具身智能

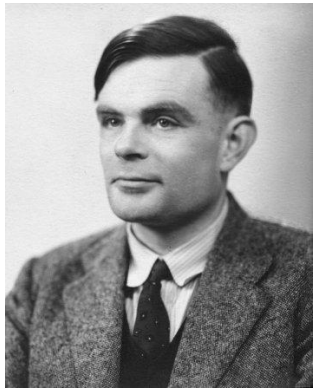
刘华平

2025年4月17日

- **背景：具身智能**
- 具身智能的体系
- 具身智能关键技术
- 探索与实践
- 面向具身智能的AIGC
- 总结

1 背景

➢ 人工智能的发展



COMPUTING MACHINERY AND INTELLIGENCE

By A. M. Turing

1. The Imitation Game

I propose to consider the question, "Can machines think?" This should begin with definitions of the meaning of the terms "machine" and "think." The definitions might be framed so as to reflect so far as possible the normal use of the words, but this attitude is dangerous, If the meaning of the words "machine" and "think" are to be found by examining how they are commonly used it is difficult to escape the conclusion that the meaning and the answer to the question, "Can machines think?" is to be sought in a statistical survey such as a Gallup poll. But this is absurd. Instead of attempting such a definition I shall replace the question by another, which is closely related to it and is expressed in relatively unambiguous words.

In the process of trying to imitate an adult human mind we are bound to think a good deal about the process which has brought it to the state that it is in. We may notice three components,

(a) The initial state of the mind, say at birth,

(b) The education to which it has been subjected,

(c) Other experience, not to be described as education, to which it has been subjected.

Instead of trying to produce a programme to simulate the adult mind, why not rather try to produce one which simulates the child's? If this were then subjected to an appropriate course of education one would obtain the adult brain. Presumably the child-brain is something like a note-book as one buys it from the stationers. Rather little mechanism, and lots of blank sheets. (Mechanism and writing are from our point of view almost synonymous.) Our hope is that there is so little mechanism in the child-brain that something like it can be easily programmed. The amount of work in the education we can assume, as a first approximation, to be much the same as for the human child.

We may hope that machines will eventually compete with men in all purely intellectual fields. But which are the best ones to start with? Even this is a difficult decision. Many people think that a very abstract activity, like the playing of chess would be best. It can also be maintained that it is best to provide the machine with the best sense organs that money can buy, and then teach it to understand and speak English. This process could follow the normal teaching of a child. Things would be pointed out and named, etc. Again I do not know what the right answer is, but I think both approaches should be tried.

We can only see a short distance ahead, but we can see plenty there that needs to be done.

我们的目光所及，只是不远的前方。

早在1950年，图灵就曾在他的经典论文《计算机械与智能》中提到，人工智能的未来将走向两条路：一条是“计算的智能体”，即像阿尔法狗一样依赖强大计算力的大模型；另一条路则是“具身智能”，机器拥有像婴儿一样的感知能力，通过与环境的互动和学习不断进化。换句话说，具身智能并不是最近才火的概念，它的源流可以追溯到人工智能的最早理论。

1 背景

➢ 什么是具身智能

身体利用感知-运动系统在与环境交互过程中产生智能

The embodiment hypothesis is the idea that intelligence emerges in the interaction of an agent with an environment and as a result of sensorimotor activity.

——Smith and Gasser

具身认知

- ✓**Be Multimodal**
- ✓**Be Incremental**
- ✓**Be Physical**
- ✓**Explore**
- ✓**Be Social**
- ✓**Learn a Language**

具身人工智能

- **多模态感知，…**
- **持续学习，…**
- **物理交互，…**
- **灵活探索，…**
- **社交模仿，…**
- **语言学习，…**

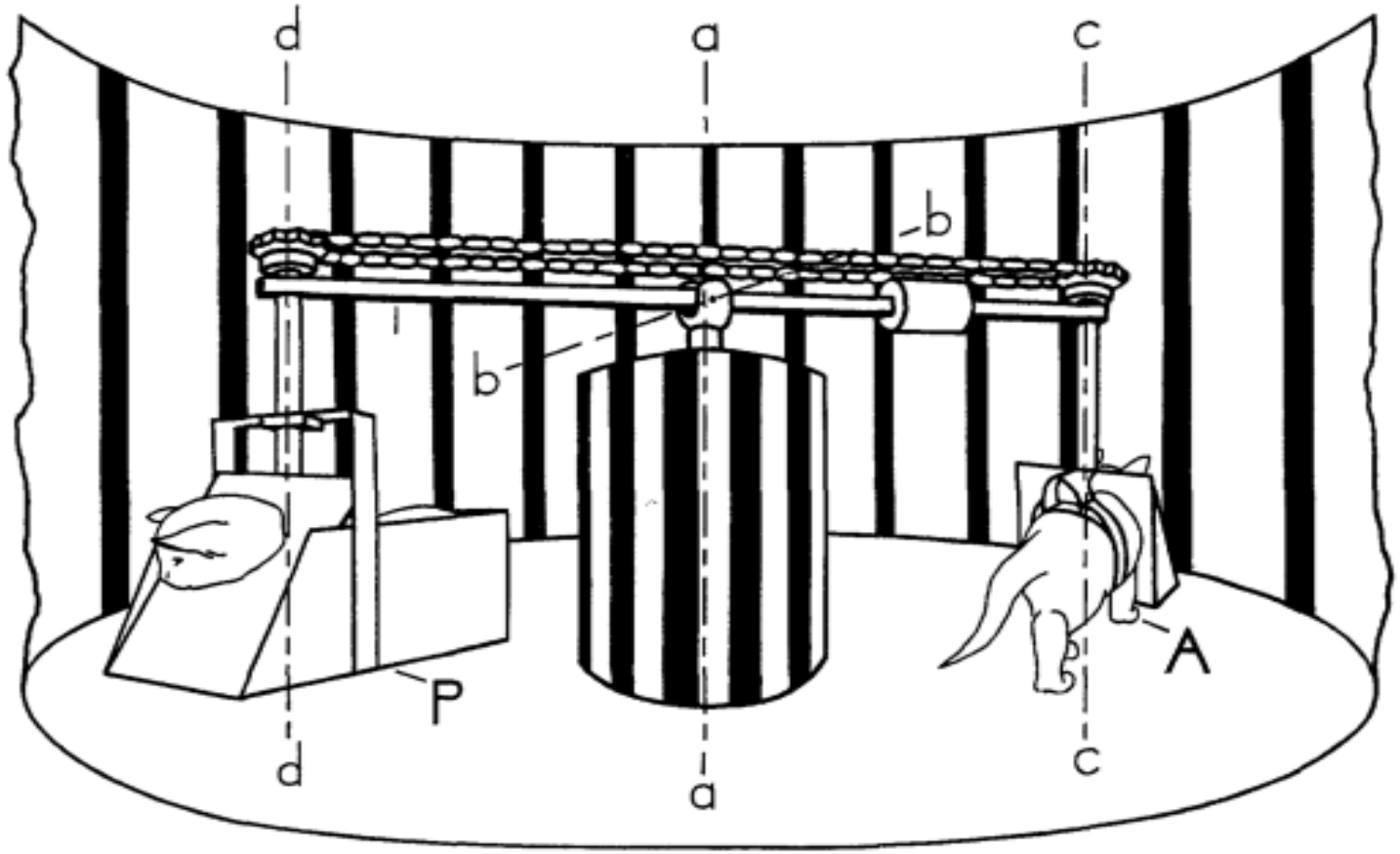
- L. Smith and M. Gasser, “The development of embodied cognition: six lessons from babies.,” Artificial Life, vol.11, no. 1-2, 2005.

1 背景

- 无处不在的具身智能

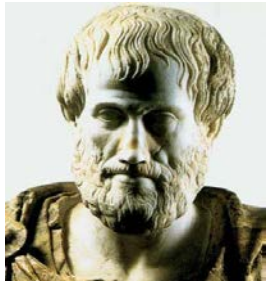
1 背景

- 极端的例子



1 背景

➢ 很长的过去



Aristotle
（公元前384-前322）

身体的主动性首先在于“通过触摸而感觉”

Charles Robert Darwin
（1809-1882）

心境和动作之间的联系就是情感的真正含义

Claude Bernard
（1813-1878）

“没有脑袋”的生理学

Walter Bradford Cannon
（1871-1945）

身体的智慧

Jean Piaget
（1896-1980）

动作是认识的源泉

James J. Gibson
(1904-1979)

感知与动作在与环境交互中密切联系

René Descartes
1596-1650

我思知我在

Martin Heidegger
（1889-1976）

《存在与时间》

比表征更基本的活动是操劳

Maurice Merleau-Ponty
（1908-1961）

知觉现象学：身体和主体其实是同一个实在

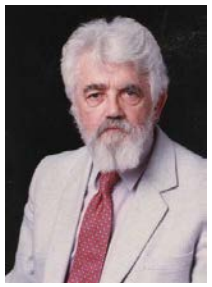
Norbert Wiener
（1894-1964）

控制论

1 背景

➢ 很短的历史

Marvin Lee Minsky
(1927-2016)

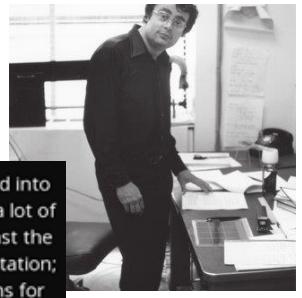


John McCarthy
(1927-2011)

达特茅斯会议
（1956）

John Haugeland
(1945-2010)

GOF AI



When David Marr at MIT moved into computer vision, he generated a lot of excitement, but he hit up against the problem of knowledge representation; he had no good representations for knowledge in his vision systems.

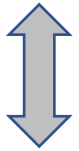
— Marvin Minsky —

David Marr
(1945-1980)

- 人类认知和思维的基本单元是符号，而认知过程就是在符号表示上的一种运算。
- 人和计算机都是物理符号系统,因此可以用计算机来模拟人的智能行为,即用计算机的符号操作来模拟人的认知过程。
- 启发式算法 → 专家系统 → 知识工程

心有余而力不足

The spirit is willing, but the flesh is weak

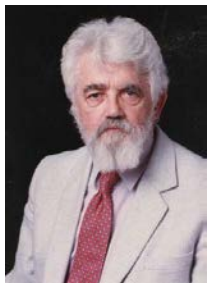


The vodka is good but the meat is spoiled

1 背景

➢ 很短的历史

Marvin Lee Minsky
(1927-2016)

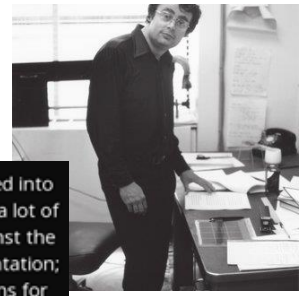


John McCarthy
(1927-2011)

达特茅斯会议
（1956）

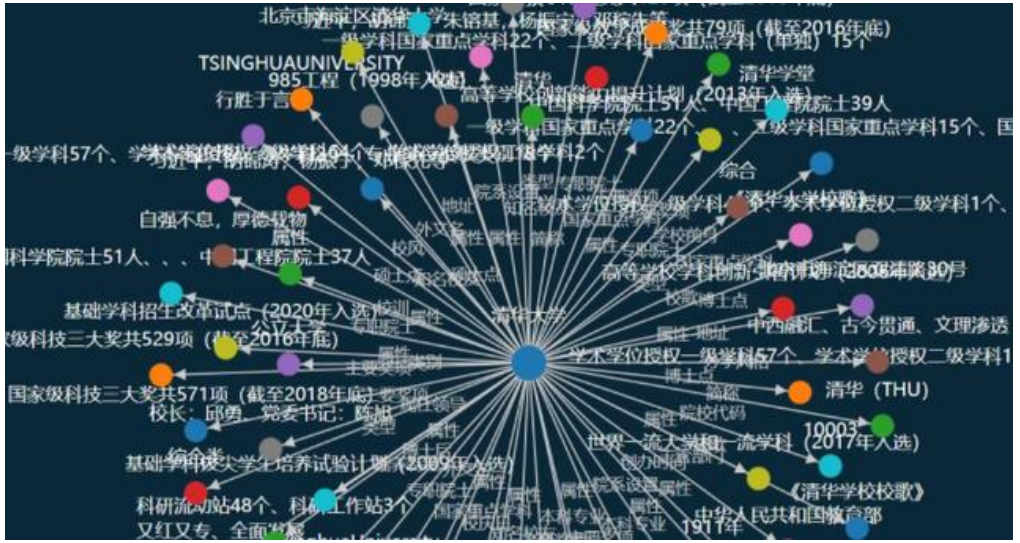
John Haugeland
(1945-2010)

GOF AI



David Marr
(1945-1980)

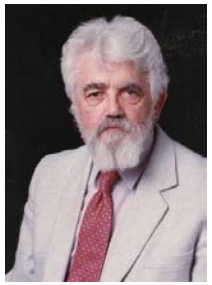
- 人类认知和思维的基本单元是符号，而认知过程就是在符号表示上的一种运算。
- 人和计算机都是物理符号系统,因此可以用计算机来模拟人的智能行为,即用计算机的符号操作来模拟人的认知过程。
- 启发式算法 → 专家系统 → 知识工程



1 背景

➢ 很短的历史

Marvin Lee Minsky
(1927-2016)



John McCarthy
(1927-2011)

达特茅斯会议
（1956）

John Haugeland
(1945-2010)

GOF AI

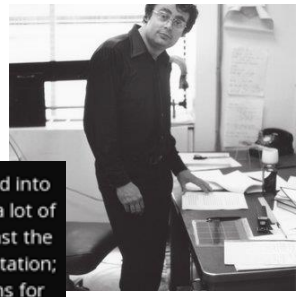
Hopfield
（1933-）

Rumelhart
（1942-2011）

Hilton
（1947-）

神经网络->反向传播->深度学习

- 数据挖掘
- 机器学习
- 模式识别



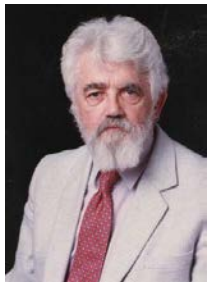
When David Marr at MIT moved into computer vision, he generated a lot of excitement, but he hit up against the problem of knowledge representation; he had no good representations for knowledge in his vision systems.

— Marvin Minsky —

David Marr
(1945-1980)

➢ 很短的历史

Marvin Lee Minsky
(1927-2016)



John McCarthy
(1927-2011)

达特茅斯会议
（1956）

John Haugeland
(1945-2010)

GOF AI

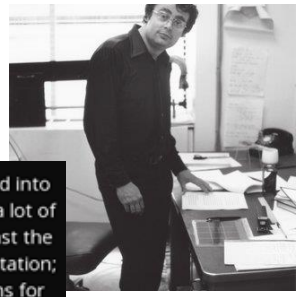
Hopfield
（1933-）

Rumelhart
（1942-2011）

Hilton
（1947-）

神经网络->反向传播->深度学习

- 数据挖掘
- 机器学习
- 模式识别

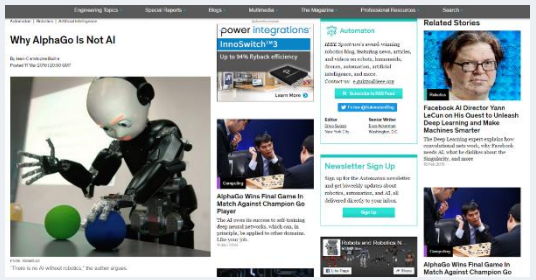


When David Marr at MIT moved into computer vision, he generated a lot of excitement, but he hit up against the problem of knowledge representation; he had no good representations for knowledge in his vision systems.

— Marvin Minsky —

David Marr
(1945-1980)

已经在“人类最后智力骄傲”上碾压人类的AlphaGo ...

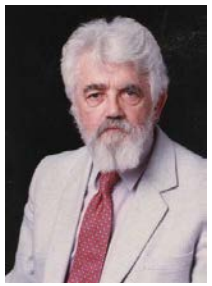


却连挪动一枚小小的棋子都需要人类帮助才能完成

1 背景

➢ 很短的历史

Marvin Lee Minsky
(1927-2016)

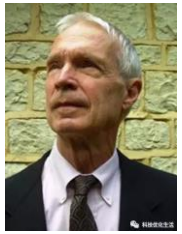


John McCarthy
(1927-2011)

达特茅斯会议
（1956）

John Haugeland
(1945-2010)

GOF AI



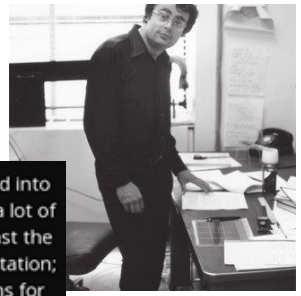
Hopfield
（1933-）

Rumelhart
（1942-2011）

Hilton
（1947-）

神经网络->反向传播->深度学习

- 模式识别
- 数据挖掘
- 机器学习



When David Marr at MIT moved into computer vision, he generated a lot of excitement, but he hit up against the problem of knowledge representation; he had no good representations for knowledge in his vision systems.

— Marvin Minsky —

David Marr
(1945-1980)

Pfeifer
(1947-)

Moravec
(1948-)

Brooks
(1954-)

Cangelosi
(1967-)

- 智能是具身化和情境化的，智能需要一个身体

- 机器人学
- 机构学
- 形态智能

- 强化

1 背景

➢ 莫拉维克悖论

要让电脑如成人般地下棋是相对容易的，但是要让电脑有如一岁小孩般的感知和行动能力却是相当困难的。

Hans Moravec

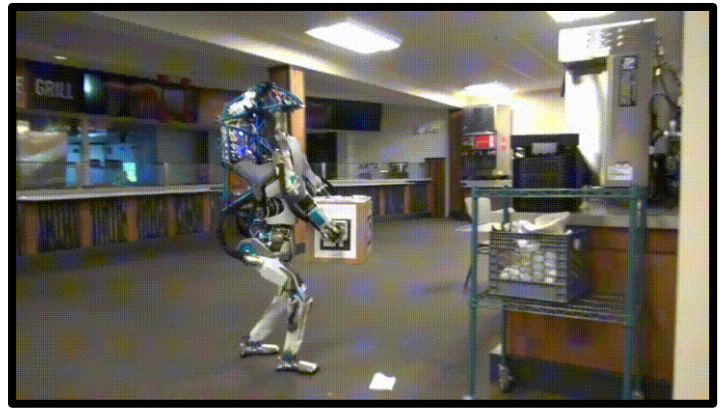
Donald Knuth

人工智能已经在几乎所有需要思考的领域超过了人类，但是在那些人类和其它动物不需要思考就能完成的事情上，还差得很远。

——Donald Knuth（2019）

1997：IBM的DeepBlue战胜卡斯帕罗夫

2016：DeepMind的AlphaGo”战胜李世石

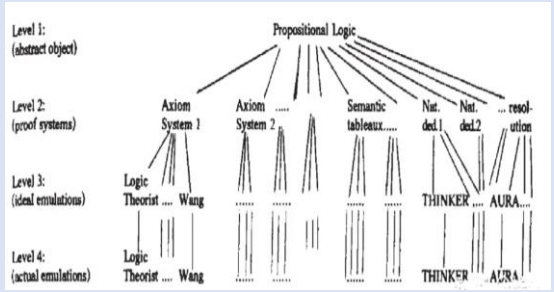


1 背景

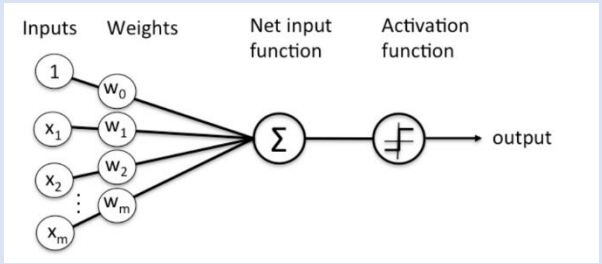
➢ 离身智能 v. s. 具身智能

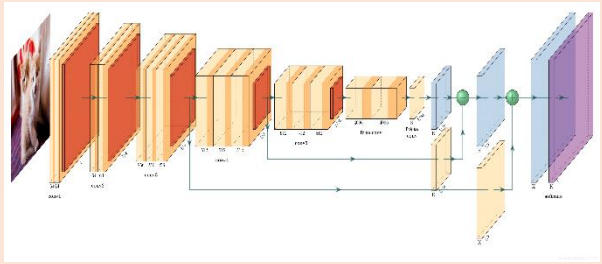
具身智能

符号主义：表示

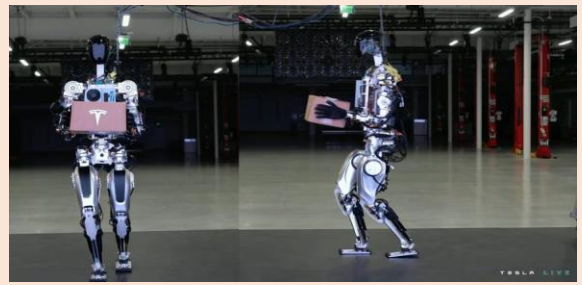


联结主义：计算





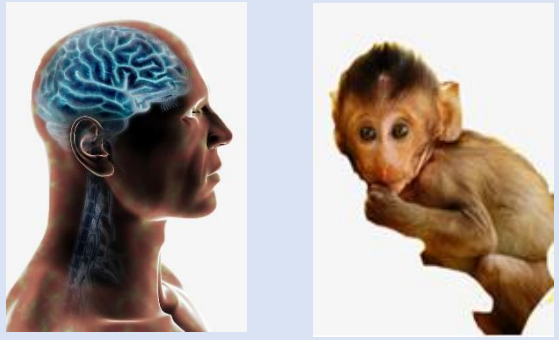
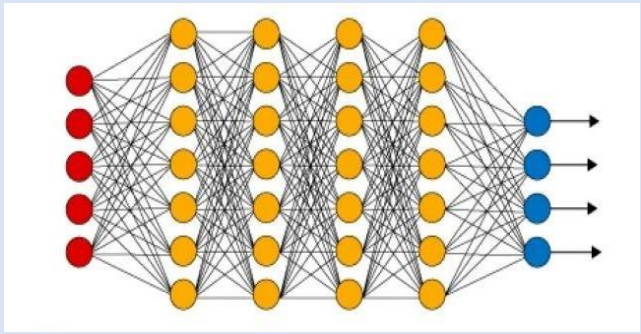
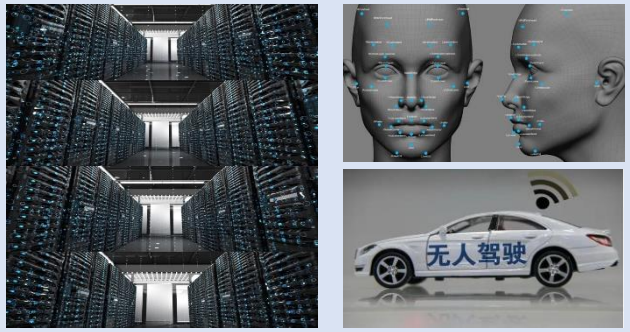
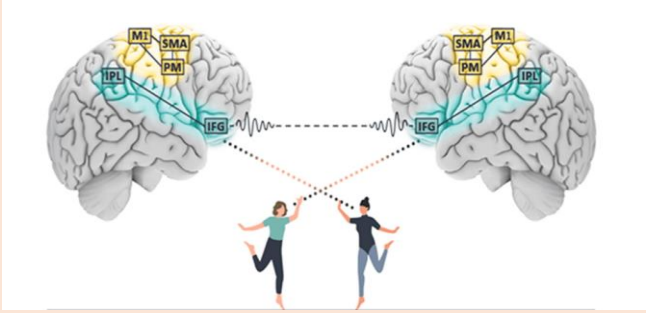
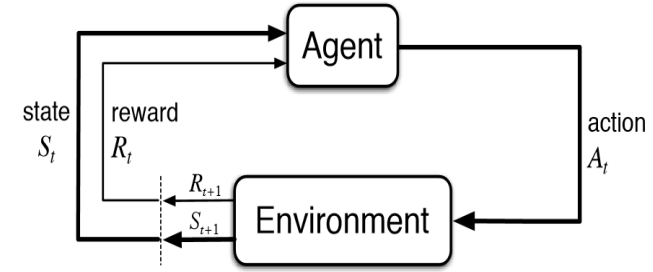
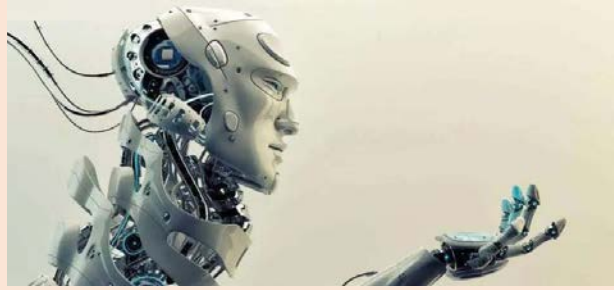
行为主义：交互



离身智能

1 背景

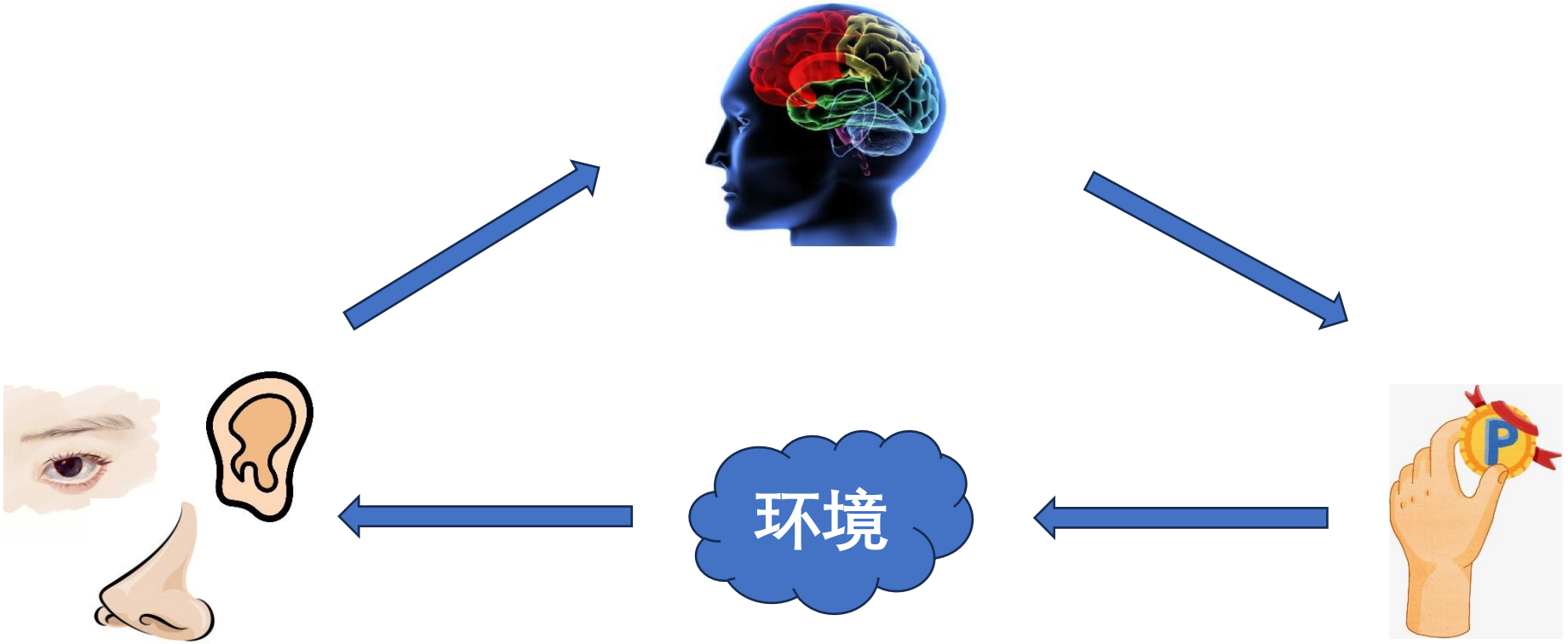
➢ 共同的目标

	理解生物系统	智能行为的抽象	设计人工智能体
离身智能			
具身智能			

1 背景

➢ 感知-动作回路才是认知的中心

一旦我们开始探索生物与环境之间的协调与交互，就很难再确定感知是什么时候结束的，而认知是什么时候开始的。

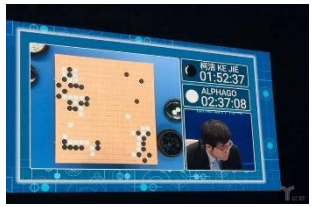


- 背景：具身智能
- 具身智能的体系
- 具身智能关键技术
- 探索与实践
- 面向具身智能的AIGC
- 总结

2 具身智能的体系

➢ 狭义与广义的具身智能

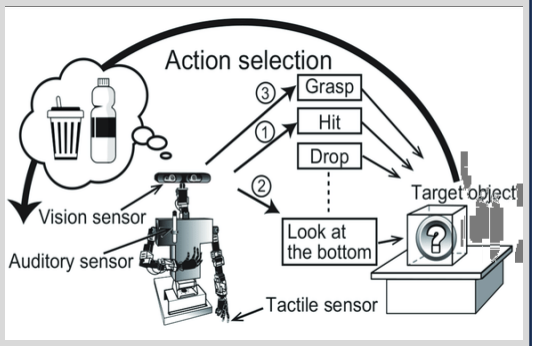
离 身 智 能　　具 身 智 能



2 具身智能的体系

➢ 具身智能的典型任务

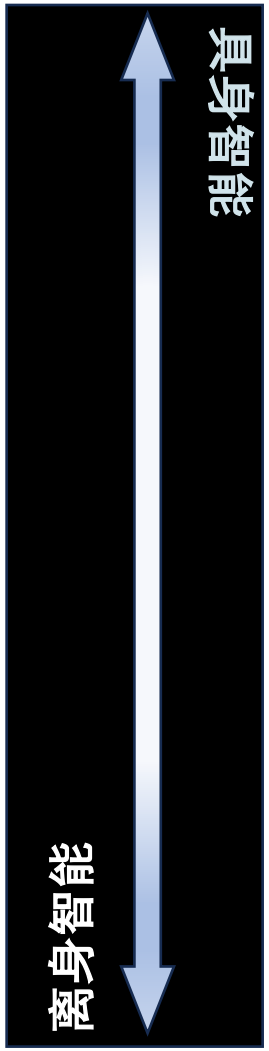
发现



学习

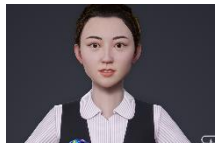
控制

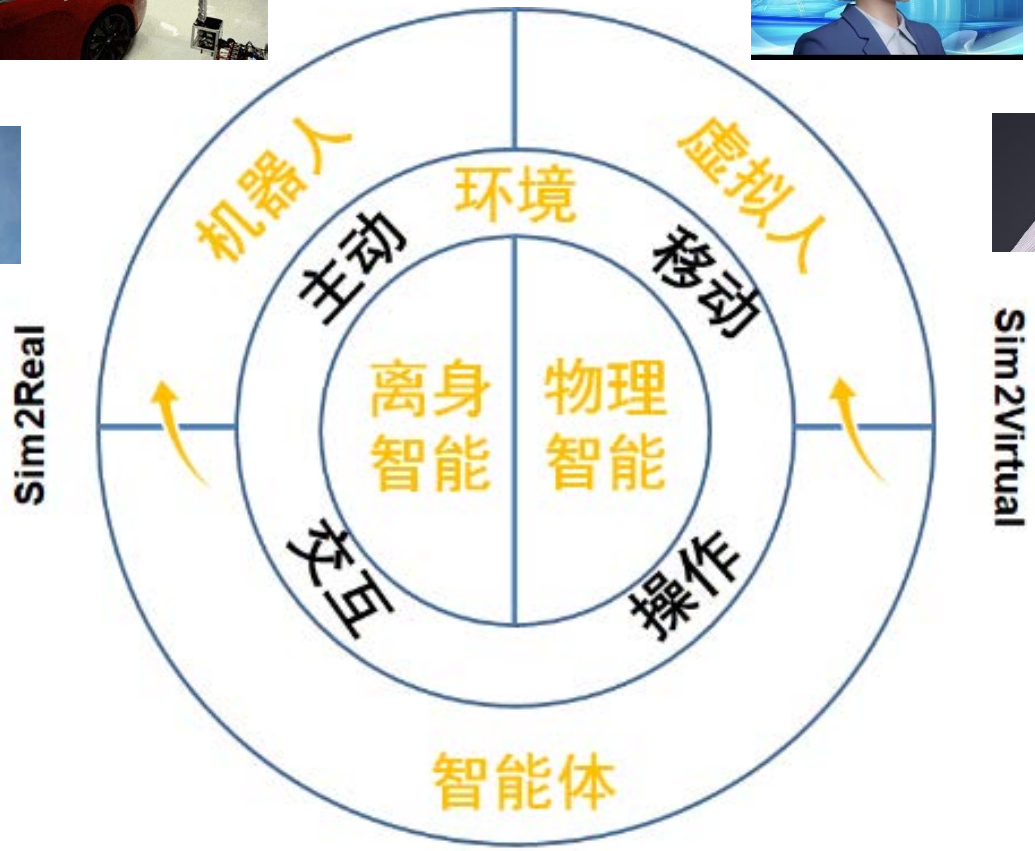
优化



2 具身智能的体系

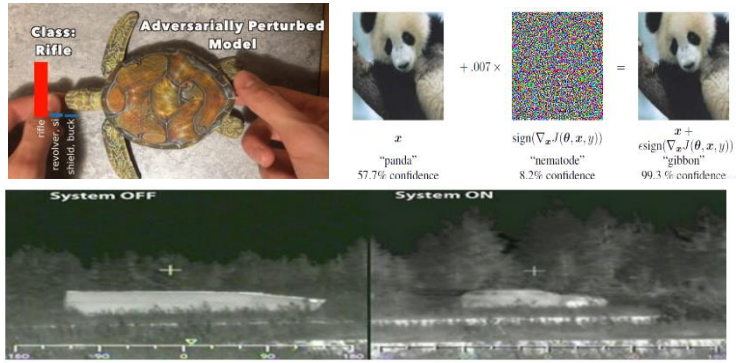
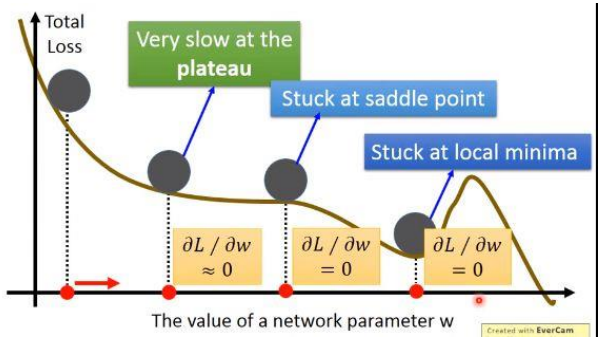
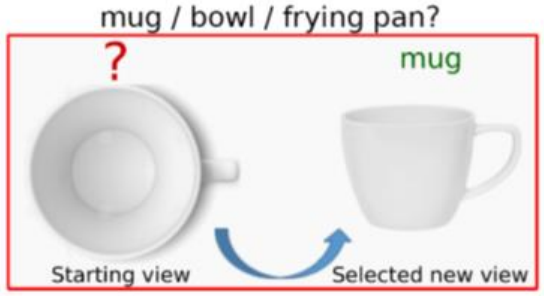
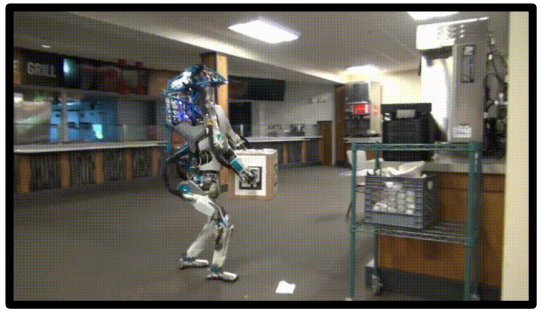
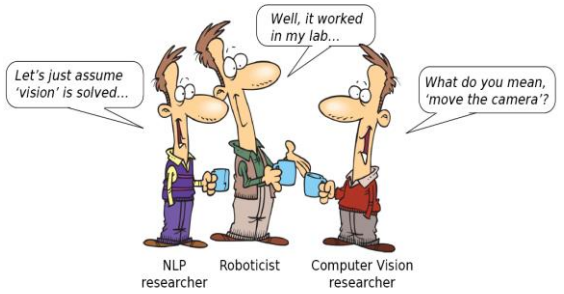
➢ 具身智能与机器人的关系





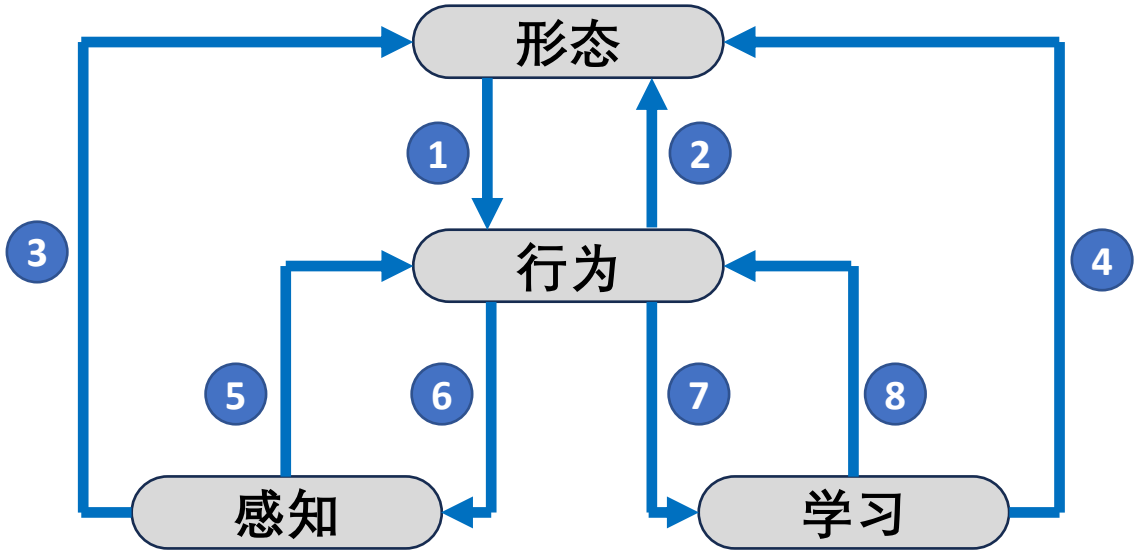
2 具身智能的体系

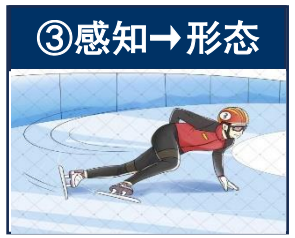
➢ 具身智能的优点、缺点与难点

	Good	Bad	Ugly
Disembodied Intelligence			
Embodied Intelligence			

2 具身智能的体系

➢ 具身智能的体系结构



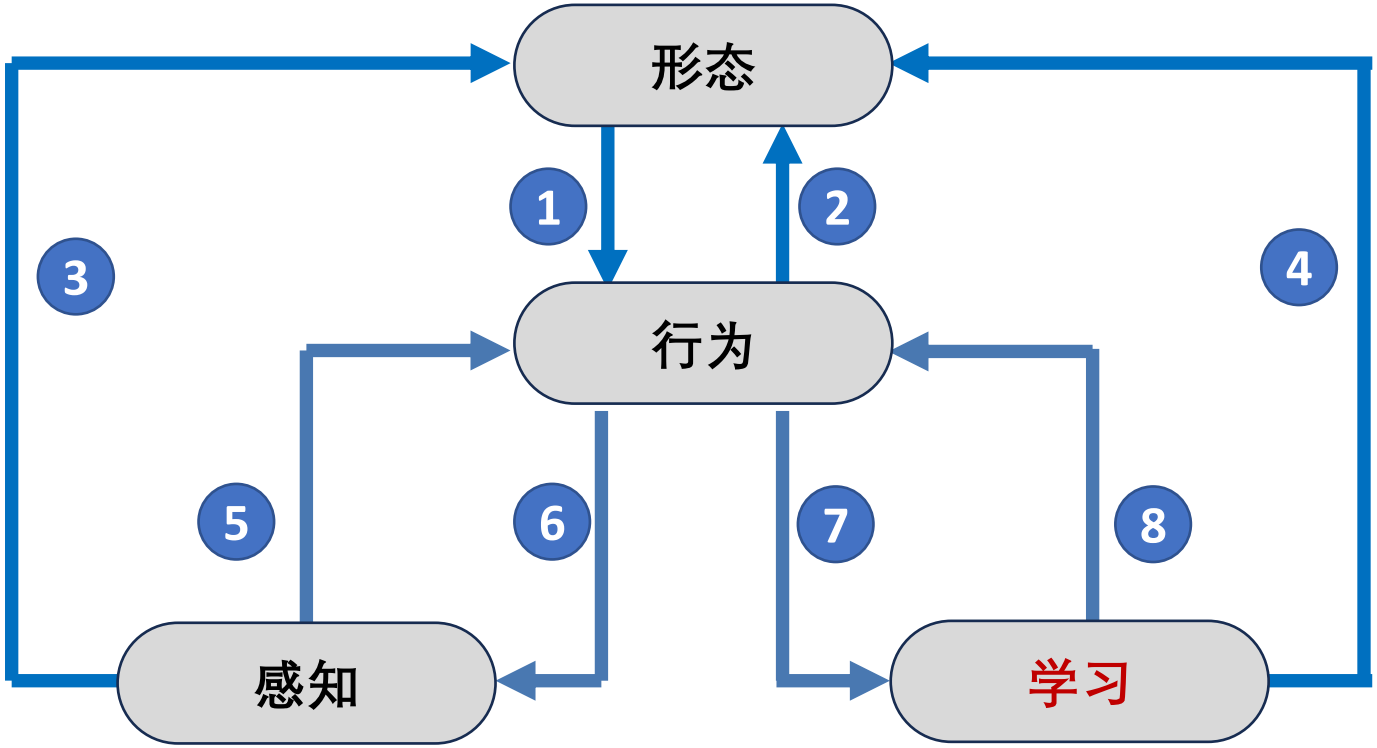


① 基于形态的行为生成　③ 基于感知的形态变换　⑤ 基于感知的行为生成　⑦ 基于行为的自主学习

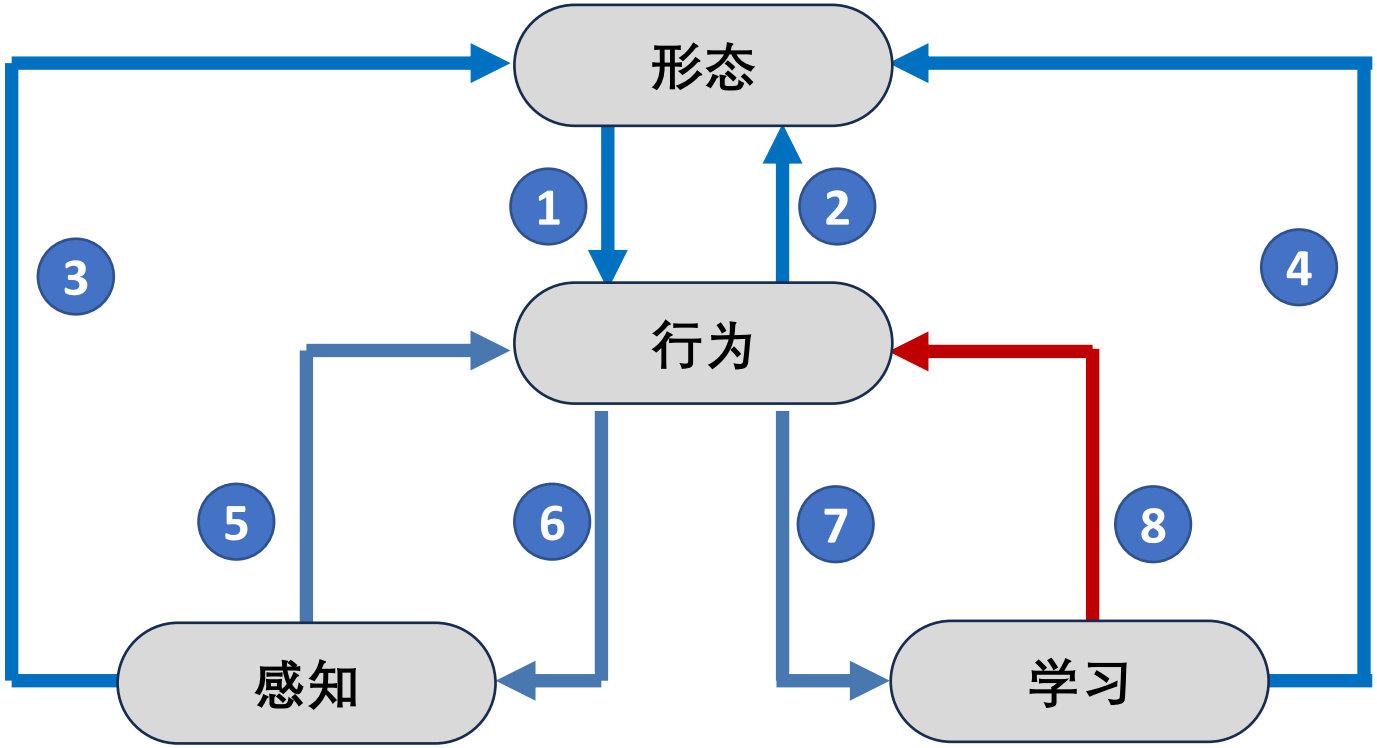
② 基于行为的形态控制　④ 基于学习的形态优化　⑥ 基于行为的主动感知　⑧ 基于学习的行为优化

- 背景：具身智能
- 具身智能的体系
- 具身智能关键技术
- 探索与实践
- 面向具身智能的AIGC
- 总结

3 关键技术



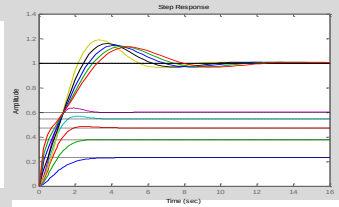
3 关键技术

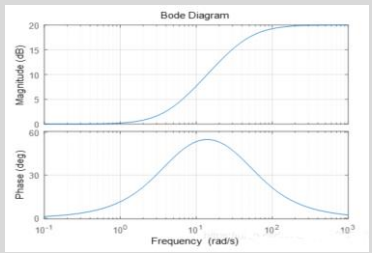


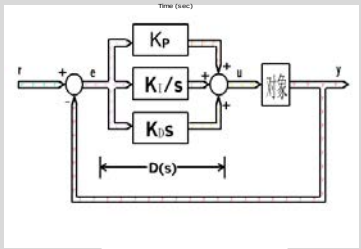
3.0 学习→行为：强化学习

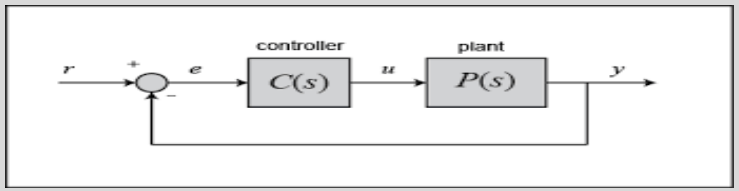
经典控制：经验

$$G_c(s) = \frac{s + \frac{1}{T}}{s + \frac{1}{\alpha T}} \quad (\alpha < 1)$$

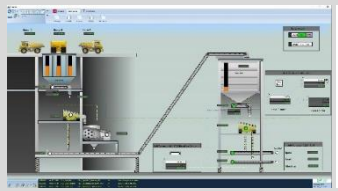


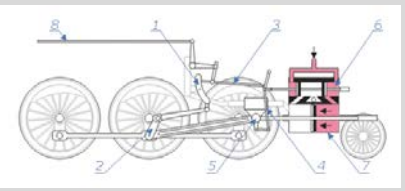






$$u_t = K_P e_t + K_I \int e_t dt + K_D \frac{de_t}{dt}$$

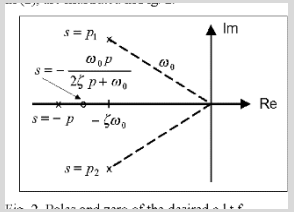




现代控制：模型

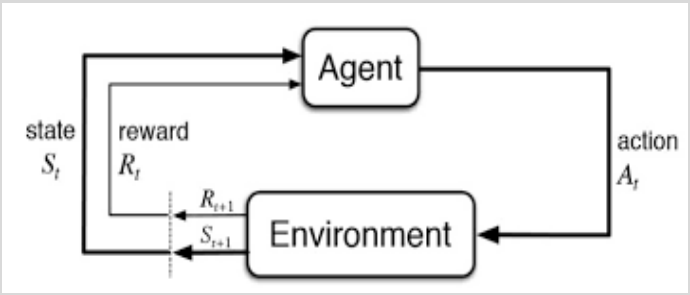
$$\begin{cases} \text{state equations} \\ \text{output equations} \end{cases} \begin{cases} \dot{x}_1 = f_1(x_1, x_2, \ldots, x_n, u_1, \ldots, u_m) \\ \vdots \\ \dot{x}_n = f_n(x_1, x_2, \ldots, x_n, u_1, \ldots, u_m) \\ y_1 = h_1(x_1, x_2, \ldots, x_n, u_1, \ldots, u_m) \\ \vdots \\ y_p = h_p(x_1, x_2, \ldots, x_n, u_1, \ldots, u_m) \end{cases}$$

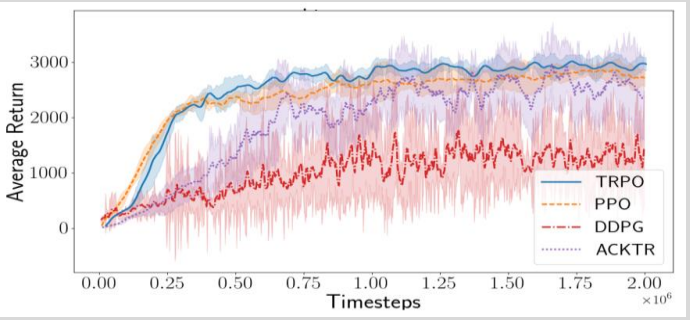
$$\begin{cases} \dot{x}_t = Ax_t + Bu_t \\ y_t = Cx_t + Du_t \end{cases}$$



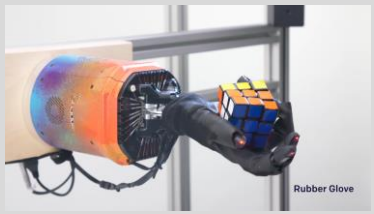
$$u_t = -Kx_t$$

智能控制：学习



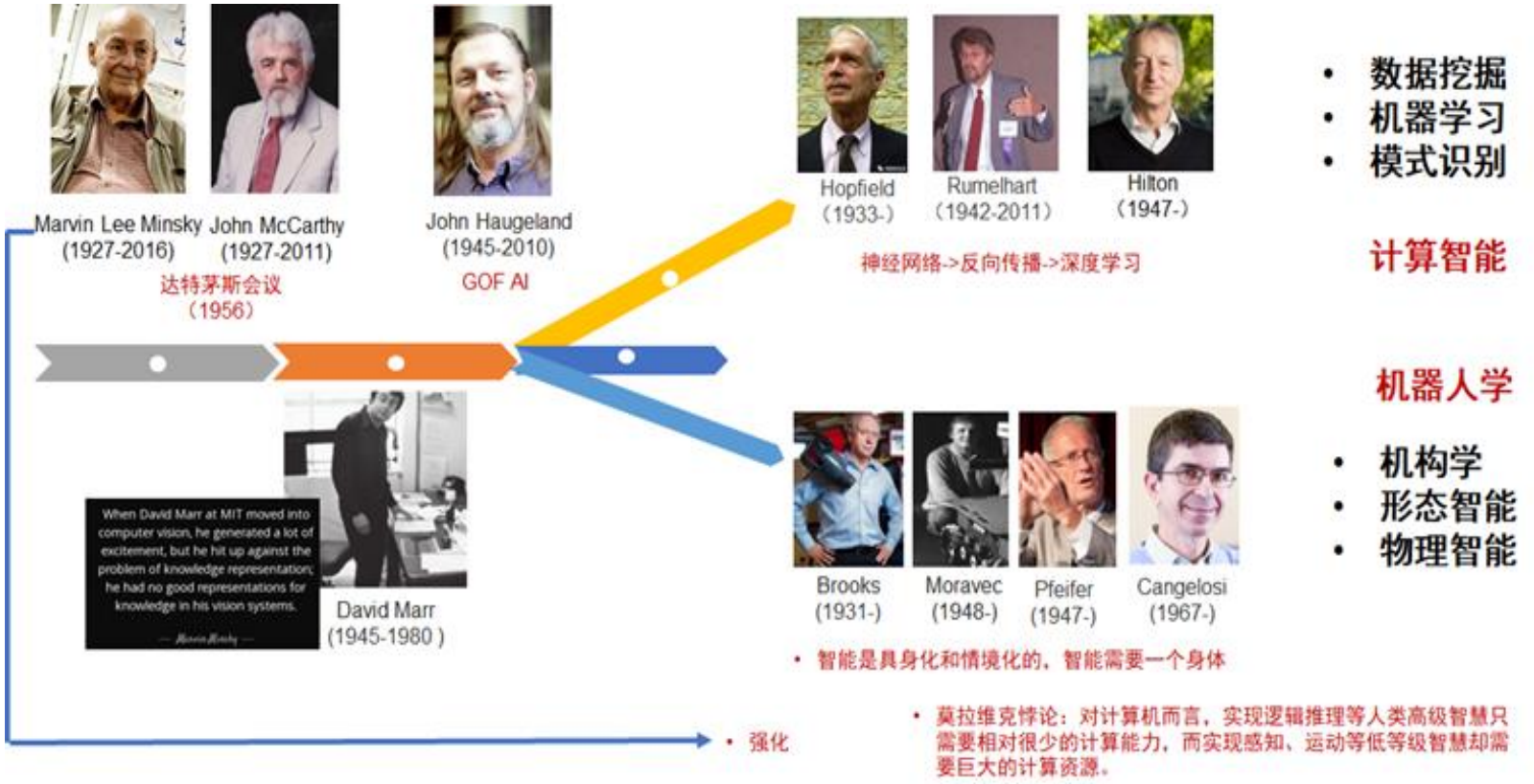


$$\pi_\theta(a_t \mid s_t)$$



3.0 学习→行为：强化学习

➢ 起源



人工神经网络
Artificial Neural Networks

Walter Pitts, 1943

深度学习
Deep Learning

Geoffrey Hinton, 2006

最优控制与动态规划
Markov Decision Process

Richard Bellman, 1957

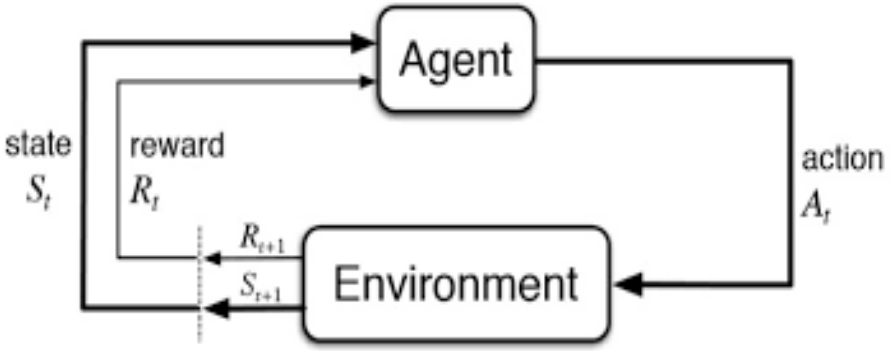
时序差分与最优控制
Q-learning

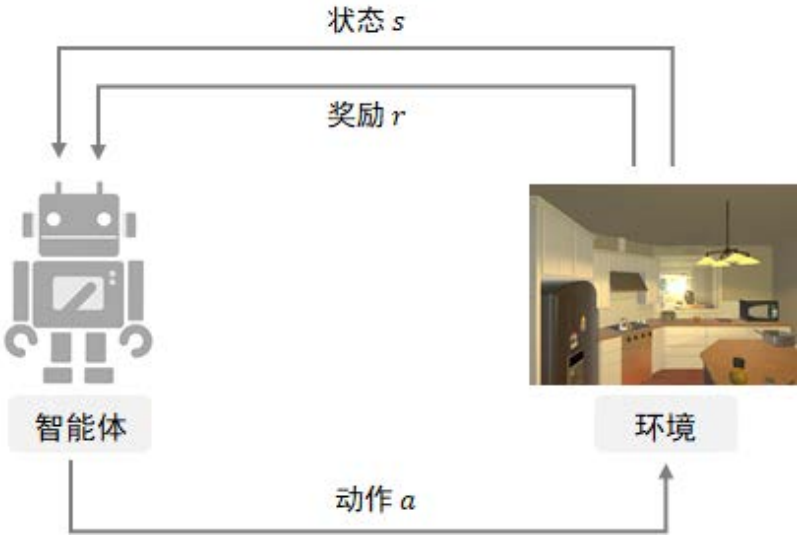
Chris Watkins, 1989

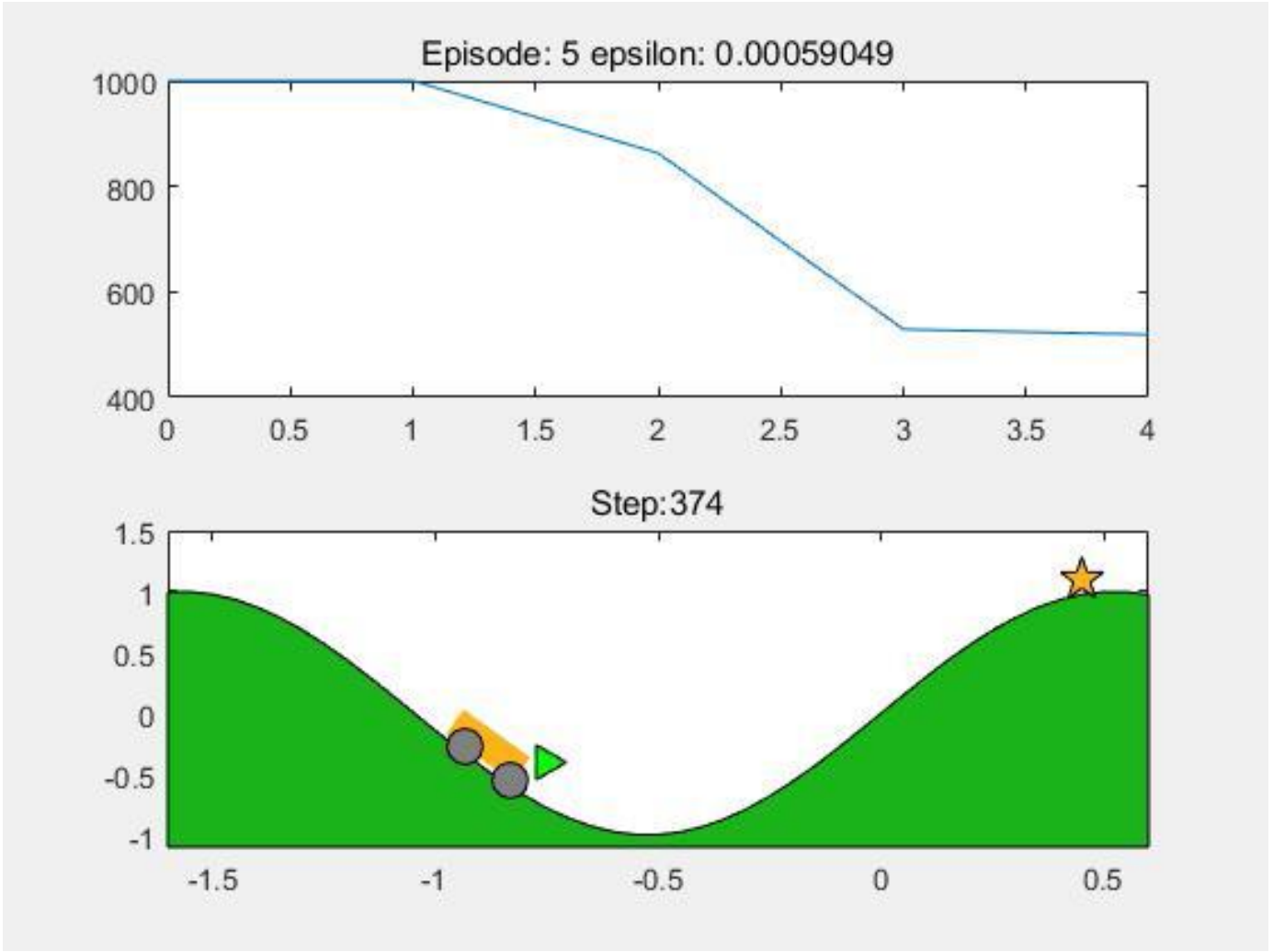
深度强化学习
DQN

DeepMind, 2015

3.0 学习→行为：强化学习


Agent
state
S_t
reward
R_t
action
A_t
R_{t+1}
S_{t+1}
Environment


状态 s
奖励 r
智能体
环境
动作 a


Episode: 5 epsilon: 0.00059049
1000
800
600
400
0
0.5
1
1.5
2
2.5
3
3.5
4
Step:374
1.5
1
0.5
0
-0.5
-1
-1.5
-1
-0.5
0
0.5

3.0 学习→行为：强化学习

➢ 两类基本算法

$$\max_{\theta} \ \mathrm{E}[\sum_{t=0}^{H} R(s_t)|\pi_\theta]$$

Q-学习算法

$$Q(s_t,a_t) = r_{t+1} + \gamma r_{t+2} + \gamma^2 r_{t+3} + \gamma^3 r_{t+4} + \cdots$$

$$Q(s_t,a_t) = r_{t+1} + \gamma Q(s_{t+1},a_{t+1})$$

$$Q(s_t,a_t) \leftarrow Q(s_t,a_t) + \alpha\left[r + \gamma \max_a Q(s_{t+1},a) - Q(s_t,a_t)\right]$$

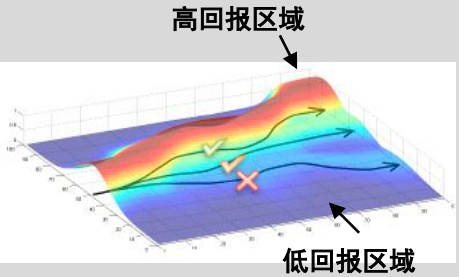
$$\pi^*(s) = \arg\max_{a\in\mathcal{A}} Q^*(s,a)$$

发展：DQN

策略梯度算法

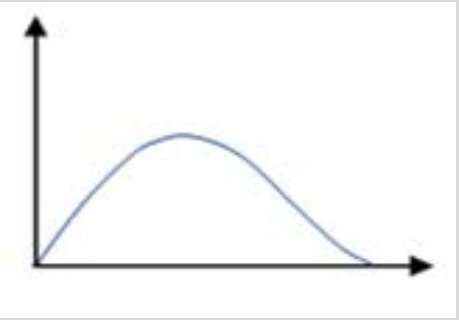
$$J(\pi) = \mathbb{E}_{\tau\sim p_\pi(\tau)}\left[\sum_{t=0}^{\infty}\gamma^t r(s_t,a_t,s_{t+1})\right]$$

$$\nabla_\theta J(\pi_\theta) \approx \frac{1}{N}\sum_{n=1}^{N}\sum_{t=0}^{H^{(n)}} R(\tau^{(n)})\nabla_\theta \log \pi_\theta(a_t^{(n)} \mid s_t^{(n)})$$



$$\theta \leftarrow \theta + \eta\nabla_\theta J$$

$$\pi_\theta(a \mid s)$$



发展：REINFORCE

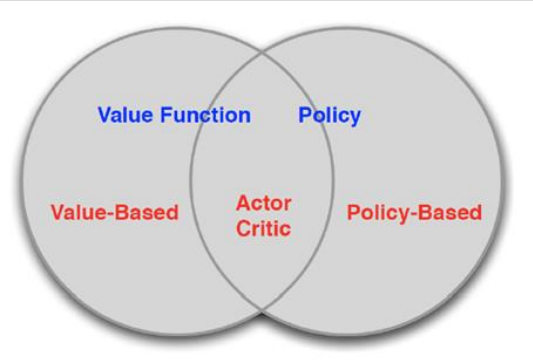
3.0 学习→行为：强化学习

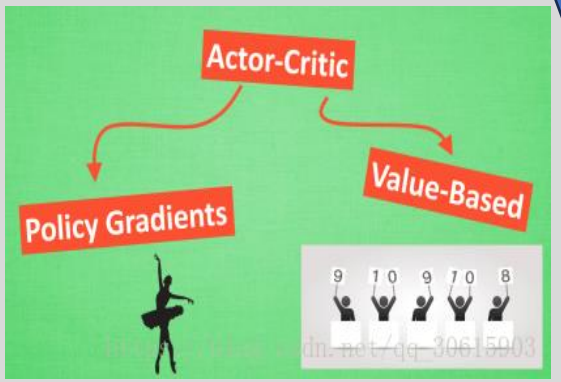
➢ 两类典型算法

Actor-Critic

$$\nabla_\theta J(\pi_\theta) = \mathbb{E}_{s\sim p_0(s)}\mathbb{E}_{a\sim\pi_\theta(\cdot|s)}\left(Q^{\pi_\theta}(s,a)\cdot\nabla_\theta \log\pi_\theta(a\,|\,s)\right)$$

$$\hat{Q}_\omega(s,a)\approx Q^{\pi_\theta}(s,a) \qquad \nabla_\theta J(\pi_\theta) = \hat{Q}_\omega(s_t,a_t)\cdot\nabla_\theta \log\pi_\theta(a_t\,|\,s_t)$$





PPO

$$A^{\pi_\theta}(s_t,a_t) = Q^{\pi_\theta}(s_t,a_t) - V^{\pi_\theta}(s_t)$$

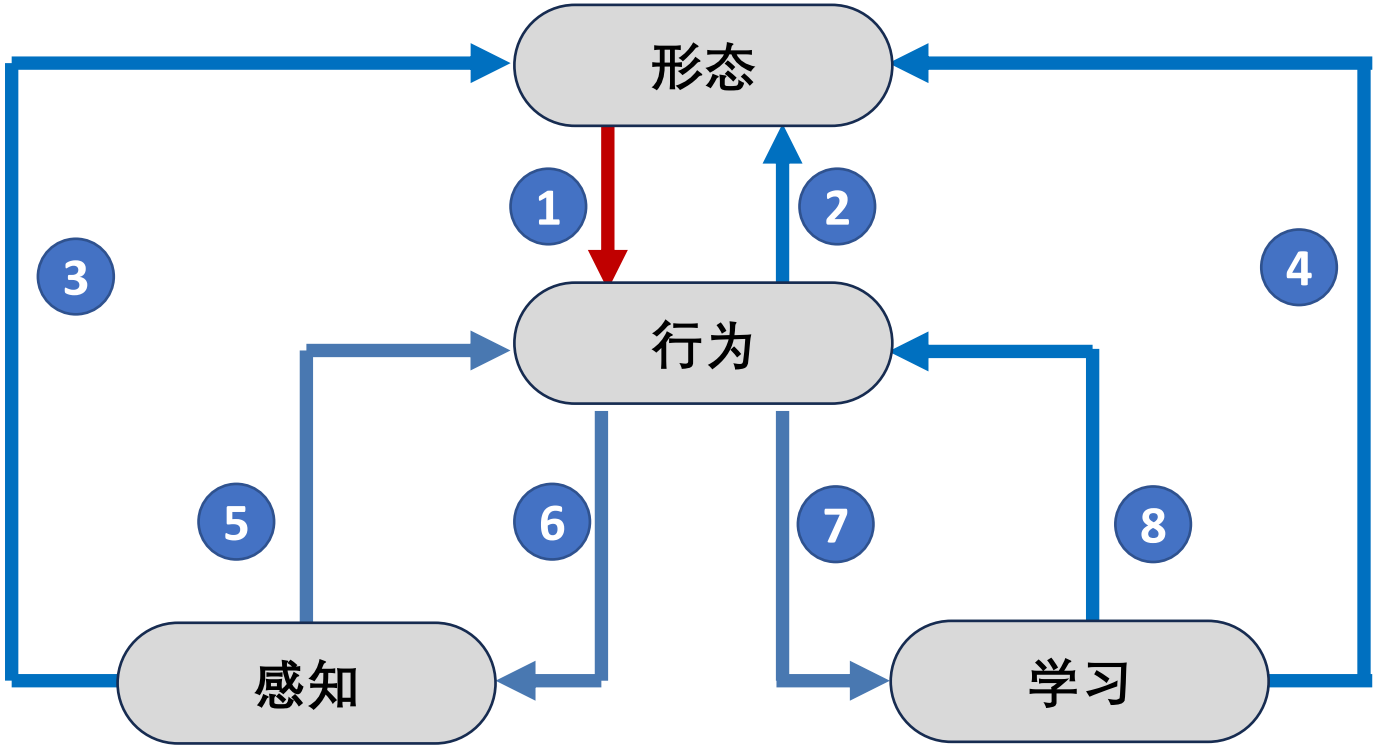
$$\mathbb{E}_{\tau\sim p_{\pi_\theta}(\tau)}\left(\sum_{t=0}^{\infty}\gamma^t A^{\pi_{\theta_{\text{old}}}}(s_t,a_t)\right)$$

$$L_{\pi_{\theta_{\text{old}}}}(\pi_\theta) = \mathbb{E}_{\tau\sim p_{\pi_{\theta_{\text{old}}}}(\tau)}\left(\sum_{t=0}^{\infty}\gamma^t \rho_t(\theta) A^{\pi_{\theta_{\text{old}}}}(s_t,a_t)\right)$$

$$L_{\pi_{\theta_{\text{old}}}}(\pi_\theta) = \mathbb{E}_{\tau\sim p_{\pi_{\theta_{\text{old}}}}(\tau)}\left(\sum_{t=0}^{\infty}\gamma^t \min\left\{(\rho_t(\theta)A^{\pi_{\theta_{\text{old}}}}(s_t,a_t), \text{clip}(\rho_t(\theta),1-\varepsilon,1+\varepsilon)A^{\pi_{\theta_{\text{old}}}}(s_t,a_t)\right\}\right)$$

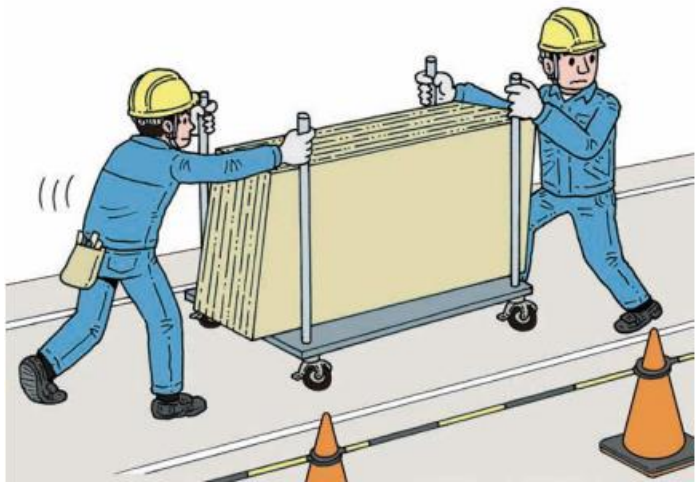
$$\text{clip}(\rho_t(\theta),1-\varepsilon,1+\varepsilon) = \begin{cases} \rho_t(\theta) & 1-\varepsilon \le \rho_t(\theta) \le 1+\varepsilon \\ 1-\varepsilon & \rho_t(\theta) < 1-\varepsilon \\ 1+\varepsilon & \rho_t(\theta) > 1+\varepsilon \end{cases}$$

3.1 形态→行为



3.1 形态→行为

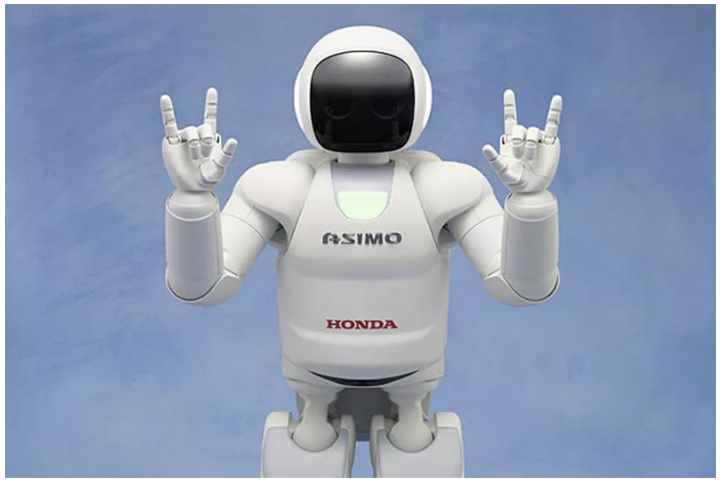
- 例子



3.1 形态→行为

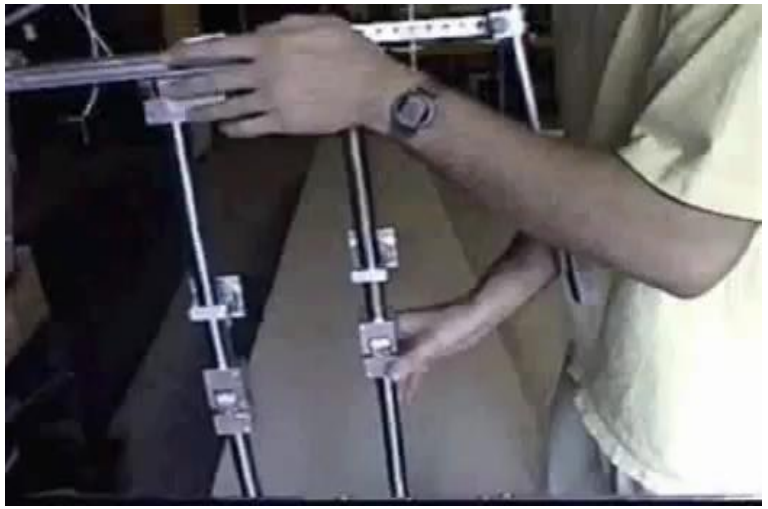
➢ 边界

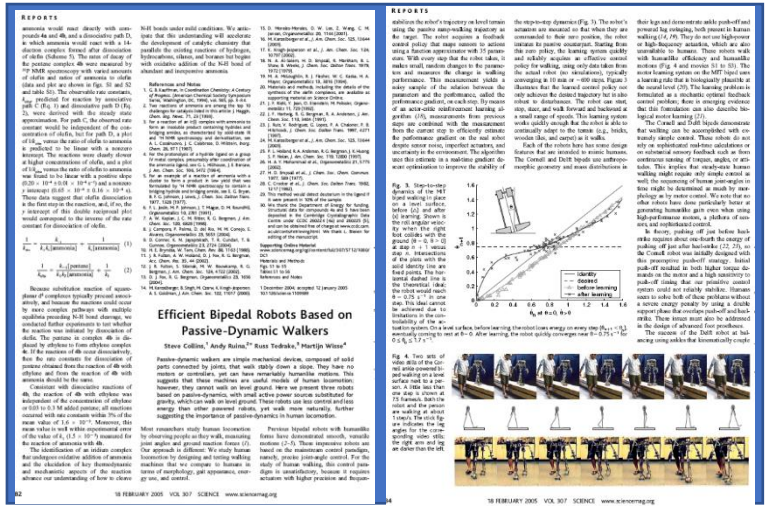
由于形态计算方面的研究与仿生机器人的研究关系非常密切，二者之间的关系会引起混淆。事实上，形态计算更关心的是利用形态来产生行为，而并非从形态上简单地模仿某些生物。很多仿生型机器人，只是形体上模仿了动物的身体，这种模仿可能能获得一些自由度上的突破，例如腿式机器人相比轮式能爬楼，但在行为控制方面，并没有充分利用形态自身的优势，而且仍需要设计复杂的控制器。本田的Asimo机器人，尽管已经充分接近人类的外形，但其每一个关节都需要用特定的算法来控制。这些情形都不属于形态计算之列。

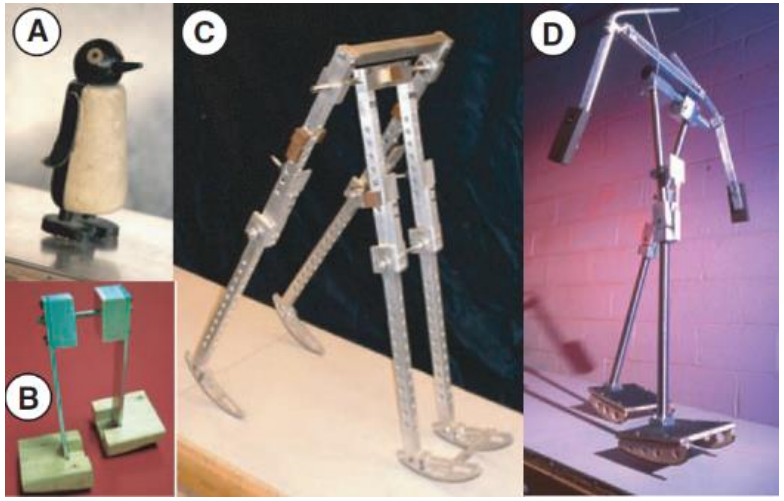


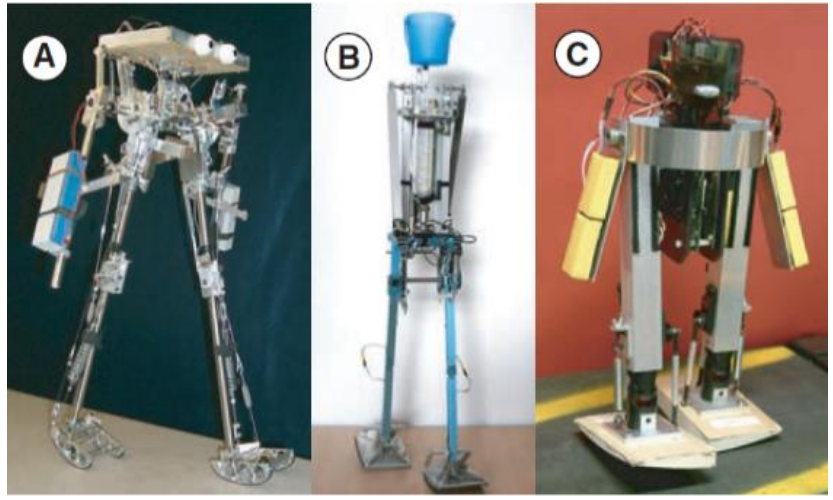
3.1 形态→行为

➢ 被动行走机器人（Passive Walking Robot）

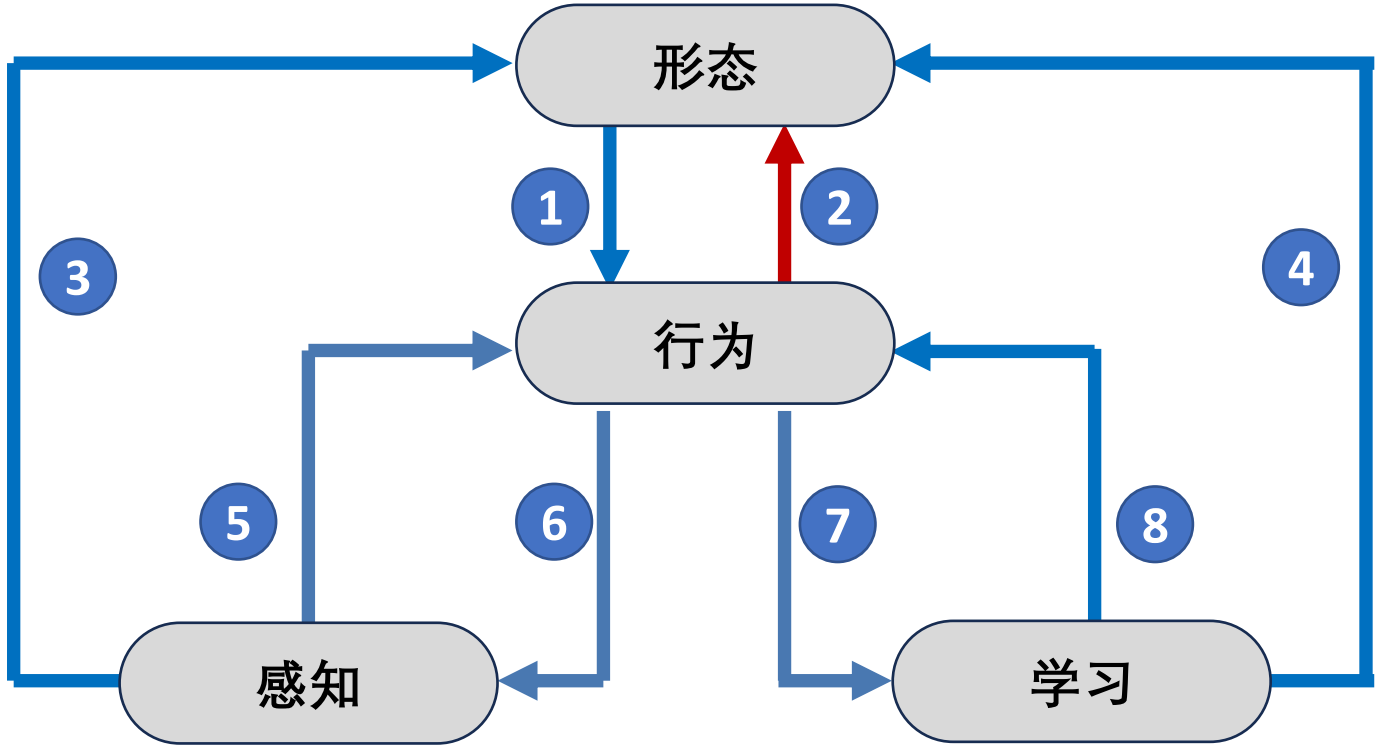








3.2 行为→形态：形态控制



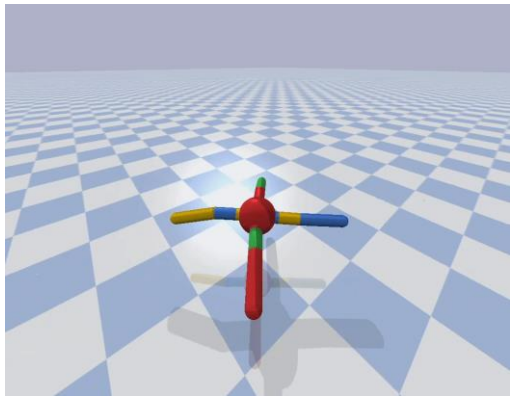
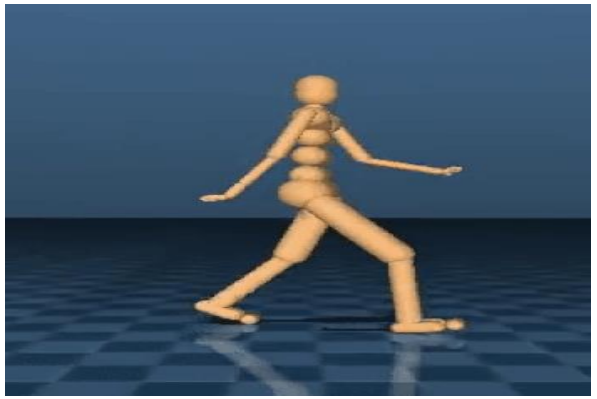
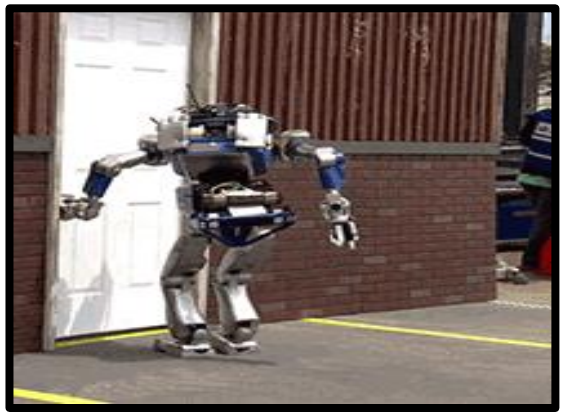
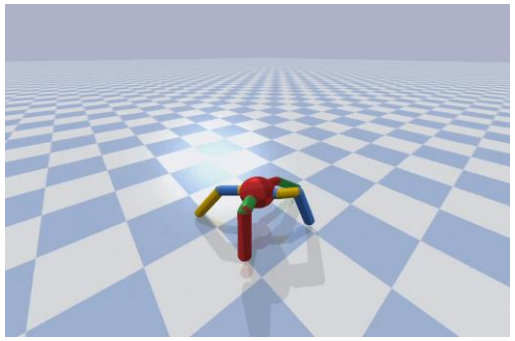
3.2 行为→形态：形态控制

- 形态约束

3.2 行为→形态：形态控制

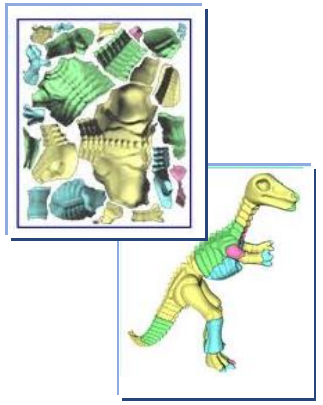
➢ 强化学习的维数灾

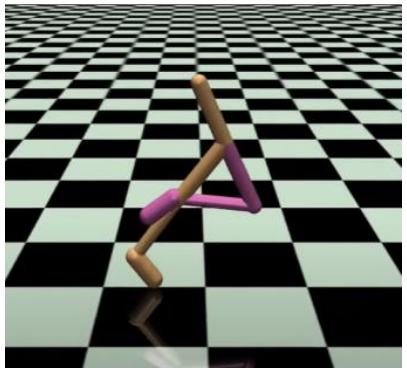
- 机器人自由度多，优化学习困难
- 机器人形态复杂，难以利用
- 在不同形态之间的迁移

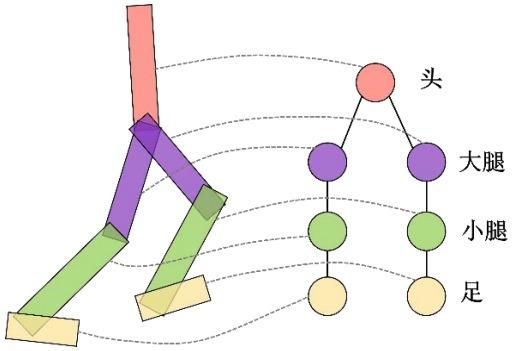





3.2 行为→形态：形态控制

- 引入结构性约束







$$\pi_\theta(a_t \mid s_t)$$

图结构表示

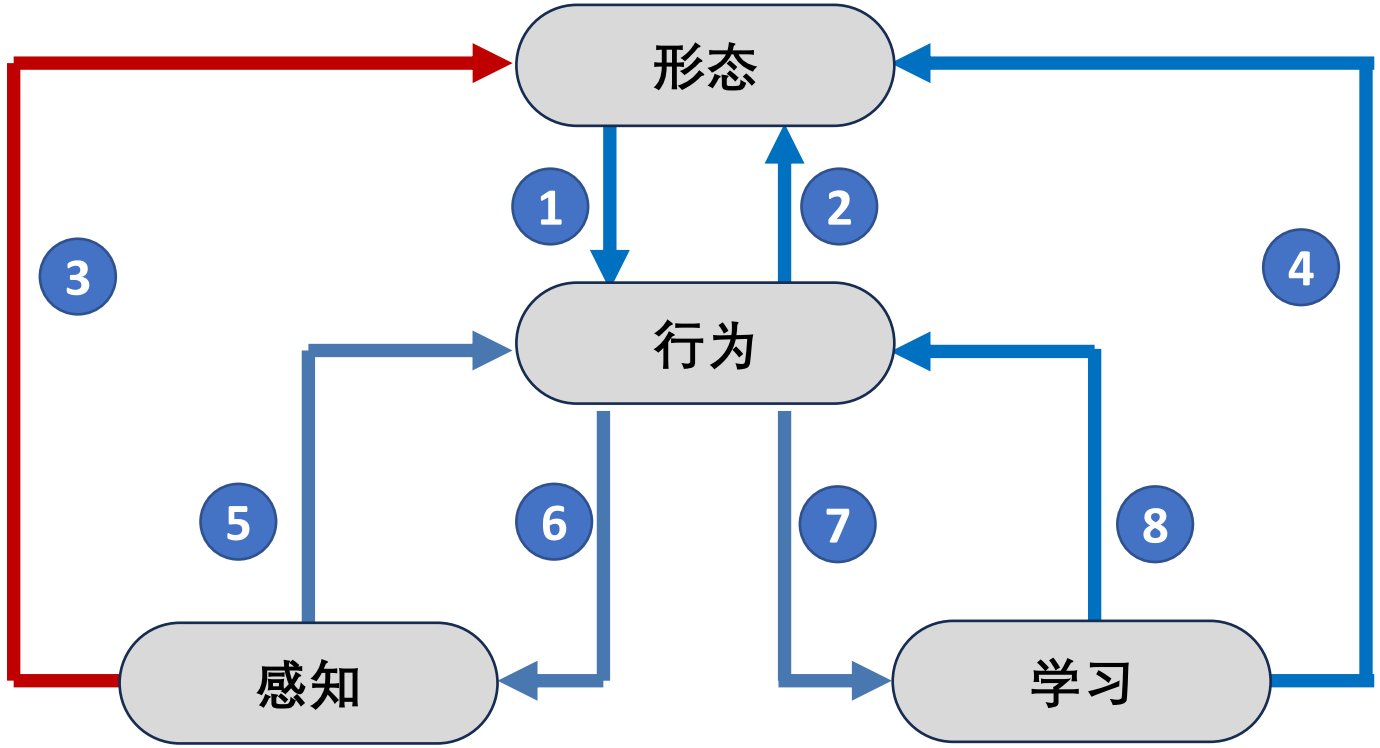
$$D=(V,E,Y)$$

$N_{out}(v)$ 所有以v作为头节点的有向边所指向的尾节点的集合。这一集合反映的是节点v的“输出节点”。

$N_{in}(v)$ 所有以v作为尾节点的边的头结点的集合。这一集合反映的是节点v的“输入节点”。

$$Y_V(v_i)\in\{1,2,...,|Y_V|\} \qquad Y_E(\{v_i,v_j\})\in\{1,2,...,|Y_E|\}$$

3.3 感知→形态：形态变换



3.3 感知→形态：形态变换

3.3 感知→形态：形态变换

➢ 地形适应的形态变换

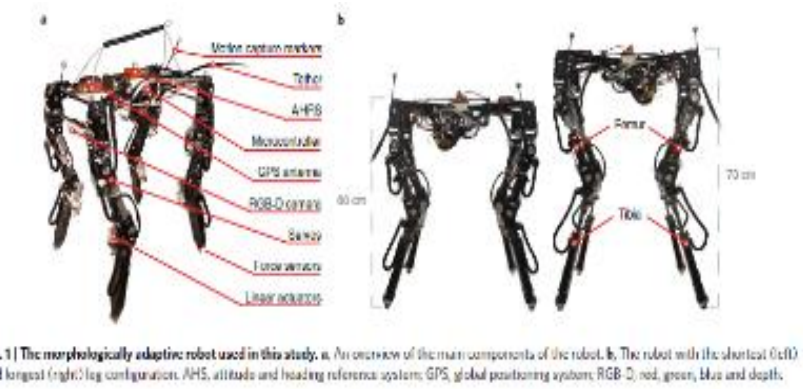
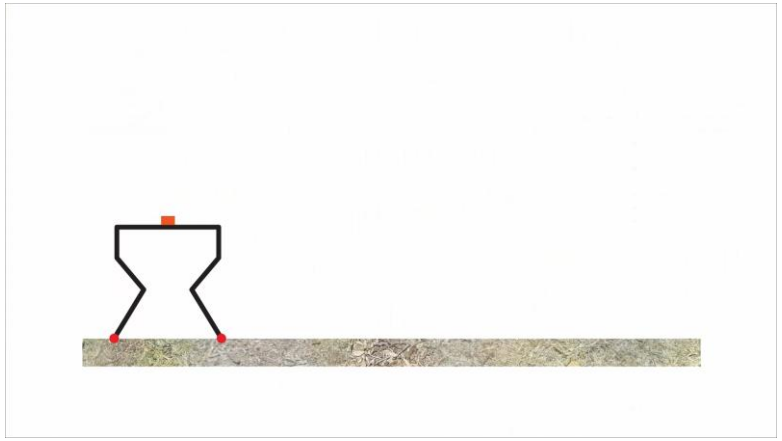
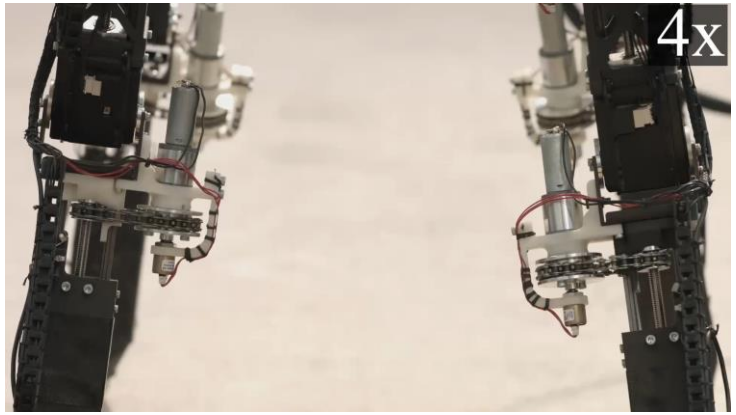
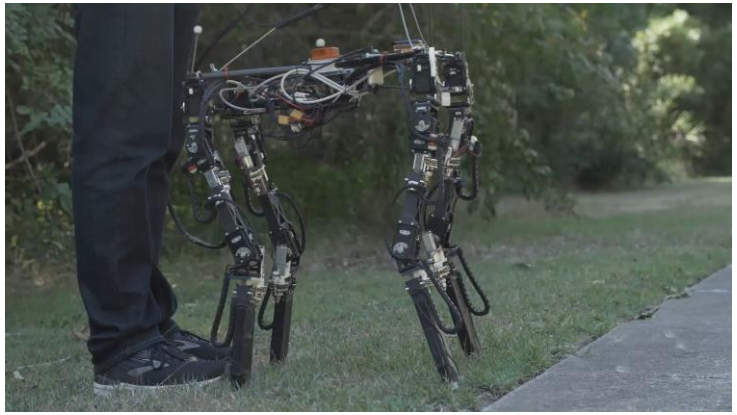


Fig. 1 | The morphologically adaptive robot used in this study. a, An overview of the main components of the robot. b, The robot with the shortest (left) and longest (right) leg configuration. AHRS, attitude and heading reference system; GPS, global positioning system; RGB-D, red, green, blue and depth.





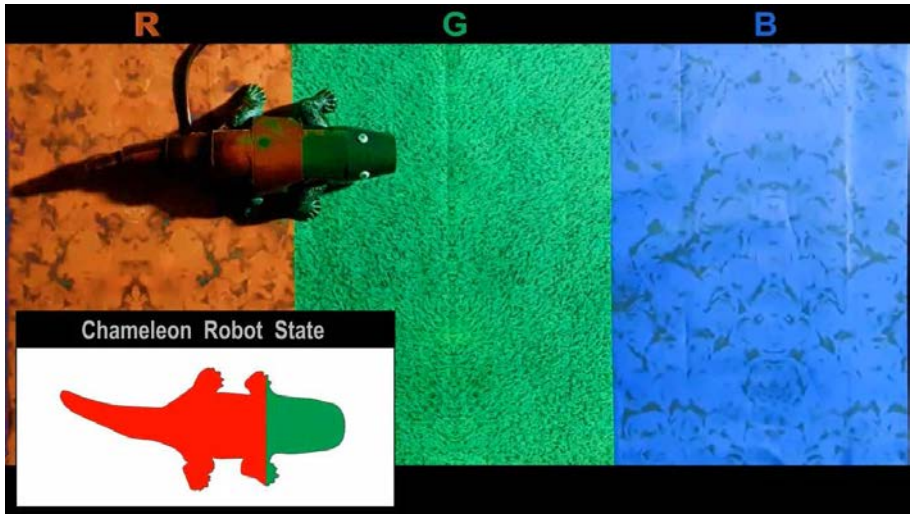


- Nygaard T F, Martin C P, Torresen J, et al. Real-world embodied AI through a morphologically adaptive quadruped robot[J]. Nature Machine Intelligence, 2021, 3(5): 410-419.

3.3 感知→形态：形态变换

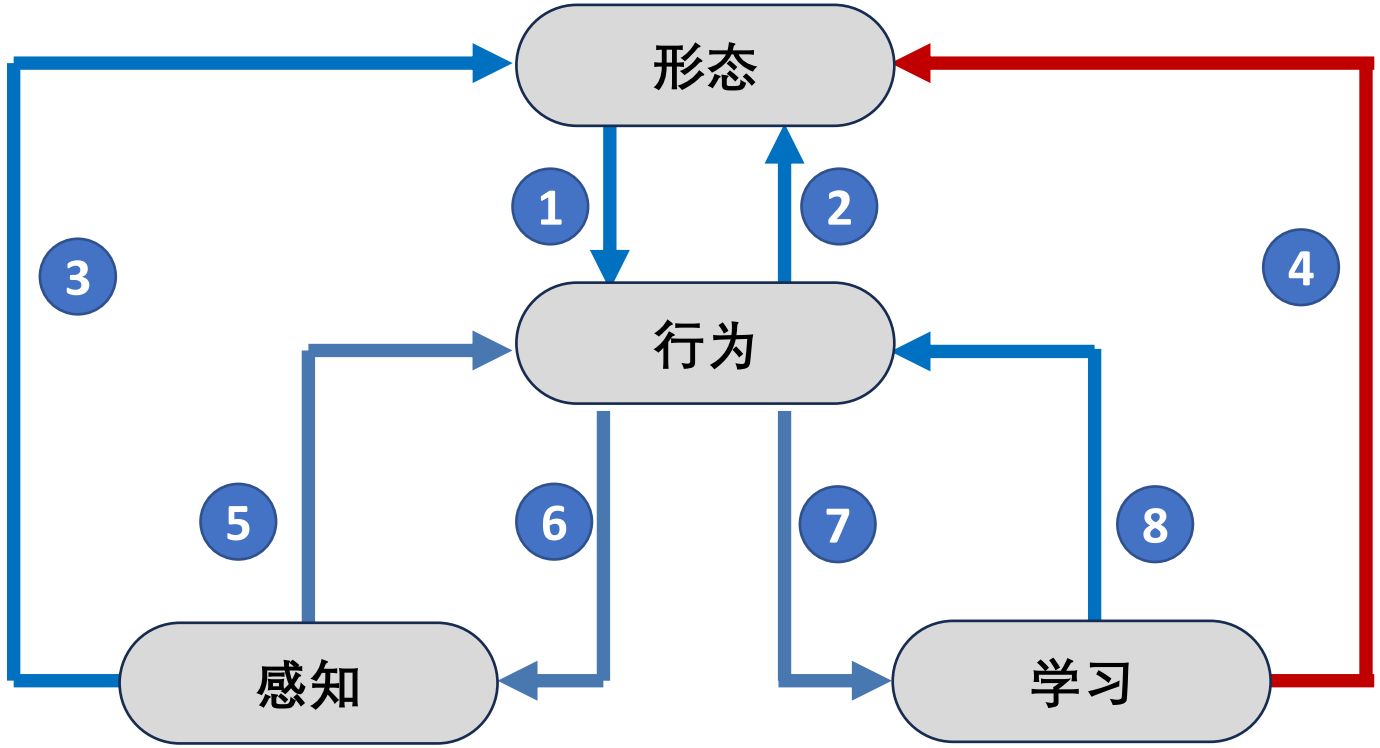
➢ 仿变色龙软体机器人

将人造变色龙皮肤应用于软体机器人上，并结合颜色传感器和反馈控制系统，从而使得这种设备级的自适应人造伪装技术能够检测到背景环境的颜色，并令变色龙软体机器人上实现实时背景颜色匹配。



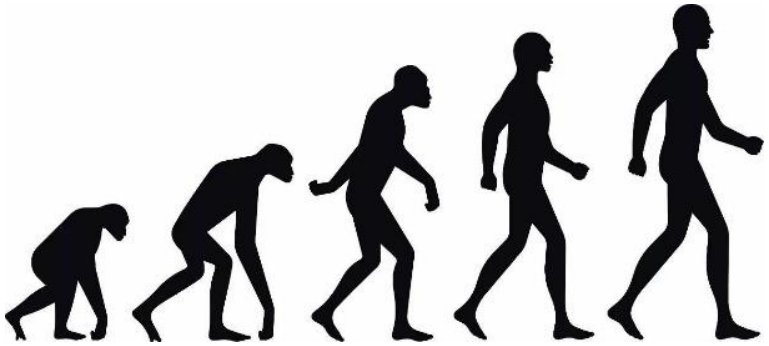
- Kim H, Choi J, Kim K K, et al. Biomimetic chameleon soft robot with artificial crypsis and disruptive coloration skin[J]. Nature communications, 2021, 12(1): 1-11.

3.4 学习→形态：形态生成



3.4 学习→形态：形态生成

- 脑-体协同进化

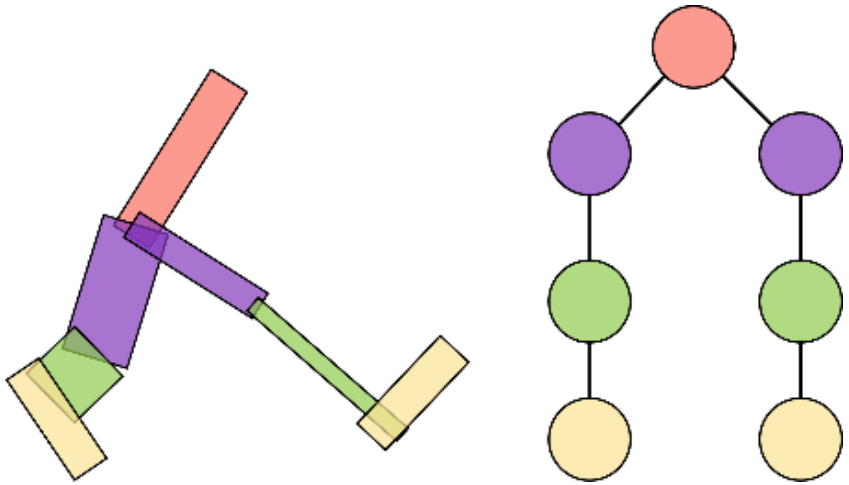


$$(D^*, \pi^*) = \arg\max_{D,\pi} J(D, \pi)$$

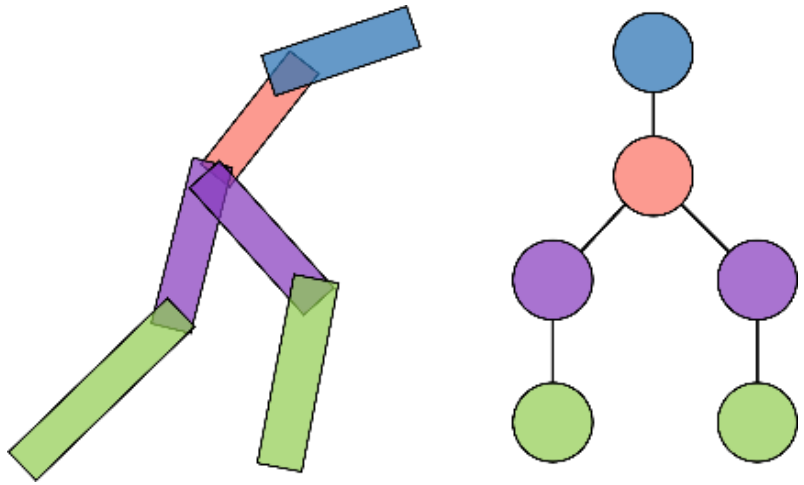
$$\pi^* = \arg\max_{\pi} J(\bar{D}, \pi)$$

3.4 学习→形态：形态生成

- 脑-体协同进化



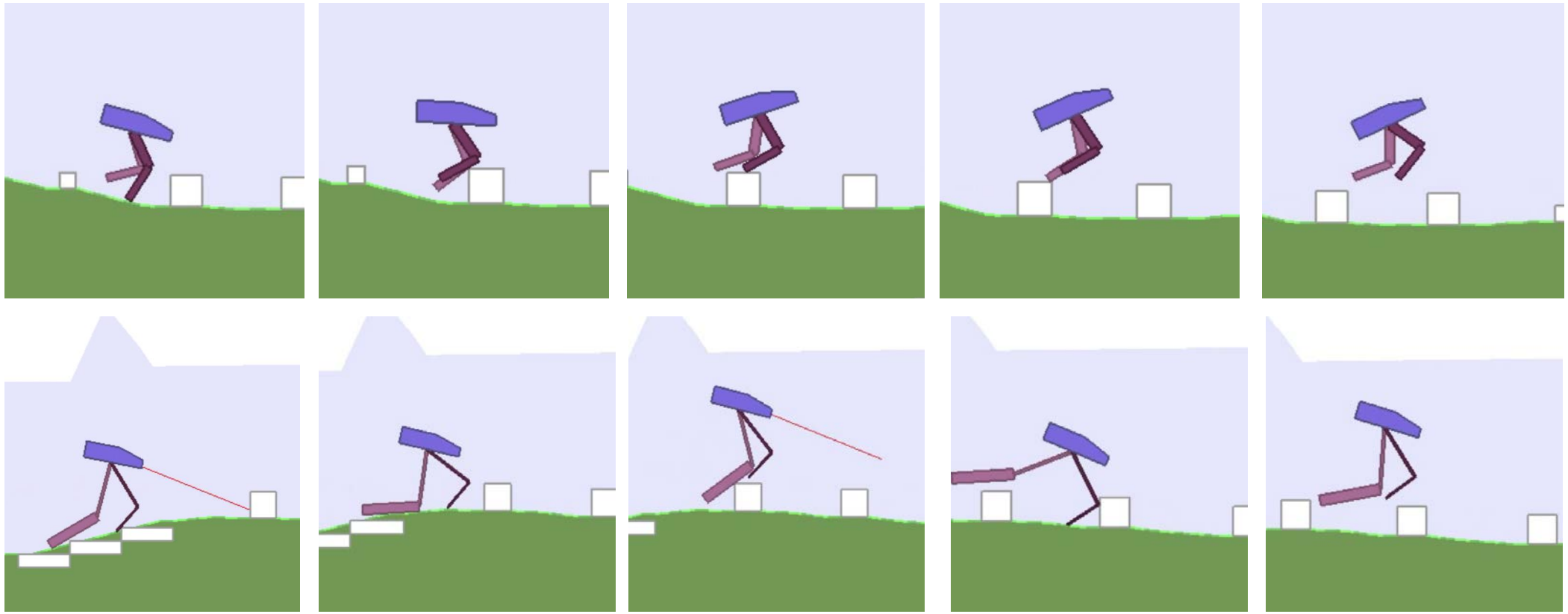
形态参数



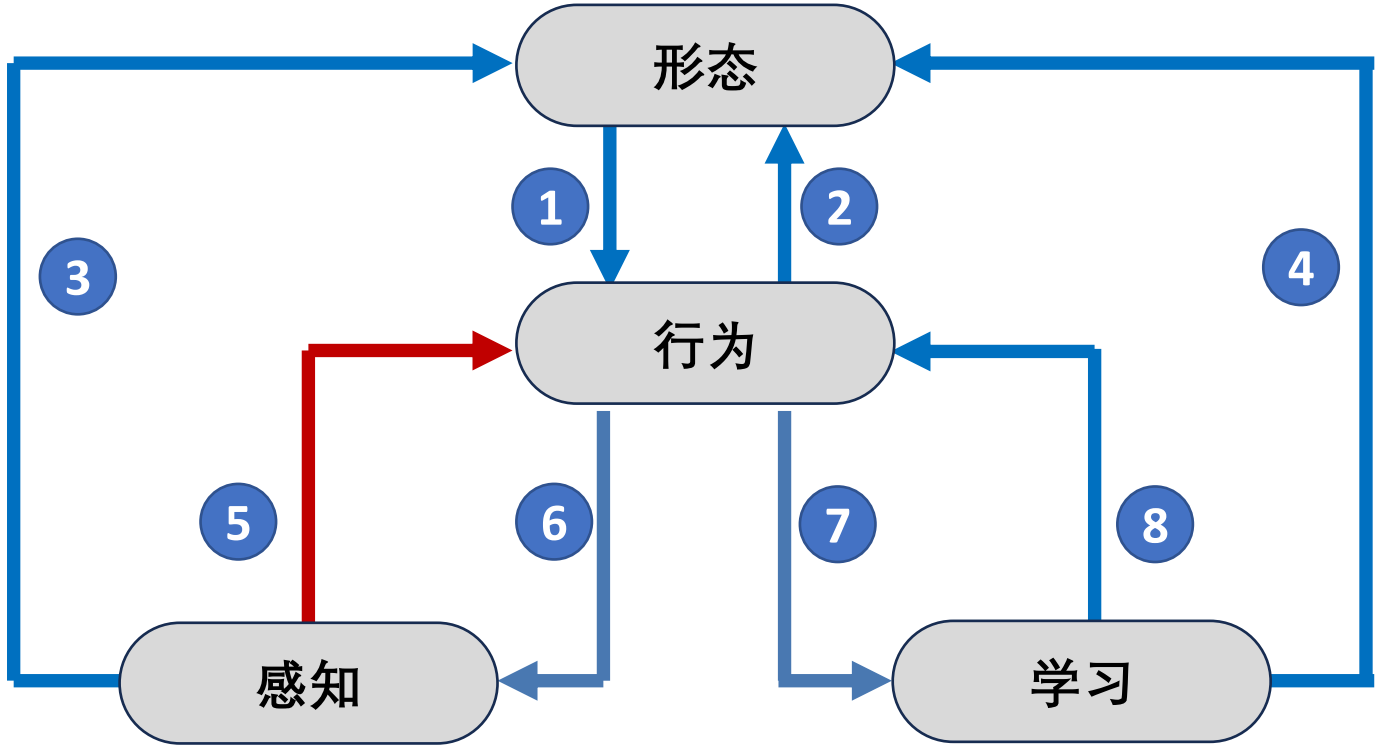
形态结构

3.4 学习→形态：形态生成

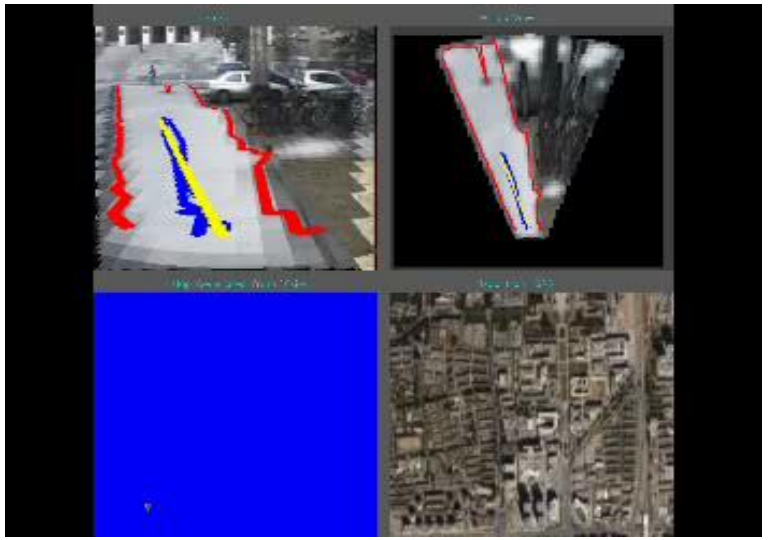
- 形态参数的优化



3.5 感知→行为：视觉导航

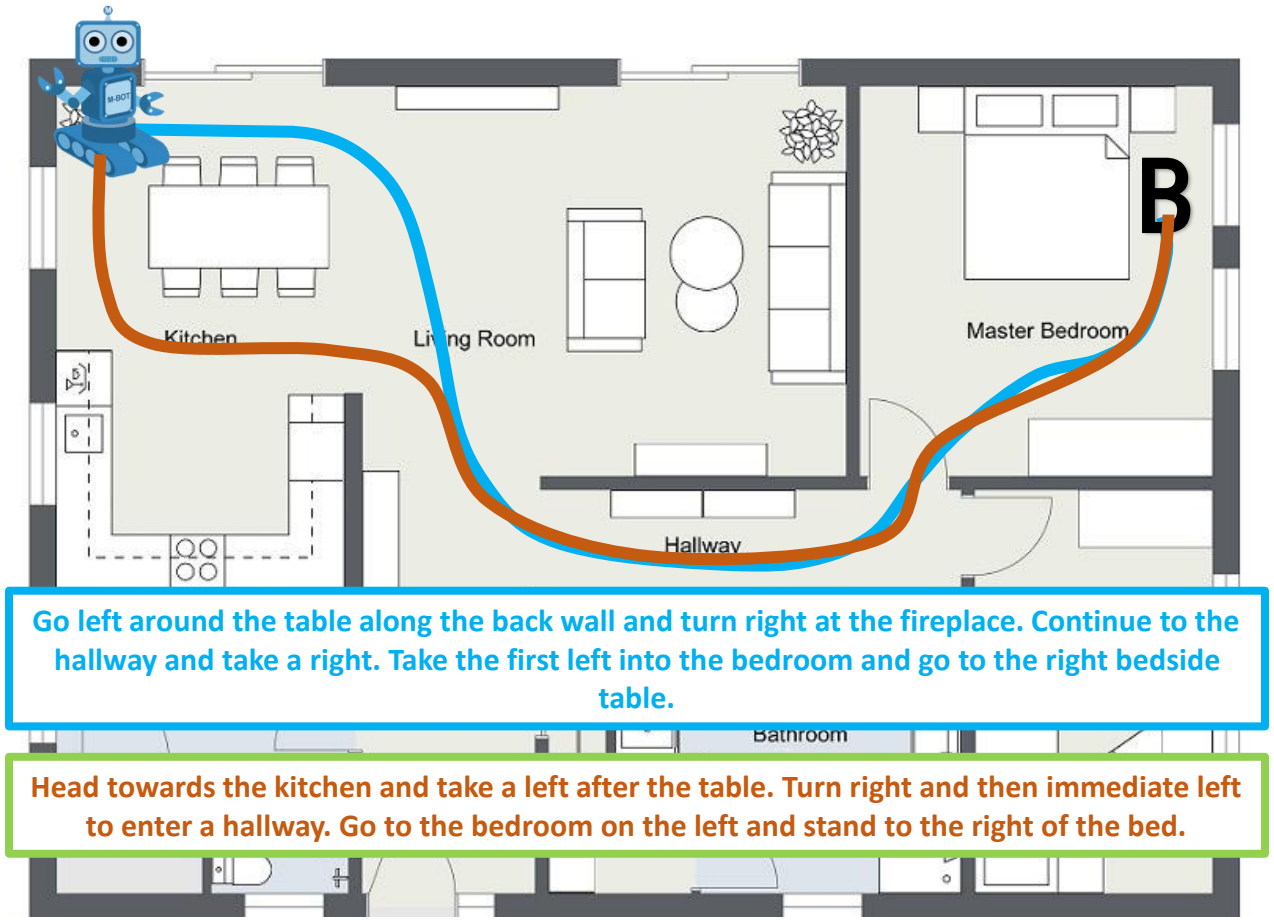


3.5 感知→行为：视觉导航



3.5 感知→行为：视觉导航

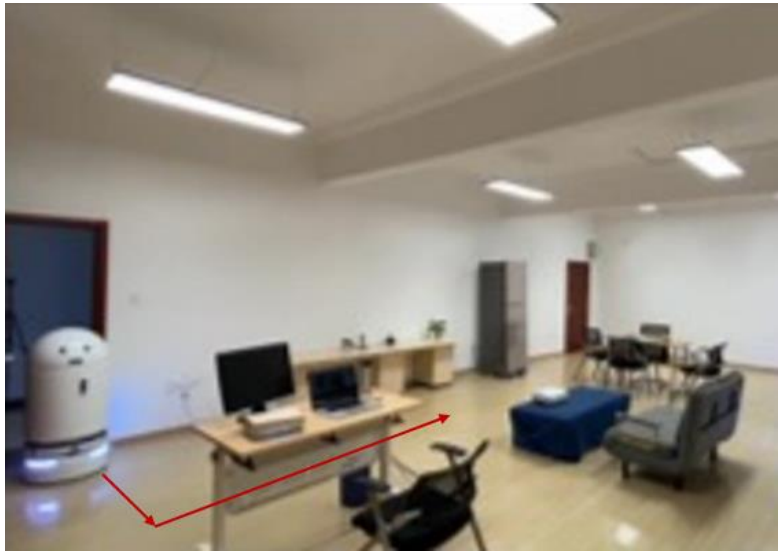
➢ 视觉语言导航



- Vision language navigation (VLN) 让智能体跟着自然语言指令进行导航，这个任务需要同时理解自然语言指令与视角中可以看见的图像信息，然后在环境中对自身所处状态做出对应的动作，最终达到目标位置。

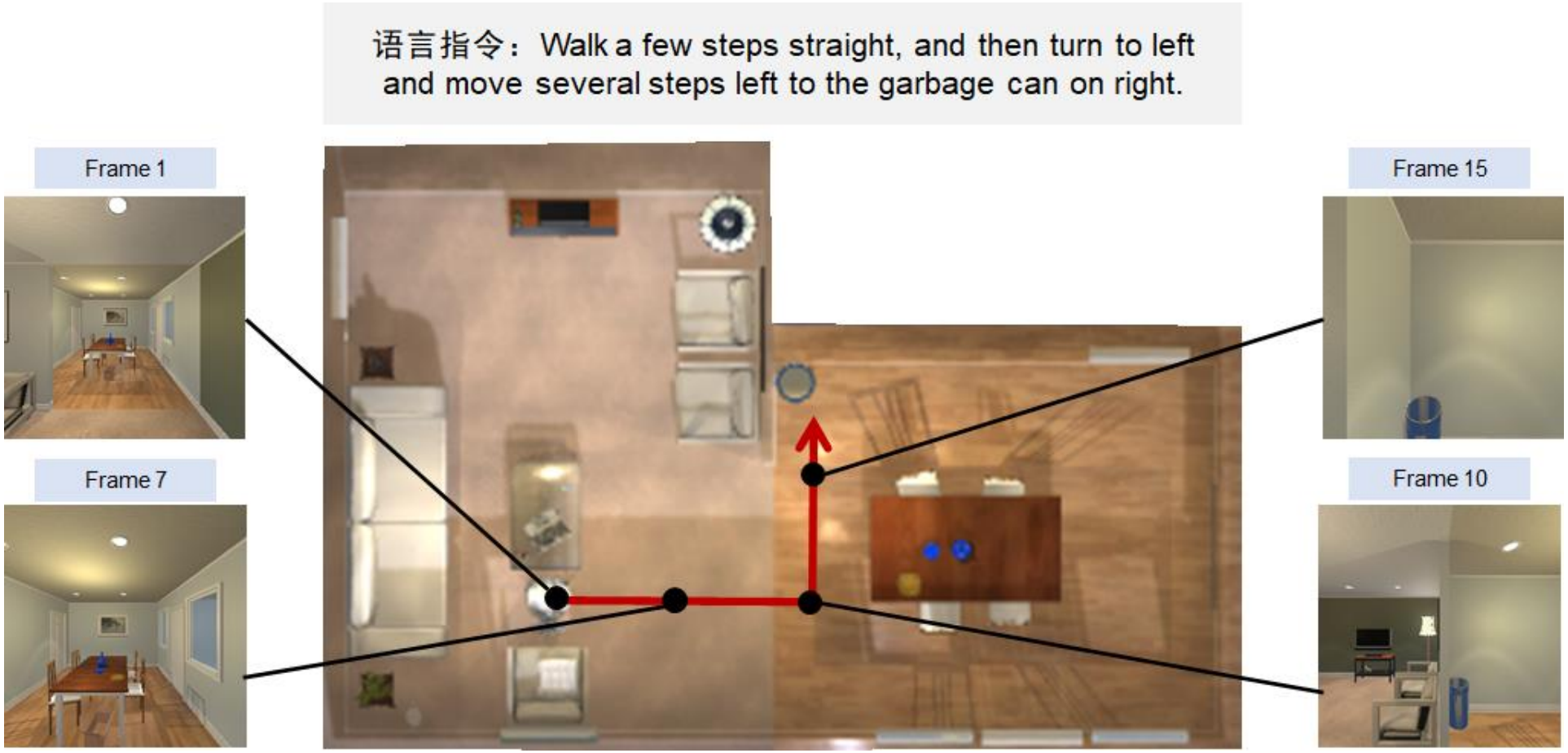
3.5 感知→行为：视觉导航

- 视觉语言导航

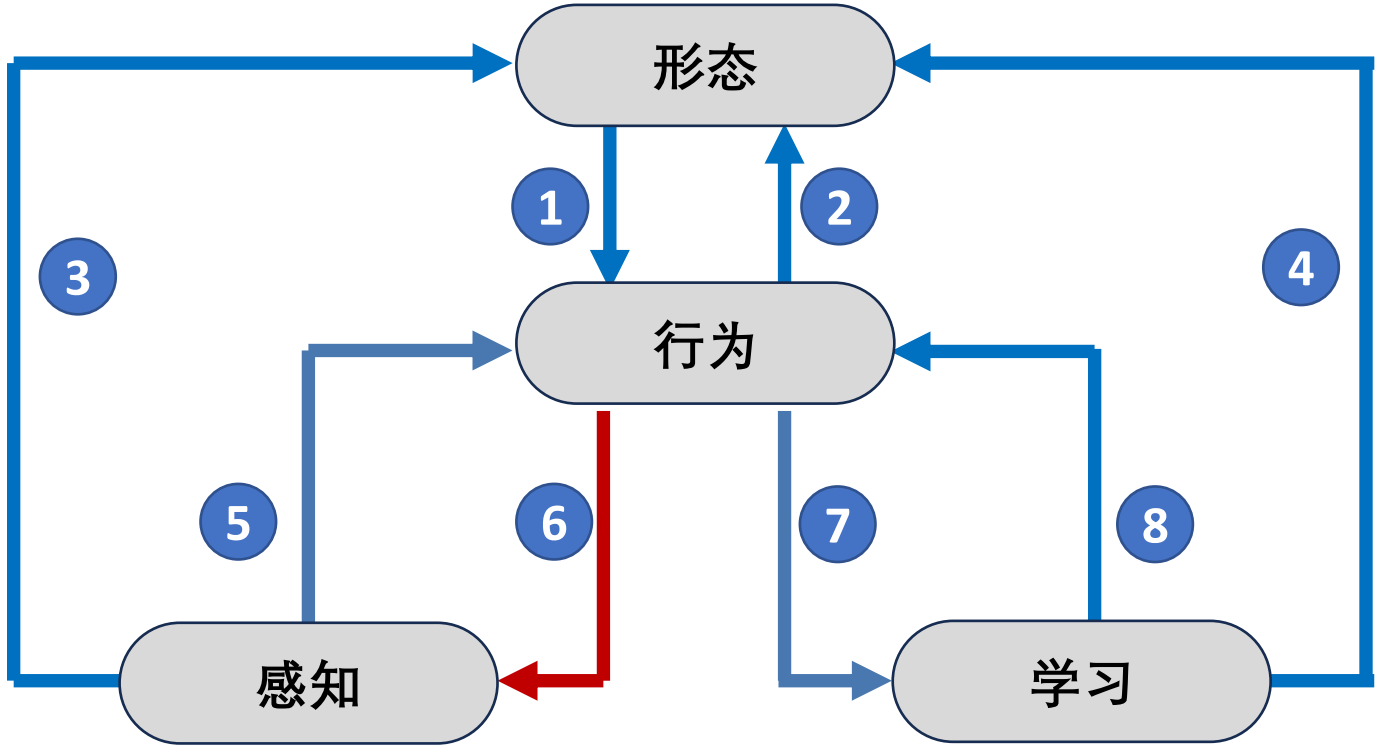


3.5 感知→行为：视觉导航

➢ 问题描述



3.6 行为→感知：主动感知

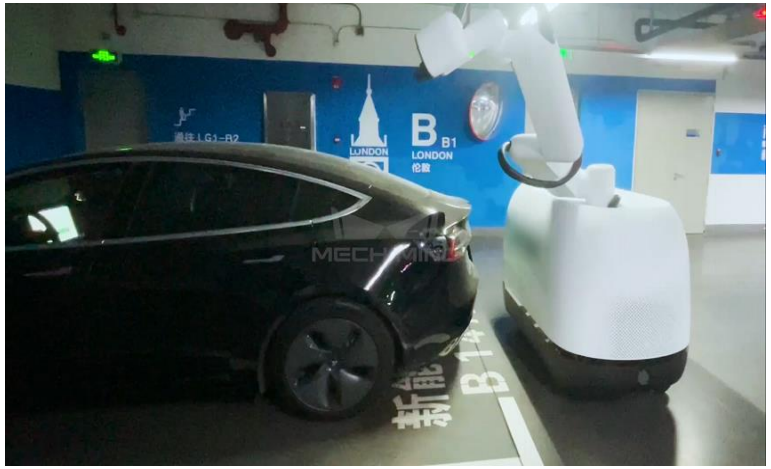


3.6 行为→感知：主动感知

➢ 生活中的主动感知

3.6 行为→感知：主动感知

- 主动感知的应用



3.6 行为→感知：主动感知

➢ 主动感知的优点

- A motivation for examining active vision is the fact that passive vision has been shown to be very problematic.
- Almost every basic problem in passive machine perception is very difficult, it is ill-posed in the sense of Hadamard.
- Problems that are ill-posed, nonlinear or unstable for a passive observer become well-posed, linear or stable for an active observer.

Active Perception

RUZENA BAJCSY, MEMBER, IEEE

Invited Paper

Active Perception (Active Vision specifically) is defined as a study of Modeling and Control strategies for perception. By modeling we mean models of sensors, processing modules and their interaction. We distinguish local models from global models by their extent of application in space and time. The local models represent procedures and parameters such as optical distortions of the lens, focal lens, spatial resolution, band-pass filter, etc. The global models on the other hand characterize the overall performance and make predictions on how the individual modules interact. The control strategies are formulated as a search of such sequence of steps that would minimize a loss function while one is seeking the most information. Examples are shown as the existence proof of the proposed theory on obtaining range from focus and stereo/vergence on 2-D segmentation of an image and 3-D shape parametrization.

I. Introduction

Most past and present work in machine perception has involved extensive static analysis of passively sampled data. However, it should be axiomatic that perception is not passive, but active. Perceptual activity is exploratory, probing, searching; percepts do not simply fall onto sensors as rain falls onto ground. We do not just see, we look. And in the course, our pupils adjust to the level of illumination, our eyes bring the world into sharp focus, our eyes converge or diverge, we move our heads or change our position to get a better view of something, and sometimes we even put on spectacles. This adaptiveness is crucial for survival in an uncertain and generally unfriendly world, as millenia of experiments with different perceptual organizations have clearly demonstrated. Yet no adequate account or theory or example of active perception has been presented by machine perception research. This lack is the motivation for this paper.

Manuscript received November 23, 1987; revised March 21, 1988. This work was supported in part by NSF Grant DCR-8410771, Air Force Grant AFOSR F49620-85-K-0018, Army/DAAC-29-84-K-0061, NSF-CER/DCR82-19196 AO2, DARPA/ONR NIH Grant NS-10939-11 as part of Cerebo Vascular Research Center, NIH 1-RO1-NS-23636-01, NSF INT85-14199, NSF DMC85-17315, ARPA N0014-85-K-0807, NATO Grant 0224/85, and by DEC Corporation, IBM Corporation, and LORD Corporation.

The author is with the Computer and Information Science Department, University of Pennsylvania, Philadelphia, PA 19104, USA.

IEEE Log Number 8822793.

II. What is Active Sensing?

In the robotics and computer vision literature, the term "active sensor" generally refers to a sensor that transmits (generally electromagnetic radiation, e.g., radar, sonar, ultrasound, microwaves and collimated light) into the environment and receives and measures the reflected signals. We believe that the use of active sensors is not a necessary condition on active sensing, and that sensing can be performed with passive sensors (that only receive, and do not emit, information), employed actively. Here we use the term active not to denote a time-of-flight sensor, but to denote a passive sensor employed in an active fashion, purposefully changing the sensor's state parameters according to sensing strategies.

Hence the problem of Active Sensing can be stated as a problem of controlling strategies applied to the data acquisition process which will depend on the current state of the data interpretation and the goal or the task of the process. The question may be asked, "Is Active Sensing only an application of Control Theory?" Our answer is: "No, at least not in its simple version." Here is why:

1) The feedback is performed not only on sensory data but on complex processed sensory data, i.e., various extracted features, including relational features.
2) The feedback is dependent on *a priori* knowledge—models that are a mixture of numeric/parametric and symbolic information.

But one can say that Active Sensing is an application of intelligent control theory which includes reasoning, decision making, and control. This approach has been eloquently stated by Tenenbaum [1]: "Because of the inherent limitation of a single image, the acquisition of information should be treated as an integral part of the perceptual process . . . Accommodation attacks the fundamental limitation of image inadequacy rather than the secondary problems caused by it." Although he uses the term *accommodation* rather than *active sensing* the message is the same.

The implications of the active sensing approach are the following:

1) The necessity of models of sensors. This is to say, first, the model of the physics of sensors as well as the noise of the sensors. Second, the model of the signal processing and

0018-9219/88/0800-0996$01.00 © 1988 IEEE

Ruzena Bajcsy

UC Berkeley

John Aloimonos

U. Maryland

Problem	Passive Observer	Active Observer
Shape for shading	Ill-posed problem. Needs to be regularized. Even then, unique solution is not guaranteed because of nonlinearity.	Well-posed problem. Unique solution. Linear equation used. Stability.
Shape from contour	Ill-posed problem. Has not been regularized up to now in the Tichonov sense. Solvable under restrictive assumptions.	Well-posed problem. Unique solution for both monocular or binocular observer.
Shape from texture	Ill-posed problem. Needs some assumption about the texture.	Well-posed problem. No assumption required.
Structure from motion	Well posed but unstable. Nonlinear constraints.	Well posed and stable. Quadratic constraints, simple solution methods, stability.
Optic flow (area based)	Ill posed. Needs to be regularized. The introduced smoothness might produce erroneous results.	Well-posed problem. Unique solution. Might be unstable.

很多困难的视觉感知问题，如果允许“动一动”，其难度会显著降低。

Revisiting active perception

Ruzena Bajcsy[1] · Yiannis Aloimonos[2] · John K. Tsotsos[3]

Auton Robot (2018) 42:177–196
https://doi.org/10.1007/s10514-017-9615-3

Received: 9 September 2016 / Accepted: 11 January 2017 / Published online: 15 February 2017
© The Author(s) 2017. This article is published with open access at Springerlink.com

Abstract Despite the recent successes in robotics, artificial intelligence and computer vision, a complete artificial agent necessarily must include active perception. A multitude of ideas and methods for how to accomplish this have already appeared in the past, their broader utility perhaps impeded by insufficient computational power or costly hardware. The history of these ideas, perhaps selective due to our perspectives, is presented with the goal of organizing the past literature and highlighting the seminal contributions. We argue that those contributions are as relevant today as they were decades ago and, with the state of modern computational tools, are poised to find new life in the robotic perception systems of the next decade.

Keywords Sensing · Perception · Attention · Control

1 Introduction

Some evening, long ago, our ancestors looked up at the night sky, just as they had done for thousands of years. But this time, it was different. For the first time, human eyes noticed patterns in the stars. They saw caricatures of living things and as they scanned the heavens more appeared. These were eventually named, once language permitted, and we now know them as constellations of stars. This perhaps represents the first instance of *connecting the dots*, in a literal as well as figurative manner. How did this come about? What is needed to enable it? That one needs the ability to find a dot, and link one dot to another creating a group, is clear. Underlying this, however, is the critical ability to hypothesize a pattern: that group of stars looks like something I know—let me see if I can find other stars to complete the pattern. This is the essence of active perception—to set up a goal based on some current belief about the world and to put in motion the actions that may achieve it.

Through the years, the topic of perception, and particularly vision, has been a great source of wonder and study by philosophers and scientists alike. This history cannot be reviewed here and the interested reader should see Pastore (1971) and Wade and Wade (2000), among others. This paper presents the history of the computational perspective on the problem of active perception, with an emphasis on visual perception, but broadly applicable to other sensing modalities. Those interested in a biological perspective on active perception should see Findlay and Gilchrist (2003) and Poletti and Rucci (2013). What follows is a brief and selective history of the birth of the computational active perception paradigm. However, there is one source of motivation for the topic that deserves mention. The early work of Gibson (1950) proposed that perception is due to the combination of the environment in which an agent exists and how that agent interacts with the environment. He was primarily interested in optic flow that is generated on the retina when moving through the environment (as when flying) realizing that it was the path of motion itself that enabled the perception of specific elements, while dis-enabling others. That path of motion was under the

This is one of several papers published in *Autonomous Robots* comprising the Special Issue on Active Perception.

All authors contributed equally to all aspects of this manuscript.

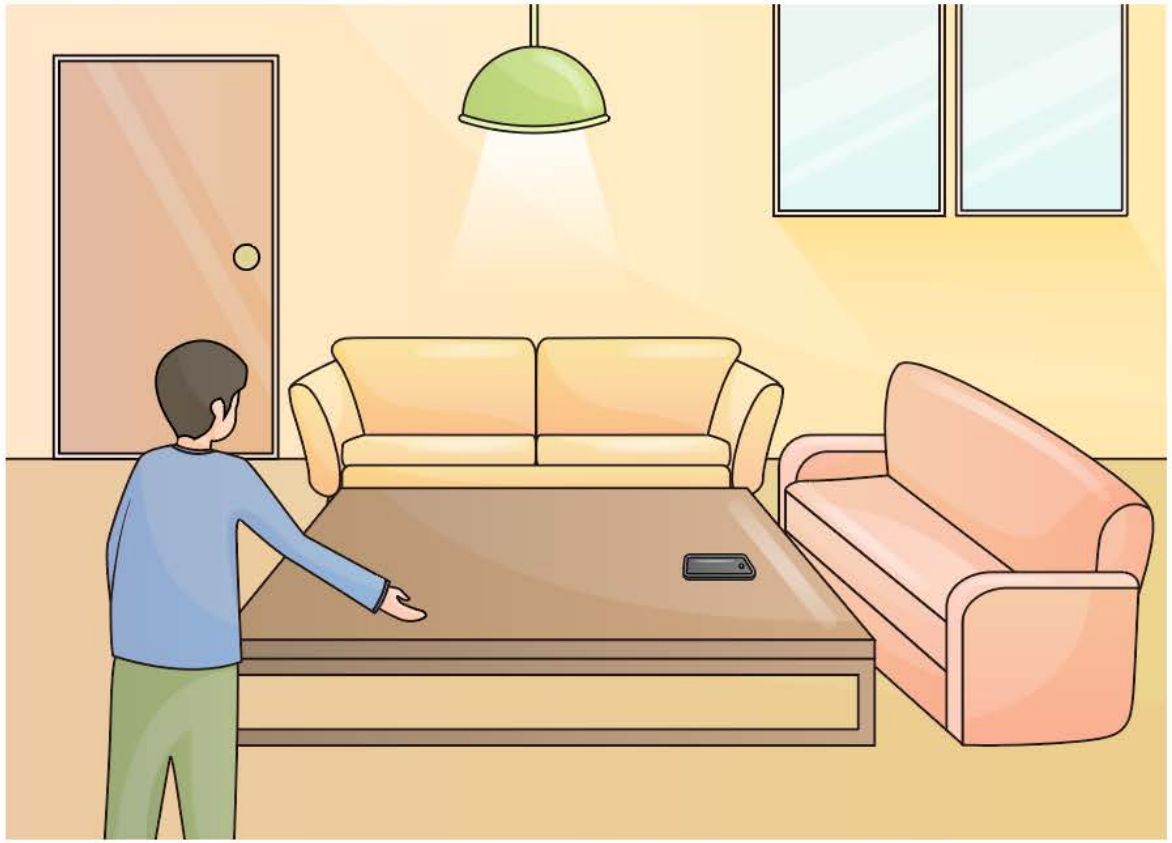
✉ John K. Tsotsos
tsotsos@cse.yorku.ca

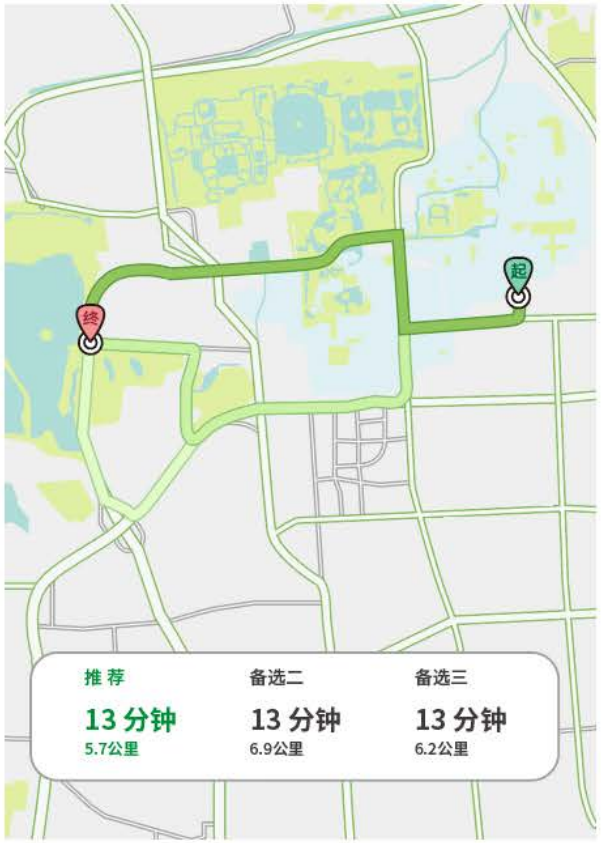
[1] Department of Electrical Engineering and Computer Sciences, University of California, Berkeley, CA, USA
[2] Department of Computer Science, University of Maryland, College Park, MD, USA
[3] Department of Electrical Engineering and Computer Science, York University, Toronto, ON, Canada

- Aloimonos, John, Isaac Weiss, and Amit Bandyopadhyay. "Active vision." International journal of computer vision 1, no. 4 (1988): 333-356.

3.6 行为→感知：主动感知

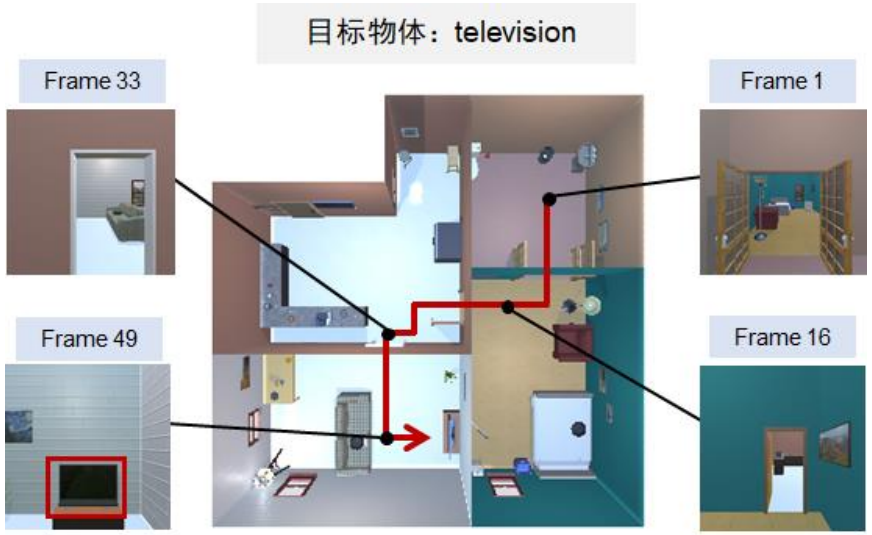
- 视觉语义导航（Visual Semantic Navigation, VSN）





3.6 行为→感知：主动感知

➢ VSN问题描述



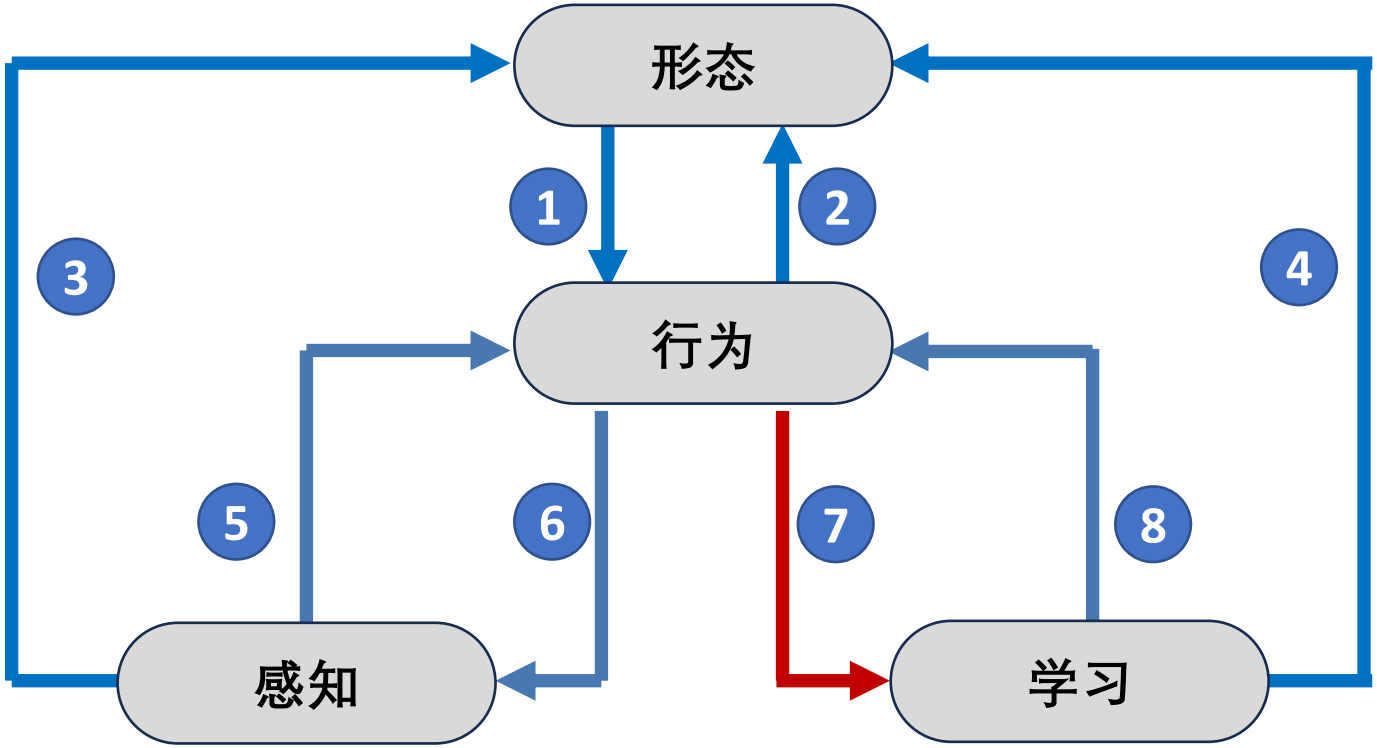
$$s_t = \varphi(I_t)$$

$$o = \phi(O)$$

$$a_t = \mathrm{Nav}(s_t, a_{t-1}, o)$$

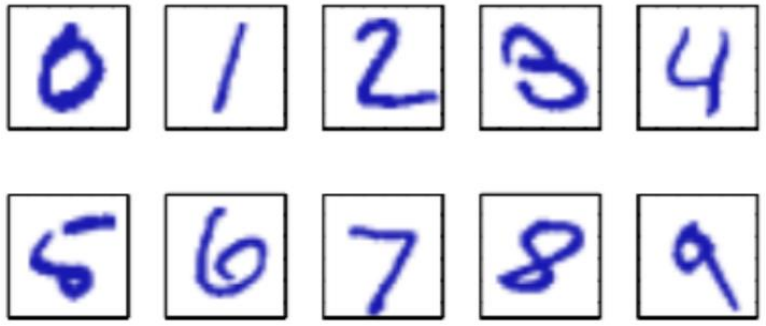
$$s_{t+1} = \mathrm{Env}(s_t, a_t)$$

3.7 行为→学习：具身学习



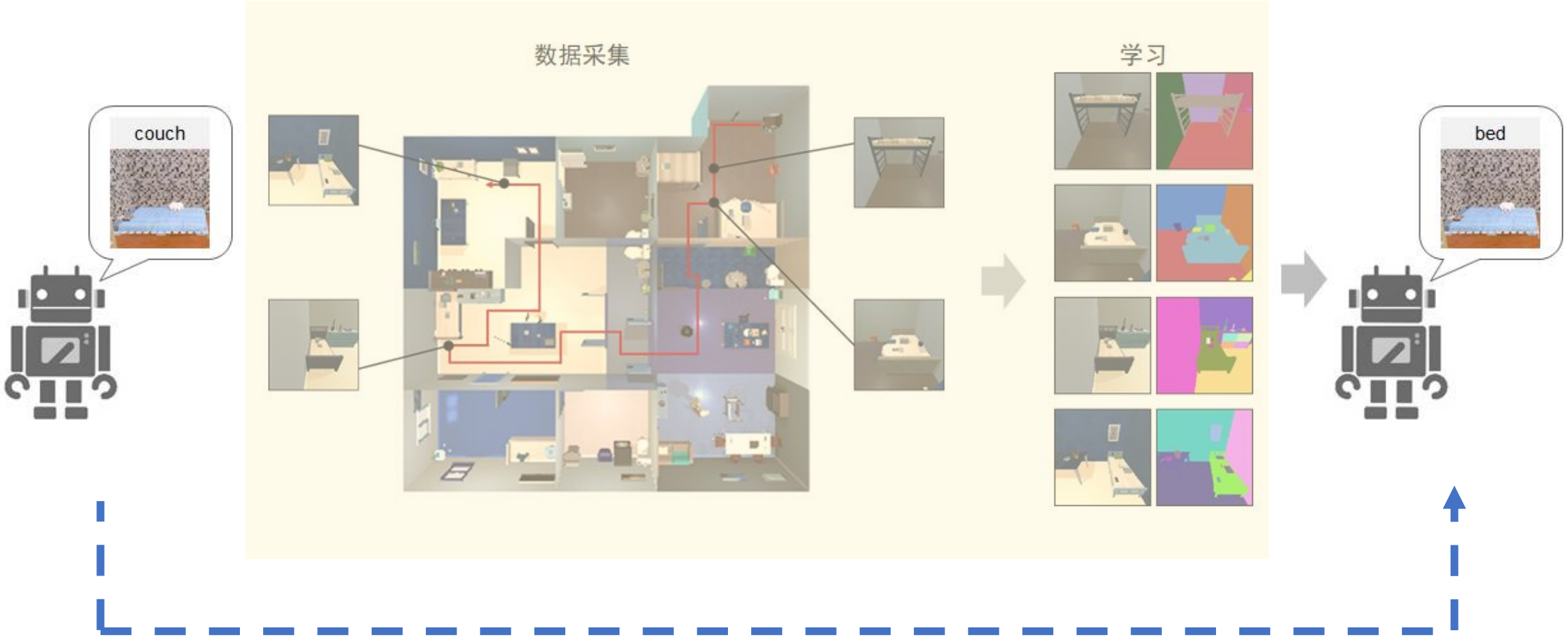
3.7 行为→学习：具身学习

- 人是如何学习的？



3.7 行为→学习：具身学习

➢ 问题描述

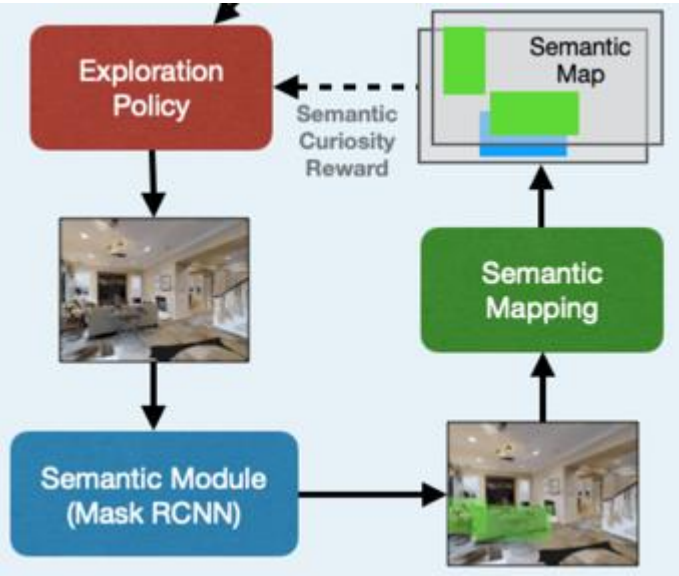


3.7 行为→学习：具身学习

➢ 问题描述

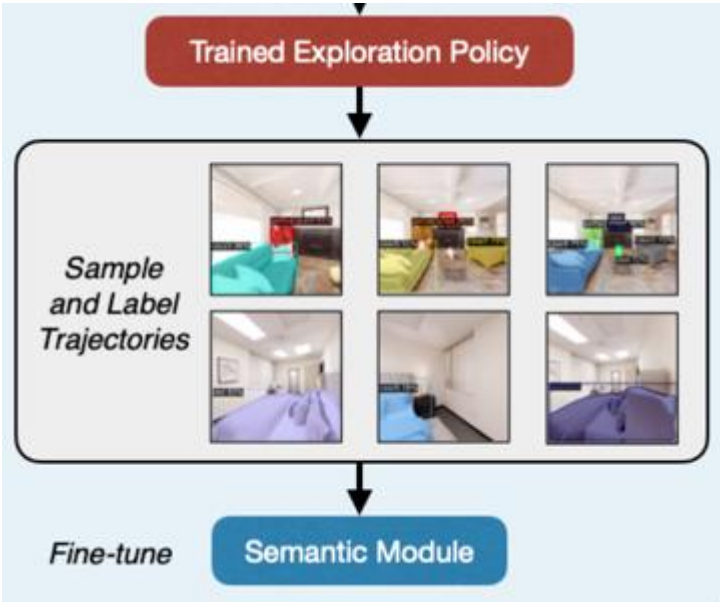
学习探索路径

学习如何探索环境才能获得更好的样本



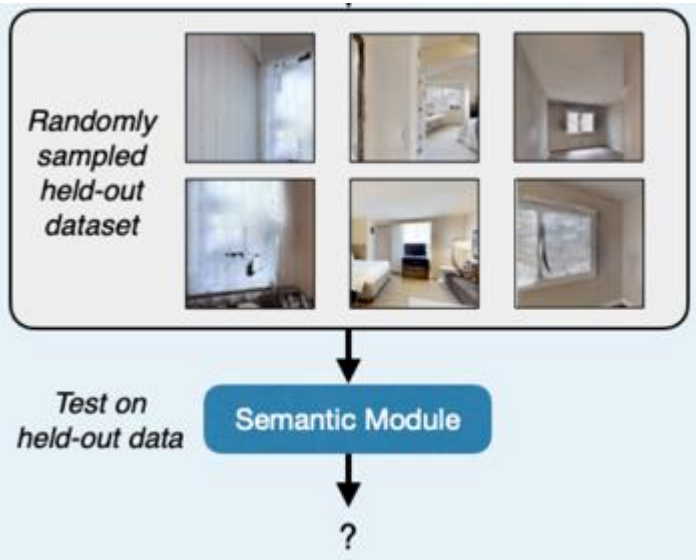
采集训练样本

利用探索策略获取样本与标签，提升感知器性能

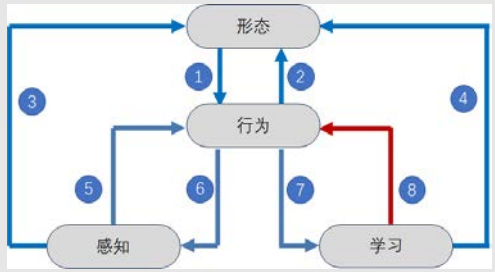


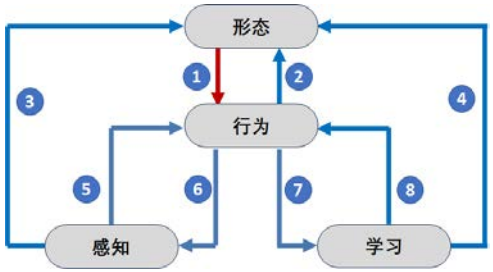
部署应用

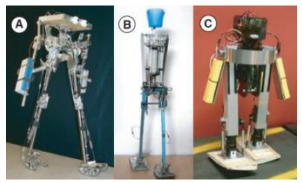
在新场景中对改进的感知器进行测试

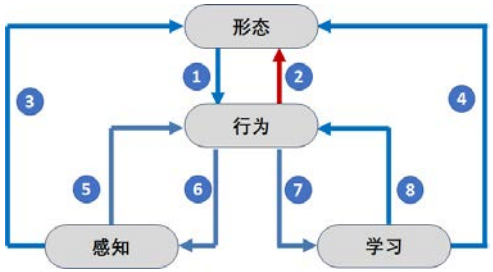


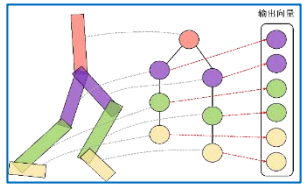
3.8 总结

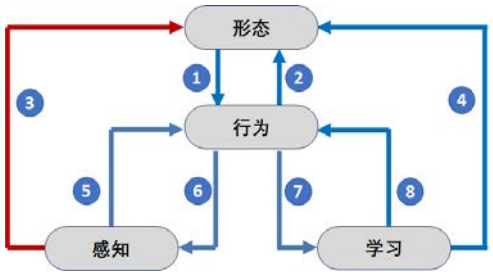

形态
行为
感知
学习
1
2
3
4
5
6
7
8
状态 s
奖励 r
智能体
环境
动作 a

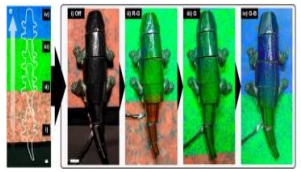

形态
行为
感知
学习
1
2
3
4
5
6
7
8

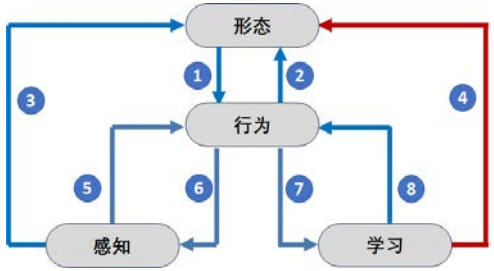

A
B
C

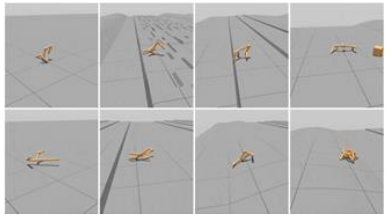

形态
行为
感知
学习
1
2
3
4
5
6
7
8

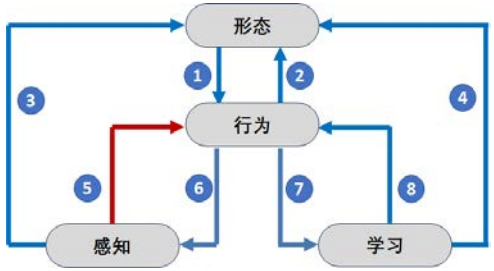

输出向量

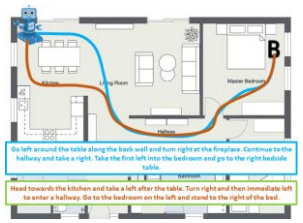

形态
行为
感知
学习
1
2
3
4
5
6
7
8

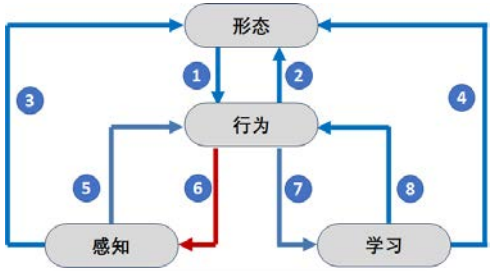


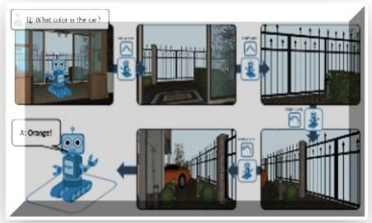

形态
行为
感知
学习
1
2
3
4
5
6
7
8

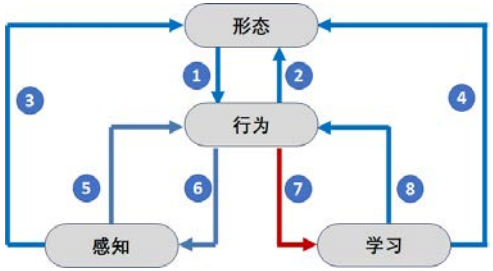


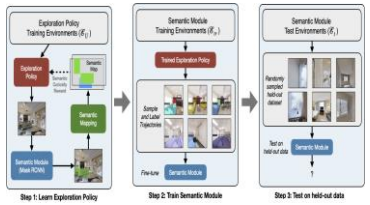

形态
行为
感知
学习
1
2
3
4
5
6
7
8


B
Go left around the table along the back wall and turn right at the fireplace. Continue to the hallway and take a right. Take the first left into the bedroom and go to the right bedside table.
Head towards the kitchen and take a left after the table. Turn right and then immediate left to enter a hallway. Go to the bedroom on the left and stand to the right of the bed.


形态
行为
感知
学习
1
2
3
4
5
6
7
8


Q: What color is the car?
A: Orange!

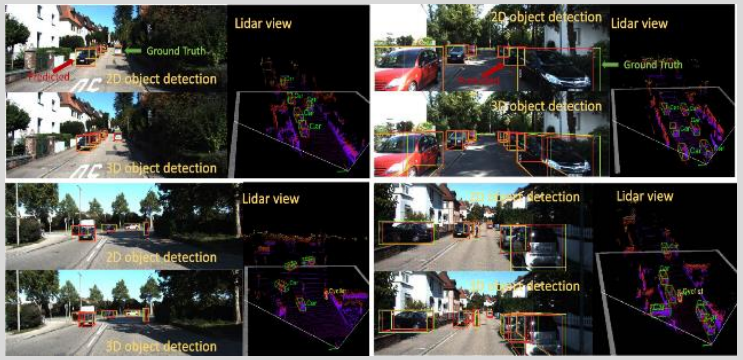

形态
行为
感知
学习
1
2
3
4
5
6
7
8

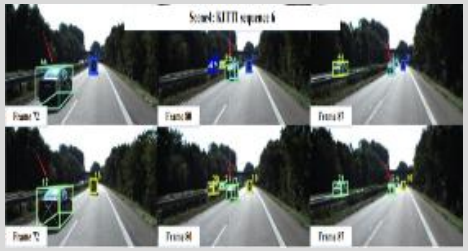

Exploration Policy
Training Environments (E_T)
Semantic Map
Semantic Mapping
Semantic Module
Shared ROMs
Semantic Module
Training Environments (E_{train})
Trained Exploration Policy
Sample and Label Trajectories
Fine-tune
Semantic Module
Semantic Module
Test Environments (E_T)
Randomly sampled held-out dataset
Test on held-out data
Semantic Module
Step 1: Learn Exploration Policy
Step 2: Train Semantic Module
Step 3: Test on held-out data

- 背景：具身智能
- 具身智能的体系
- 具身智能关键技术
- 探索与实践
- 面向具身智能的AIGC
- 总结

脑、身体与环境的交互中涌现智能

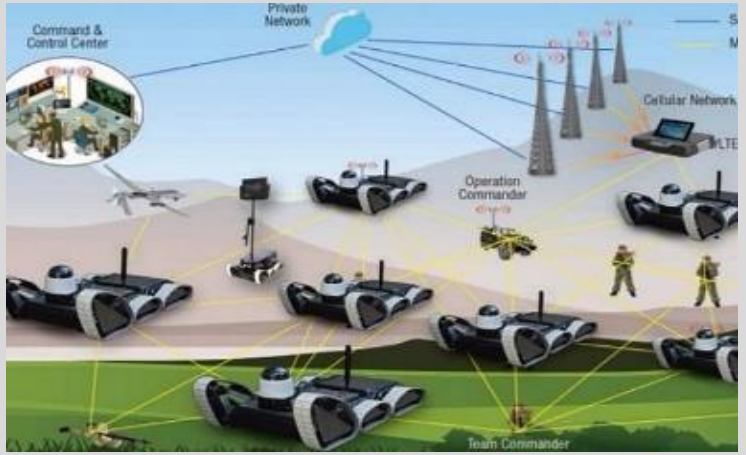
物理环境





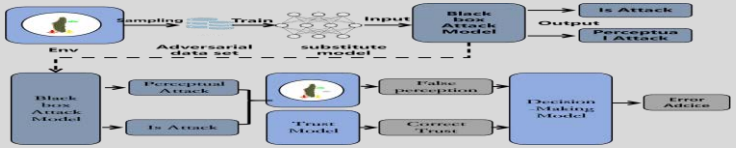
主动交互感知

智能体



异构随遇协同

人



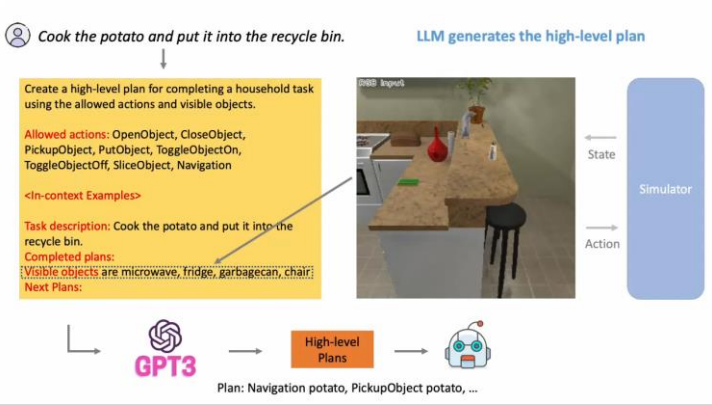
信任持续演进

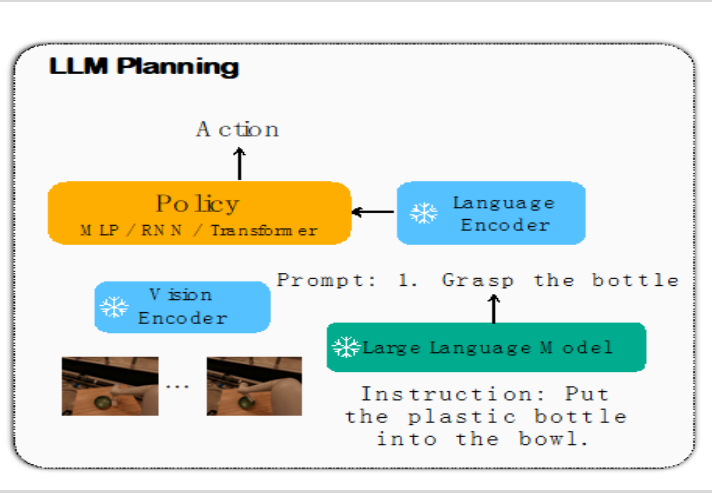
- 背景：具身智能
- 具身智能的体系
- 具身智能关键技术
- 探索与实践
- 面向具身智能的AIGC
- 总结

5 面向具身智能的AIGC

➢ 大模型

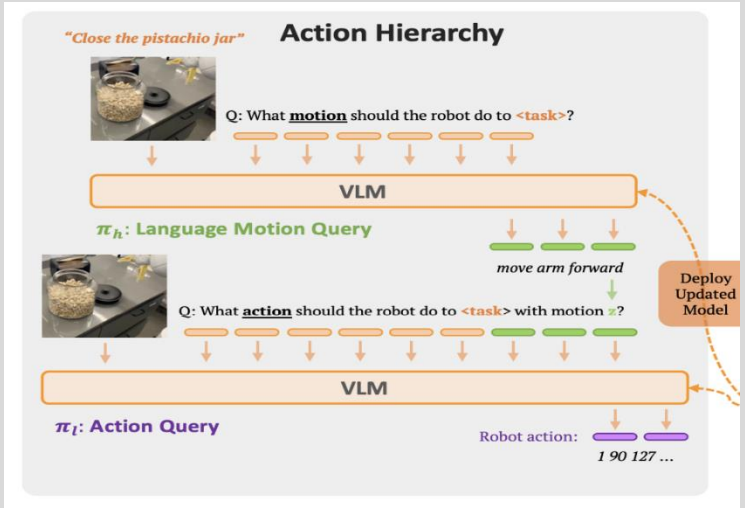
Planning





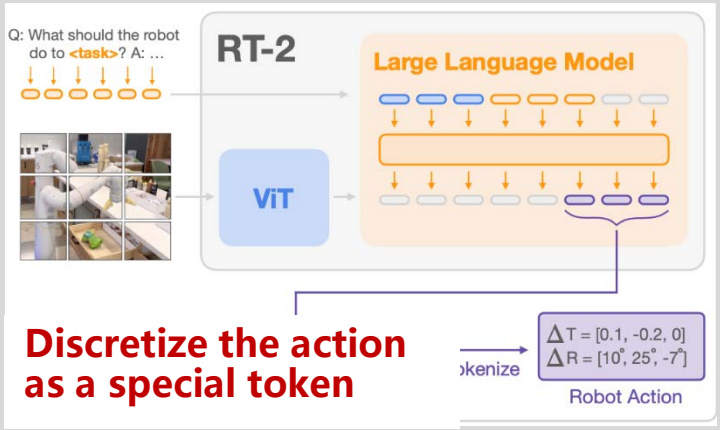
Hierarchy

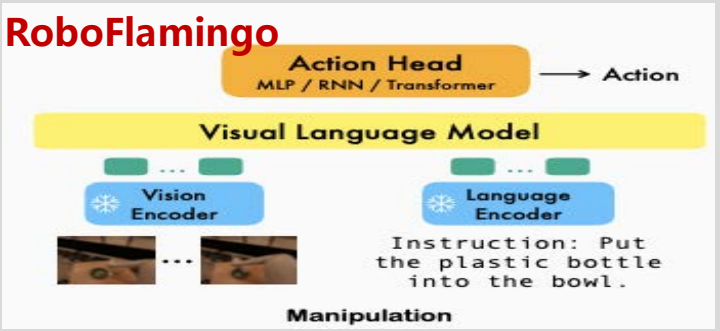
- Input image, state, 3D representation
- VLM output subgoal text tokens, language motions
- Low-level Action Model take these to output action



VLA

- VLM directly generate action token

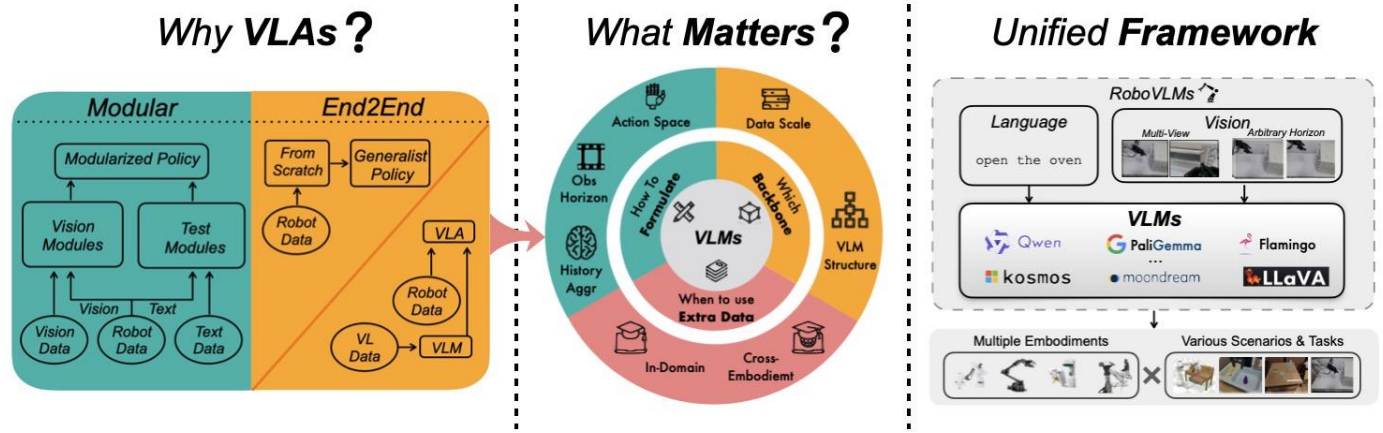


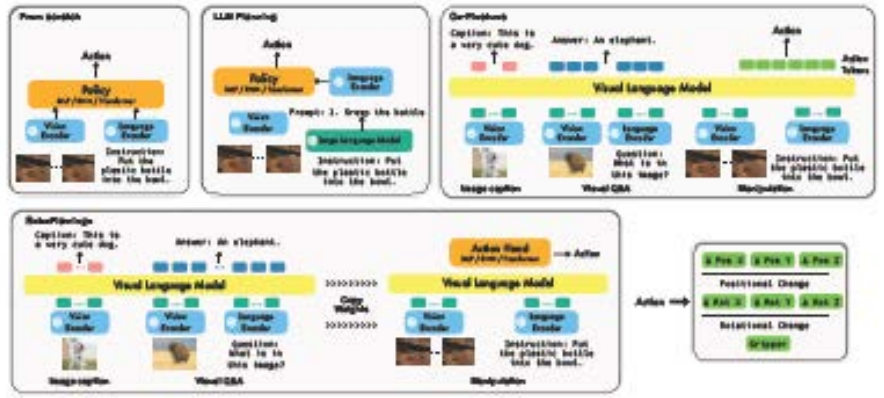


引入policy head，降低微调模型的成本

5 面向具身智能的AIGC

➢ 具身大模型





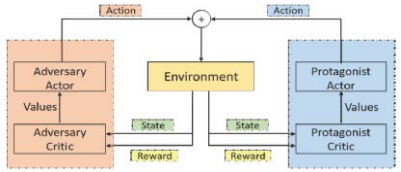
- Vision-language foundation models as effective robot imitators, ICLR, 2024

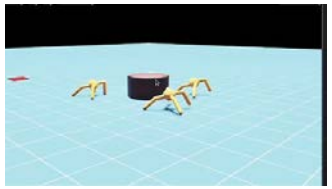
5 面向具身智能的AIGC

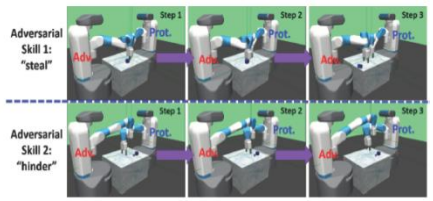
➢ 基于竞争、对抗的多体具身交互学习方法

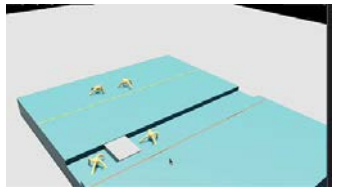
对抗行为学习

- 通过两个智能体之间的行为对抗（“偷走”与“抢回”），得到了鲁棒性更强的智能体行为策略。





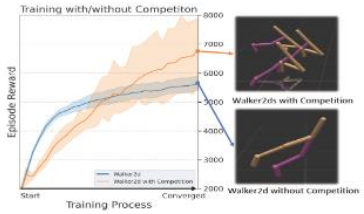


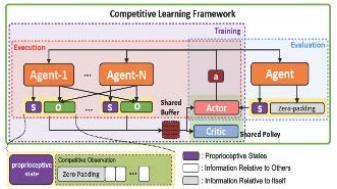


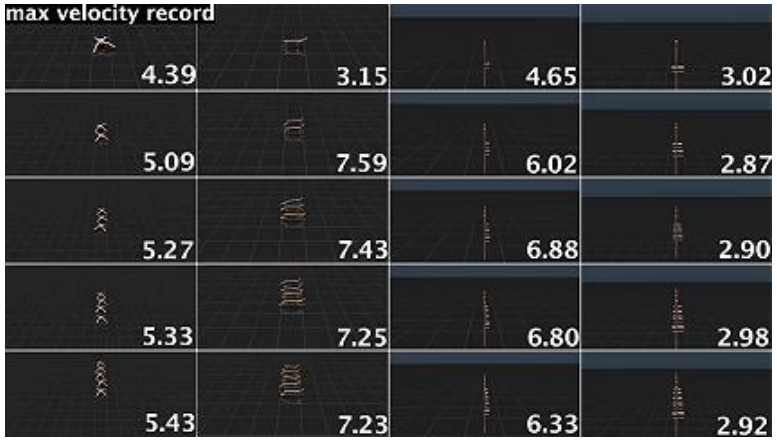
- Adversarial Skill Learning for Robust Manipulation,ICRA, 2021
- Learning a distributed hierarchical locomotion controller for embodied cooperation, CoRL, 2024

竞争行为学习

- 通过在智能体独立训练过程中引入竞争信息，得到了比单一训练更好的结果，揭示了“竞争促进学习”的机制。



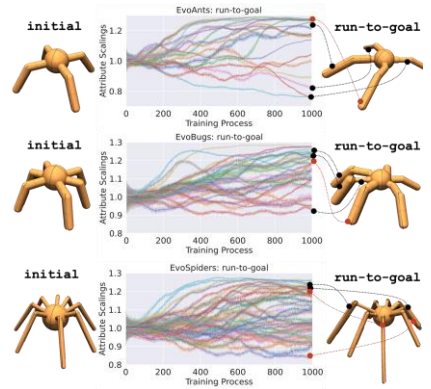


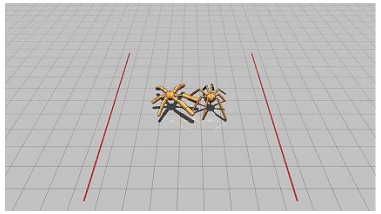


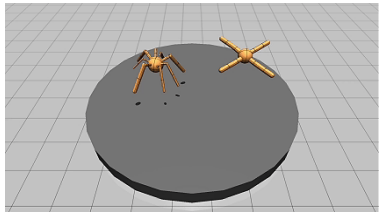
- Stimulate the Potential of Robots via Competition, ICRA, 2024
- Adversarial decision making against intelligent targets in cooperative multi-agent systems, IEEE T-CDS, 2023

对抗形态学习

- 模拟人类身体发育机理，探索了智能体在对抗过程中可改变形态的方法，为形态优化提供了新的解决思路。





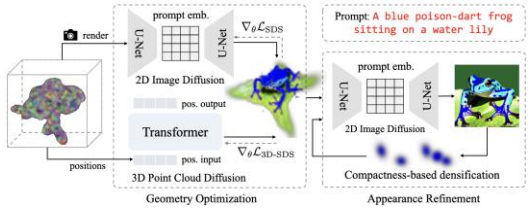


CompetEvo: Towards morphological evolution from competition, IJCAI, 2024

5 面向具身智能的AIGC

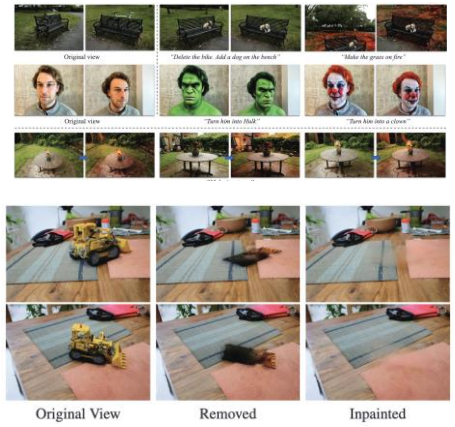
➢ 基于AIGC的材料-结构一体化学习

GSGen



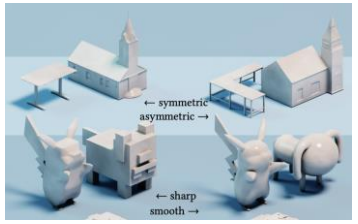
Text-to-3D using Gaussian Splatting

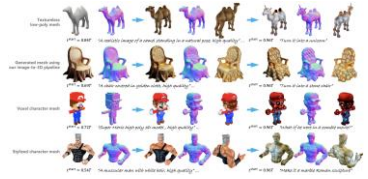
GaussianEditor



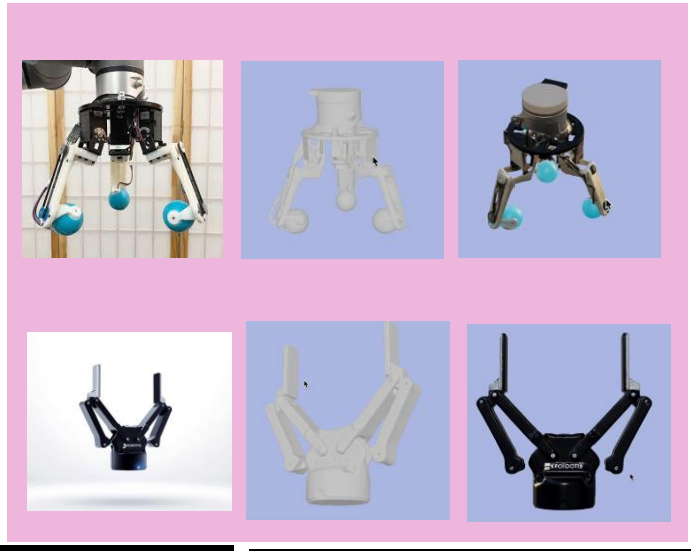
Editing any 3D Gaussians within minutes

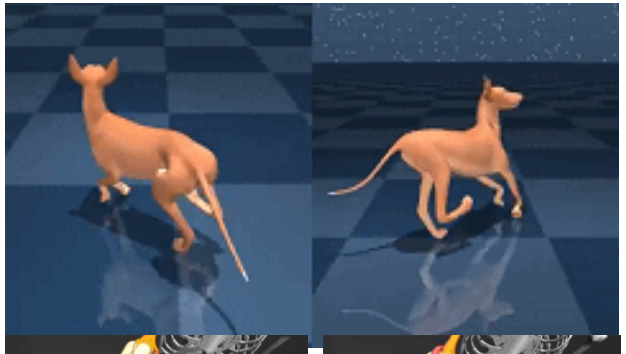
MeshEdit

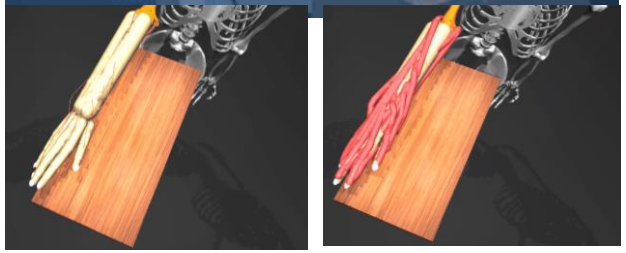


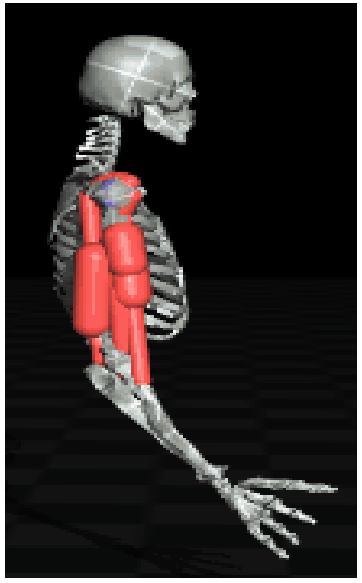


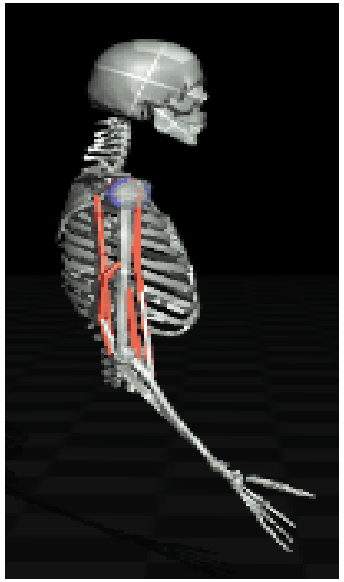
Mesh Editing with textual instruction and complex demands

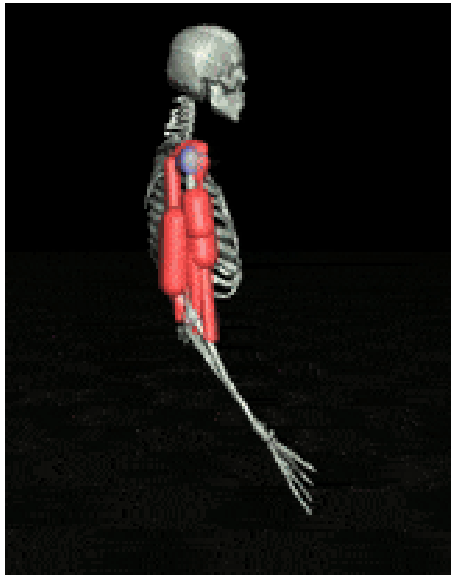




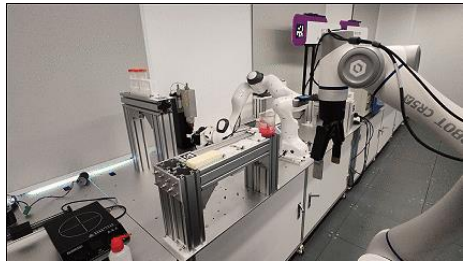


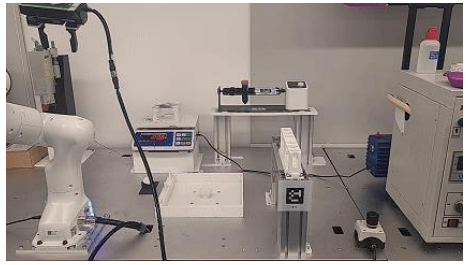


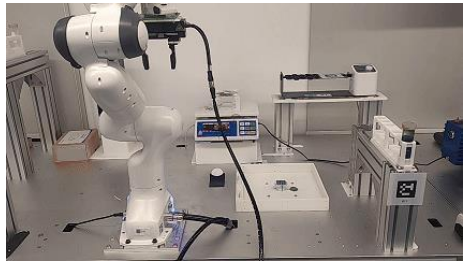




➢ 大模型驱动的具身智能科学实验室

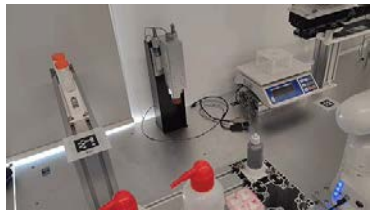






获取原料及模具

递送材料到操作区



获取试管

去除试管盖

放置试管到电子秤

添加硅胶

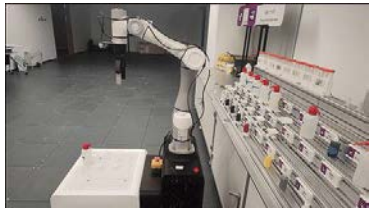
添加磁粉

添加试管盖

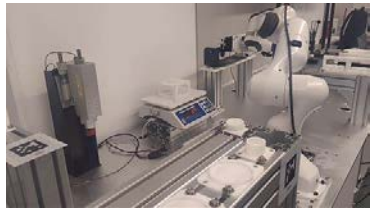
放置试管，开始搅拌

放置模具到电子秤

从工作区取回材料



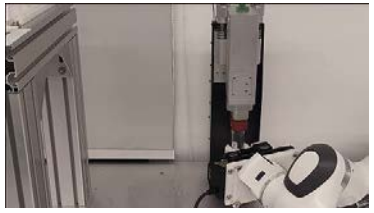
放回材料到材料区



从搅拌机取下试管

去除试管盖

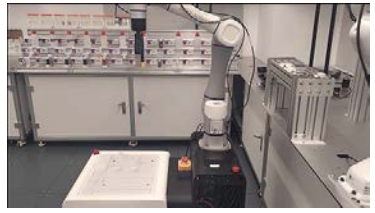
浇筑模型



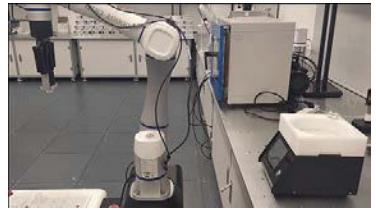
添加试管盖

放置废弃试管到废料区

取下浇筑好的模具



移动机器人获取模具

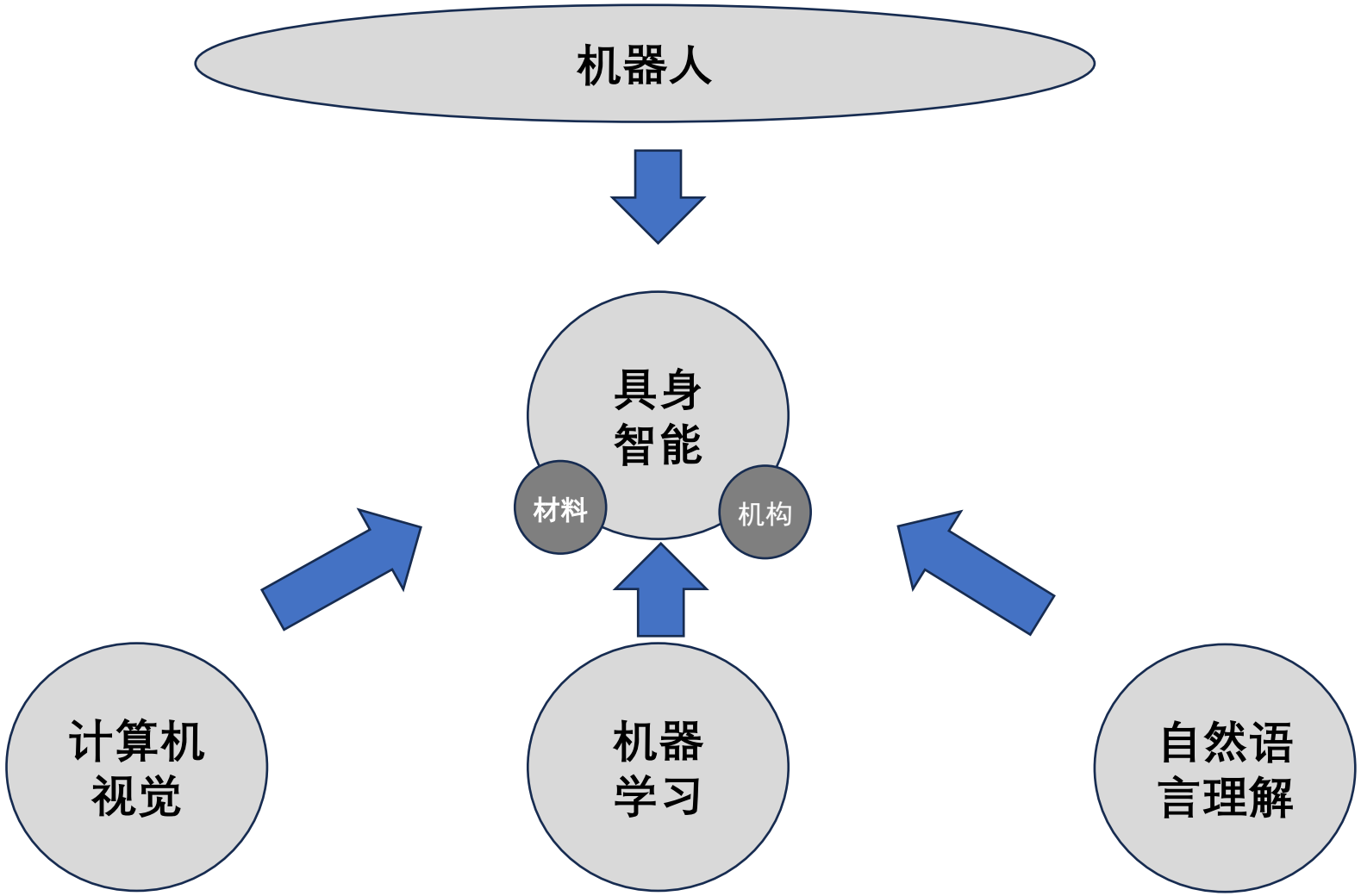


将模具放置到匀胶机

- 背景：具身智能
- 具身智能的体系
- 具身智能关键技术
- 探索与实践
- 面向具身智能的AIGC
- 总结

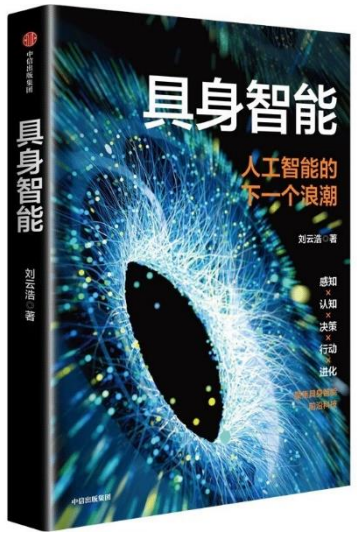
总结

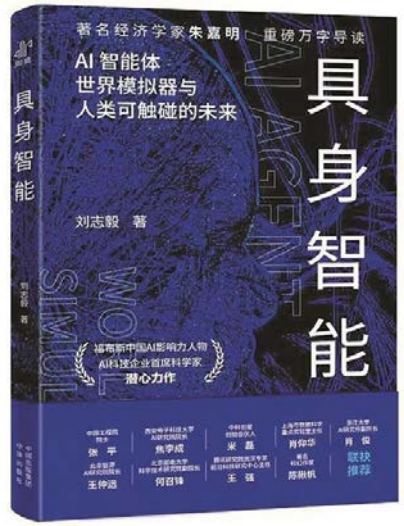
- 多学科交叉



教材


具身智能导论
刘华平 郭迪 孙富春 编著
工业和信息化部"十四五"规划教材
中国教育出版传媒集团
高等教育出版社


具身智能
人工智能的下一个浪潮
刘云浩 著
感知×认知×决策×行动×进化
具身智能
刘云浩 著


著名经济学家朱嘉明 重磅万字导读
AI 智能体
世界模拟器与
人类可触碰的未来
具身智能
刘志毅 著
AI AGENT
WORLD SIMULATOR
福布斯中国AI影响力人物
AI科技企业首席科学家
潜心力作
具身智能
张平
王仲选
焦李成
何召锋
宋晨
王强
肖仰华
郑纳机
肖俊
联袂推荐


具身智能机器人系统
Broadview
具身智能
机器人系统
甘一鸣 俞波 万梓泉 刘少山 著
中国工信出版集团
电子工业出版社


未可知人工智能研究院院长
《AIGC，智能创作时代》作者
全新力作
人形机器人、自动驾驶、智能家居
多场景设备
人机共生新世界
具身智能 从虚拟AI到现实AI
具身智能
从虚拟AI到现实AI 杜雨 张孜铭 著
中国出版集团
中译出版社


浅觉终来得上纸
行躬要事以知绝

谢　谢！