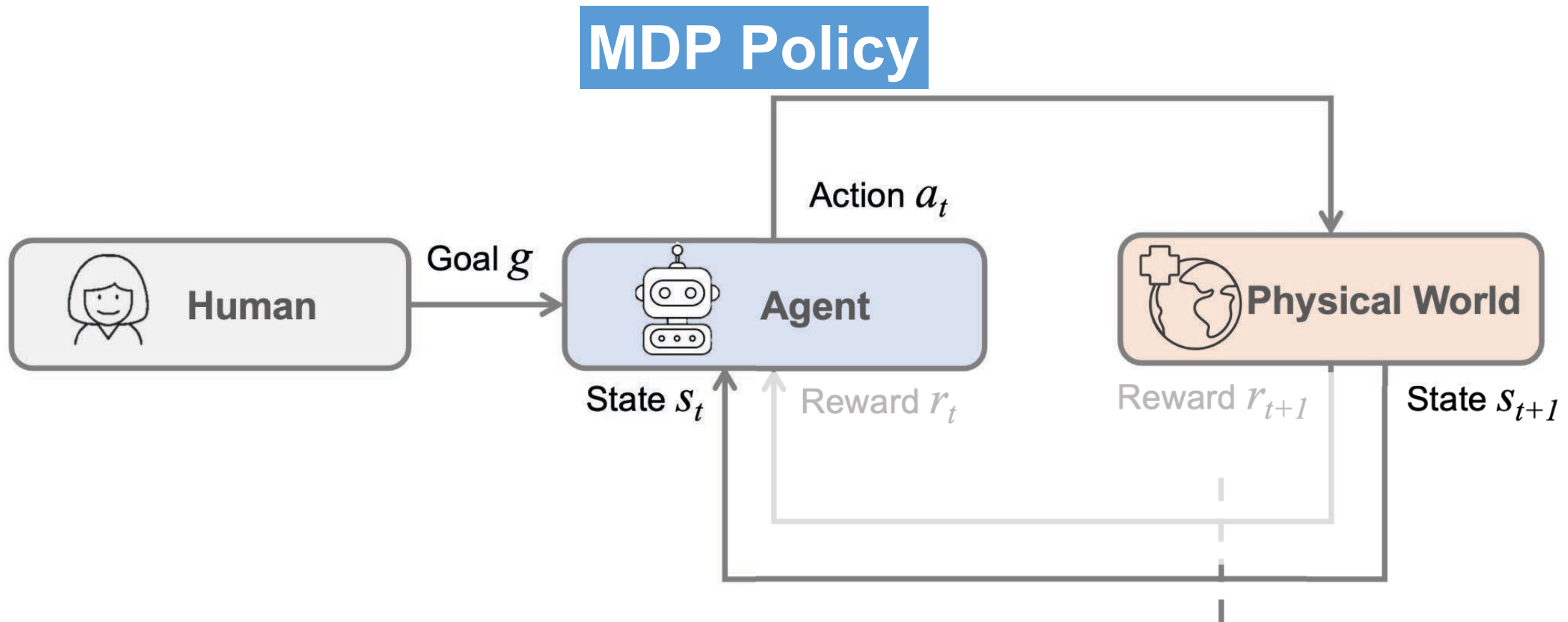
- **apple** in fridge
- **burger** in fridge
- **water bottle** in fridge
- **paper bag** on counter



Input → Goal Interpretation



Let us go back to MDPs (Markov Decision Processes)



Policy

$$\pi(o, g) \rightarrow a$$

Input: Preserving food



... inside (strawberry, pan) ...



... cooked (strawberry) ...



BEHAVIOR



Action Trajectory

- ...
- A7 OPEN (oven)
- A8** RIGHT_GRASP(pan)
- A9** RIGHT_PLACE_INSIDE (oven)
- A10** CLOSE(oven)
- A11** COOK(strawberry)
- ...

LLM Output



This video is for demonstration only. There're no actual controller-level actions. For action execution examples, visit our repository: <https://github.com/embodyed-agent-interface/embodyed-agent-interface>.

Input: Preserving food



... inside (strawberry, pan) ...



... cooked (strawberry) ...



BEHAVIOR



Action Trajectory

- ...
- A7 OPEN (oven)
- A8 RIGHT_GRASP(pan)
- A9 RIGHT_PLACE_INSIDE (oven)**
- A10 CLOSE(oven)
- A11 COOK(strawberry)
- ...

LLM Output



This video is for demonstration only. There're no actual controller-level actions. For action execution examples, visit our repository: <https://github.com/emodied-agent-interface/emodied-agent-interface>.

Input: Preserving food



... inside (strawberry, pan) ...



... cooked (strawberry) ...



BEHAVIOR



Action Trajectory

- ...
- A7 OPEN (oven)
- A8 RIGHT_GRASP(pan)
- A9 RIGHT_PLACE_INSIDE (oven)
- A10** CLOSE(oven)
- A11** COOK(strawberry)
- ...

LLM Output



This video is for demonstration only. There're no actual controller-level actions. For action execution examples, visit our repository: <https://github.com/emodied-agent-interface/emodied-agent-interface>.

Input: Preserving food



... inside (strawberry, pan) ...



... cooked (strawberry) ...



BEHAVIOR



Action Trajectory

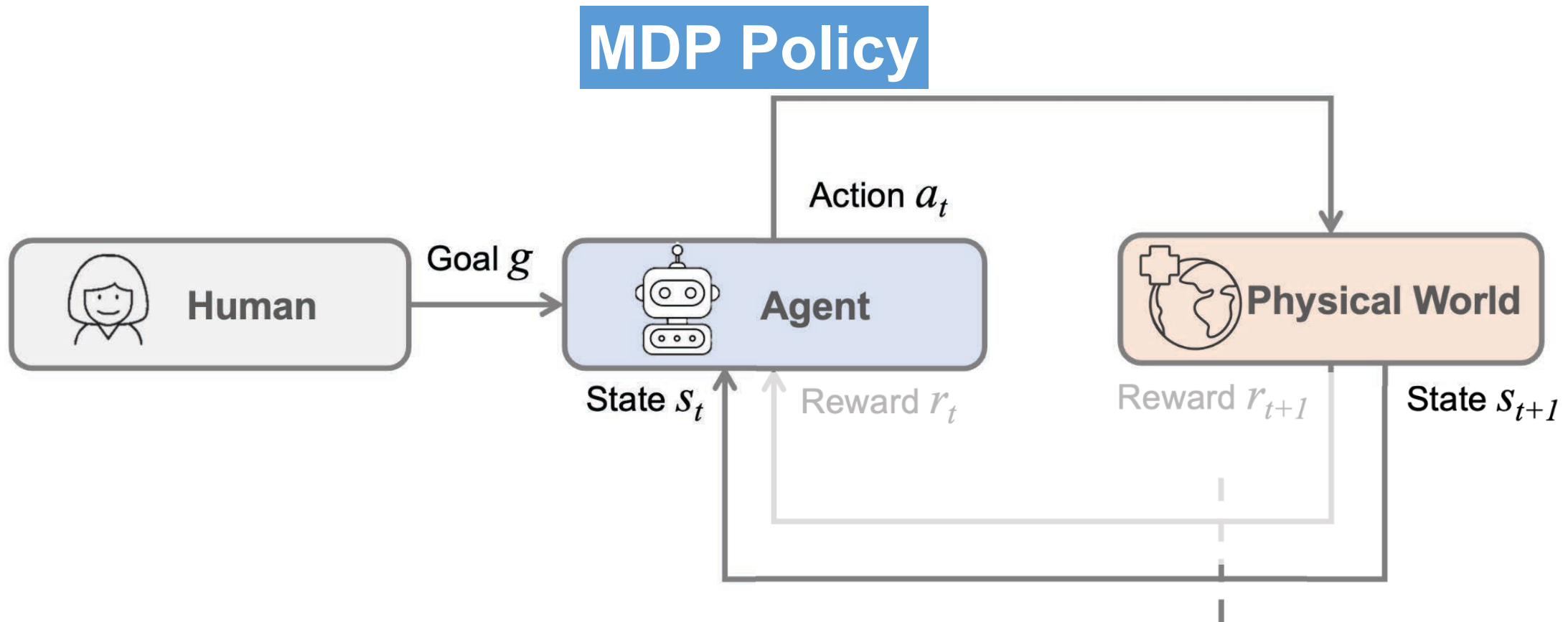
- ...
- A7 OPEN (oven)
- A8 RIGHT_GRASP(pan)
- A9 RIGHT_PLACE_INSIDE (oven)
- A10 CLOSE(oven)
- A11 COOK(strawberry)**
- ...

LLM Output

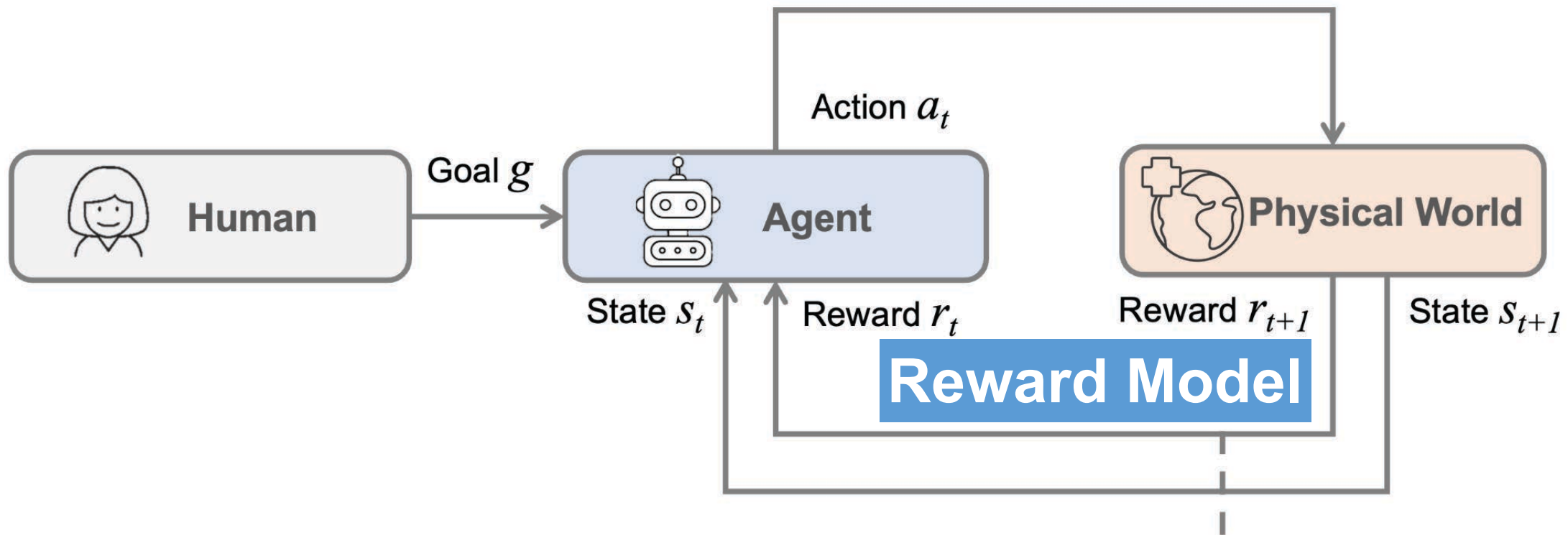


This video is for demonstration only. There're no actual controller-level actions. For action execution examples, visit our repository: <https://github.com/emodied-agent-interface/emodied-agent-interface>.

Let us go back to MDPs (Markov Decision Processes)

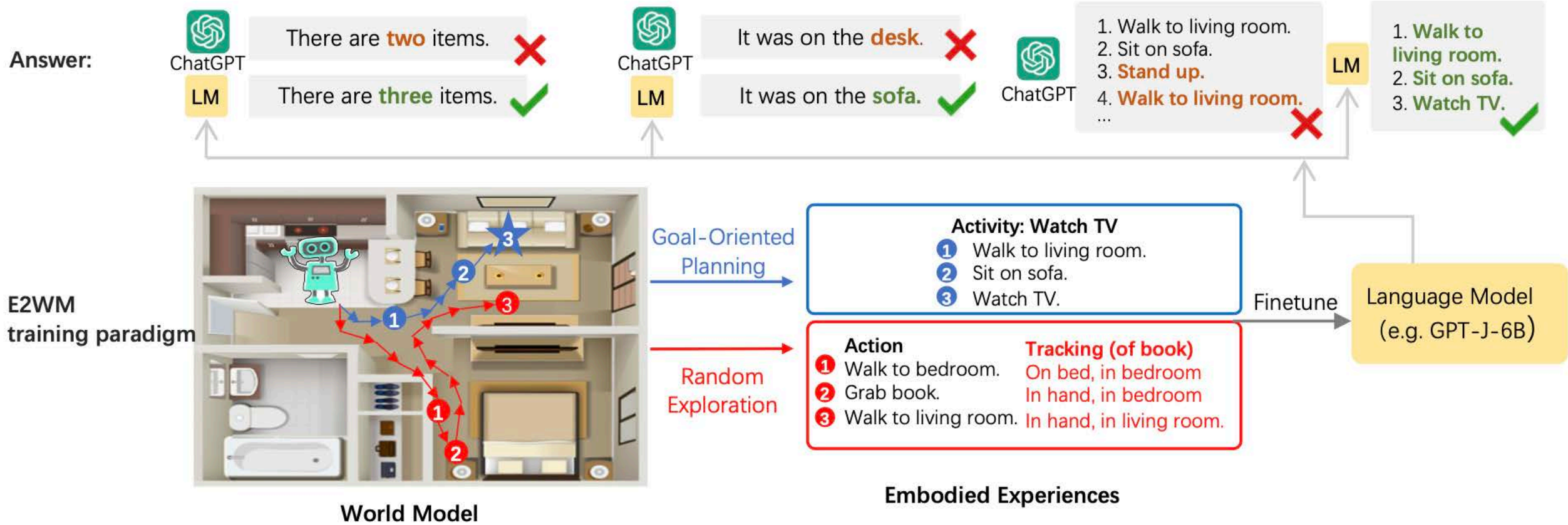


Let us go back to MDPs (Markov Decision Processes)

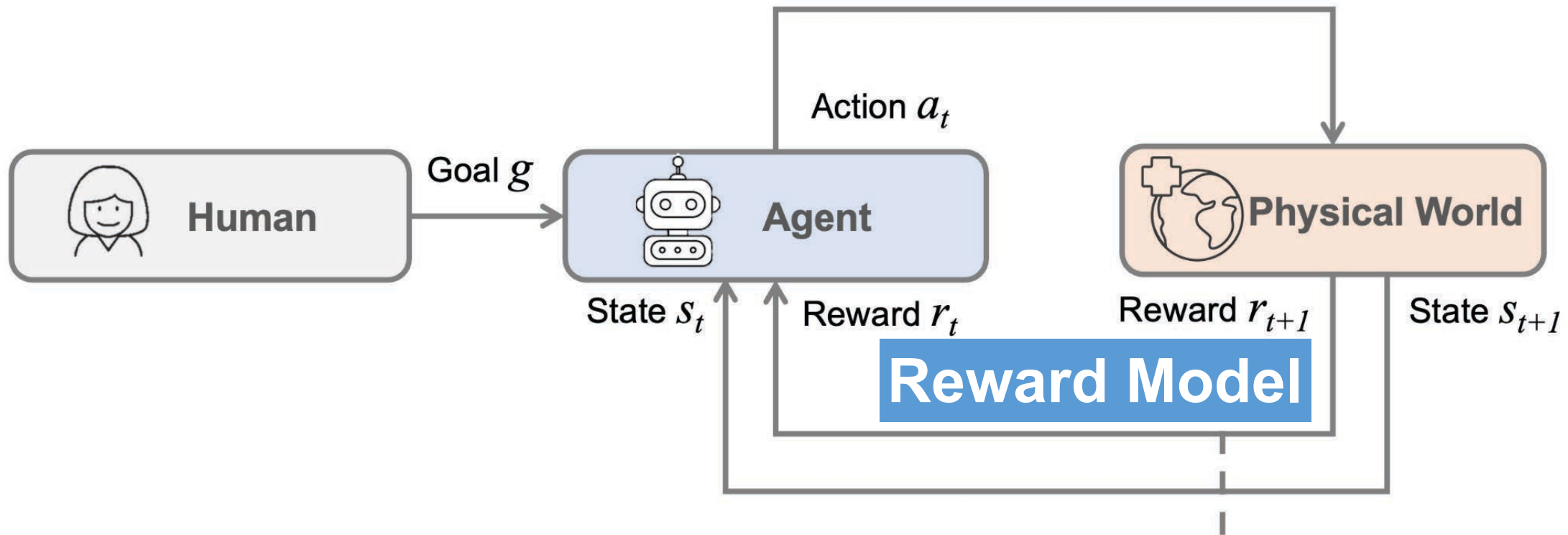


Reward Model

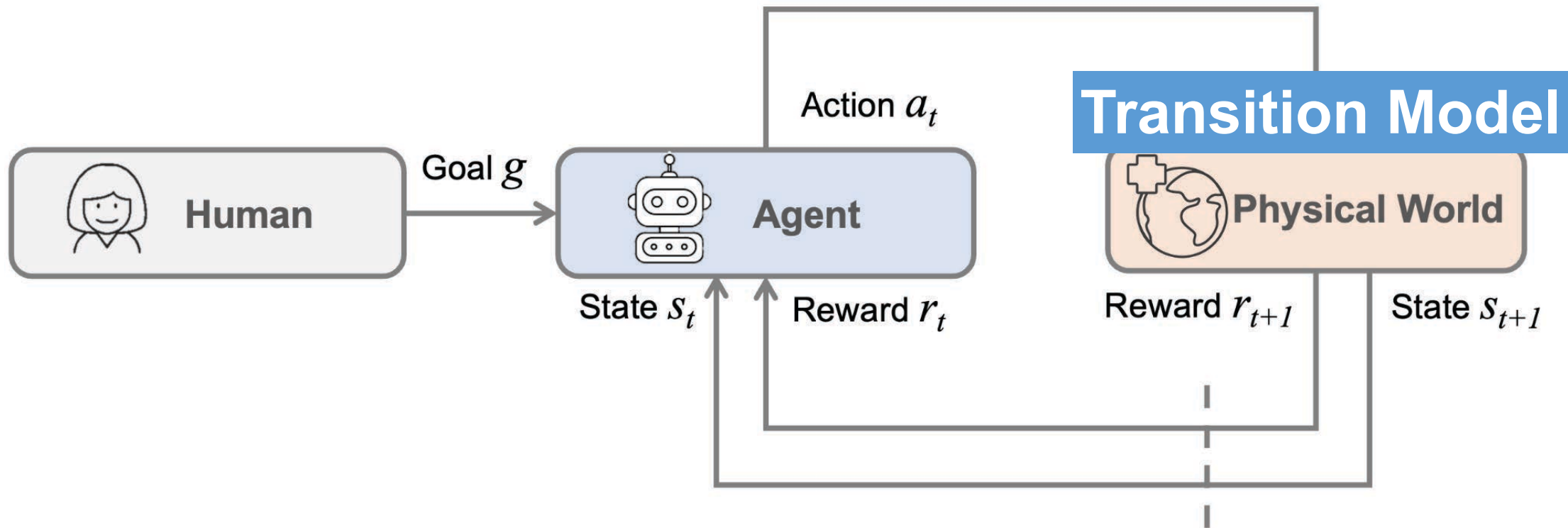
$$o, a \rightarrow r$$



Let us go back to MDPs (Markov Decision Processes)



Let us go back to MDPs (Markov Decision Processes)



Transition Model

$$O_t, a \rightarrow O_{t+1}$$

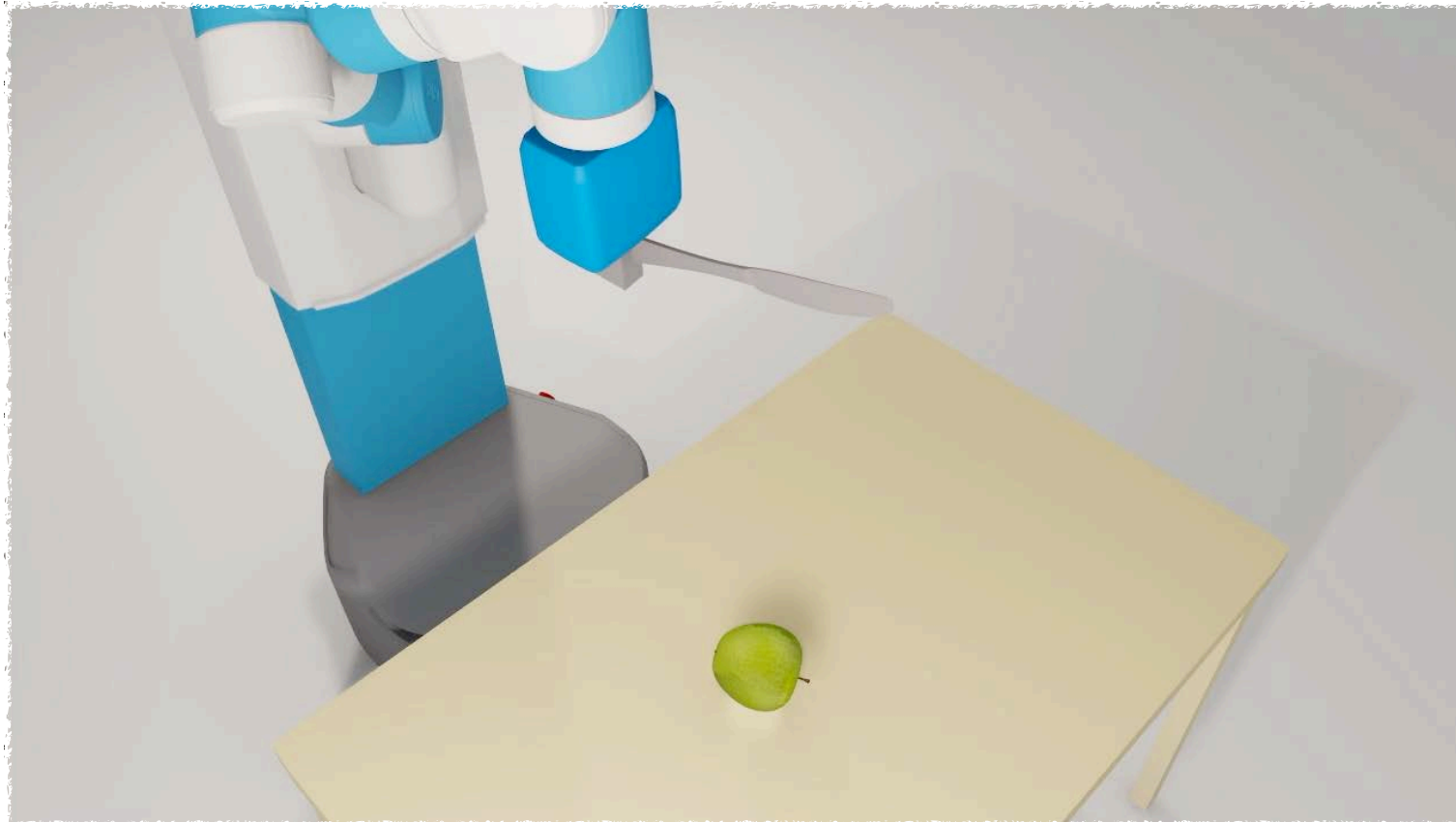
"World Modeling"



OmniGibson

Physics Transition Examples

```
class SlicingRule:
```





OmniGibson

Physics Transition Interface

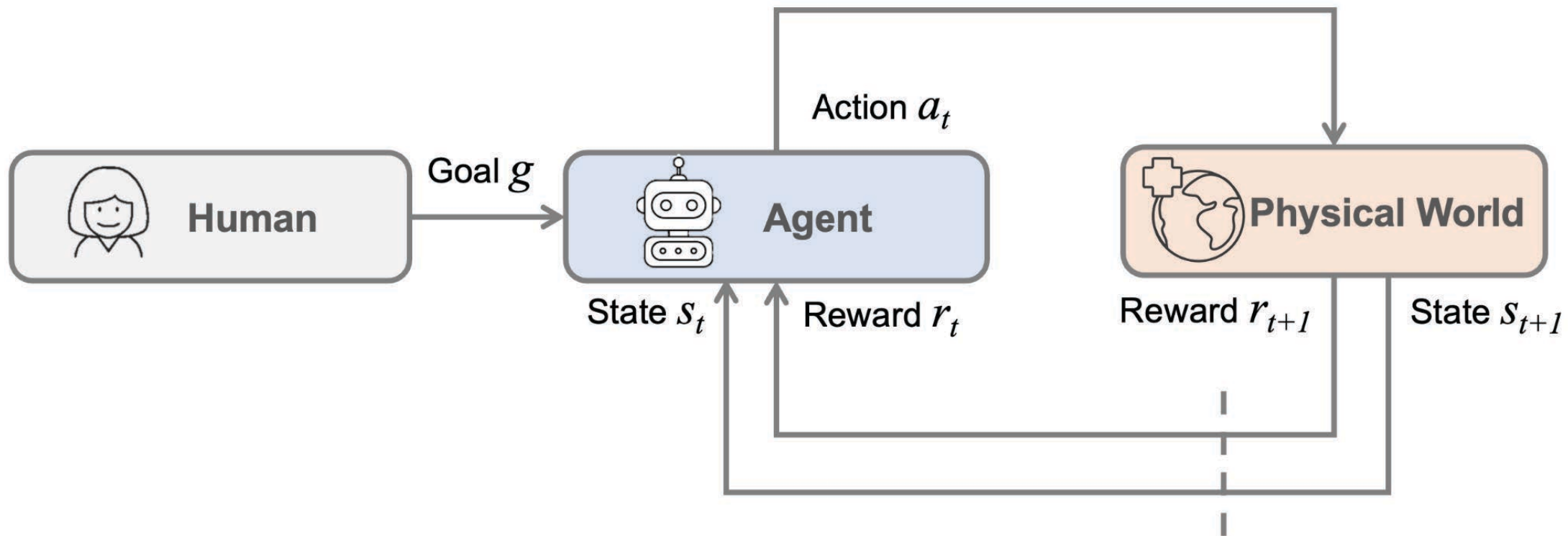
```
class TransitionRule:  
    def condition(self, *args) -> bool  
    def transition(self, *args)
```

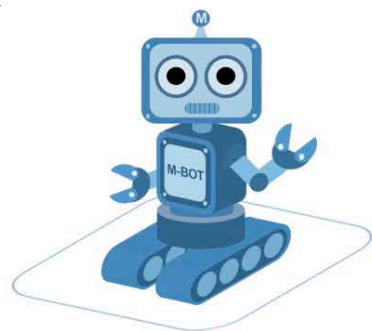
Determines whether a transition should occur

What should happen when a transition is triggered

Allows us to capture arbitrarily complex **physical phenomena!**

Let us go back to MDPs (Markov Decision Processes)





Different Instantiations of MDP



Perfect actuations

Perfect odometry

RGB only perception

-
-
-

Noisy actuations

Noisy odometry

Multiple perceptual modalities

-
-
-

Content	Time	Presenter
1. Motivation and Overview	15min	Manling Li
2. Foundation Models meet Virtual Agents	45min	Manling Li
3. Foundation Models meet Physical Agents		
Overview & Perception	25min	Jiayuan Mao
High-level & Low-level Decision Making	50min	Wenlong Huang
Break		
4. Robotic Foundation Models	30min	Yunzhu Li
5. Remaining Challenges	15min	Yunzhu Li
QA	30min	