Ladder of Intentions

Unifying agent architectures for explainability and transferability

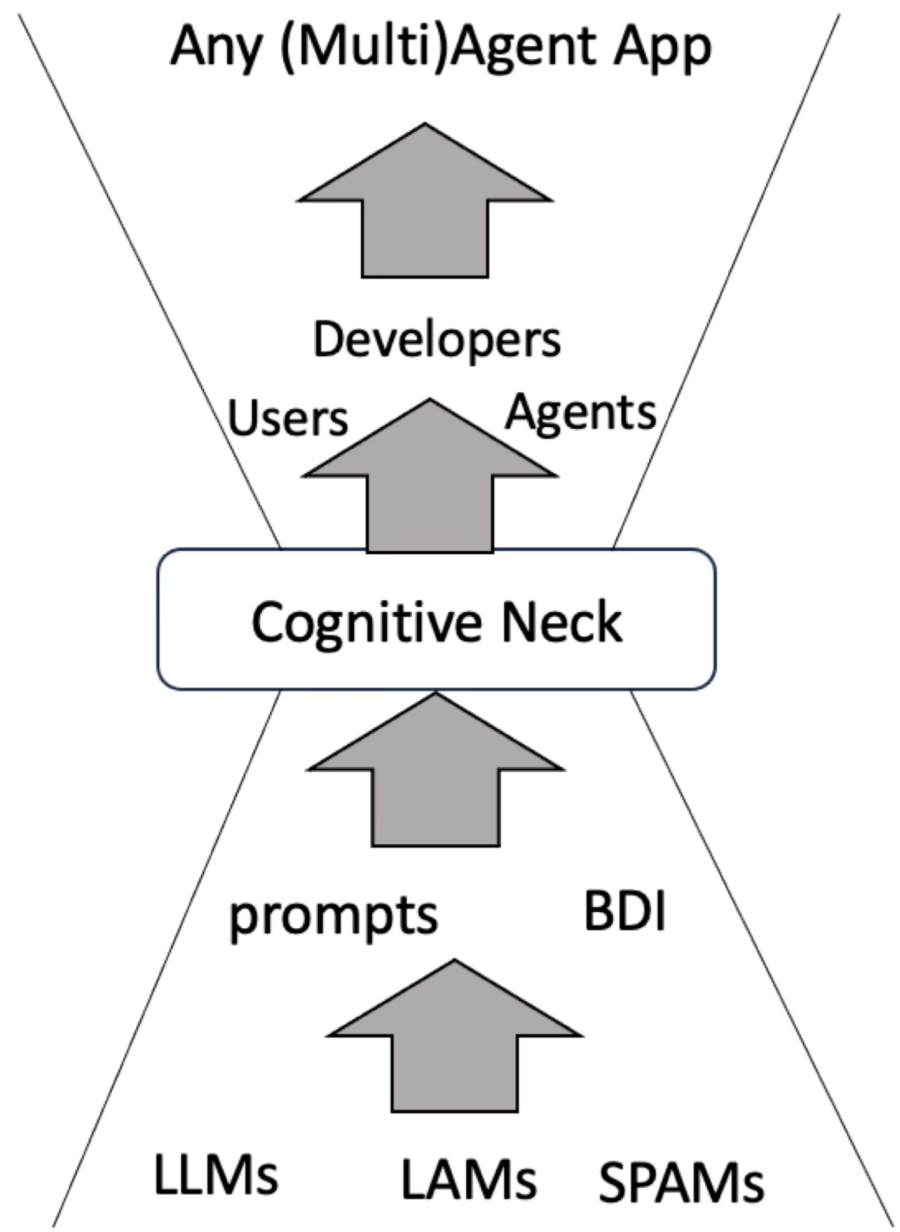
Victor Gimenez-Abalos, Adrian Tormos, Filip Edström, ***Sergio Alvarez-Napagao***, Javier Vázquez-Salceda, Mattias Brännström, John Lindqvist

Preface

From *The Cognitive Hourglass: Agent Abstractions in the Large Models Era*

Ricci, A., Mariani, S., Zambonelli, F., Burattini, S., & Castelfranchi, C. (2024, January). The Cognitive Hourglass: Agent Abstractions in the Large Models Era. In *AAMAS* (Vol. 24, pp. 2706-2711).

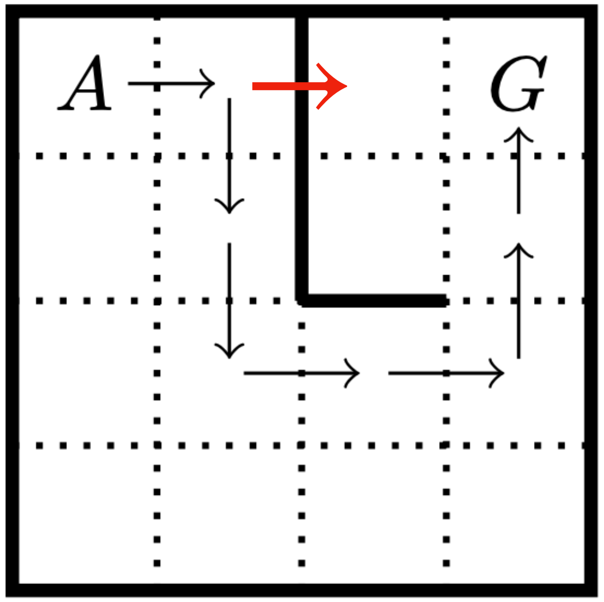
*"**Cognitive concepts** that are **pillars for the understanding and engineering of agent systems** constitute the indispensable neck of the cognitive hourglass, that is, **the fundamental human-compatible level of abstraction necessary for humans** to understand/design/govern agents and MAS at the **application level regardless of the specific AI technologies** adopted at the implementation level"*



Motivation and context

We actually **need a common vocabulary of XAI for agents**!

- shared/implementable **across any architecture**...
- ...so that all XAI can speak **in a similar manner**
  - to humans
  - or to other machines.



Why did the agent ram into the wall?

Motivation and context

Why is this **important**?

- Homogenising types of answers means **decoupling** the two processes:

Generating a truthful answer

Generating human-interpretable answers

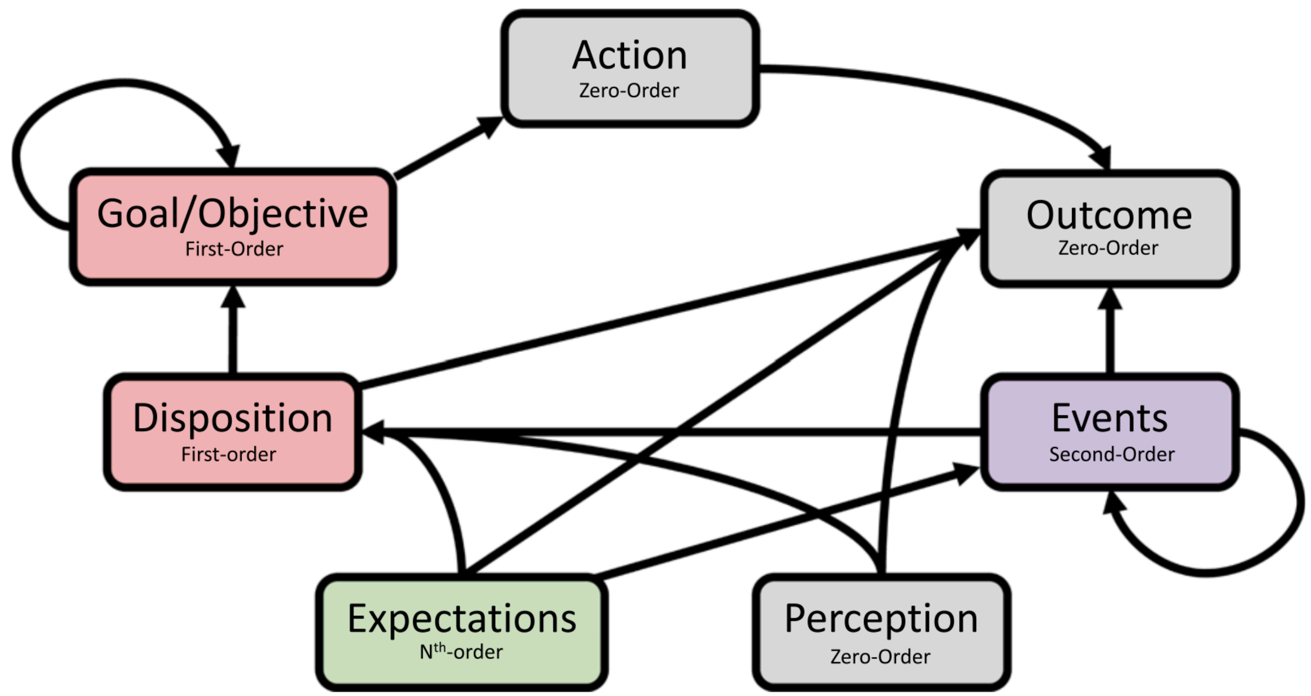
- This should help **reuse findings in the second one** for novel architectures

Motivation and context

- Agent architectures:
    - Policy-based / reinforcement learning (Q-learning, REINFORCE), BDI, Voyager, ReAct, SOAR, ACT-R, ...
    - **First-order explanations vary!**
- Therefore finding such a vocabulary is hard given that agent reasoning is **extremely heterogeneous**, ranging from trivial to extremely complex
    - Even for simple action choice in single-agent environments!

Background

- Classifying agent explainability in terms and levels is already explored in the literature
- However, for some agents, first-order explanations can be
  - **very complex**, *e.g. Voyager*
  - **very different**, *e.g. REINFORCE*



Dazeley, R., Vamplew, P. & Cruz, F. Explainable reinforcement learning for broad-XAI: a conceptual framework and survey. Neural Comput & Applic 35, 16893–16916 (2023). https://doi.org/10.1007/s00521-023-08423-1

- But **humans tend to explain via intentions and beliefs** (Malle, Bratman)
  - Is there any way to reconcile this?

Our proposal

- This paper is a first attempt at finding common ground between architectures...
  - via building a *meta-architecture*
  - **an optic from which to see existing architectures**
  - stratifying behaviour using Intentions, and based on Beliefs
    - **Intentions are imperative routines** (goal-directed behaviour)
    - **Beliefs are statements in the chosen formalism of the architecture**
- Both artifacts can be given or learnt, in a way that explanations at a level refer to the same concepts and look similar across architectures

Our proposal

- Informally: our target is to be able to *"make BDI"* with PDDL, Q-learning and Voyager comparable architectures
- We do this by **building a Structural Causal Model**
  - Albeit one with very complex variables
  - This model can be used to trace causality through the graph

Our proposal

Key insight

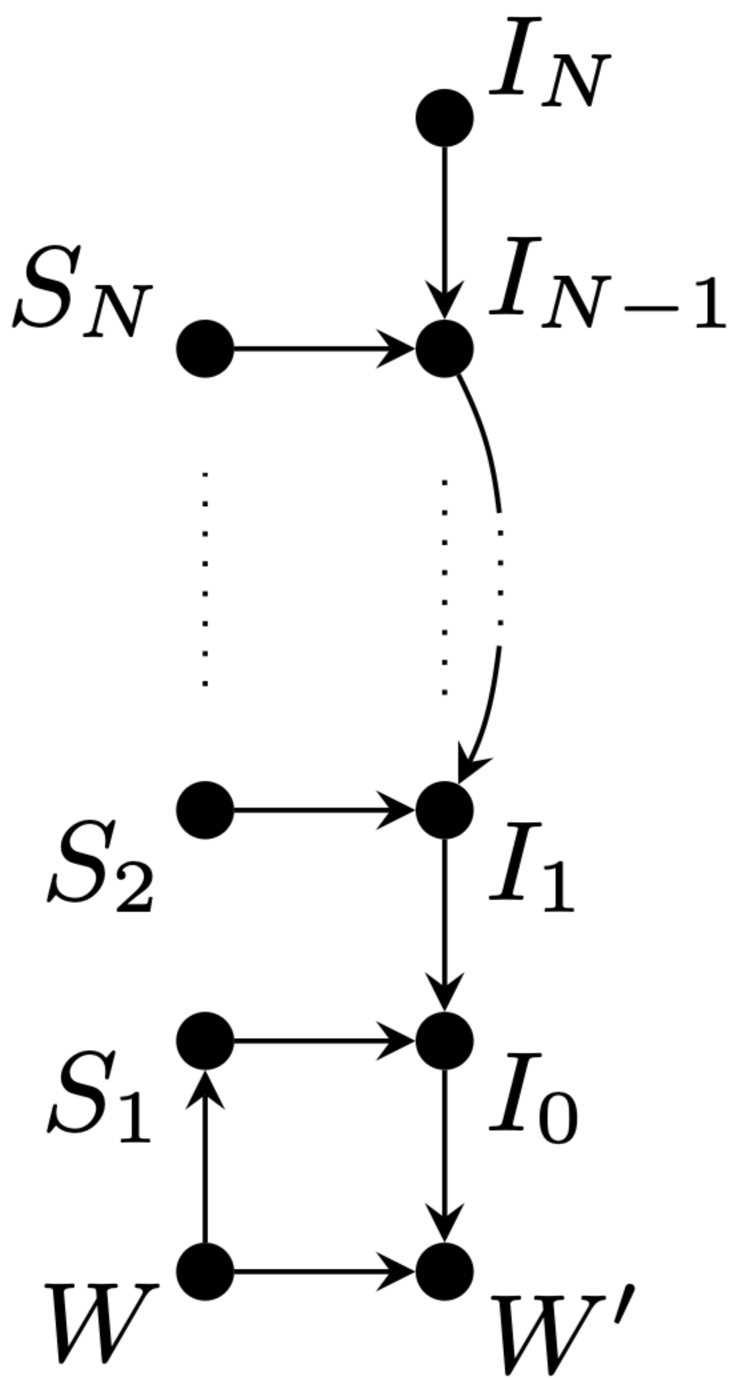
- Any action (simple or complex) is caused by:

State + Policy

- Generally, the focus of XAI is on the state, but... why is the policy as it is?
    - ***Q:*** *"What was the cause of this policy?"* ***A:*** *"It was trained"*
    - If there is a learning process, there is a method to use 'experience' to determine the policy. Furthermore, there are reasons for that learning process, and so on.
    - ***Q:*** *"What was the cause of this training?"* ***A:*** *"It was the designer intent"*
- We call this causal chain **a ladder of intentions**

The Ladder of Intentions

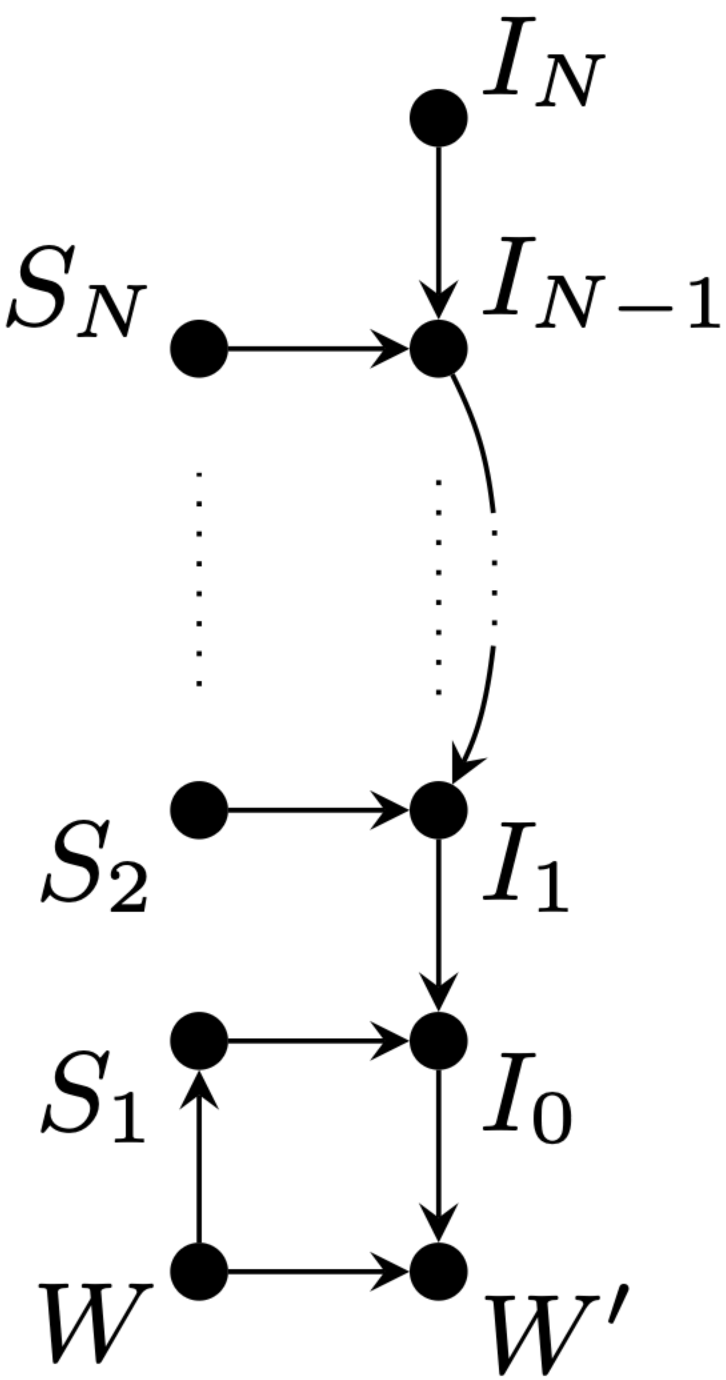
Static view



- Any explanation can be a chain of *explanandums*
  - Referring to *explanans* of a previous sentence
  - Until the explainee is satisfied or there is no further explanation possible:
    - observations *(some observed quality of the environment)*, or
    - designer-choice *(this was so because someone made it so)*

The Ladder of Intentions

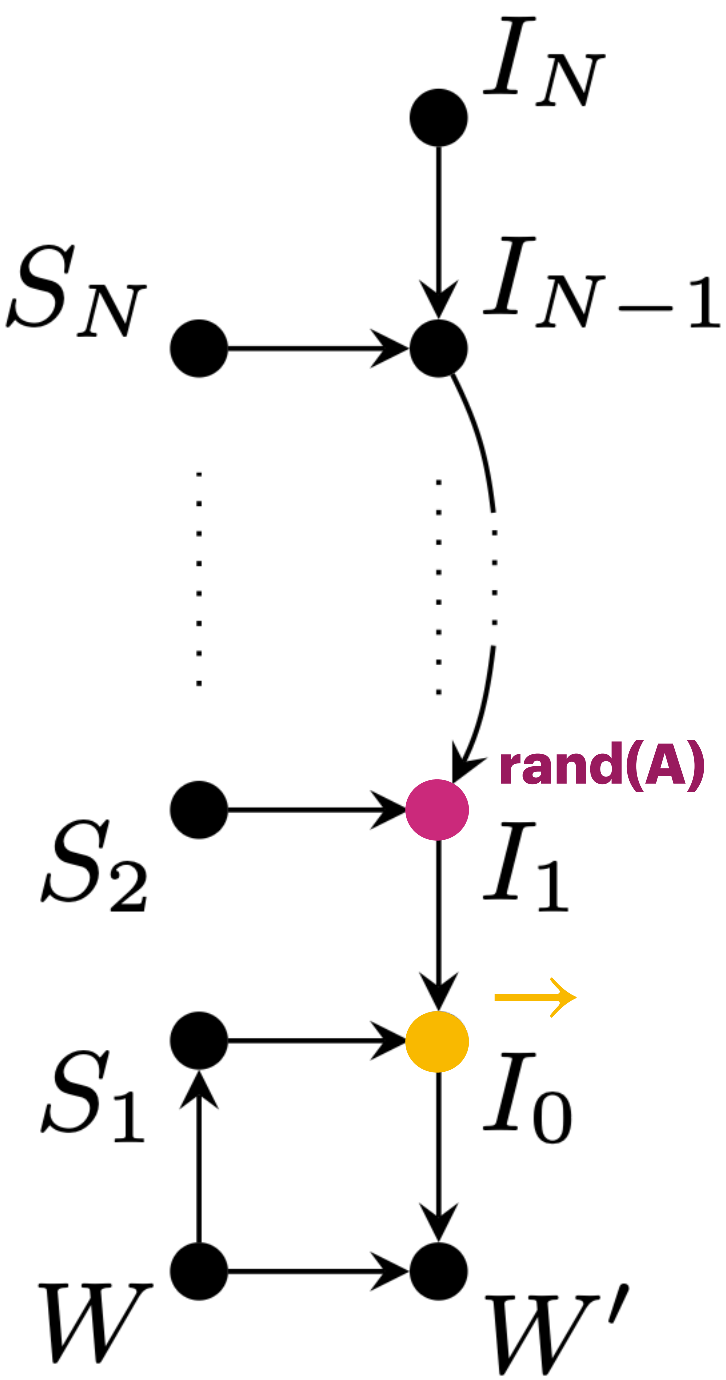
Static view

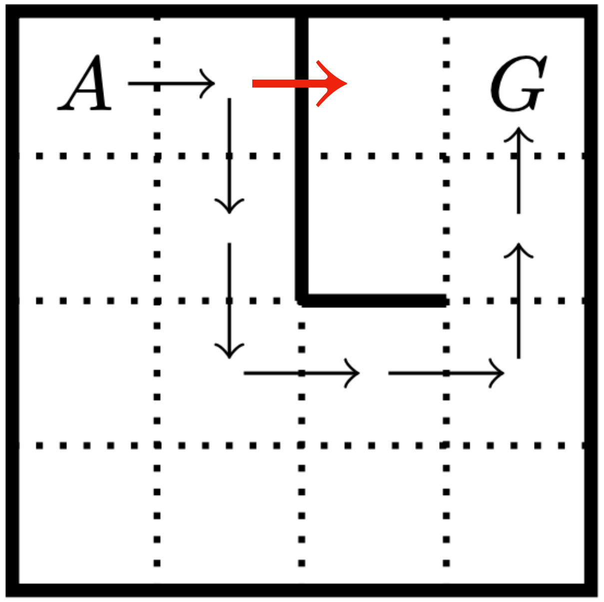


- **Going UP is questioning the *desire* of an intention**
  - Resulting in another, higher-level intention
- **Going sideways is questioning the *beliefs* on how that desire is to be achieved**

The Ladder of Intentions

Static view: **Q-Learning (exploration)**

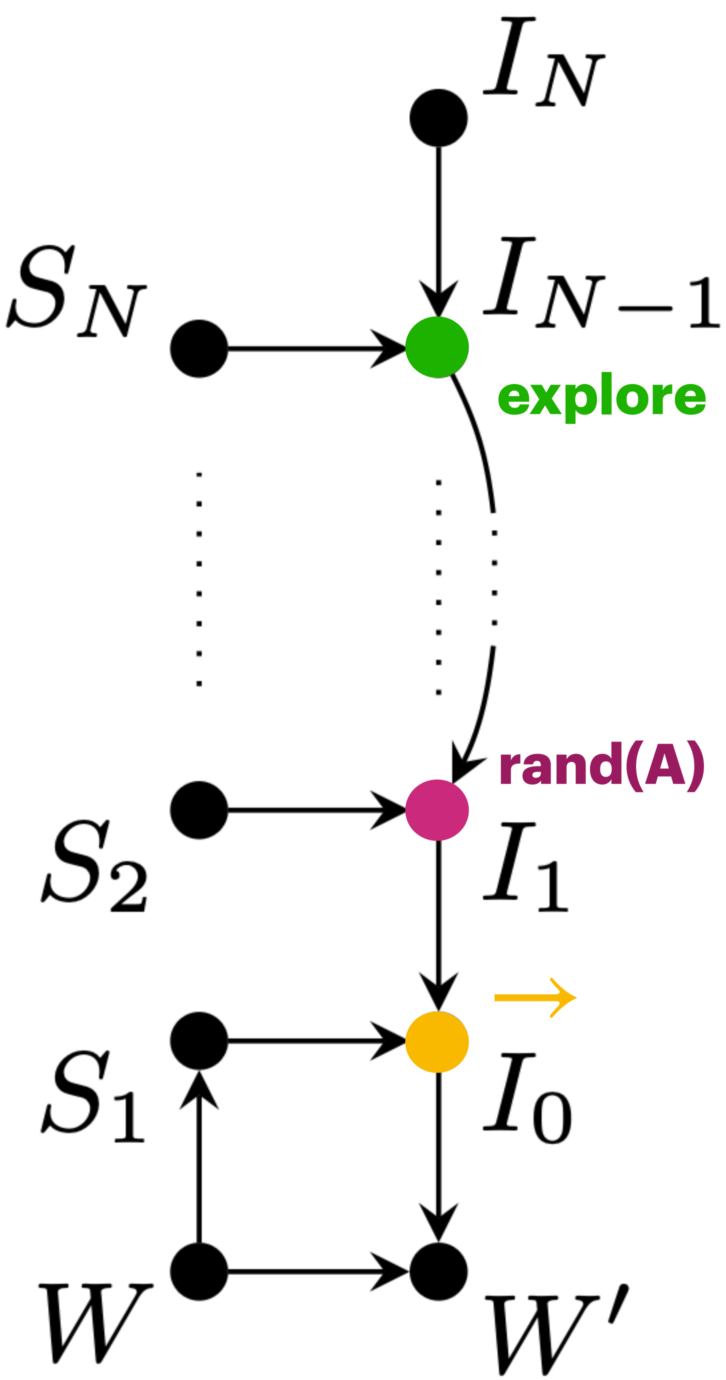


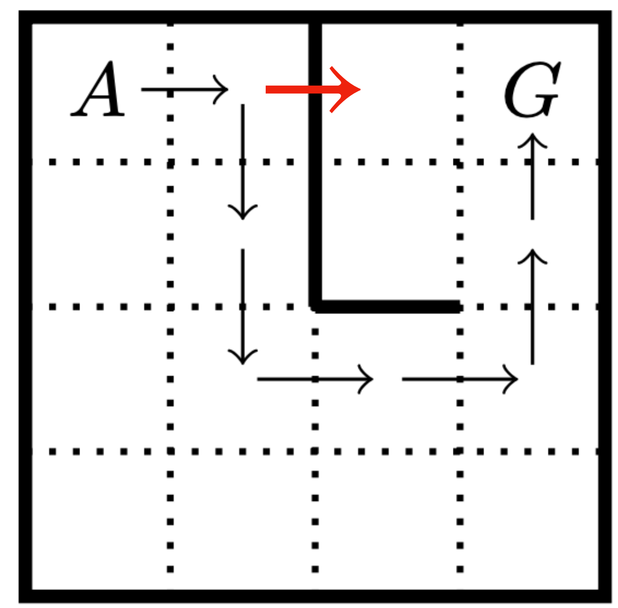


- Why did you ram into the wall (I_0) at t_1?
  - I wanted to pick a random action (I_1) so I did →

The Ladder of Intentions

Static view: **Q-Learning (exploration)**

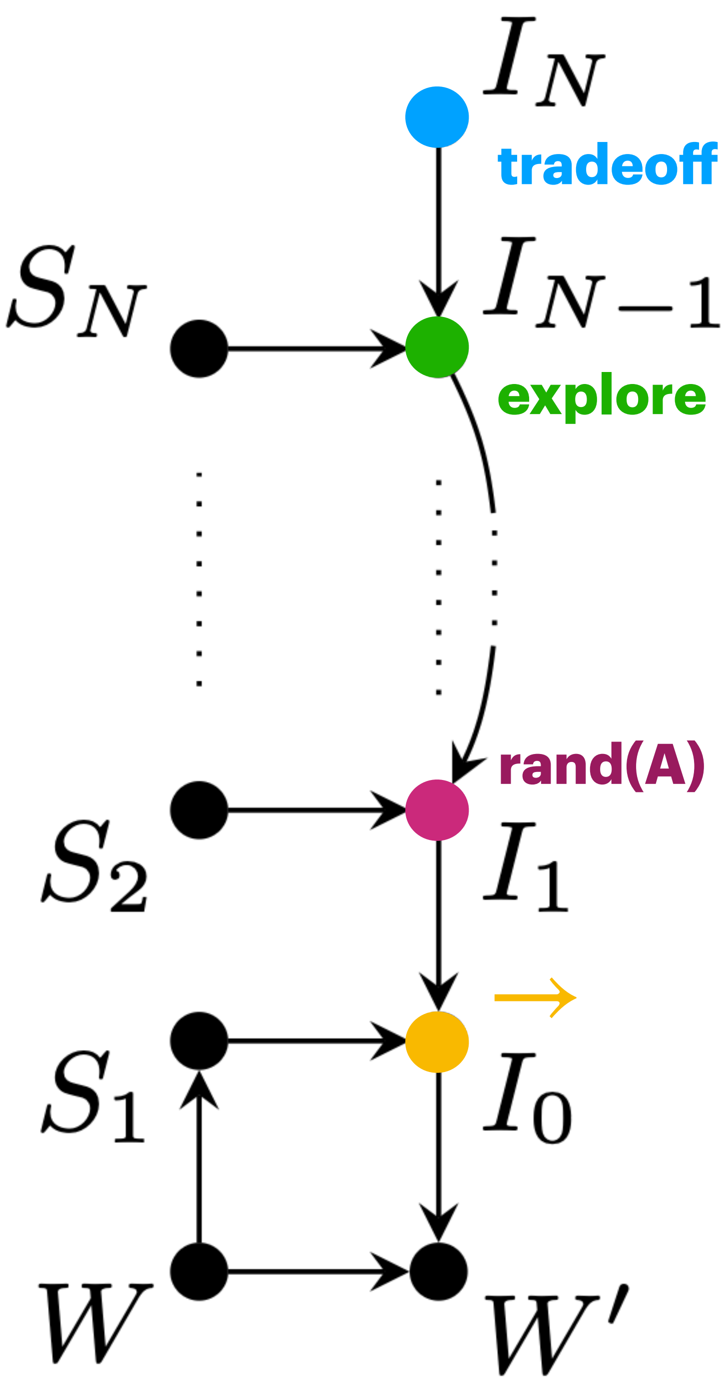


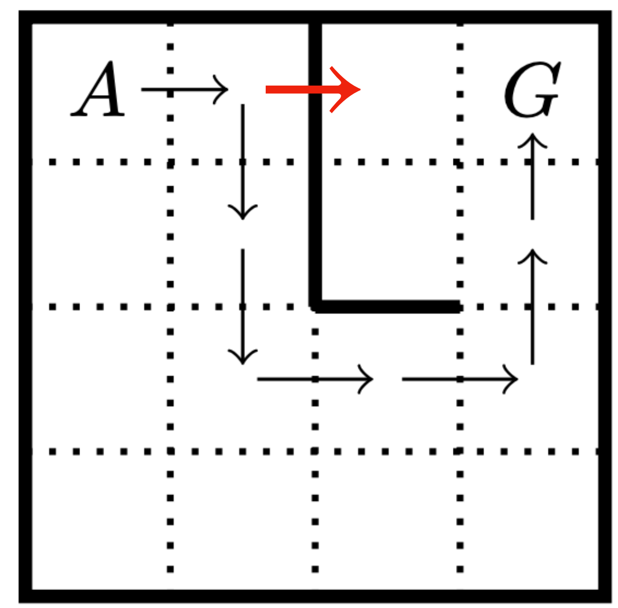


- Why did you do a random pick (I_1) at t_1?
  - Because I wanted to explore (I_2)

The Ladder of Intentions

Static view: **Q-Learning (exploration)**

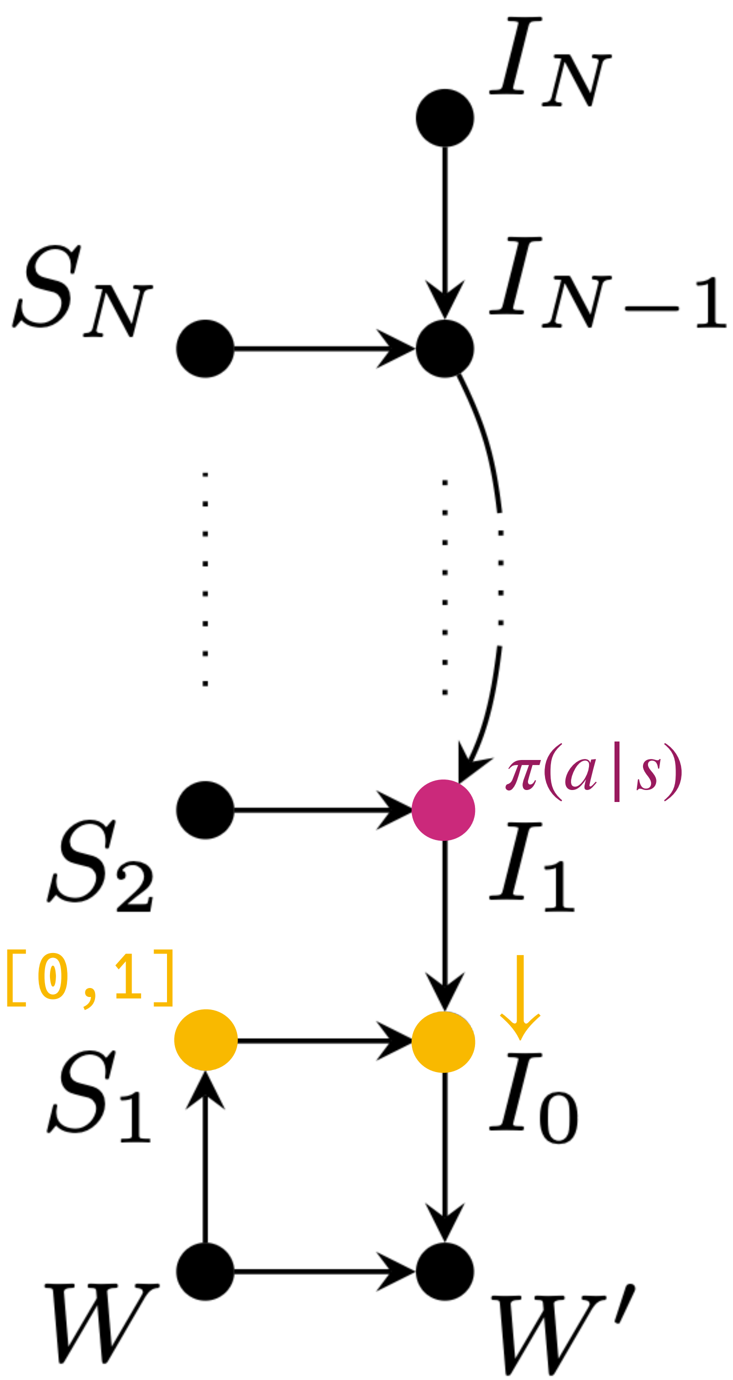


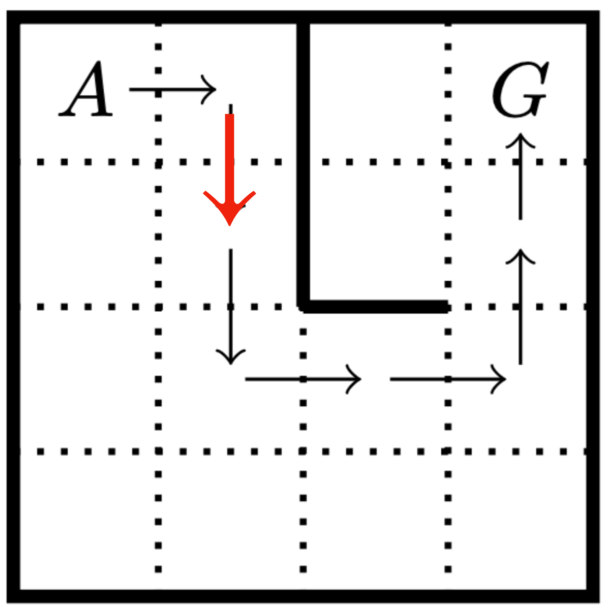


- Why did you explore (I_2) at t_1?
  - Because I want to get to the goal as fast as possible and to do that I need to trade-off exploring and exploiting what I know (I_3)

The Ladder of Intentions

Static view: **Q-Learning (exploitation)**

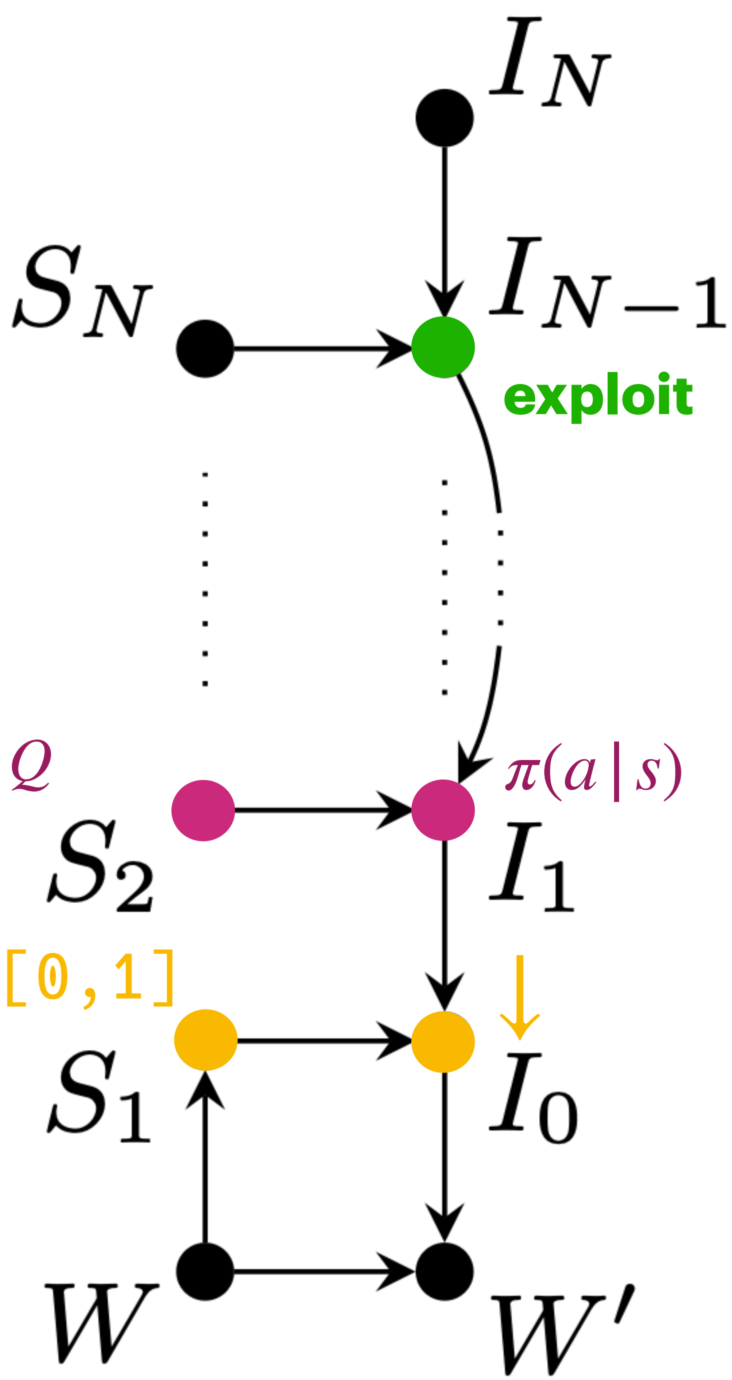


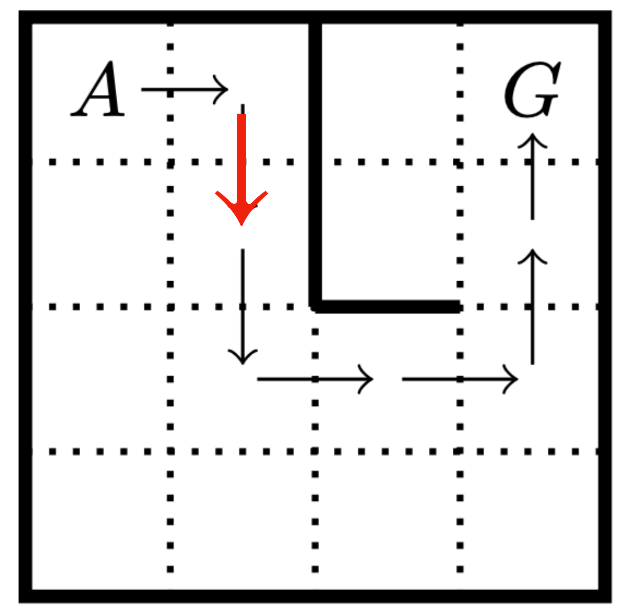


- Why did you move around the wall (I_0) at t_1?
  - Because I believed I was in `pos=[0,1]` (S_1) and wanted to follow the policy (I_1) so I did ↓

The Ladder of Intentions

Static view: **Q-Learning (exploitation)**

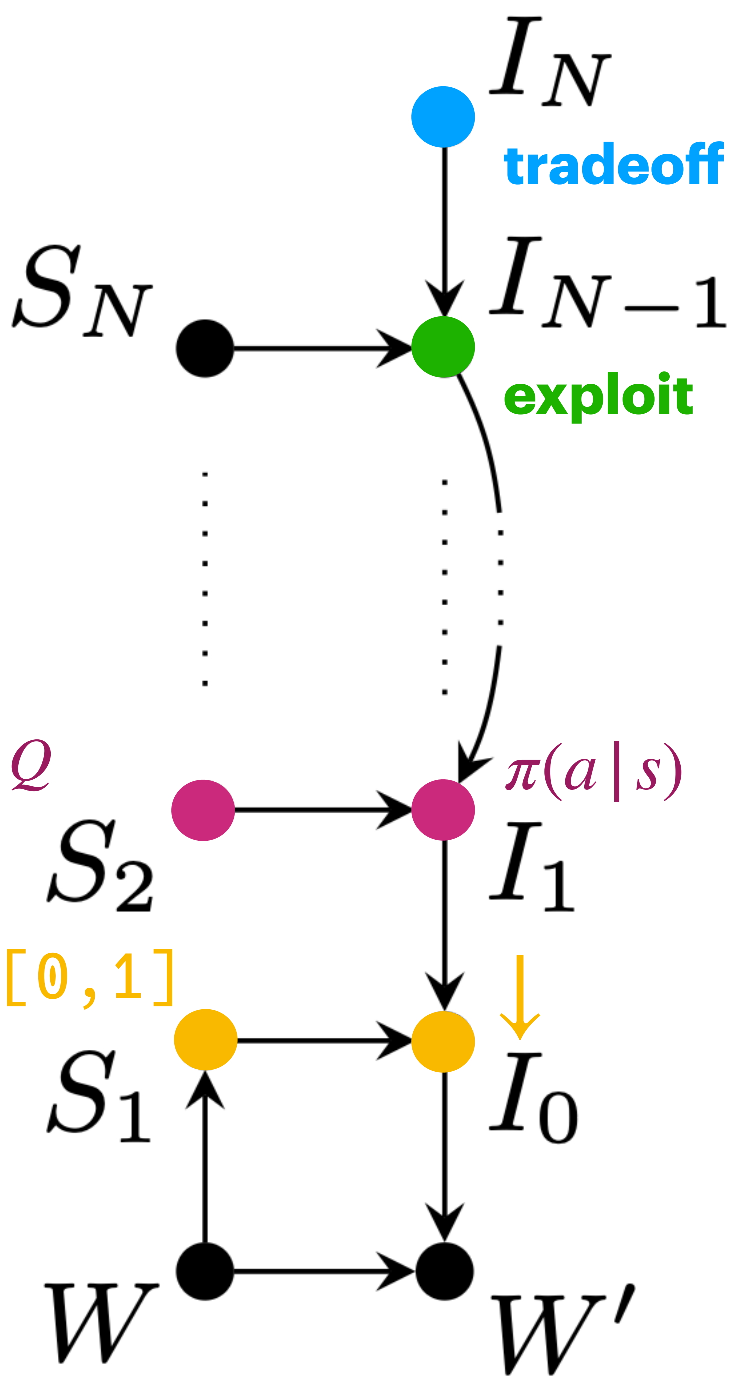


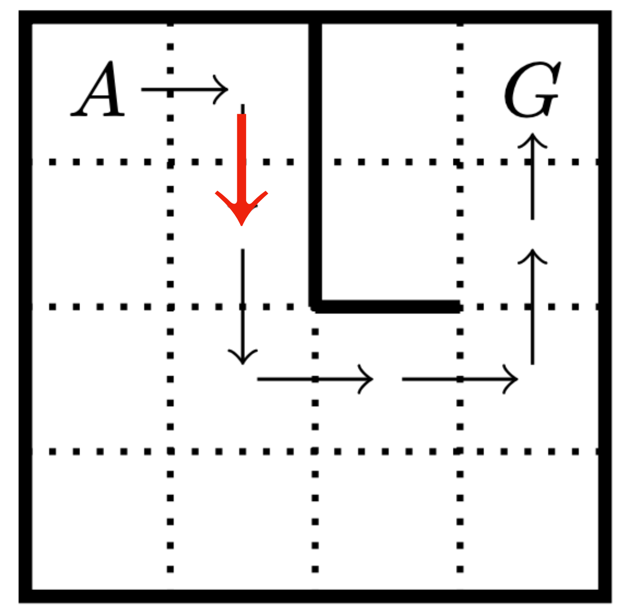


- Why did you follow this policy (I_1) at t_1?
  - Because I believed in this `Q(s,a)` which, maximising, makes me go to the goal (S_2) and I wanted to exploit it to go to the goal (I_2)

The Ladder of Intentions

Static view: **Q-Learning (exploitation)**

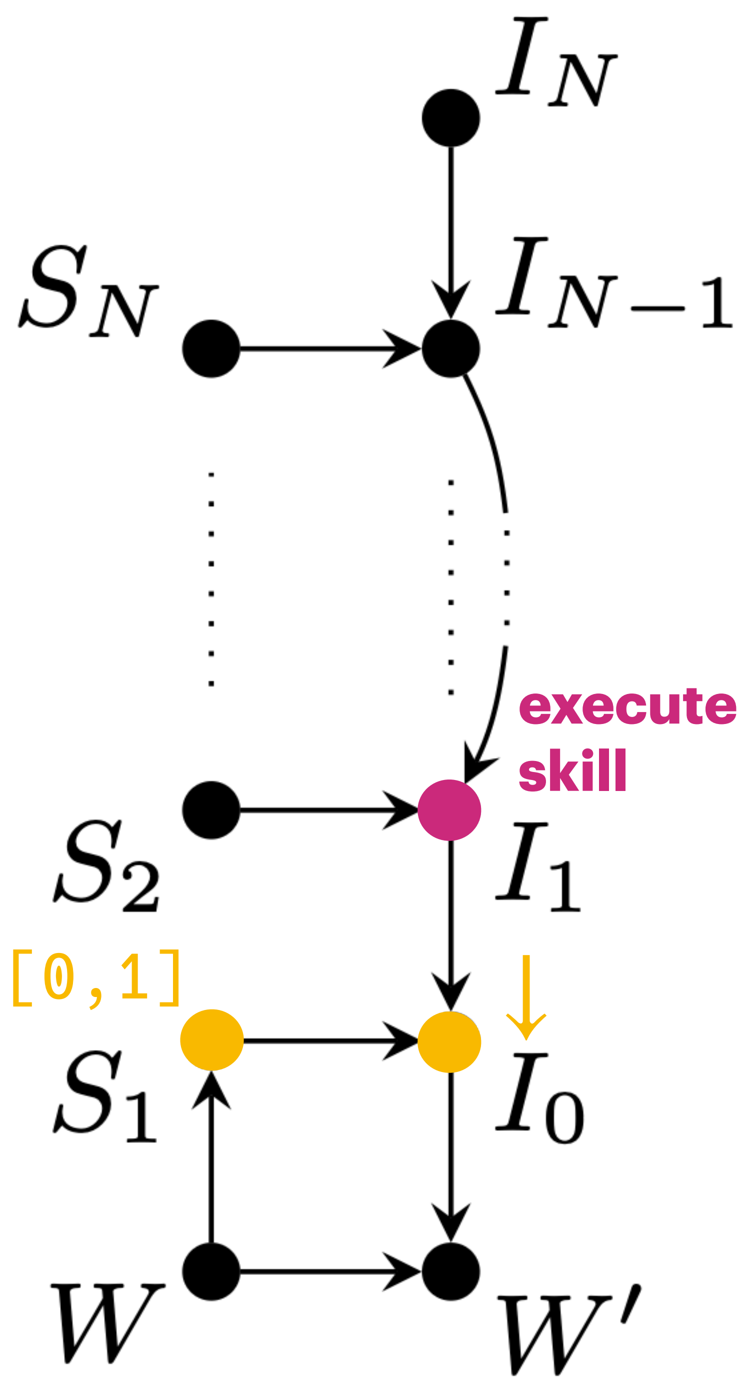


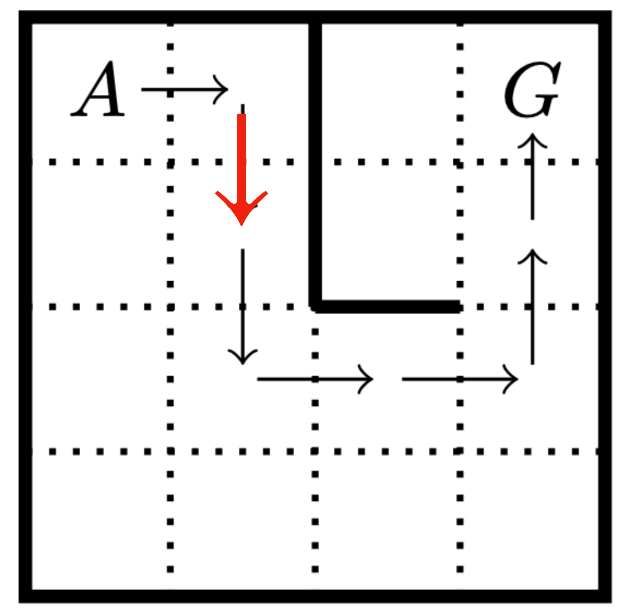


- Why did you explore (I_2) at t_1?
  - Because I want to get to the goal as fast as possible and to do that I need to trade-off exploring and exploiting what I know (I_3)

The Ladder of Intentions

Static view: **Voyager**



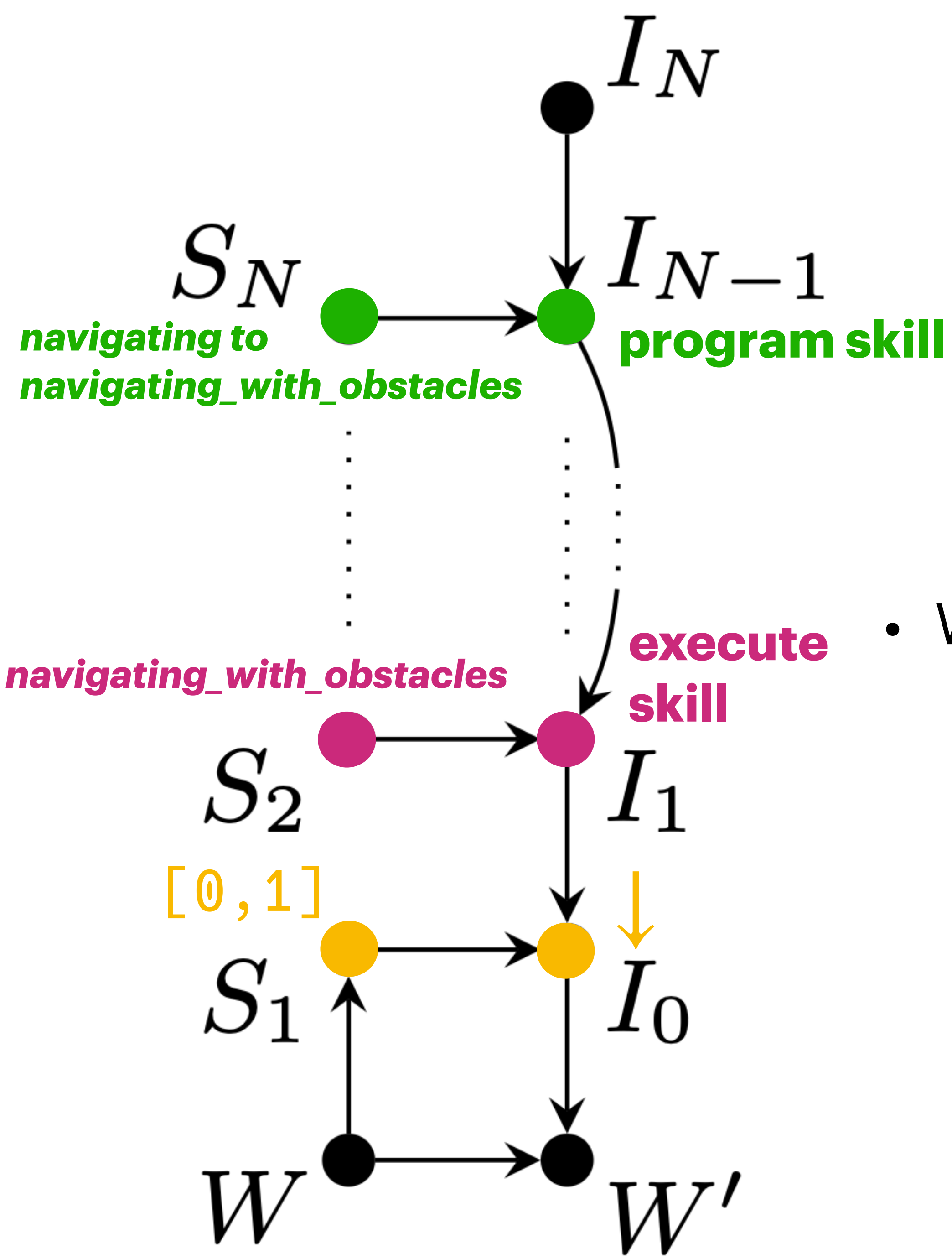


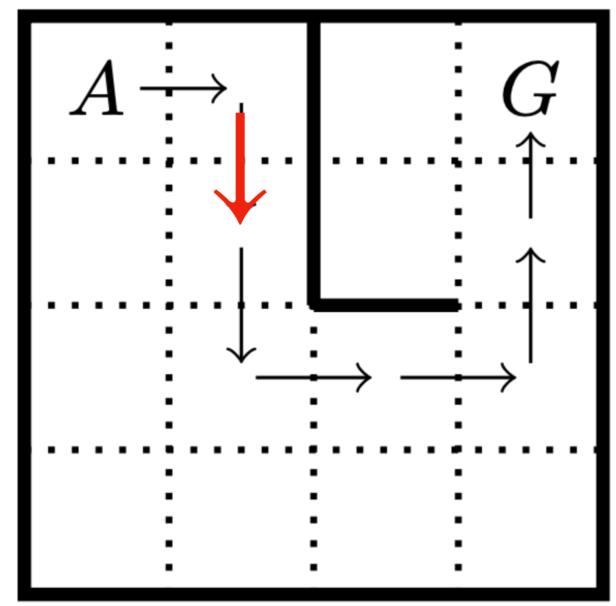
- Why did you move around the wall (I_0) at t_1?

  - I believed I was in `pos=[0,1]` (S_1) and I was executing the skill *navigate_with_obstacles* (I_1) so I did ↓

The Ladder of Intentions

Static view: **Voyager**

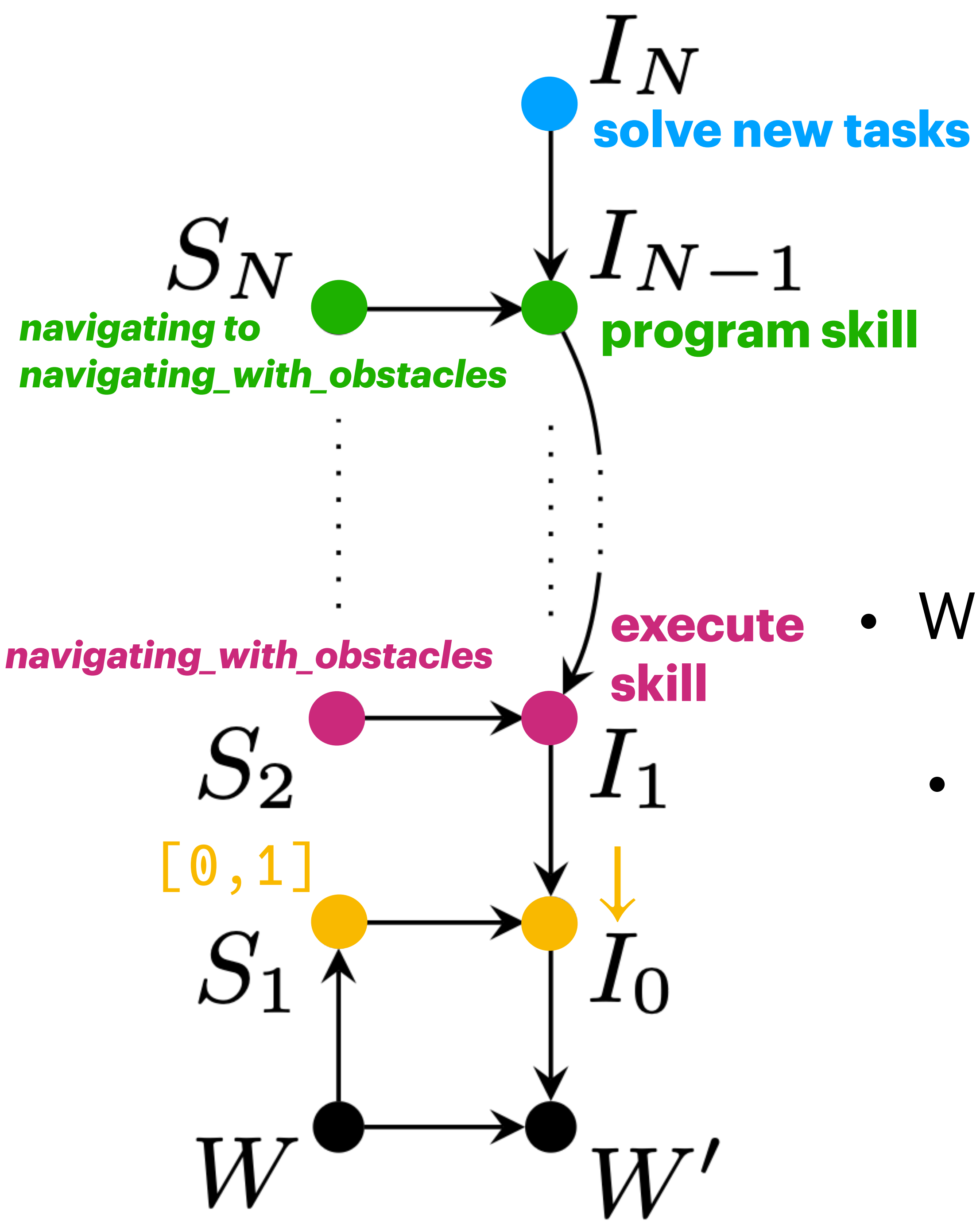


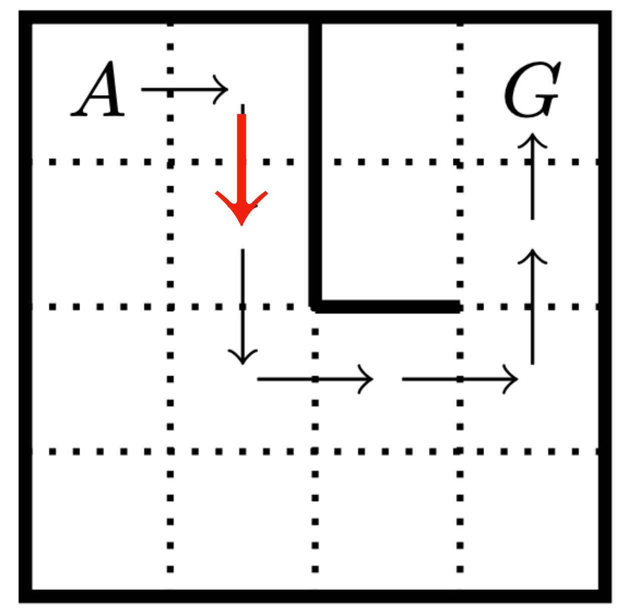


- Why did you execute this skill (I_1) at t_1?
  - At t_0 I believed I was in `position=[0,0]` and could use *navigate*, but environment feedback (*an obstacle impeded me from going right*) showed it didn't work, so I programmed a new skill to *navigate_with_obstacles* (S_2) which corrects the previous one and is chosen to go to the goal (I_3)

The Ladder of Intentions

Static view: **Voyager**

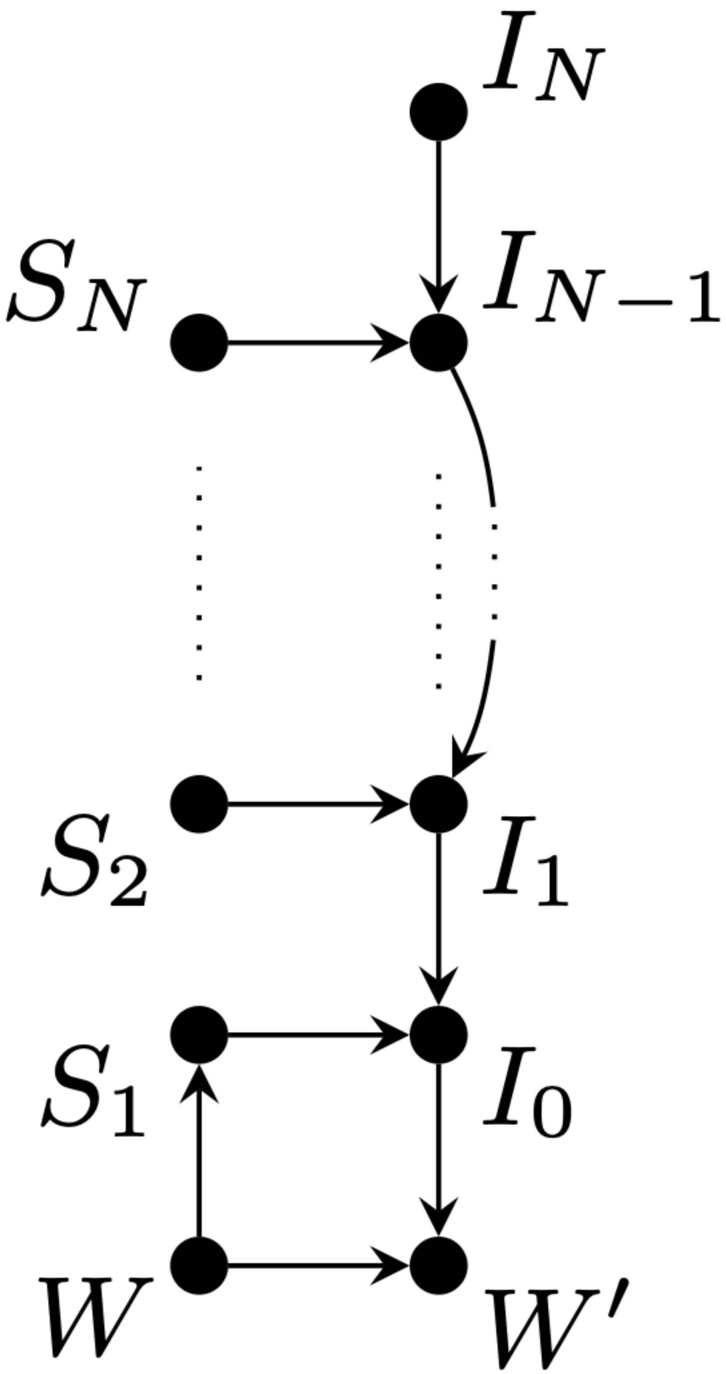




- Why did you program a new skill (I_2) at t_1?
  - Given feedback ($S_2 \subset S_3$) it seemed like a new skill was needed to solve the newly identified task of navigating with obstacles (S_3), and I want to solve new tasks (I_3)

The Ladder of Intentions

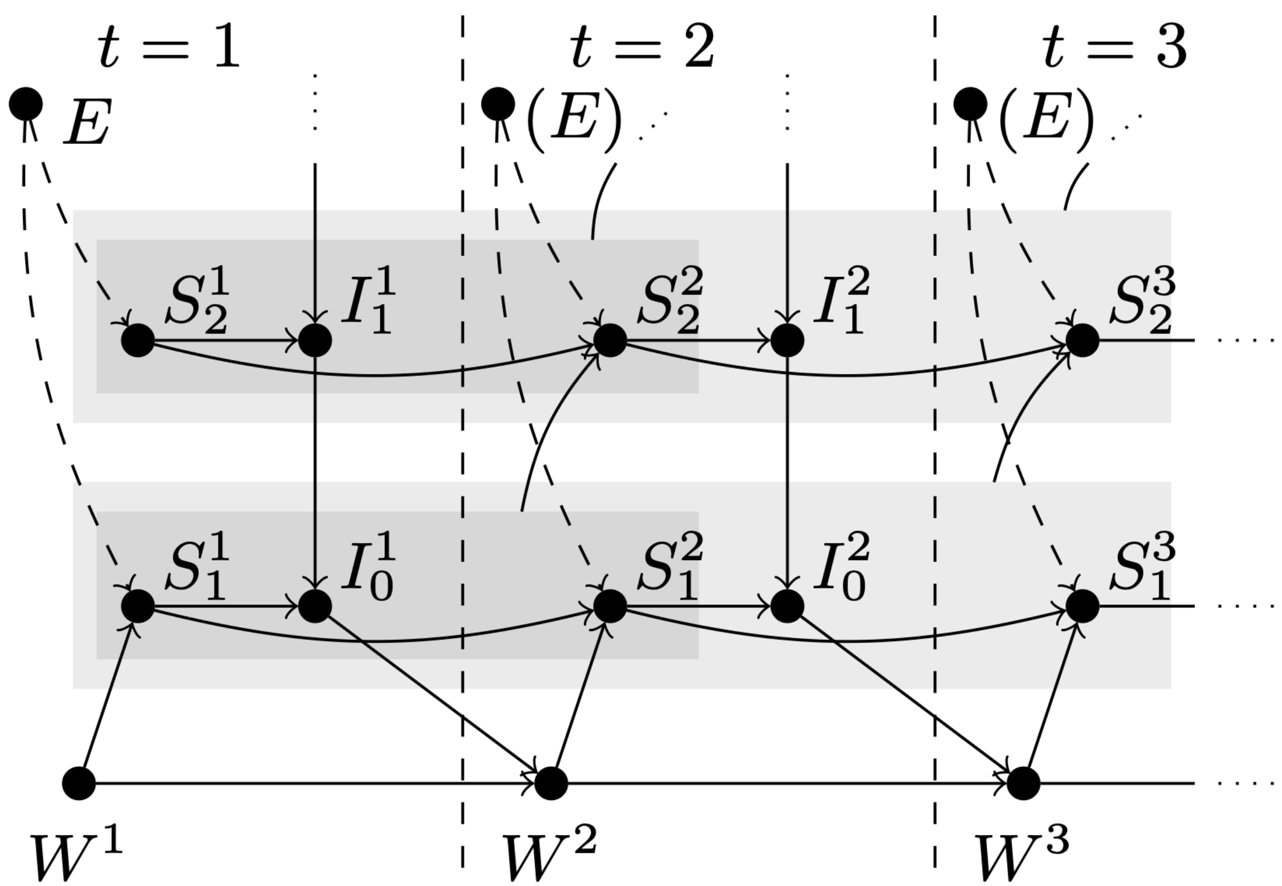
Static view



- The main issue is **choosing a non-arbitrary separation that will continue to work for new architectures**
  - We chose the idea of **statements that reify or include other statements** as being the separator, and starting at observations of the environment and actions
  - Environmental observations belong on the 1st level, whilst a statement referring to how observations would change when taking actions (ie consequences of action) will belong on the second, and statements referring to how changing a course of action will affect how I learn about consequences will belong on the third
- This seems overcomplicated, but when using the language of an architecture it is more easy to determine

The Ladder of Intentions

Dynamic view

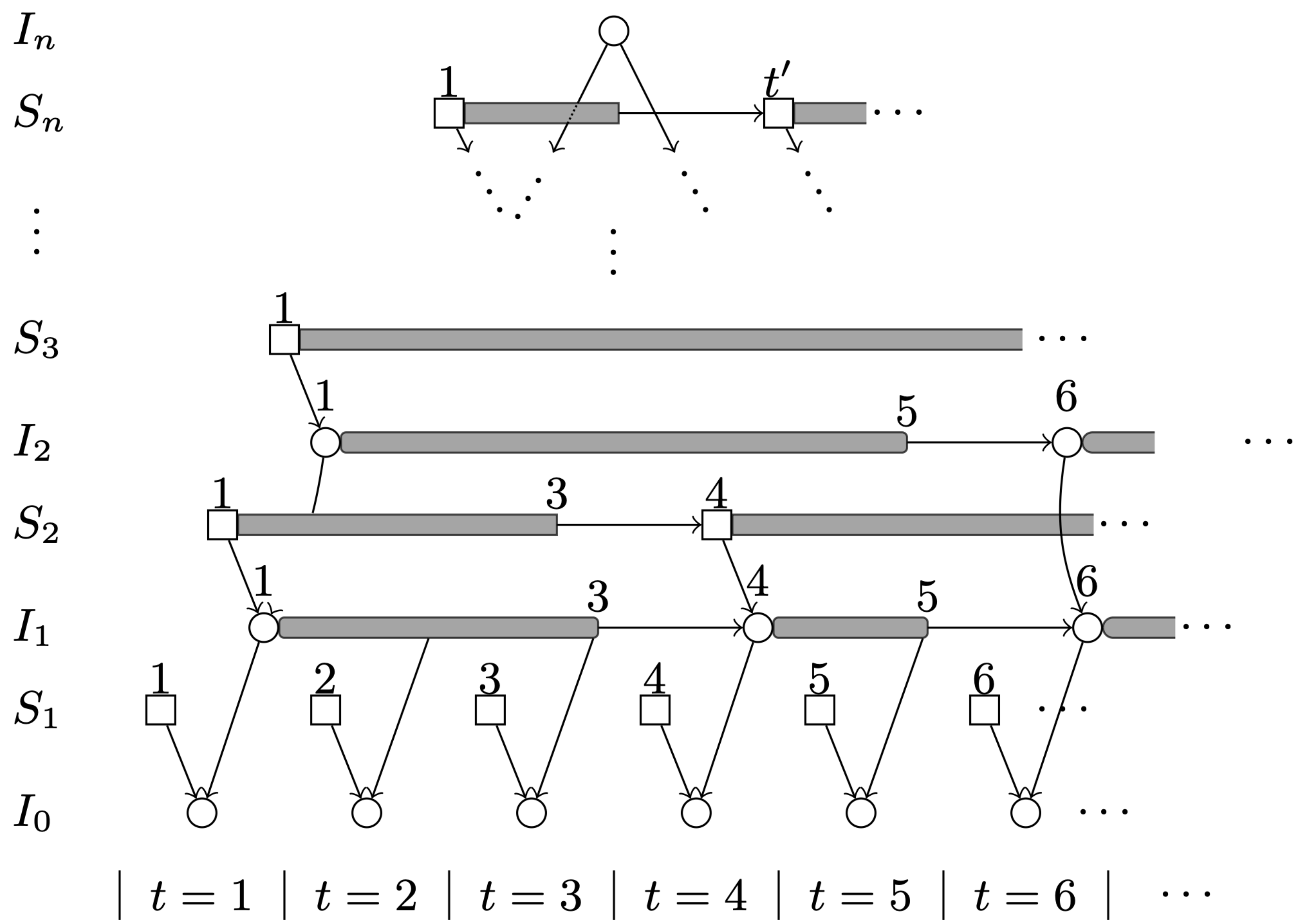


- If the model has learning, it is the case that ***observations* of a lower level cause some changes on upper levels**, e.g. seeing an unexpected observation may make us reconsider consequences of actions, and so on.

- This means that ***learning* statements are generated by compiling experiences of lower levels**

The Ladder of Intentions

Dynamic view



- In consequence: **intentions and beliefs are fluents**
- They hold at some times, until some experience from a lower level forces us to reconsider...
  - ...by updating statements...
  - ... thus producing a cascade of changes in intentions downward

n	REINFORCE [41]	Q-learning [40]	BDI [3]	FB Representation [37]	Voyager [38]
1	Standard+Reward Policy ($a \sim P^{\pi}(a\|s)$)	Standard+Reward Policy ($argmax_a Q(s,a)$)	Standard Plan	Standard+Reward $argmax_a max_z F(s,a,z)B(z,s')$	Standard + Errors + API (Mindflayer) Program/skill
2	Empirical $v = Q(s,a)$, $\nabla_\theta log \pi_\theta(s,a) v$ Policy training algorithm	Estimated Q-function ($Q(s,a)$) Action-sampling policy generator	World model (*e.g.* PDDL domain file), desires Means-ends reasoning to solve the goal	Successor Functions (F,B), desires/rewards of states FB explorer (off-line); FB exploiter	Available skills, Possible tasks, LLM[8], Feedback Skill generator/corrector to solve a task
3		$\varepsilon = P(Q(s,a) < Rand\, a)$ Explore/exploit mechanism	Desire prioritisations Deliberation (goal selection)	Given current goal Goal selector	Task list priorisation[9], directive prompt Automatic curriculum planner loop
4			Values over desire prioritisation (when used) Value reasoner (*e.g.* water tanks [13])		

Thanks for attending!

Any questions?

This paper

XAI in AVs

Intentional policy graphs (main track)

Look for us at Poster/Technical Session 3 of the main conference (paper 999) for more on explainability!