

Foundations of Artificial Intelligence

A4. Introduction: Rational Agents

Malte Helmert

University of Basel

February 18, 2026

Foundations of Artificial Intelligence

February 18, 2026 — A4. Introduction: Rational Agents

A4.1 Systematic AI Framework

A4.2 Example

A4.3 Rationality

A4.4 Summary

Introduction: Overview

Chapter overview: introduction

- ▶ A1. Organizational Matters
- ▶ A2. What is Artificial Intelligence?
- ▶ A3. AI Past and Present
- ▶ **A4. Rational Agents**
- ▶ A5. Environments and Problem Solving Methods

A4.1 Systematic AI Framework

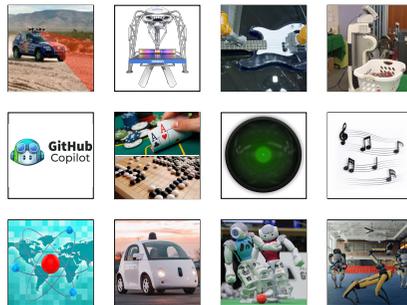
Systematic AI Framework

so far we have seen that:

- ▶ AI systems act rationally



- ▶ AI systems applied to wide variety of challenges



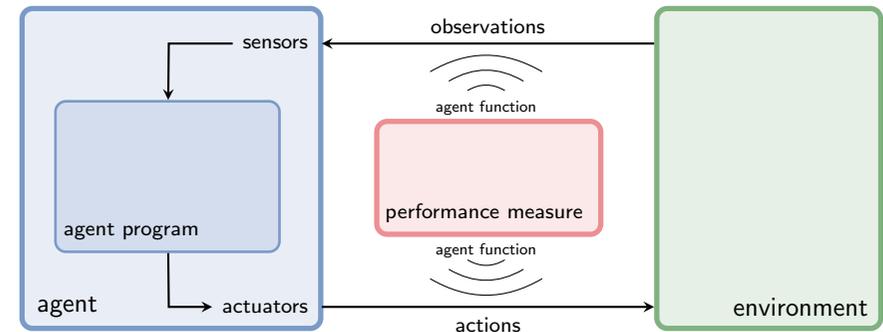
now: describe a **systematic framework** that

- ▶ captures this **diversity of challenges**
- ▶ includes an entity that **acts** in the environment
- ▶ determines if the agent acts **rationally** in the environment

Systematic AI Framework

so far we have seen that:

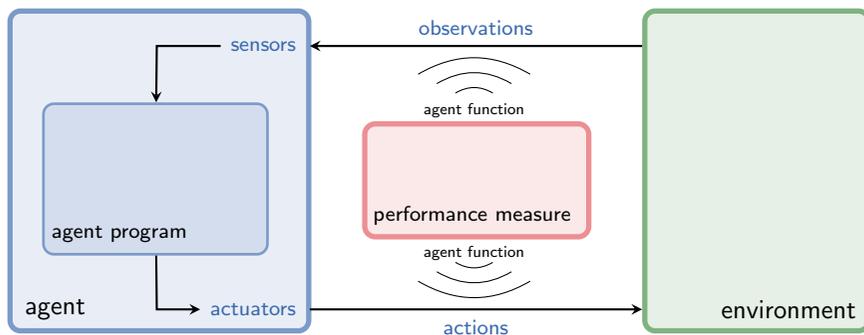
- ▶ AI systems act rationally



now: describe a **systematic framework** that

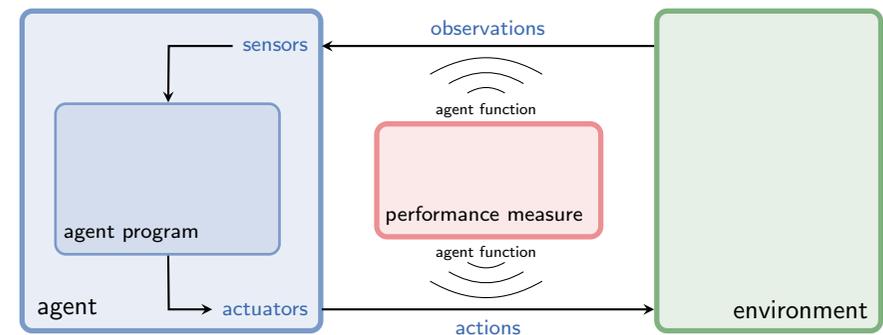
- ▶ captures this **diversity of challenges**
- ▶ includes an entity that **acts** in the environment
- ▶ determines if the agent acts **rationally** in the environment

Agent-Environment Interaction



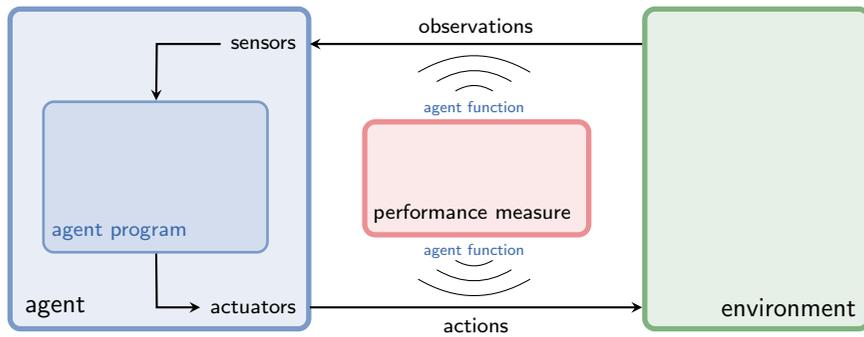
- ▶ **sensors**: physical entities that allow the agent to **observe**
- ▶ **observation**: data perceived by the agent's sensors
- ▶ **actuators**: physical entities that allow the agent to **act**
- ▶ **action**: abstract concept that affects the state of the environment

Agent-Environment Interaction



- ▶ **sensors** and **actuators** are not relevant for the course (↔ typically covered in courses on **robotics**)
- ▶ **observations** and **actions** describe the agent's capabilities (the **agent model**)

Formalizing an Agent's Behavior



1 as agent program:

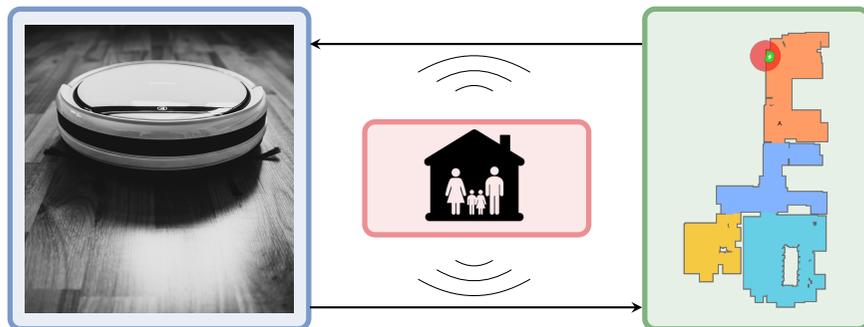
- ▶ internal representation
- ▶ specifics possibly *unknown* to outside
- ▶ takes *observation* as input
- ▶ outputs an *action*
- ▶ computed on physical machine (the *agent architecture*)

2 as agent function:

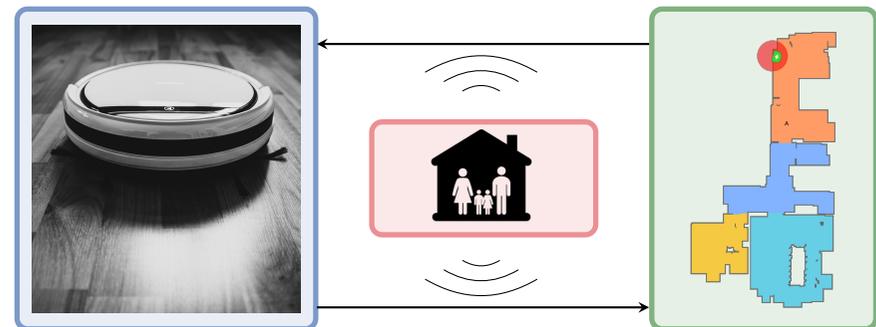
- ▶ external characterization
- ▶ maps *sequence of observations* to (probability distribution over) *actions*
- ▶ abstract mathematical formalization

A4.2 Example

Vacuum Domain

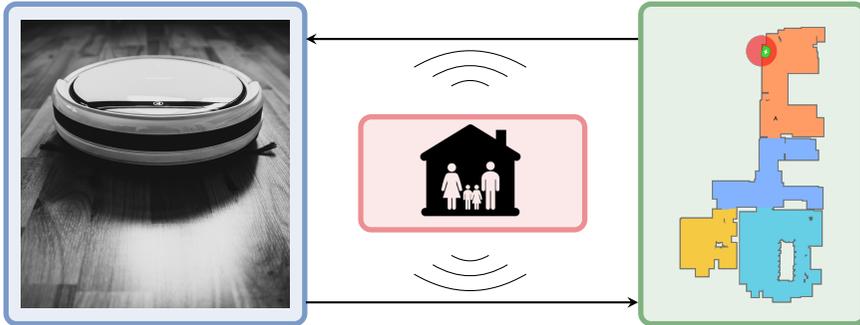


Vacuum Agent: Sensors and Actuators



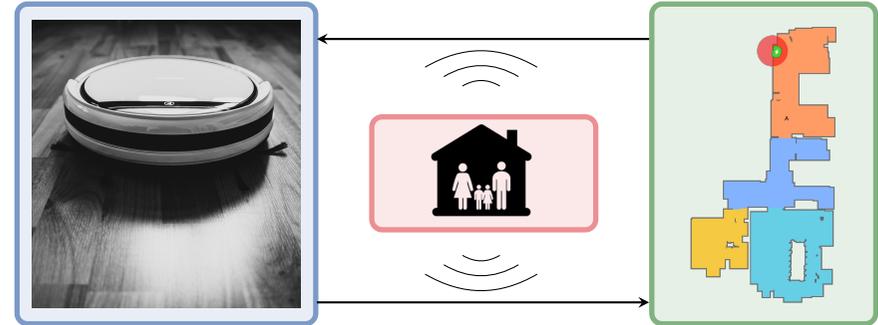
- ▶ **sensors:** cliff sensors, bump sensors, wall sensors, state of charge sensor, WiFi module
- ▶ **actuators:** wheels, cleaning system

Vacuum Agent: Observations and Actions



- ▶ **observations:** current location, dirt level of current room, presence of humans, battery charge
- ▶ **actions:** move-to-next-room, move-to-base, vacuum, wait

Vacuum Agent: Agent Program

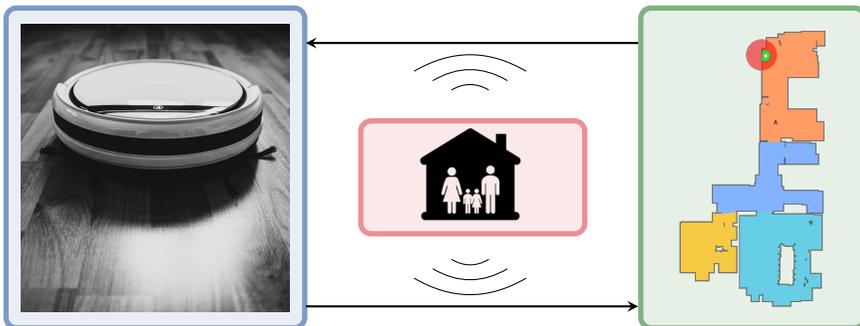


```

1 def vacuum-agent([location, dirt-level, owner-present, battery]):
2   if battery ≤ 10%: return move-to-base
3   else if owner-present = True: return move-to-next-room
4   else if dirt-level = dirty: return vacuum
5   else: return move-to-next-room

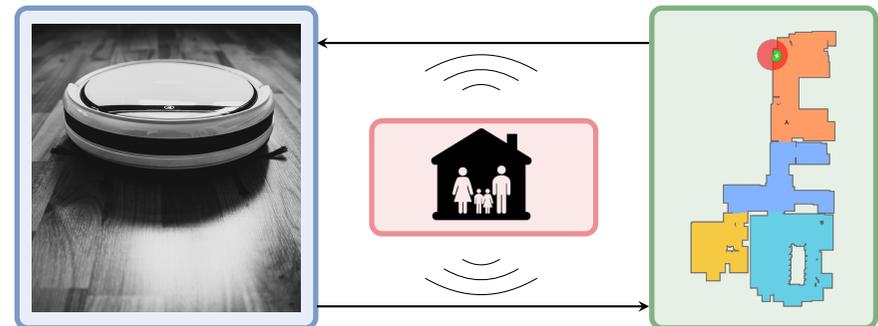
```

Vacuum Domain: Agent Function



observation sequence	action
⟨[blue, clean, False, 100%]⟩	<i>move-to-next-room</i>
⟨[blue, dirty, False, 100%]⟩	<i>vacuum</i>
⟨[blue, clean, True, 100%]⟩	<i>move-to-next-room</i>
...	...
⟨[blue, clean, False, 100%], [blue, clean, False, 90%]⟩	<i>move-to-next-room</i>
⟨[blue, clean, False, 100%], [blue, dirty, False, 90%]⟩	<i>vacuum</i>
...	...

Vacuum Domain: Performance Measure



potential influences on **performance measure**:

- ▶ dirt levels
- ▶ noise levels
- ▶ energy consumption
- ▶ safety

A4.3 Rationality

Evaluating Agent Functions



What is the **right** agent function?

Rationality

rationality of an agent depends on performance measure (often: utility, reward, cost) and environment

Perfect Rationality

- ▶ for each possible observation sequence
- ▶ select an action which maximizes
- ▶ expected value of future performance
- ▶ given available information on observation history
- ▶ and environment

Perfect Rationality of Our Vacuum Agent

Is our vacuum agent **perfectly rational**?



depends on performance measure and environment, e.g.:

- ▶ Do actions reliably have the desired effect?
- ▶ Do we know the initial situation?
- ▶ Can new dirt be produced while the agent is acting?

Performance Measure

- ▶ specified by designer
- ▶ sometimes clear, sometimes not so clear
- ▶ significant impact on
 - ▶ desired behavior
 - ▶ difficulty of problem



Performance Measure

- ▶ specified by designer
- ▶ sometimes clear, sometimes not so clear
- ▶ significant impact on
 - ▶ desired behavior
 - ▶ difficulty of problem



Perfect Rationality of Our Vacuum Agent

consider performance measure:

- ▶ +1 utility for cleaning a dirty room

consider environment:

- ▶ actions and observations reliable
- ▶ world only changes through actions of the agent

our vacuum agent is perfectly rational

Perfect Rationality of Our Vacuum Agent

consider performance measure:

- ▶ -1 utility for each dirty room in each step

consider environment:

- ▶ actions and observations reliable
- ▶ yellow room may spontaneously become dirty

our vacuum agent is not perfectly rational

Rationality: Discussion

- ▶ perfect rationality \neq omniscience
 - ▶ incomplete information (due to limited observations) reduces achievable utility
- ▶ perfect rationality \neq perfect prediction of future
 - ▶ uncertain behavior of environment (e.g., stochastic action effects) reduces achievable utility
- ▶ perfect rationality is rarely achievable
 - ▶ limited computational power \rightsquigarrow bounded rationality

A4.4 Summary

Summary (1)

common metaphor for AI systems: **rational agents**

agent interacts with **environment**:

- ▶ sensors perceive **observations** about state of the environment
- ▶ actuators perform **actions** modifying the environment
- ▶ formally: **agent function** maps observation sequences to actions

Summary (2)

rational agents:

- ▶ try to maximize **performance measure (utility)**
- ▶ **perfect rationality**: achieve maximal utility in expectation given available information
- ▶ for “interesting” problems rarely achievable \rightsquigarrow **bounded rationality**