北京理工大學

# 机器人前沿技术课程报告

# 基于强化学习的仿生机器鼠行为交互系统设计

**Design of Robotic Rat Behavior Interaction System Based on Reinforcement Learning**


学　　院：机电学院

专　　业：机械工程

学生姓名：谢宏钊

学　　号：3120200167

指导老师：福田敏男


2021 年 5 月 13 日

# 第 1 章 绪论

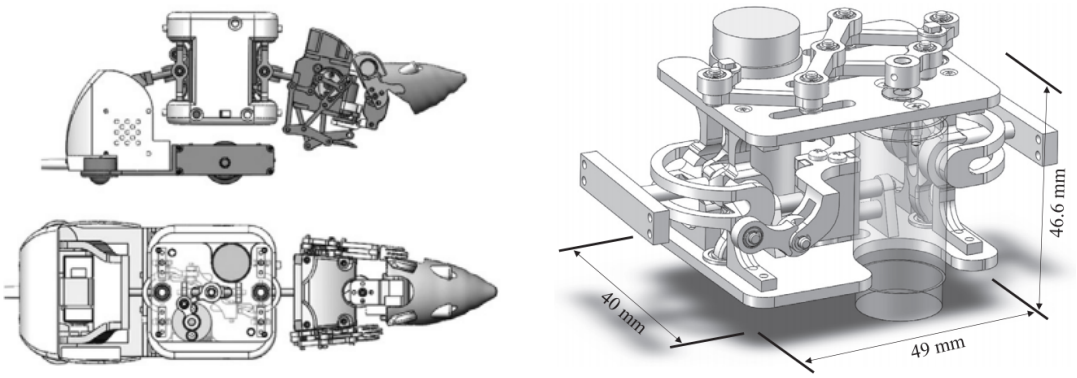
仿生机器人是机器人研究的一大分支，旨在通过研究生物系统的结构、性状、原理、行为以及相互作用，从而为机器人设计、控制和决策提供新的设计思想、工作原理和系统构成[1]，是一门集生命科学、物质科学、数学与力学、信息科学、工程技术以及系统科学等学科的交叉学科[2]。生物体结构合理，运动灵活，具有良好的环境适应特性和良好的生存能力，仿生机器人因其部分模仿生物结构和性状特点而继承了一部分这些特性，受到研究者的青睐[3]。生物鼠是被广泛使用的实验动物之一[4]，对其行为模式的研究收到生物学家的广泛关注，但由于生物鼠行为随机、难以预测，相关的实验开展存在困难。作为仿生机器人的一个基本实例, 设计精巧的仿生机器鼠可以模拟生物鼠各个关节的运动，从而产生与生物鼠相似的基本行为，并以此引发生物鼠的特异性反应[5]。因此，利用仿生机器鼠与生物鼠进行行为交互，探究交互过程中生物鼠的反应与仿生机器鼠行为的联系，对研究生物鼠的行为模式和机器人的控制策略均有重要意义[6]。

但在实践中，仿生机器鼠与生物鼠的行为交互仍然面临诸多困难。首先，两者交互时行为的一致性有赖于双方，特别是仿生机器鼠反应的快速性[7]。而生物鼠身体构造精巧，动作灵活，反应机敏，模仿其身体特点设计的仿生机器鼠往往具有复杂的结构和较多的自由度，给仿生机器鼠的动作规划带来了挑战。在产生仿鼠动作过程中，仿生机器鼠的控制系统往往需要耗费大量时间计算正、逆运动学问题，使得仿生机器鼠反应迟钝，难以满足交互实验的需求。其次，生物鼠行为表现具有普遍的个体差异[8]，这导致难以用特定的方法对其行为加以预测或控制，给仿生机器鼠的行为生成方式提出了挑战。单一机制的行为生成算法在经过调试后，往往只能适用于某一特定的生物鼠，当切换交互对象后，相应的算法往往需要进行调整，实验的可重复性大大降低。最后，生物鼠在实验环境中往往表现出渐进的环境适应性，这意味着同一生物鼠在实验之初和实验进行一段时间后表现出的行为模式存在较明显的差异。而单一的仿生机器鼠行为生成机制无法适应生物鼠的这一行为变化，这导致了部分控制策略在初期表现优秀，但随着时间推移，其可用性往往大不如前。上述现实困难使得现有的机器鼠与生物鼠的交互实验存在着可重复性差、应用场景单一和持续时间较短等不足。

# 第 2 章 仿生机器鼠及其动作规划

## 2. 1 仿生机器鼠模型

经过数十年的研究，现有的仿生机器鼠模型在尺寸和结构上已与生物鼠相近，且能够产生绝大数多数与生物鼠相似的交互行为，这为本文的研究提供了坚实的基础。我们设计并优化的仿生机器鼠（后称机器鼠模型，图2-1）具有优异的产生仿鼠运动的能力[9]，本文将利用这一模型开展研究。该模型的主要特点包括：




(a) 机器鼠模型　　(b) 腰部机构

图 2-1 机器鼠模型及其腰部机构[9]

（1）轮式驱动。

（2）腰部灵活。

（3）质量较小，运动灵活。

此外，机器鼠模型躯干部分拥有 7 个主动自由度，其头颈部活动范围广，响应速度快，能够执行嗅探、探索等高频动作。是较为理想的开展行为交互实验的媒介。其部分关键属性与生物鼠比较如表2-1所示。

表 2-1 方式机器鼠关键属性

| 质量 | 体长 | 体宽 | 初始姿态高度 | 自由度数目 | 最大速度 |
|---|---|---|---|---|---|
| 400 $g$ | 195 $mm$ | 56 $mm$ | 74 $mm$ | 11 | 1.5 $ms$ |

## 2.2 仿生机器鼠动作规划

行为（Behaviour）是指由个体、器官、系统或者人造物做出的一系列动作序列，这些动作往往是对自身或者包含自身和其他个体、器官或者系统的物理环境的反应[10]。尽管在生物学中，科学家对如何精确定义行为仍然存在分歧，但一种常见解释为，行为是包含其全部器官的内部协调反应，而这种反应是对其内部或外部刺激做出的[11]。由此可见，行为应当是一系列动作的有意义组合，这一“意义”应当是其目标的完成情况。对生物鼠而言，其行为应当是以引起其交互对象适当反应为目的，而开展行为交互实验的机器鼠也应当遵守这一准则。

一直以来，研究者们对生物鼠行为做了较为丰富、详细的研究。统计表明，生物鼠不同行为出现的频率存在较大差异。当两只生物鼠处于同意环境中时，仍然会产生大量不具有交互意义的行为，以时间度量，这些非交互性的行为占到所有行为的60% 左右[12]，同时那些交互性的行为所占时间比例也不同。

虽然生物鼠的行为受到多种因素影响，例如食物、水源等，但在行为交互实验中对生物鼠行为影响最为显著的因素为其交互伙伴的相关属性，这些属性包括其气味、性别、年龄等实验中的不变特性，也包括其动作、叫声等在实验过程中经常变化的特性[13]。Barnett 对生物鼠的行为进行的研究表明，当两只生物鼠交互时，它们将倾向于表现征服、服从、梳理和嗅探四种行为模式，这四种模式具有一定的个体对称性，例如，当其中一只生物鼠处于征服状态时，另一只生物鼠通常表现为服从状态[8]。而当两只生物鼠处于未交互状态时，其表现行为包括探索、直立等个体行为。

根据上述分析，对仿生机器鼠的动作设计应当遵循相似性、紧密性和典型性的原则，对相应各选取原则的解释如下。

（1）与生物鼠相应动作具有相似性。

（2）与交互紧密相关。

（3）具有典型性。

根据上述原则，本文选定直行、后退、左转、右转、直立、嗅探、梳理、被梳理、匍匐和攀爬 10 种动作作为仿生机器鼠的基本动作。

# 第 3 章 仿生机器鼠行为交互仿真平台

ROS 利用 URDF（Unified Robot Description Format）描述机器人模型，URDF 通过 xml 格式对机器人系统的模型、驱动器、传感器和场景等进行组织，开发者只需添加相应的标签对上述属性进行定义。同时，该文件提供了一套便于控制程序与 Gazebo 进行信息交流的机制。为提高系统的鲁棒性、可扩展性，本文根据 ROS 松耦合的特性，建立了基于关节控制、动作执行和行为生成三层系统架构的仿真平台。

## 3.1 关节控制层

如图3-1所示，关节控制层的主要作用为完成对 Gazebo 中仿生机器鼠模型各关节控制方式的抽象，并为上层控制器提供相应的程序接口，提高复用率和仿真平台的稳定性。

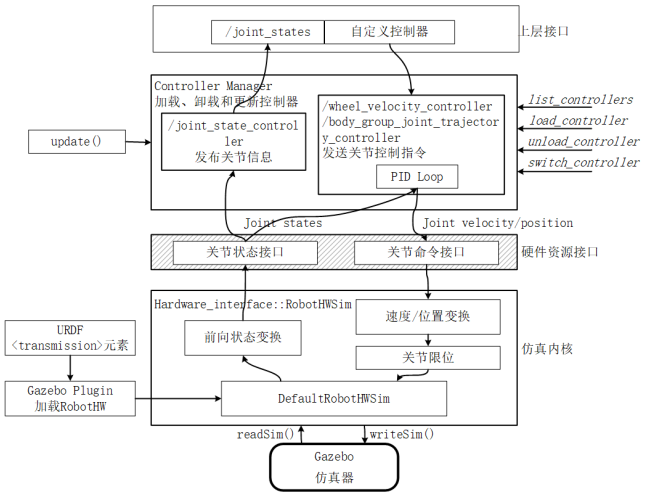



图 3-1 关节控制层数据流图

## 3.2 动作执行层

动作即为一个或数个关节相互配合的运动，在关节控制层已搭建的基础上，仿真系统将依靠调用相关的控制接口实现特定动作。

仿生机器鼠的动作分为两大主要板块:轮部运动和躯干运动。在关节控制层中,机器鼠轮部关节控制器为速度控制器，订阅机器鼠所命名空间下的/left_wheel_joint_velocity 和/right_wheel_joint_velocity 话题。因此在动作执行层中，只需向上述话题发布速度指令即可完成控制。

动作执行层数据流为图3-2。

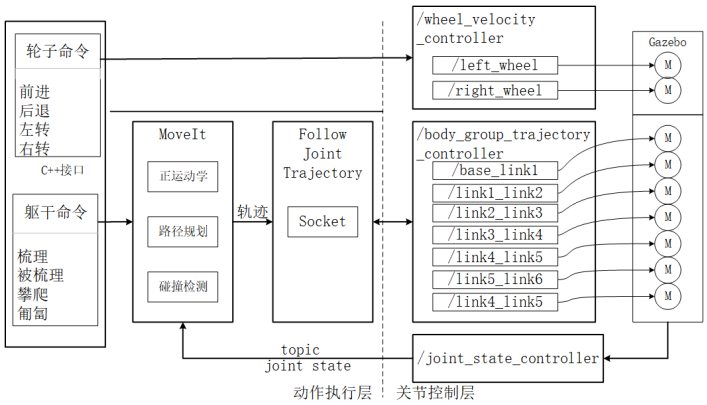



图 3-2 动作执行层数据流图

## 3.3 行为生成层

研究者对生物鼠的行为表现进行了统计并进行了数据分析，结果表明生物鼠不同行为的表现频率不同，并建立了相应的生物鼠行为表现的数据库[12]。其研究提供了丰富的关于生物鼠行为表现的统计资料，本文将根据这一研究已揭示的生物鼠不同行为表现概率设计仿真平台行为生成层的行为决策机制。

根据不同状态设定不同的动作序列，并从 Gazebo 仿真环境中获取相应的反馈信息是必要的，相应的仿生机器鼠状态判断机制（图3-3）可以通过 Gazebo 提供的相关话题和服务实现。

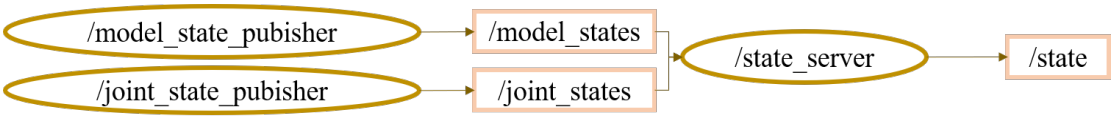



图 3-3 仿生机器鼠状态判断机制

根据上述分析，本文建立的行为生成层数据流为图3-4。

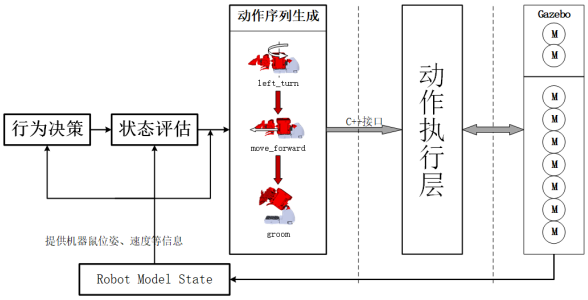



图 3-4 行为生成层数据流图（规则鼠）

# 第 4 章 基于强化学习的仿生机器鼠行为控制方法

## 4. 1 强化学习简介

强化学习（Reinforcement Learning）这一名词来源于心理学。俄国心理学家 Pavlov 在 1927 年发表的《*Conditioned Reflexes*》中用“强化”（reinforcement）一词描述特定刺激使生物趋于采取某种策略的现象，相应的刺激称为“强化物”（reinforcer）[14]。在后续的研究中，心理学家将强化分为正强化（positive reinforcement）和负强化（negative reinforcement），其中正强化使得生物趋于做出能够获取更多利益的行为，而负强化则使其避免损害[15]。

生物鼠行为交互过程中下一状态只与当前状态有关，具有马尔可夫性。本文根据生物鼠交互的行为特性，划分有限的状态和动作集合，利用 Q-学习算法进行训练，在仿真平台中进行了测试。Q-学习的流程用伪代码表示为算法4-1。

**算法 4-1** Q-学习流程

以任意方式初始化 $Q(s,a)$，其中 $Q(terminal-state,\cdot)=0$;
**for each** episode **do**
    初始化状态 $S$;
    **while** $S$ **is not** terminal-state **do**
        根据 $S$ 和 $Q(S,\cdot)$ 选择动作 $A$;
        执行动作 $A$，观测奖励 $R$ 和下一状态 $S'$;
        $Q(S,A)\leftarrow Q(S,A)+\alpha\left[R+\gamma\max_a Q(S',a)-Q(S,A)\right]$;
        $S\leftarrow S'$;
    **end while**
**end for each**

## 4. 2 基于 Q 学习的行为决策机制

### 4. 2. 1 Q 表初始化

实现 Q-学习最简易的途径为通过 Q 表，该表格存储所有状态下采取各种动作后的奖励值，因而 Q 表的大小受限于状态和动作空间的数目。

在仿真平台中，基于强化学习控制的机器鼠（学习鼠）在产生行为时与规则鼠一样需要调用其动作执行层的接口函数，因此其动作空间为相应接口集合的子集。考虑到不同动作对行为交互的影响[8]，为降低 Q 表的大小，训练的复杂度，本文将前进、后退、左转、右转、梳理、被梳理、攀爬、匍匐共 8 种动作作为 Q 学习中的动作集合。

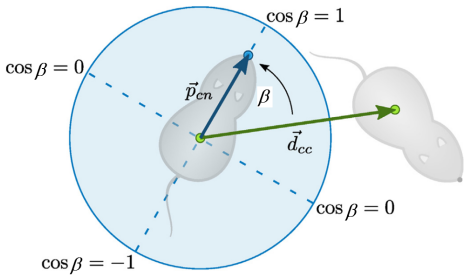



图 4-1 状态分类依据[12]

对行为交互中过程的分类应当充分考虑生物鼠的行为决策依据。如图4-1，Lorbach et al. 在研究生物鼠交互行为时利用相互间的相对位置确定所处状态，图中，$d_{cc}$ 表示两者中心点的距离，$\beta$ 为自身中心点与鼻端连线和自身中心点与交互对象中心点连线构成的夹角，考虑到角度具有对称性，用 $\cos\beta$ 和 $\sin\beta$ 的值进行划分[12]。

同时考虑交互对象执行的动作对状态的影响，在两者距离较近时用动作进行状态分类。由此得到的行为交互中的状态共有背后、左侧、右侧、远距、梳理、被梳理、攀爬、匍匐、其他共 9 种。

据此建立的 Q 表大小为 $9 \times 8$，在程序中用二维数组表示。根据 Q-学习算法（算法4-1），Q 表初始值可以是任意的，需要保证最终状态下所有动作的 Q 值为 0（$Q\left(terminal-state,\cdot\right)=0$）[16]，据此，本文将 Q 表中所有值均初始化为 0。

### 4. 2. 2 动作选择策略

常见的 Q-学习动作选择策略包括 $\epsilon-$ 贪心算法（$\epsilon-greedy$）、置信区间上界算法（$UCB$）、汤普森采样（$Thompson\ sampling$）等，$\epsilon-greedy$ 算法实现较为简单，在各种情况下均有较好的适应性，能够较好地平衡“探索”和“利用”行为，因此本文选定其作为学习鼠的动作选择策略。

$\epsilon-greedy$ 算法的数学表达为式4-1，式中，$\pi\left(a^{*}|s\right)$ 表示学习鼠在状态 $s$ 条件下做出动作 $a^{*}$ 的概率，$\epsilon$ 为贪心率，取值范围为 $[0,1]$，$m$ 为动作空间的大小。

$$\pi\left(a^{*}|s\right)=\begin{cases}\dfrac{\epsilon}{m}+1-\epsilon, & 当\ a^{*}=arg\max\limits_{a\in\mathbb{A}}Q(s,a)\\ \dfrac{\epsilon}{m}, & 其他.\end{cases} \tag{4-1}$$

# 第 5 章 仿真结果与分析

## 5. 1 训练结果

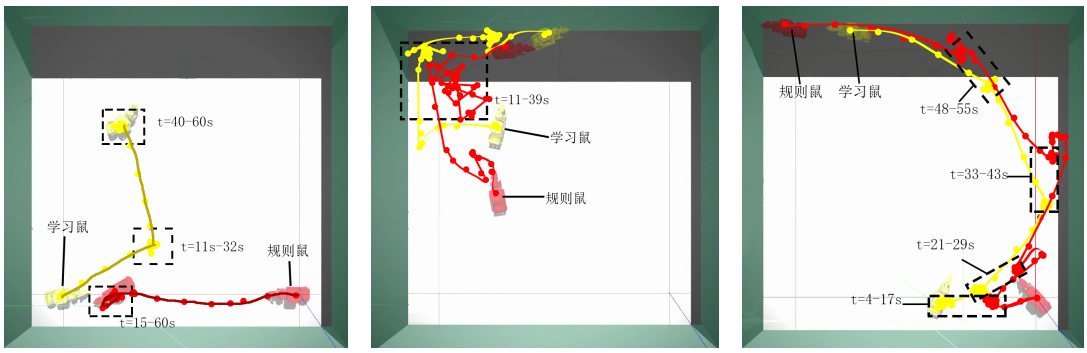
在训练进行约 100 h 后，Q 表逐渐收敛，此时学习鼠行为模式能够适应规则鼠的变化。

图5-1展示了在实验中各阶段机器鼠在 1 $min$ 内的活动区域及运动轨迹，阶段划分标准为：

（1）初期：实验启动期，迭代次数 $<$ 1000，Q 表中至少有一处状态-动作组合对应的值未完成更新，即保持初始值 0。这一阶段大约在实验启动后的 10 $min$ 内结束。

（2）中期：1000 $<$ 迭代次数 $<$ 10000，Q 表全部状态-动作组合对应的值完成了至少一次更新，即学习鼠已经对所有动作进行了尝试。

（3）后期：迭代次数 $>$ 10000，Q 表趋于收敛，经过实验，这一阶段通常为训练 13 $h$ 之后。




(a) 初期　　(b) 中期　　(c) 后期

图 5-1 实验各阶段机器鼠活动区域热点图及运动轨迹

## 5. 2 数据分析

距离是评价机器鼠行为交互的重要指标之一，图5-2为图5-1中各阶段 1 $min$ 内距离变化情况。该图显示，在训练初期，二者距离变化较少，这与图5-1(a)中存在的两处学习鼠停留较长位置相符。

在进入训练中期后，学习鼠能够更频繁地进入与规则鼠开展有效交互地距离（$d_{cc} < 0.3\ m$）范围（13 $\sim$ 22 $s$，35 $\sim$ 39 $s$ 和 27 $\sim$ 29 $s$ 等），这表明了学习鼠已经学习到部分有利于行为交互的规则。

在训练后期，学习鼠能够频繁进入规则鼠的有效交互距离以内，同时，当两者距

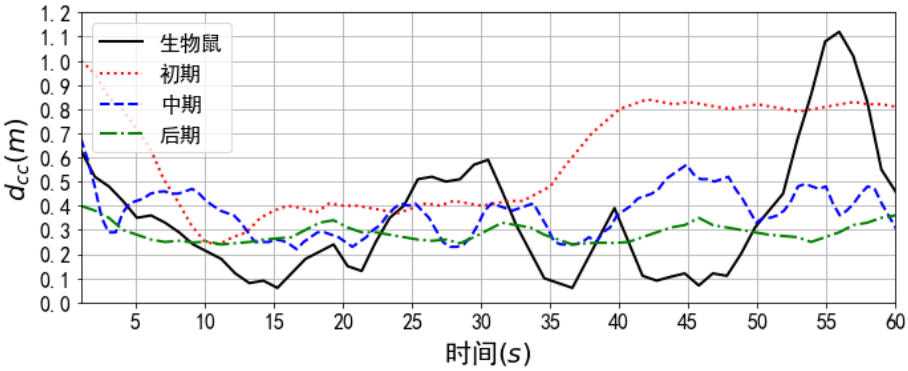



图 5-2 机器鼠间的距离变化（1 $min$）

离超出这一范围时，由于学习鼠经训练获得的跟随特性，其能够迅速重新建立行为交互的联系。但与生物鼠相比，机器鼠的距离变化跟平缓，且运动范围明显更小：生物鼠之间的距离最远超过 1 $m$，最近距离仅为 0.06 $m$，机器鼠距离始终在 0.25 ~ 0.4 $m$ 之间，这表明机器鼠的活跃程度与生物鼠存在差异。造成这种差异的原因是由于 Q-学习目标的单一性（本文设定学习鼠训练目标为与规则鼠开展有效交互），学习鼠必须时刻靠近规则鼠。同时由于状态划分以 $d_{cc} = 0.3\ m$ 为界限，使得学习鼠在进入这一范围后不再靠近规则鼠。

上述分析表明了基于强化学习的仿生机器鼠行为交互具有良好的适应规则鼠随机性的能力，仿生机器鼠能够在仿真平台中表达与生物鼠相似的行为交互。

# 参考文献

[1] 孙久荣, 戴振东. 仿生学的现状和未来 [J]. 生物物理学报, 2007(02): 109-115.

[2] 王国彪, 陈殿生, 陈科位, 等. 仿生机器人研究现状与发展趋势 [J]. 机械工程学报, 2015, 51(13): 27-44.

[3] 沈惠平, 马小蒙, 孟庆梅, 等. 仿生机器人研究进展及仿生机构研究 [J]. 常州大学学报 (自然科学版), 2015, 27(01): 1-10.

[4] 孔琪, 夏霞宇, 秦川. 实验动物品系数据库的建立 [J]. 中国比较医学杂志, 2015, 25(04): 78-83.

[5] Gao Z, Shi Q, Fukuda T, et al. An Overview of Biomimetic Robots with Animal Behaviors [J]. Neurocomputing, 2019, 332: 339-350.

[6] Frohnwieser A, Murray J C, Pike T W, et al. Using Robots to Understand Animal Cognition: Robots in Animal Cognition [J]. Journal of the Experimental Analysis of Behavior, 2016, 105(1): 14-22.

[7] Klein B A, Stein J, Taylor R C. Robots in the Service of Animal Behavior [J]. Communicative & Integrative Biology, 2012, 5(5): 466-472.

[8] Barnett S. The Rat: A Study in Behavior [M]. Revised Edition. Canberra: Australian National University Press, 1976.

[9] Li C, Shi Q, Gao Z, et al. Design and Optimization of a Lightweight and Compact Waist Mechanism for a Robotic Rat [J]. Mechanism and Machine Theory, 2020, 146: 103723.

[10] Minton E A, Kahle L R. Belief Systems, Religion, and Behavioral Economics [M]. First Edition. New York, USA: Business Expert Press, 2013.

[11] Levitis D A, Jr W Z L, Freund G. Behavioural Biologists Do Not Agree on What Constitutes Behaviour [J]. Animal Behaviour, 2009, 78: 103-110.

[12] Lorbach M, Kyriakou E I, Poppe R, et al. Learning to Recognize Rat Social Behavior: Novel Dataset and Cross-Dataset Application [J]. Journal of Neuroscience Methods, 2018, 300(SI): 166-172.

[13] Whishaw I Q, Kolb B. The Behavior of the Laboratory Rat: A Handbook with Tests [M]. First Edition. Oxford, New York: Oxford University Press, 2005.

[14] Pavlov P I. Conditioned Reflexes: An Investigation of the Physiological Activity of the Cerebral Cortex [M]. First Edition. Oxford: Oxford University Press, 1927.

[15] Michael J. Positive and Negative Reinforcement, a Distinction That Is No Longer Necessary; Or a Better Way to Talk about Bad Things [J]. Journal of Organizational Behavior Management, 1975, 3(1): 33-44.

[16] Sutton R S, Barto A G. Reinforcement Learning: An Introduction [M]. Second Edition. Cambridge, USA: MIT Press, 2018.