Verifying Autonomous Agents in Dynamic Environment

Mengwei Xu

University of Glasgow
Autonomous Agents

Definition:

An entity

which **perceives** its environment,
which **deliberates** accordingly,
which **takes actions** autonomously,

in order to achieve some objectives
An entity which **perceives** its environment, which **deliberates** accordingly, which **takes actions** autonomously, in order to achieve some objectives.
Autonomous Agents

Reasoning Cycle

- **Perceive**
  - What is the world like now

- **Deliberate**
  - What should I do now

- **Act**
  - Action to be done

**Sensors** → **Percepts** → **Environment**

**Actuators** ↔ **Actions**
Autonomous Agents
Reasoning Cycle

while true:
  Perceive
  Deliberate
  Act
end
**Autonomous Agents**

Beliefs-Desires-Intentions (BDI) Framework

1. **Event Selection**: Sensors gather information on the current state of the world.
2. **Plan Selection**: Based on the beliefs about the current world state, the agent selects relevant plans.
3. **Intention Selection**: From the selected plans, the agent chooses the most appropriate intentions.

**What is the world like now**

**What should I do now**

**Execute Intentions**: The agent executes the chosen intentions to effect changes in the world.

**Action to be done**

**Beliefs**

- **Check Context**

**Pending Events**

- **New**

**Relevant Plans**

- **New**

**Applicable Plans**

- **New**

**Plan Library**
(a) simulation: one run of agent behaviour in one environment;
(b) existing verification approaches: all possible agent behaviours in one environment
(c) our proposed approach: verify all possible agent behaviours in all possible environments
Autonomous Agents

Verification Framework

agent

Sensors

What is the world like now

BDI Semantics

Action to be done

Actuators

dynamics

percepts

dynamics

actions

environment
Autonomous Agents

Verification Framework

while true:
  environment update
end

normal environment changes such as from p to ¬ p

while true:
  environment update
end

Autonomous Agents

Verification Framework

while true:
  environment update
end

normal environment changes such as from p to ¬ p
Autonomous Agents
Verification Framework

while true:
    environment update
    perceive

end
Autonomous Agents
Verification Framework

**while** true:
  
environment update
  perceive
    **while** true:
      one agent semantic step
    end
end
Autonomous Agents
Verification Framework

while true:
    environment update
    perceive
    while true:
        one agent semantic step
    end
    act
end
while true:
    environment update
    perceive
while true:
    one agent semantic step
end
act
end
Autonomous Agents

Examples

1. Plan library
2. e_patrol_init : true ← goal(detection, e_patrol_task, false); return
3. e_patrol_task : true ← goal(harsh_weather, e_patrol, false); e_pause
4. e_patrol : true ← patrol
5. e_pause : harsh_weather ∧ ¬parked ← activate_parking; wait
6. e_pause : harsh_weather ∧ parked ← wait

7. initial environment state
8. $\Theta_0 = \{-a, -b, -c, -d, e\_patrol\_init\}$

9. environment transition function

$$\delta(\Theta) = \begin{cases} 
\{\Theta, (\Theta \setminus \{a\}) \cup \{a\}, (\Theta \setminus \{b\}) \cup \{b\}, (\Theta \setminus \{a, -b\}) \cup \{a, b\}\} & \text{if } -a \land -b \in \Theta \\
\{\Theta, (\Theta \setminus \{a\}) \cup \{a\}\} & \text{if } -a \land b \in \Theta \\
\{\Theta, (\Theta \setminus \{b\}) \cup \{b\}\} & \text{if } a \land -b \in \Theta \\
\{\Theta\} & \text{if } a \land b \in \Theta \\
\{((\Theta \setminus \{b, c\}) \cup \{-b, -c\}\} & \text{if } b \land c \in \Theta 
\end{cases}$$

where $a = \text{detection}$, $b = \text{harsh\_weather}$, $c = \text{waited}$ (the effect of action wait) and $d = \text{returned}$ (the effect of action return).
Autonomous Intelligent Agents

Examples

<table>
<thead>
<tr>
<th></th>
<th>Design in Fig. 5</th>
<th>Design in Fig. 6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Safety Property</td>
<td>False</td>
<td>True</td>
</tr>
<tr>
<td>Completion Property</td>
<td>False</td>
<td>False</td>
</tr>
<tr>
<td>Response Property</td>
<td>True</td>
<td>True</td>
</tr>
<tr>
<td>Commitment Property</td>
<td>True</td>
<td>True</td>
</tr>
<tr>
<td>States</td>
<td>167</td>
<td>282</td>
</tr>
<tr>
<td>Transitions</td>
<td>242</td>
<td>373</td>
</tr>
<tr>
<td>Build time (s)</td>
<td>54.05</td>
<td>128.89</td>
</tr>
<tr>
<td>Rule applications</td>
<td>1306</td>
<td>2152</td>
</tr>
</tbody>
</table>

Table I: Properties checked: where safety property is $\neg E[F(\varphi_1 \land \neg \varphi_2 \land (XX\varphi_2))]$, completion property $A[F\varphi_3]$, response property $A[\varphi_4 \implies F\varphi_5]$, and commitment property $A[\varphi_5 \implies F\varphi_6]$.

$\varphi_1 = \text{harsh\_weather}$  $\varphi_2 = \text{returned}$
Autonomous Agents

Future Work

while true:
    environment update
    perceive
    while true:
        one agent semantic step
        yet to be implemented
    end
    act
end

1. normal environment changes such as from $p$ to not $p$
2. the request of new events
3. the command of event status changes
Questions