基底神经节强化学习的 *Actor-Critic* 模型:从自然大鼠到人工大鼠

Actor–Critic Models of Reinforcement Learning in the Basal Ganglia: From Natural to Artificial Rats

Mehdi Khamassi[1,2], Loïc Lachèze[1], Benoît Girard[1,2], Alain Berthoz[2], Agnès Guillot[1]

[1]AnimatLab, LIP6, Paris, France

[2]LPPA, CNRS – Collège de France, Paris, France

（Song Jian，translate）

自 1995 年以来，许多针对强化学习的 *Actor-Critic* 架构被提出作为大鼠基底神经节类多巴胺强化学习机制的模型。然而，这些模型通常在不同的任务中进行测试，因此很难比较它们对自主“动作规划者”（animat）的效率。在这里，我们将比较一个“动作规划者”中的四个架构，因为它执行相同的“奖励-寻找”（reward-seeking）任务。这将说明不同的假设对不同的 *Actor* 子模块和 *Critic* 单元的管理的后果，以及它们或多或少自主决定的协调。我们说明了，“专家”混合协调模块的经典方法，根据每个模块的性能，不允许解决我们的任务。然后，我们讨论了如何有效地应用哪种原理来组合这些单元。最后从我们的 Psikharpax 项目的角度讨论了 *Critic* 模型的改进和自然任务的 *Actor-Critic* 模型的准确性。该项目是一只人工老鼠，必须在不可预测的环境中自主生存。

关键词：animat 方法；TD 学习；Actor-Critic 模型；S-R 任务；分类单元导航

1 引言

这项工作的目的是在本期文章中，在 Girard, Filliat, Meyer, Berthoz 和 Guillot (2005)引入的行动选择（Action selection）架构中增加学习能力。这一架构将在人工老鼠 Psikharpax 中实现，该机器人将至少展示出其天然对应物的一些自主和适应能力（Filliat 等人，2004）。这一学习过程利用了 *Actor-Critic* 架构，该架构已被提出作为大鼠基底神经节类多巴胺强化学习机制的模型（Houk、Adams 和 Barto，1995）。在这样的模型中，*Actor* 网络学会选择 *Actor*，以便最大限度地增加未来奖励的加权和，就像另一个网络 *Critic* 在线计算的那样。*Critic* 通过时间差异（TD）学习规则将其对回报的估计与实际回报的估计进行比较来预测这个总和，其中两个连续预测之间的误差用于更新突触权重（Sutton 和 Barto，1998）。最近对自 1995 年以来基于这一原则建立的许多计算模型的回顾，强调了由于 *Actor* 和 *Critic* 模块的详细实现与已知的基底神经节解剖和生理学不一致所引起的几个问题（Joel、Niv 和 Ruppin，2002）。在本文的第一部分中，我们将考虑一些主要问题，这些问题将随着解剖学和神经生理学知识的更新而更新。在第二部分中，我们将通过比较执行相同经典操作条件反射学习（S-R 任务）的“动作规划者”来说明关于不同 *Actor-Critic* 设计的替代假设的结果。在测试中，动物在一个正迷宫中自由移动，奖励放在一只迷宫通道的末端。奖励地点在每次试验开始时随机选择，它是指特定地点的局部刺激。动物必须自主地学习将连续的感官信息与一定的奖励价值联系起来，并选择一系列的 *Actor*，使其能够从迷宫中的任何地方达到目标。这个实验比其他用来验证 *ActorCritic* 模型的实验更现实，这些模型的特点是刺激和奖励之间存在先验的固定时间间隔（例如，Suri 和 Schultz，1998），在试验中奖励位置不变（例如，Strosslin，2004），或者是离散状态空间（例如，Baldassarre，2002）。

在这项任务中，我们将比较四个不同的原则，这些原则是由试图解决第一部分中提到的问题的 *Actor-Critic* 模型所启发的。第一个是 Houk 等人提出的开创性模型。（1995），它使用一个 *Actor* 和一个预测单元（模型 AC：一个 *Actor*，一个 *Critic*），这应该能够在整个环境中诱导学习。第二个原则实现了一个 *Actor* 和几个 *Critic*（模型 AMC1：一个 *Actor*，多个 *Critic*）。*Critic* 是由混合的“专家”组成的，他们使用一个门控网络来决定在环境的每个区域中使用哪一个 *Critic*，这取决于该区域的性能。“专家”混合的原则是从几个现有的

模型（Jacobs，Jordon，Nowlan 和 Hinton，1991；Baldassarre，2002；Doya，Samejima，Katagari 和 Kawato，2002）中得到启发的。第三个是受 Suri 和 Schultz（2001）的启发，也使用了一个 *Actor* 和几个 *Critic*“专家”。然而，哪位“专家”应该在环境的每个子区域工作的决定与“专家”的表现无关，而是取决于“动作规划者”感知的感官空间的划分（模型 AMC2：一个 *Actor*，多个 *Critic*）。第四个原则（模型 MAMC2：多个 *Actor*，多个 *Critic*）提出了与前一个 *Critic* 相同的原则，结合了多个 *Actor*，后一个原则是 Doya 等人（2002）模型的特征之一。特别是为连续任务而设计，也是 Baldassarre 模型（2002）的一个特征。在这里，我们在四个模型中实现这些原则，对每个 *Actor* 组件使用相同的设计。比较了学习速度和将学习扩展到整个实验环境的能力。

本文最后一部分在对人工和自然啮齿动物强化学习任务的知识掌握的基础上，对实验结果进行了讨论。

2Actor-Critic 设计：问题

Actor-Critic 模型的两个主要原则是（i）实施 TD 学习规则，从而逐步翻译强化信号。从奖励发生的时间到奖励之前的环境情境，以及（ii）将模型分为两个不同的部分：一个用于根据当前的感官输入(*Actor*)选择运动行为(Action)，另一个用于通过多巴胺信号驱动学习过程(*Critic*)。

Schultz 对猴子多巴胺神经元电生理学的研究表明，多巴胺释放模式类似于 TD 学习规则（参见 Schultz，1998 年的综述）。此外，基底神经节是多巴胺神经元的主要输入，也是这些神经元发送强化信号的一个特殊目标（Gerfen、Herkenham 和 Thibault，1987）。此外，基底神经节似乎是由两个不同的子系统,与纹状体的两个不同部分的主要输入核基底神经节丘脑投射到运动区，另一个突出的多巴胺神经元，影响这些神经元的放电模式，至少在某种程度上（Joel 和 Weiner，2000）。

这些特性导致基底神经节的第一个 *Actor-Critic* 模型提出纹状体的基质构成 *Actor*，而这一结构的纹状体是 *Critic*（Houk 等人，1995，图 1）。模型中使用了从纹状体到多巴胺能系统的“直接”和“间接”路径的经典分离（SNc：黑质致密部和 VTA：腹侧被盖区；Albin、Young 和 Penney，1989）来解释多巴胺神经元放电的时序特征。

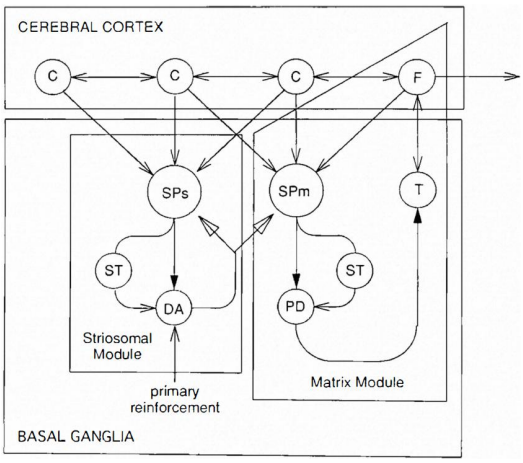

图 1：Houk 等人(1995)提出的模型中基底神经节的模块化组织(包括纹状小体和基质模块)与 *Actor-Critic* 结构之间的对应关系示意图。附加缩写：F：前额皮质的柱状结构；C:其他皮质柱;SPs:纹状体的多棘神经元纹状小体间室;SPm：纹状体基质模块的多棘神经元；ST：底丘脑侧环路；DA:黑质致密部多巴胺神经元;PD：苍白球神经元；T：丘脑神经元。(改编自 Houk 等人，1995)

提出了许多模型来改进和完善 Houk 等人的模型。然而，这些计算模型中的大多数都存在神经生物学的不一致性和不完全性，这与最近的基底神经节解剖假设有关（Joel 等人，2002）。

一个重要的缺点是，与已知的基底神经节解剖结构相比，这些模型的 *Actor* 部分往往过于简单化，并且没有考虑到纹状体的重要解剖和生理特征。例如，最近的研究表明纹状体中的神经元具有不同的多巴胺受体（D1 受体或 D2 受体；Aizman 等人，2000）。这意味着 *Actor* 中至少有两种不同的途径，在这两种途径上，强直多巴胺具有相反的作用，超越了纹状体中“直接”和“间接”途径的经典功能分离（Gurney、Prescott 和 Redgrave，2001a，b）。

同样，一些来自纹状体解剖的约束限制了 *Critic* 网络的可能架构。特别是，纹状体仅由一层中等多棘神经元组成，中间神经元占 5%（Houk 等人，1995）。因此，*Critic* 模型不能由复杂的多层网络组成，用于奖励预测计算。这种解剖学上的限制导致了几位作者将 *Critic* 建模为一个神经元（Houk 等人，1995；Montague、Dayan 和 Sejnowski，1996），这在相对简单的任务中工作得很好。对于更复杂的任务，几个模型为任务的每个子部分分配一个单一的 *Critic* 神经元。这些模型在用于协调这些神经元的计算机制上有所不同。Baldassarre（2002）和 Doya 等人（2002）建议采用混合“专家”法协调 *Critic* 模块：在任务执行期间某个时间表现最佳的模块成为该子部分学习过程中的“专家”。另一个模型提出了“专家”与任务子部分（如刺激或事件）的先验联合，独立于每个“专家”的表现（Suri 和 Schultz，2001）。它仍然需要评估每个原则的效率，因为它们在不同的任务（例如，Wisconsin 卡片分类测试、离散导航任务、操作条件反射）中一直在工作。

这些模型还质疑“直接”和“间接”通路中基底神经节的功能分离（参见 Joel 等人，2002 年的综述）。这些反对意见建立在电生理数据（回顾见 Bunney、Chiodo 和 Grace，1991）和解剖数据（Joel 和 Weiner，2000）的基础上，这些数据表明这两种途径无法产生解释多巴胺神经元放电模式所需的时间动态。这些发现导致人们质疑 *Critic* 在背侧纹状体条纹体中的定位，一些模型利用了其在腹侧纹状体中的应用（Brown、Bullock 和 Grossberg，1999；Daw，2003）。这些研究得到了最近人类 fMRI 数据的支持，显示了背侧纹状体作为 *Actor* 和腹侧纹状体作为 *Critic* 之间的功能分离（O'Doherty 等人，2004），但它们作为电生理数据（Thierry、Gioani、Degenetais 和 Glowinski，2000）表明腹侧纹状体（伏隔核核心）的一个重要部分并不广泛投射到大鼠大脑的多巴胺系统。

我们可以得出结论，*Critic* 的精确实施仍然是一个悬而未决的问题，如果我们也考虑到一个最近的模型，假设一个新的功能区别纹状小体的背侧纹状体基于 GABA-A 和 GABA-B 受体在多巴胺神经元解释时间动态的预期（Frank，Loughry 和 O'Reilly，2001）。

除了这些神经生物学上的不一致之外，许多 *Actor-Critic* 模型所关注的一些计算要求似乎对于自然的奖励-寻求任务来说是不必要的。例如，由于 Houk 等人的模型不能解释多巴胺神经元放电模式的时间特征，因此大多数替代模型都集中在模拟多巴胺抑制的精确时间，当最终没有出现奖励时，预期会出现奖励。为此，他们集中在刺激描述的时间分量的实现上，刺激描述是在模型之外计算的，并通过皮质投射发送给模型（Montague 等人，1996；Schultz、Dayan 和 Montague，1997）。这些模型在 Schultz、Apicella 和 Ljungberg（1993）选择的相同任务中进行测试，以记录猴子的多巴胺神经元，在刺激和奖励之间使用固定的时间仓。然而，在啮齿动物需要寻找食物或任何其他类型奖励的自然情况下，任务的时间特征很少固定，而是取决于动物的 *Actor* 和环境的变化/进化。

3 方法

本研究的目的是评估现有的基于基底神经节的 *Actor-Critic* 模型在相同的自主人工系统中实现时的效率。主要解决的问题如下：

◎实施一个详细的 *Actor*，其结构将更接近背侧纹状体的解剖结构，评估强化学习在该结构中是否仍然可行。

◎比较一个 *Critic* 单元的功能，与几种协调不同 *Critic* 模块的替代方法，以解决单个神经元不足的复杂任务。

◎在涉及分类单元导航的自然任务中对模型的测试，其中的事件不是由固定的时间箱预先确定的。相反，动物在它的运动中感知到一个连续的感觉流，并且必须反应性地切换它的动作，以达到一个奖励。

3.1 模拟环境和任务

图 2 显示了模拟的实验装置，包括一个简单的二维十字迷宫。尺寸相当于 5 m×5 m 的环境，1 m 的大走廊。在此环境中，墙由 256 灰度的段组成。不模拟照明条件的影响。迷宫的每面墙都是黑色的（亮度=0），除了每只迷宫通道末端和迷宫中心的墙外，其他的墙都用特定的颜色表示：中间的十字架是灰色的（191），迷宫通道末端的三面墙是深灰色的（127），第四面墙是白色的（255），显示奖励位置（相当于一个水槽提供两滴-非即时奖励，动物不知道先验）。

十字迷宫任务模拟神经生物学和 *Actor* 研究，将作为模型的未来验证（Albertin、Mulder、Tabuchi、Zugaro 和 Wiener，2000）。在这项任务中，在每次试验开始时，随机选择一只迷宫通道末端来提供奖励。相关的墙是白色的，而其他三个末端的墙是深灰色的。动物必须学会在靠近白墙（距离<30 厘米）并面对白墙（角度<45°）时选择“饮酒”的动作会给它奖励。在这里，我们假设 n 次迭代（n=2）的奖励为 1，而不考虑如何确定此奖励的“娱乐度”。我们期望动物学习一系列特定于背景事件的 *Actor*，这样它就可以从迷宫中的任何起点到达奖励站点：

◎当看不到白墙时，面向迷宫中心并向前移动。

◎到达中心后（动物可以看到白色墙壁），转向白色刺激。

◎向前移动，直到接近奖励位置。

◎饮酒。

当奖励被消耗时，试验结束：奖励位置的墙壁颜色变为深灰色，随机选择一个新的迷宫通道末端来提供奖励。然后，动物必须再次执行所学的 *Actor* 序列。注意，两个连续的试验之间没有中断：试验依次进行。

“动作规划者”执行上述 *Actor* 序列的效率和流畅性越高，获得奖励所需的时间就越少。因此，选择验证模型的标准是目标时间，沿着实验绘制模型的学习曲线。

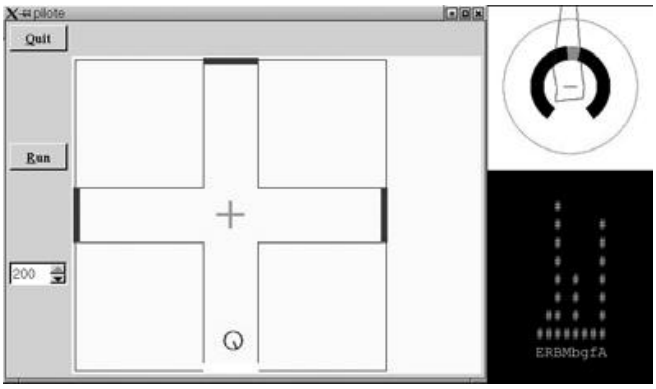

图 2：左:迷宫环境下的机器人。白色迷宫通道表示奖励位置。其他的迷宫通道末端不提供任何奖励，并显示在黑色。右上角:机器人的视觉感知。右下角:模型中不同通道的激活程度。

3.2“动作规划者”

“动作规划者”由一个圆（直径 30 cm）表示。其平移和旋转速度分别为 40 cm/s–1 和 10°/s。其模拟传感器区域如下：

◎全向线性摄像头，每 10°提供最近感知片段的颜色；这就产生了 36 色表构成“动作规划者”的视觉感知（见图 2）；

◎8 个 5 米范围的声纳，指向不确定度为 5°，测量误差为 10 厘米。

声纳被一个低水平的障碍回避反射所使用，当“动作规划者”太接近障碍物时，种反射会推翻 *Actor-Critic* 模型所做的任何决定。

“动作规划者”提供了一个视觉系统，在每个时间步计算 36 个颜色表中的 12 个输入变量（$\forall i \in [1, 12]$，$0<var_i<1$）。这些感官变量构成了 *Actor-Critic* 的状态空间，因此将作为 *Actor* 和模型 *Critic* 部分的输入（图 3）。变量计算如下：

◎ *seeWhite(*255*)*（特别的，*seeGray(*191*), seeDarkGray(*127*)*）=1，如果颜色表包含值 255（特别的 191，127），否则为 0。

◎*angleWhite*、*angleGray*、*angleDarkGray*=（颜色表中“动作规划者”头部方向与所需颜色之间的框数）/18。

◎*distanceWhite,distanceGray, distanceDarkGray* =（颜色表中包含所需颜色的最大连续框数）/18。

◎*nearWhite*（特别的 *nearGray*，*nearDarkGray*）=1–*distanceWhite*（特别的 *nearGray*，*nearDarkGray*）。

用这样的连续变量表示环境意味着模型永久地接收一个感官信息流，并且必须自主地学习与任务解决相关的事件（感官背景事件）。

“动作规划者”有 6 个动作：饮酒，向前走，变白，变灰，变暗灰，等待。这些动作构成了 *Actor* 模型（如下所述）的输出，以及对低级模型的输入，将其转换为“动作规划者”引擎的适当顺序。

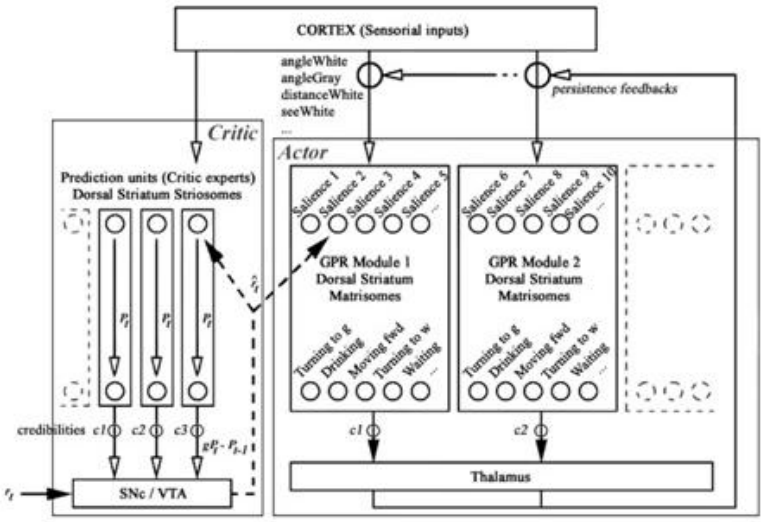

图 3：本文测试模型的总体方案。*Actor* 是一组 GPR 模块，其中显著性作为输入，动作作为输出。*Critic*（涉及纹状体背侧，黑质致密（SNc））向 *Actor* 传播一个对所选动作触发的瞬间增强的估计。该方案的特殊性在于将 *Actor* 和 *Critic* 的几个模块结合起来，并将 *Critic* 的预测和 *Actor* 模块的决策与可信度进行权衡。这些可信度可以通过门控网络（模型 AMC1）或背景事件相关方式（模型 AMC2 和 MAMC2）计算。

3.3 模型：对 Actor 角色的描述

Actor-Critic 模型的灵感来自于大鼠基底神经节。如第 2 节所述，*Actor* 可以假设在基底神经节的基质部分实现，而纹状体背面的纹体被认为是 *Critic* 的解剖对应体。*Critic* 产生类多巴胺的强化信号，帮助他学会在任务期间预测奖励，并使 *Actor* 学会在任务期间经历的每个感官环境中选择适当的 *Actor*。

Actor 中实施的架构是最近由 Gurney、Prescott 和 Redgrave（2001a，b）提出的一个模型，因此被称为 GPR 模型，它取代了简单的“Winner-takes-all”，通常由 *Actor* 模型组成，从生物学角度来看，这似乎更合理。

与其他 *Actor* 一样，GPR 模型由一系列并行通道组成，每个通道代表一个操作（在我们的实现中，我们使用 6 个通道对应于用于任务的 6 个操作）。这种结构构成了第 1 节讨论的将基底神经节主要功能分离为“直接”和“间接”路径的另一种观点（Gurney 等人，2001a，b）。所有这些通道通过背侧纹状体由两个不同的回路组成：第一个是“选择”通路，通过周围网络上的前馈偏离中心正确地实现动作选择，并由背侧纹状体中具有 D1 型受体的细胞介导。第二个是“控制”通路，由同一区域具有 D2 型受体的细胞介导。它的作用是通过增强通道间的选择性来调节选择，并控制 *Actor* 内部的全局活动。此外，模型中的皮层-基底神经节-丘脑回路允许它考虑每个通道在选择过程中的持续性（有关模型的详细描述和数学实现，请参见 Gurney 等人，2001a，b）。后一个特征显示了一些阻止机器人进行 *Actor* 振荡的有趣特性（Montes Gonzalez、Prescott、Gurney、Humphries 和 Redgrave，2000；Girard、Cuzin、Guillot、Gurney 和 Prescott，2003）。

在我们的实现中，*Actor* 模型的输入值是显著性的——即，从 12 个感官变量中计算出的给定动作的强度、实现偏倚的常量以及对于在前一个时间点选择的动作的持续因子等于 1（图 3）。.在每个时间点 *t*（在我们的模拟中时间点被 1-s 的 bin 分隔），具有最高显著性的动作被“动作规划者”选择执行，动作 *i* 的显著性是

$$sal_i(t)=\left[\sum_{j=1}^{13}\mathrm{var}_j(t)\bullet w_{i,j}(t)\right]+persist_i(t)\bullet w_{i,14}(t) \quad (1)$$

其中 $var_{13}(t)$=1，$\forall t$，并且 $w_{i,j}(t)$是每个动作 *i* 的突触权重，与输入变量 *j* 的关联强度。这些权重是随机启动的（$\forall$ i,j,–0.02<$w_{i,j}(t=0)$<0.02），学习过程的目标是找到一组权重，使“动作规划者”能够有效地执行任务。

添加了一个探索函数，允许“动作规划者”在给定的背景事件中尝试一个动作，即使 *Actor* 的权重没有给出在考虑的背景事件中执行该动作的足够倾向。为此，我们引入了一个时钟，在两种不同的情况下触发探索：

◎当“动作规划者”在被模型评估为负的情况下(当 *Critic* 计算的奖励预测 *P(t)*低于一个固定阈值时），“动作规划者”已经被陷入大量的时间步长（*时间*优于固定阈值α)的情况评估负的模型）。

◎当“动作规划者”把 *P (t)*很高的情况保持很长一段时间，但这种预测没有增加那么多（| *P (t + n)- P (t)* | <ε），且没有奖励。

如果这两个条件中有一个是正确的，则会触发探索：随机选择 6 个动作中的一个。其显著性设置为 1（请注意，当探索（*exploration*）=错误（*false*）时，$sal_i(t)$<1，$\forall i,t,w_{i,j}(t)$），并保持 15 个时间步长（“动作规划者”180°转弯或从迷宫中心到单臂末端所需的时间）。

3.4 模型：对模型的 Critic 部分进行描述

对于模型的 *Critic* 部分，测试了基于现有技术的不同原则。其目的是首先检验单个 *Critic* 单元的假设，同时也为 *Critic* 提供足够的计算能力，使其能够正确估计整个任务环境中的价值函数。换言之，*Critic* 将不得不处理几个不同的感官环境，走廊、迷宫中心、迷宫通道末端等，相当于不同的刺激，并将正确的奖励预测与这些环境相关联。

一个明显的可能性是一个多层感知器有几个隐藏层，但是，如第 2 节所述，有一些解剖学上的限制，阻止我们采用这种选择：我们的 *Critic* 被认为是位于背侧纹状体的纹状小体中，它的结构是由只有一层中等多棘神经元（Houk 等人，1995）。因此，我们需要一种更通用的方法，将多个 *Critic* 模块组合在一起，每个 *Critic* 模块由一个神经元组成，并处理问题空间的特定部分。这里所采用的方法是“专家”的混合，提出将一个非线性可分问题划分成一组线性可分问题，并对每个考虑的子问题影响不同的“专家”（Jacobs、Jordan、Nowlan 和 Hinton，1991）。

本文中的 *Critic* 主要在以下两个方面存在差异：

◎第一个（模型 AMC1）实现了“专家”的混合，其中一个门控网络用于确定每个区域使用的是哪一个“专家”；

◎第二个（模型 AMC2）实现了“专家”的混合，其中根据视觉感知的分类手动确定的环境分区来决定每个子区域中的“专家”工作。

此外，由于“动作规划者”必须在连续状态空间中解决一个任务，不同的 *Critic*“专家”发送给同一个 *Actor* 的强化信号之间可能存在干扰。这样一来，一个模型将只使用一个 *Actor*（模型 AMC2），而另一个模型将使用与每个“专家”关联的一个 *Actor* 模块（模型 MAMC2）。图 3 显示了采用不同模块的总体方案，如本文所示的模型所示。

将比较 AMC1、AMC2 和 MAMC2 模型的性能，以及 Houk 等人（1995）提出的由基底神经节激发的具有开创性的 *Actor-Critic* 模型。使用一个单细胞 *Critic* 和一个 *Actor*（模型 AC）。

我们从描述最简单的 *Critic* 开始，这个 *Critic* 属于 AC 模型。

3.4.1 模型 AC

在这个模型中，在每个时间步长，*Critic* 都是一个单一的线性单元，它基于与 *Actor* 相同的输入变量(持久性变量除外)计算对奖励的预测

$$P(t)=\sum_{j=1}^{13}\mathrm{var}_j(t)\bullet w'_j(t) \tag{2}$$

其中，$w'_j(t)$是 *Critic* 的突出权重。

然后利用该预测方法，利用 TD-规则对强化信号进行计算：

$$\hat{r}(t)=r(t)+gP(t)-P(t-1) \tag{3}$$

其中 *r(t)*为“动作规划者”实际获得的奖励，*g* 为折扣因子($0<g<1$)，决定在未来奖励的总和中考虑未来期望奖励的程度。

最后，利用这个强化信号分别根据下面的方程更新 *Actor* 和 *Critic* 的突触权值：

$$w_{i,j}(t)\leftarrow w_{i,j}(t-1)+\eta\bullet\hat{r}(t)\bullet\mathrm{var}_j(t-1) \tag{4}$$

$$w_j(t)\leftarrow w'_j(t-1)+\eta\bullet\hat{r}(t)\bullet\mathrm{var}_j(t-1) \tag{5}$$

其中$\eta>0$ 是学习率（learning rate）。

3.4.2 模型 AMC1

当这个 *Critic* 实现 *N* 个专家时，每个“专家 *k*”计算自己在第 *t* 步时的奖励预测：

$$P(t)=\sum_{j=1}^{13}w'_{k,j}(t)\bullet\mathrm{var}_j(t) \tag{6}$$

其中，$w'_{k,j}(t)$是“专家 *k*”的突出权重。

然后 *Critic* 的全局预测是“专家”预测的加权和：

$$P(t)=\sum_{k=1}^{N}cred_k(t)\bullet p_k(t) \tag{7}$$

其中 $cred_k(t)$是“专家 *k*”在 *t* 时刻的可信度。这些可信度是通过一个门控网络计算出来的，这个门控网络学会在每种感觉环境下将最佳可信度与预测误差较小的“专家”联系起来。根据 Baldassarre(2002)的描述，门控网络由 *N* 个线性单元组成，这些单元接收的输入变量与“专家”的输入变量相同，并由此计算出一个输出函数：

$$o_k(t)=\sum_{j=1}^{13}w''_{k,j}(t)\bullet\mathrm{var}_j(t) \tag{8}$$

其中，$w''_{k,j}(t)$是门控单元的突出权重。

将“专家 *k*”的可信度计算为 $o_f(t)$输出的 softmax 激活函数：

$$cred_k(t)=\frac{o_k(t)}{\sum_{f=1}^{N}o_f(t)} \tag{9}$$

关于学习规则，公式 3 用于确定发送给 *Actor* 的全局强化信号，而每个 *Critic* 的“专家”根据自己的预测误差都有一个特定的强化信号：

$$\hat{r}_k(t) = r(t) + gP(t) - p_k(t-1) \quad (10)$$

每个“专家 k”的突触权值按下式更新：

$$w''_{k,j}(t) \leftarrow w''_{k,j}(t-1) + \eta \bullet \hat{r}_k(t) \bullet \mathrm{var}_j(t-1) \bullet h_k(t) \quad (11)$$

其中 $h_k(t)$为“专家 k”对 *Critic* 全局预测误差的贡献，定义为

$$h_k(t) = \frac{cred_k(t-1) \bullet corr_k(t)}{\sum_{f=1}^{N} cred_f(t-1) \bullet corr_f(t)} \quad (12)$$

其中 $corr_k(t)$是“专家 k”的正确性的度量，定义为：

$$corr_k(t) = \exp\left(\frac{-\hat{r}_k(t)^2}{2\sigma^2}\right) \quad (13)$$

其中σ是一个尺度参数取决于“专家”的平均误差(见在附录的参数表)。

最后，为了更新门控网络的权值，我们使用如下公式：

$$w''_{k,j}(t) \leftarrow w''_{k,j}(t-1) + m \bullet diff(t) \bullet \mathrm{var}_j(t-1) \quad (14)$$

而且 $diff(t) = h_k(t) - cred_k(t-1)$其中 m 为特定于门控网络的学习率。

因此，“专家 k”在特定的感官情境下的可信度取决于它在这种情境下的表现。

3.4.3 模型 AMC2

该 *Critic* 还实施了 N 个“专家”。然而，在计算每个“专家”的可信度方面，它不同于 AMC1 模型。

我们希望在这里实现的原则是将“专家”的可信度与他们的表现分离开来。相反，“专家”被分配到环境的不同的子区域（这些区域被计算为感知空间中的窗口），永远被他们的关联区域所吸引，并逐步学习以提高其在实验中的表现的准确性。为了改进他们的模型，这一原则被 Houk 等人（1995）采用，假设不同的纹状小体可以专门处理不同的 *Actor* 任务。Suri 和 Schultz（2001）在使用多个 TD 模型时实现了这一观点，每个模型仅计算模拟范式中发生的一个事件（刺激或奖励）的预测。

为了检验这一原理，我们手动确定的环境分区（例如，感官空间的粗略表示）代替了门控网络：在时间步 t，当前区域β依赖于视觉系统计算的 12 个感官变量。示例：如果（*seeWhite*=1，*angleWhite*<0.2，*distanceWhite*>0.8），则 *zone*=4（例如，β=4）。则所有其他“专家”的 $cred_\beta(t)$=1，$cred_k(t)$=0，“专家”β必须计算出 12 个连续感官变量中的奖励预测。“专家”的预测和强化信号由与模型 AMC1 的 *Critic* 相同的方程确定。

这是对所考虑原则进行测试的第一步。事实上，我们假设另一个大脑区域，如顶叶皮层或海马体，将根据当前的感官感知来确定该区域（感官结构）（McNaughton，1989；Burgess，Jeffery 和 O'Keefe，1999），并将其发送给基底神经节的 *Actor-Critic* 模型。在这里，环境被划分成 N=30 个区域，每个区域都有一个“专家”。该方案与 Suri 和 Schultz 采用的方案的主要区别在于，在他们的工作中，每个子区域的“专家”培训是分阶段进行的，只有在对所有“专家”进行培训后，才能对整个任务进行全局模型的测试。在这里，“专家”们在一个单一的实验中同时接受训练。

最后，应该注意的是，这种方法不同于对状态空间进行粗略的编码，后者构成了对 *Actor* 和 *Critic* 的输入（Arleo 和 Gerstner，2000）。在这里，我们实现了可信度空间的粗略编码，以确定在给定的感官配置中哪位“专家”最可信，并将 12 个连续的感官变量加上上面描述的常量作为强化学习过程的状态空间。这意味着，在给定的区域内，相关“专家”必须学习

根据不同的输入变量来近似一个连续的奖励值函数。

3.4.4 模型 MAMC2

该模型的 *Critic* 与模型 AMC2 中的 *Critic* 相同，并且仅在关联 *Actor* 方面有所不同。我们不使用单个 *Actor*，而是实现了 n 个不同的 *Actor* 模块。每个 *Actor* 模块与第 3.4 节中描述的简单 *Actor* 具有相同的结构，由六个通道组成，代表任务的六个可能动作。区别在于，只有与“动作规划者”当前所在区域相关联的 *Actor* 的动作才能竞争以确定“动作规划者”的当前动作。因此，如果“动作规划者”在时间 *t* 处于*β*区并执行了动作 *i*，则 *Critic* 在下一个时间步计算的增强信号 $\hat{r}(t+1)$将仅用于根据以下方程式更新 *Actorβ*的动作 i 的权重：

$$w_{k,i,j}(t) \leftarrow w'_{k,i,j}(t-1) + \eta \bullet \hat{r}(t) \bullet \mathrm{var}_j(t-1) \qquad (15)$$

其他方程与 AMC2 模型的 *Critic* 使用的方程相同。如上所述，这一原则（对 *Actor-Critic* 模型的每个模块使用特定的控制器或特定的 *Actor*）受到 Doya 等人（2002）的工作的启发。

3.5 结果

为了比较四个模拟模型的学习曲线，以评估哪些模型能够有效地解决任务，我们采用了以下标准：经过 50 次训练（每项实验 100 次）后，“动作规划者”必须达到与手动控制的模型相同的性能，而手动控制的模型已经可以解决这个任务（表 1）。为此，我们用适当的手动控制的突触权重新模拟了 GPR 动作选择模型，并且没有任何学习过程，这样“动作规划者”就可以像已经学习一样解决任务。在这个模型中，“动作规划者”进行了 50 次实验，每次实验平均执行 142 次迭代。正如上面提到的，由于每次迭代持续了大约 1 秒，所以每次试验都要花费 2 分钟多一点的时间，这个手动控制的“动作规划者”才能获得奖励。

表 1：每个模型的性能

模型	GPR	AC	AMC1	AMC2	MAMC2
性能（迭代次数）	142	587	623	3240	97

表 1 显示了每个模型的性能，以每次试验后的平均迭代次数 50 来衡量。图 4 说明了在二维环境中执行的四个实验的结果，每个模型一个。*x* 轴代表试验期间的连续试验。对于每个试验，*y* 轴显示“动作规划者”获得奖励和消耗奖励所需的迭代次数。图 4a 显示了模型 AC 的学习曲线，可以看出，模型在试验 7 之前迅速提高了性能，并在试验 25 时稳定下来。然而，在试验 50 之后，试验的平均持续时间仍然是 587 次迭代，比选择的标准高出近 4 倍。我们可以通过模型 AC 只包含 *Critic* 中的一个神经元来解释这个限制，它只能解决线性可分离问题。因此，模型只能学习任务的一部分（奖励地点附近的区域），无法将学习扩展到迷宫的其他部分。因此，“动作规划者”学会了在奖励区域中选择适当的 *Actor*，但它仍然在环境的其余部分执行随机 *Actor*。

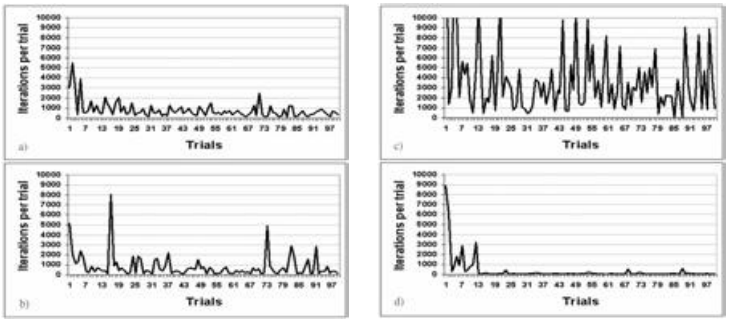

图 4: 4 个模型的二维十字迷宫任务模拟学习曲线，模拟了 100 多次实验，*x* 轴代表试验期间的连续试验。对于每个试验，*y* 轴显示“动作规划者”获得奖励和消耗奖励所需的迭代次数。(为了更好的可读性，将其截断为 10000 次)。(a)模型 AC，(b)模型 AMC1， (c)模型 AMC2， (d)模型 MAMC2。

AMC1 模型的设计是为了减轻 AC 模型的计算局限性，因为它意味着由一个门控网络控制的几个临界单元。图 4b 显示了在十字迷宫任务中模拟后的学习曲线。在实验开始时，该模型还设法减少了每次试验的运行时间。然而，可以看出，学习过程比前一个更不稳定。此外，在第 50 次试验后，该模型的迭代性能达到 623 次，并不比 AC 模型好，而且，该模型也不能将学习扩展到整个迷宫。我们可以通过这样一个事实来解释这种失败：门控网络没有在任务的不同子部分中专门化不同的“专家”。作为一个例子，图 5 显示了每个 *Critic* 的“专家”在最后一次实验中计算出的奖励预测。可以注意到，在整个试验过程中，第一位“专家”（暗曲线）的预测最高。这是因为它是唯一一个门控网络已经学会认为可信的，它的可信度在整个实验过程中保持在 90%以上。因此，只有一个“专家”参与学习过程，模型在计算上等同于 AC 模型：它不能将学习扩展到整个迷宫，这一点在感知到图 5 中的奖励位置(刺激发生)之前没有任何奖励预测，这一点得到了证实。

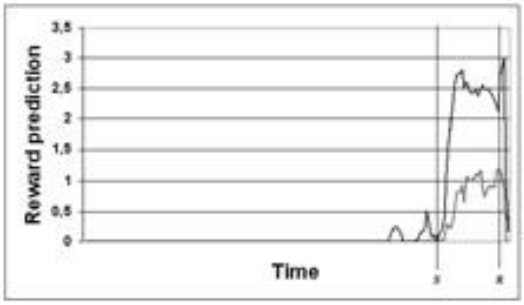

图 5: 实验第 100 次实验期间，每个 *Critic* 的 AMC1 模型专家计算的奖励预测。时间 0 表示试验开始。S:“动作规划者”对刺激(白墙)的感知。R:开始发放奖励。暗曲线代表专家 1 的预测。其他专家的预测被融入到光的曲线中，或者等于 0。

图 4c 显示了 AMC2 模型的学习曲线，它实现了“专家”协调的另一个原则。由于每个“专家”都是特定环境领域的先验“专家”，因此该模型不受与 AMC1 模型相同的限制。因此，它很快把学习扩展到了整个迷宫。然而，这一过程的结果是在 *Actor* 的计算中产生干扰：同一个 *Actor* 接收到所有“专家”的教学信号，并且仍然无法在强化 *Actor* 之间正确切换。例如，当动作“饮酒”得到加强时，即使“动作规划者”离奖励地点很远，*Actor* 开始永久地选择这个动作。这些干扰解释了 AMC2 模型所获得的非常差的性能。

最后一个模拟模型（模型 MAMC2）表现最好。它的学习曲线如图 4d 所示。这个模型实现了几个 *Actor* 模块（一个 *Actor* 模块连接到每个 *Critic*“专家”）。因此，它避免了学习过程中的干扰，并迅速收敛到每次测试 97 次迭代的性能。这种良好的性能不能只与多 *Actor* 一起实现；我们尝试将多个 *Actor* 模块组合到模型 AMC1 中，并在每次测试中获得 576 次迭代的性能。因此，任务的完成意味着多角色和“专家”的良好专业化的结合。

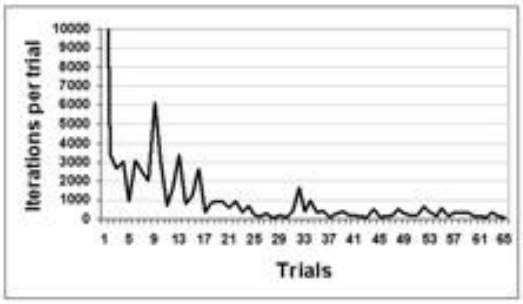

图 6：三维环境下的学习曲线:x 轴表示试验次数;y 轴表示每次试验的迭代次数。

为了检查模型 MAMC2 在更真实的条件下学习相同任务的能力，我们在 3D 环境中对其进行了模拟，实时工作并实现了物理动态（图 7）。这项实验涉及到一个中间步骤，有利于实现真正的 Pekee 机器人（Wany Robotics）。在这种环境下，“动作规划者”仍然能够学习任务，并在 35 次试验后获得良好的表现(图 6;试验 35～65 次之间对应的“动作规划者”平均性能：每次试验迭代 284 次)。

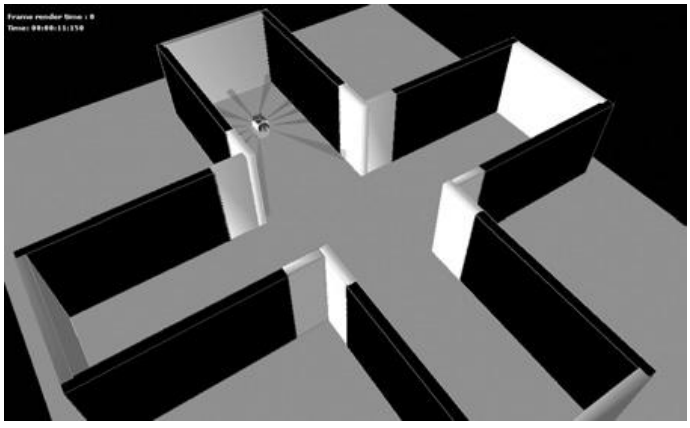

图 7：三维环境下的十字迷宫任务仿真。就像 2D 环境一样，一个随机的迷宫通道末端是白色的，并提供奖励。动物必须执行分类导航，以便找到和消费这个奖励。由“动作规划者”的身体产生的灰色条纹代表了它的声纳传感器，用于它的低水平避障反射。

4 讨论和未来工作

在这项工作中，我们比较了基于不同原则的基底神经节的几个 *Actor-Critic* 模型在 S-R 任务上的学习能力。AC、AMC1、AMC2 和 MAMC2 模型的仿真结果表明：

◎单个组件 *Critic* 无法解决任务（模型 AC）；

◎由门控网络控制的多个 *Critic* 模块（模型 AMC1）无法提供良好的专业化，且任务仍未解决；

◎几个 *Critic* 模块先验与任务的不同部分(模型 AMC2)和连接到单个 *Actor*(*Actor* 组件组成的 6 通道 GPR)允许学习从奖励位置延伸到遥远的地区，但仍然受到不同 *Critic* 向同一 *Actor* 发出的信号之间的干扰。

MAMC2 模型将多个 *Critic* 模块与 AMC2 模型的原理相结合，实现了多个 *Actor* 组件，在任务中产生了更好的效果，将学习传播到整个迷宫中，缩短了学习时间。然而，关于该模型的生物学合理性和泛化能力，还有一些问题需要提出。

4.1 所提议模型的生物学合理性

当使用单个 GPR*Actor* 时，每个行动表示仅在一个通道中表示——每个动作由一个通道组成的 *Actor* 模块（Gurney 等人，2001a，b）——并且结构信贷分配问题（当获得一个奖励时，动作会加强）可以简单地解决：最显著的作用是通过 D1 纹状体的局部复发抑制回路抑制相邻的纹状体(Brown 和 Sharp，1995)。因此，只有一个通道在 *Actor* 将有足够的突触前和突触后的活动，以符合加强。

当使用多个 *Actor* 模块时，此属性不再为真：即使在给定的时间内每个 *Actor* 模块只能激活一个通道，每个 *Actor* 模块都将有自己的激活通道，并且多个同时发生的突触将有资格在全局 *Actor* 中进行强化。为了解决这个问题，我们在工作中考虑到，在给定的时间内，整个 *Actor* 中只有一个频道是合格的。然而，这意味着基底神经节具有以下两个特征之一：要么纹状体中的 *Actor* 模块之间存在非局部抑制，要么在多巴胺增强信号中存在某种选择性，以便即使有几个通道激活后，只有目标模块中的那些接收多巴胺信号。

据我们所知，在基底神经节中没有发现这些特征，一些研究倾向于反驳多巴胺的选择性（Pennartz，1996）。

4.2 计算问题

还需要解决几个计算问题。首先，本文的研究结果表明，学习过程不受使用一个详细描述基底神经节动作选择过程的 *Actor* 的影响。这个角色有能力考虑到皮质-基底神经节-丘脑-皮质环所提供的一些持久性。在本研究中，我们并没有深入研究这种持续性对学习过程的影响。然而，我们怀疑持久性可能会挑战不同的 *Actor* 与 *Critic* 的“专家”的互动方式，因为在这种模式下，动作之间的切换并不完全遵循感觉运动环境中的切换。这个问题应该在以后的工作中加以研究。

4.2.1 多模块 Actor 的泛化能力

另一个需要解决的问题是本实验中使用的多模块 *Actor* 模型的泛化能力。事实上，MAMC2 模型避免了对 *Actor* 的干扰，因为手动确定的迷宫子区域是绝对不相交的。换句话说，一个特定区域内的“学习刺激-反应关联”不能在另一个区域内执行，并且不干扰学习过程的是第二个区域，即使与每个区域相关的视觉环境非常相似。然而，这也导致无法从一个区域归纳到另一个区域：即使我们在两个区域之间所做的区分似乎与迷宫任务相关，如果这两个区域是相似的，并且暗示着在另一个任务中有相似的运动反应，“动作规划者”将不得不在每个区域学习相同的感觉运动联系两次，每区域一次。因此，我们在这个工作中设置的分区是与任务相关的。

或者，该模型需要一种划分方法，该方法能够独立于任务自主地对感觉背景事件进行分类，能够检测两个不同背景事件之间的相似性，并且能够将学习到的 *Actor* 从第一个经验背景事件归纳到第二个经验背景事件。

4.2.2 关于奖励发放的准确时间

在这里介绍的工作中，奖励传递的时间完全取决于“动作规划者”的 *Actor*，这与其他用于验证基底神经节的 *Actor-Critic* 模型的 S-R 任务不同。在这些任务中，刺激和奖励之间有一个恒定的持续时间，并且已经设计了一些 *Actor-Critic* 模型来描述这类任务中多巴胺能神经元的精确时间动态（Montague 等人，1996）。因此，许多 *Actor-Critic* 模型都关注于刺激表征时间成分的实现，有几项工作利用这种时间表征将基底神经节强化学习的 *Actor-Critic*

模型应用于机器人（PereZ-Uribe，2001 年；Sporns 和 Alexander，2002）。我们需要将这样一个组件添加到我们的模型中，以便能够将它应用到特定类型的自然任务或生存任务中吗？

在这里提出的实验中，我们不需要这样一个刺激的时间表示，因为“动作规划者”在移动过程中感知到的连续的感觉流中有足够的信息，这样模型就可以动态地适应它的奖励预测，正如 Baldassarre 和 Parisi（2000）观察到的那样。例如，当“动作规划者”处于迷宫的中心，感知白墙（刺激预测奖励）并向奖励位置移动时，后者在“动作规划者”的视野中变得更大，模型可以学习增加其奖励预测，如图 8 所示。我们的目的并不是解释当奖励没有发生时多巴胺神经元的放电率的降低；然而，我们能够观察到这种现象，在“动作规划者”接近奖励位置，即将消耗它，但最终转离它（图 8 中的 R 事件）。

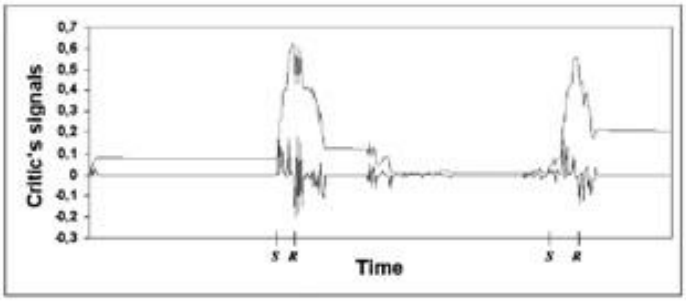

图 8：三维环境下 MAMC2 模型 *Critic* 计算的奖励预测(光曲线)和多巴胺强化信号(暗曲线):*x* 轴表示时间;*y* 轴表示 *Critic* 的信号。*S* 表示“动作规划者”对刺激的感知(白墙);*R* 表示“动作规划者”错过了奖励。

4.2.3 使用依赖于或独立于表演的 Critic

在我们的实验中，AMC1 模型实现了“专家”可信度计算的门控网络，并没有解决任务。我们在第 2 节中看到，在模拟过程中，一个“专家”迅速成为最可信的，这迫使模型只使用一个神经元来解决任务。门控网络在混合“专家”方法框架下的使用已经受到了 *Critic*（Tang、Heywood 和 Shepherd，2002）。这些作者认为，这种方法在由不相交区域组成的问题上很好地工作，但在区域边界上受到影响，不能很好地推广。

在我们的案例中，我们通过观察来解释“专家”对模型 AMC1 专业化的失败，直到模型开始学习任务，从而可以将教学信号传播到迷宫的其余部分，只有奖励位置才有价值。因此，它是唯一一个门控网络尝试培训“专家”的领域，而“专家”培训迅速达到了很高的可信度。然后，随着奖励价值开始扩展到一个新的领域，这个“专家”在获得不良表现的同时仍然拥有最好的信誉。其他“专家”由于还没有经过培训，而且新领域和第一个领域并没有出现脱节，所以他们的表现没有明显的提高。因此，它们仍然是不可预测的，并且模型开始有糟糕的性能。

Baldassarre（2002）设法获得了“专家”的良好专业化。这可以部分解释为，他的任务涉及三种不同感官环境中的三种不同奖励。模拟机器人必须从任务开始时就交替访问所有奖励。这可能有助于门控网络将良好的可信度归因于几个“专家”。然而 Baldassarre 任务中的奖励地点并非完全不一致，这导致了一个困难的专业化：三个奖励中的两个最可信的是其中一个“专家”（见 Baldassarre，2002）。

另一个模型（Tani 和 Nolfi，1999）提出了一种不同的混合“专家”，其中门控网络被“专家”可信度的动态计算所取代。他们的模型成功地将模拟机器人在运动过程中感知到的“感觉-运动”流进行了分类。然而，他们的方法并没有使用任何关于“专家”可信度和任务期间经历的不同背景之间关联的记忆。因此，与 Baldassarre 的门控网络相比，“专家”

的专业化更依赖于每个“专家”的表现，当应用于我们的十字迷宫任务中的强化学习时，与我们在实验（未发表的工作）中发现的相同限制。

4.2.4 结合自组织映射与混合专家

为了检验“专家”诚信与绩效分离的原则，我们将环境划分为若干子区域。然而，这种方法是临时的，缺乏自主性，并且在环境发生变化或变得更复杂时，会受到泛化能力的影响。我们目前正在实施自组织映射（SOMs），作为一种将用于确定这些区域的不同感官环境进行自主聚类的方法。请注意，这一主张不同于传统的使用 SOMs 将状态空间输入集群到“专家”或 *Actor-Critic* 模型（Smith，2002；Lee 和 Kim，2003）。这是 Tang 等人（2002）最近提出的信任空间聚类。我们还想将 SOMs 的使用与定位细胞（place cells）的使用进行比较。事实上，海马定位细胞模型已经被用于对 *Actor* 和 *Critic* 的输入状态空间进行粗略编码（Arleo 和 Gerstner，2000；Foster、Morris 和 Dayan，2000；Strosslin，2004），但在我们的案例中，我们希望使用定位细胞来确定“专家”的可信度。

4.3 未来工作

文献中经常提到的，如本工作所确认的，*Actor-Critic* 体系结构在连续任务中的应用比在离散任务中的应用更困难。关于这个问题，还做了其他几项工作（Doya，2000）。然而，这些体系结构仍然需要改进，以减少它们的学习时间。

尤其是，我们的“动作规划者”的学习性能似乎仍然远没有真正的老鼠在同一任务中能够达到的学习速度（Albertin 等人，2000），即使我们在模型中使用的高时间常数还不允许进行严格的比较（见附录中的参数表）。这至少可以部分解释为，我们只实施 S-R 学习（或习惯学习），而最近人们知道，大鼠有两种不同的学习系统，与不同的大脑皮层-基底神经节-丘脑环路相关：一种习惯学习系统和目标导向学习系统（Ikemoto 和 Panksepp，1999；Cardinal、Parkinson、Hall 和 Everitt，2002）。后者将是快速的，用于学习的早期阶段，并暗示了奖励目标的明确表示或行动结果或有事项的内部表示。前者将非常缓慢，并利用后者，当“动作规划者”达到良好的表现，并能够解决与反应策略（S-R）的任务（Killcross 和 Coutureau，2003；Yin，Knowlton 和 Balleine，2004）。

一些理论工作已经开始将 *Actor-Critic* 模型扩展到这一功能差异（Dayan，2001）。在我们的人工大鼠的实际案例中，这两种系统可以以两种不同的方式使用。

首先，它可能有助于升级探索功能。这个函数可以明确表示环境的不同位置，特别是奖励站点。然后，当动物第一次得到奖励时，探索功能会引导它，尝试让它到达明确记忆的奖励位置的 *Actor*。该函数还可以记住哪些 *Actor* 在不同的领域已经尝试过，但没有成功，因此在探索的情况下，选择未尝试的 *Actor* 而不是随机的 *Actor*。这将加强探索过程，并有望提高动物的学习速度。

目标导向 *Actor* 组件的第二个可能的用途是表示动物正在工作的奖励类型。当一个动物必须处理不同的奖励（食物、饮料）以满足不同的动机（饥饿、口渴）时，这是有用的。在这种情况下，明确选择“动作规划者”作为目标的当前奖励的组件可以选择 *Actor* 的子模块，这些子模块专门用于导致考虑的奖励的行为序列。当人工大鼠 Psikharpax 必须在更自然的环境中生存，满足同时发生的动机时，这种改进将成为更现实的验证。

附录

表 2：参数

符号	值	描述
Δt	*1s*	时间常数：模型两次连续迭代之间的时间。
α	*40 次迭代*	触发探测功能的时间阈值。
g	*0.98*	时间差学习规则的折扣因子。
η	*0.01*	*Actor* 和 *Critic* 模块的学习率。
N	*30*	*Critic* 模型 AMC1、AMC2 和 MAMC2 的“专家”数量。
σ	*2*	AMC1 模型混合“专家”的标度参数。
m	*0.1*	AMC1 型门控网络的学习率。

致 谢

这项研究得到了法国国家科学研究中心的 LIP6 和机器人与人工实体项目(ROBEA)的支持。感谢 Angelo Arleo, Gianluca Baldassarre, Francesco Battaglia, Etienne Koechlin 和 Jun Tani 的有益讨论。

Appendix

Table 2 Parameters.

Symbol	Value	Description
Δt	1 s	Time constant: Time between two successive iterations of the model.
α	40 iterations	Time threshold to trigger the exploration function.
g	0.98	Discount factor of the temporal difference learning rule.
η	0.01	Learning rate of the Actor and Critic modules.
N	30	Number of experts in the Critic of models AMC1, AMC2 and MAMC2.
σ	2	Scaling parameter in the mixture of experts of model AMC1.
m	0.1	Learning rate of the gating network in model AMC1.

Acknowledgments

This research has been supported by the LIP6 and the Project *Robotics and Artificial Entities* (ROBEA) of the Centre National de la Recherche Scientifique, France. Thanks for useful discussions go to Angelo Arleo, Gianluca Baldassarre, Francesco Battaglia, Etienne Koechlin and Jun Tani.

References

Aizman, O., Brismar, H., Uhlen, P., Zettergren, E., Levey, A. I., Forssberg, H., Greengard, P., & Aperia, A. (2000). Anatomical and physiological evidence for D1 and D2 dopamine receptors colocalization in neostriatal neurons. *Nature Neuroscience*, *3*(3), 226–230.

Albertin, S. V., Mulder, A. B., Tabuchi, E., Zugaro, M. B., & Wiener, S. I. (2000). Lesions of the medial shell of the nucleus accumbens impair rats in finding larger rewards, but spare reward-seeking behavior. *Behavioral Brain Research*, *117*(1–2), 173–183.

Albin, R. L., Young, A. B., & Penney, J. B. (1989). The functional anatomy of basal ganglia disorders. *Trends in Neuroscience*, *12*, 366–375.

Arleo, A., & Gerstner, W. (2000). Spatial cognition and neuromimetic navigation: A model of the rat hippocampal place cell activity. *Biological Cybernetics*, Special Issue on Navigation in Biological and Artificial Systems, *83*, 287–299.

Baldassarre, G. (2002). A modular neural-network model of the basal ganglia's role in learning and selecting motor behaviors. *Journal of Cognitive Systems Research*, *3*(1), 5–13.

Baldassarre, G., & Parisi, D. (2000). Classical and instrumental conditioning: From laboratory phenomena to integrated mechanisms for adaptation. In J.-A. Meyer, A. Berthoz, D. Floreana, H. L. Roitblat, and S. W. Wilson (Eds.), *From animals to animats 6: Proceedings of the Sixth International Conference on Simulation of Adaptive Behavior*, supplement volume (pp. 131–139). Cambridge, MA: The MIT Press.

Brown, J., Bullock, D., & Grossberg, S. (1999). How the basal ganglia use parallel excitatory and inhibitory learning, or incentive salience? *Brain Research Reviews*, *28*, 309–369.

Brown, L., & Sharp. F. (1995). Metabolic mapping of rat striatum: Somatotopic organization of sensorimotor activity. *Brain Research*, *686*, 207–222.

Bunney, B. S., Chiodo, L. A., & Grace, A. A. (1991). Midbrain dopamine system electrophysiological functioning: A review and new hypothesis. *Synapse*, *9*, 79–84.

Burgess, N., Jeffery, K. J., & O'Keefe, J. (1999). Integrating hippocampal and parietal functions: A spatial point of view. In N. Burgess, K. J. Jeffery, and J. O'Keefe (Eds.), *The hippocampal and parietal foundations of spatial cognition* (pp. 3–29). Oxford: Oxford University Press.

Cardinal, R. N., Parkinson, J. A., Hall, J., & Everitt, B. J. (2002). Emotion and motivation: The role of the amygdala, ventral striatum and prefrontal cortex. *Neuroscience Biobehavioral Reviews*, *26*(3), 321–352.

Dayan., P. (2001). Motivated reinforcement learning. In T. G. Dietterich, S. Becker, & Z. Ghahramani (Eds.), *Proceedings of NIPS 14* (pp. 11–18). Cambridge, MA: The MIT Press.

Daw, N. D. (2003). *Reinforcement learning models of the dopamine system and their behavioral implications*. Ph.D. thesis, Carnegie Mellon University, Pittsburgh, PA.

Doya, K. (2000). Reinforcement learning in continuous time and space. *Neural Computation*, *12*, 219–245.

Doya, K., Samejima, K., Katagiri, K., & Kawato, M. (2002) Multiple model-based reinforcement learning. *Neural Computation*, *14*(6), 1347–1369.

Filliat, D., Girard, B., Guillot, A., Khamassi, M., Lachèze, L., & Meyer, J.-A. (2004). State of the artificial rat Psikharpax. In S. Schaal, A. Ijspeert, A. Billard, S. Vijayakumar, J. Hallam, and J.-A. Meyer (Eds.), *From animals to animats 8: Proceedings of the Eighth International Conference on Simulation of Adaptive Behavior* (pp. 2-12). Cambridge, MA: The MIT Press.

Foster, D., Morris, R., & Dayan, P. (2000). Models of hippocampally dependent navigation using the temporal difference learning rule. *Hippocampus*, *10*, 1–16.

Frank, M. J., Loughry, B., & O'Reilly, R. C. (2001). Interactions between frontal cortex and basal ganglia in working memory: A computational model. *Cognitive, Affective and Behavioral Neuroscience*, *1*(2), 137–160.

Gerfen, C. R., Herkenham, M., & Thibault, J. (1987). The neostriatal mosaic: II. Patch- and matrix-directed mesostriatal dopaminergic and non-dopaminergic systems. *Journal of Neuroscience*, *7*, 3915–3934.

Girard, B., Cuzin, V., Guillot, A., Gurney, K., & Prescott, T. (2003). A basal ganglia inspired model of action selection evaluated in a robotic survival task. *Journal of Integrative Neuroscience*, *2*(22), 179–200.

Girard, B., Filliat, D., Meyer, J.-A., Berthoz, A., & Guillot, A. (2005). Integration of navigation and action selection functionalities in a computational model of cortico-basal-thalamo-cortical loops. *Adaptive Behavior*, 13 (2), 115–130.

Gurney, K. N., Prescott, T. J., & Redgrave, P. (2001a). A computational model of action selection in the basal ganglia: I. A new functional anatomy. *Biological Cybernetics*, *84*, 401–410.

Gurney, K. N., Prescott, T. J., & Redgrave, P. (2001b). A computational model of action selection in the basal ganglia: II. Analysis and simulation of behavior. *Biological Cybernetics*, *84*, 411–423.

Houk, J. C., Adams, J. L., & Barto, A. G. (1995). A model of how the basal ganglia generate and use neural signals that predict reinforcement. In J. C. Houk, J. L. Davis, and D. G. Beiser (Eds.), *Models of information processing in the basal ganglia*, Cambridge, MA: The MIT Press.

Ikemoto, S., & Panksepp., J. (1999). The role of the nucleus accumbens dopamine in motivated behavior: A unifying interpretation with special reference to reward-seeking. *Brain Research Reviews*, *31*, 6–41.

Jacobs, R. A., Jordan, M. I., Nowlan, S. J., & Hinton, G. E. (1991). Adaptive Mixture of Local Experts. *Neural Computation*, *3*, 79–87.

Joel, D., Niv, Y., & Ruppin, E. (2002). Actor–critic models of the basal ganglia: New anatomical and computational perspectives. *Neural Networks*, *15*, 535–547.

Joel, D., & Weiner, I. (2000). The connections of the dopaminergic system with striatum in rats and primates: An analysis with respect to the functional and compartmental organization of the striatum. *Neuroscience*, *96*, 451–474.

Killcross, A. S., & Coutureau, E. (2003). Coordination of actions and habits in the medial prefrontal cortex of rats. *Cerebral Cortex*, *13*(4), 400–408.

Lee, J. K., & Kim, I. H. (2003). Reinforcement learning control using self-organizing map and multi-layer feed-forward neural network. In *Proceedings of the International Conference on Control Automation and Systems, ICCAS 2003* (pp. 142–145). Gyeongju, South Korea.

McNaughton, B. L. (1989). Neural mechanisms for spatial computation and information storage. In L. Nadel, L. A. Cooper, P. Harnish, and R. M. Colicover (Eds.), *Neural Connections, Mental Computations* (chapter 9, pp. 285-350). Cambridge, MA: MIT Press.

Montague, P. R., Dayan, P., & Sejnowski, T. J. (1996). A framework for mesencephalic dopamine systems based on predictive Hebbian learning. *Journal of Neuroscience*, *16*, 1936–1947.

Montes-Gonzalez, F. Prescott, T. J., Gurney, K. N., Humphries, M., & Redgrave, P. (2000). An embodied model of action selection mechanisms in the vertebrate brain. In J.-A. Meyer, A. Bethoz, D. Floreana, H. L. Roitblat, and S. W. Wilson (Eds.), *From animals to animats 6: Proceedings of the Sixth International Conference on Simulation of Adaptive Behavior* (pp.157–166). Cambridge, MA: The MIT Press.

O'Doherty, J., Dayan, P., Schultz, J., Deichmann, R., Friston, K., & Dolan, R. (2004). Dissociable roles of dorsal and ventral striatum in instrumental conditioning. *Science*, *304*, 452–454.

Pennartz, C. M. A. (1996). The ascending neuromodulatory systems in learning by reinforcement: Comparing computational conjectures with experimental findings. *Brain Research Reviews*, *21*, 219–245.

Perez-Uribe, A. (2001). Using a time-delay actor–critic neural architecture with dopamine-like reinforcement signal for learning in autonomous robots. In S. Wermter J. Austin, and D. Willshaw (Eds.), *Emergent neural computational architectures based on neuroscience: A state-of-the-art survey* (pp. 522–533). Berlin: Springer.

Schultz, W. (1998). Predictive reward signal of dopamine neurons. *Journal of Neurophysiology*, *80*(1), 1–27.

Schultz, W., Apicella, P., & Ljungberg, T. (1993). Responses of monkey dopamine neurons to reward and conditioned stimuli during successive steps of learning a delayed response task. *Journal of Neuroscience*, *13*(3), 900–913.

Schultz, W., Dayan, P., & Montague, P. R. (1997). A neural substrate of prediction and reward. *Science*, *275*, 1593–1599.

Smith, A. J. (2002). Applications of the self-organizing map to reinforcement learning. *Neural Networks*, *15*(8-9), 1107–1124.

Sporns, O., & Alexander, W. H. (2002). Neuromodulation and plasticity in an autonomous robot. *Neural Networks*, *15*, 761–774.

Strösslin, T. (2004). *A connectionist model of spatial learning in the rat*. Ph.D thesis, EPFL, Swiss Federal Institute of Technology.

Suri, R. E., & Schultz, W. (2001). Temporal difference model reproduces anticipatory neural activity. *Neural Computation*, *13*, 841–862.

Sutton, R. S., & Barto, A. G. (1998). *Reinforcement learning: An introduction*. Cambridge, MA: The MIT Press.

Tang, B., Heywood, M. I., & Shepherd, M. (2002). Input partitioning to mixture of experts. In IEEE/INNS *International Joint Conference on Neural Networks* (pp. 227–232), Honolulu, Hawaii (pp. 227–232).

Tani, J., & Nolfi, S. (1999). Learning to perceive the world as articulated: An approach for hierarchical learning in sensory-motor systems. *Neural Networks, 12*(7-8), 1131–1141.

Thierry, A.-M., Gioanni, Y., Dégénetais, E., & Glowinski, J. (2000). Hippocampo-prefrontal cortex pathway: Anatomical and electrophysiological characteristics. *Hippocampus, 10*, 411–419.

Yin, H. H., Knowlton, B. J., & Balleine, B. W. (2004). Lesions of dorsolateral striatum preserve outcome expectancy but disrupt habit formation in instrumental learning. *European Journal of Neuroscience, 19*(1), 181–189.

About the Authors

Mehdi Khamassi is working as a Ph.D. student in cognitive science both at the AnimatLab of the Laboratoire d'Informatique de Paris 6 (LIP6) and at the Laboratoire de Physiologie de la Perception et de l'Action (LPPA–CNRS, Collège de France). He trained as an engineer and received a master's degree in cognitive science from the University Pierre and Marie Curie (UPMC Paris 6). His current research interests include electrophysiology experiments and computational modeling of learning processes in the rat brain.

Loïc Lachèze is working at the AnimatLab as a Ph.D. student. He received a master's degree in computer science from the University of Paris 6 in 2002. He currently contributes to the *Psikharpax* project and works on the robotic integration of visual process and control architectures of navigation and action selection. *Address*: AnimatLab, LIP6, 8 rue du capitaine Scott, 75015 Paris, France. E-mail: loic.lacheze@lip6.fr

Benoît Girard was trained as an engineer at the Ecole Centrale de Nantes (ECN), he received a Ph.D. in computer science (2003) from the University Pierre and Marie Curie (UPMC Paris 6). He is now a Post-Doc Fellow at the Laboratoire de Physiologie de la Perception et de l'Action (LPPA–CNRS, Collège de France) where he works on models of the primate saccadic circuitry. His research is focused on biomimetic neural network models of navigation, action selection and motor execution. *Address*: LPPA, Collège de France, 11 place Marcellin Berthelot, 75005 Paris, France. E-mail: benoit.girard@college-de-france.fr

Alain Berthoz was trained as an engineer. He graduated in human psychology and received a Ph.D. in biology. He is Professor at the Collège de France, where he heads the Laboratoire de Physiologie de la Perception et de l'Action (LPPA). He is a member of numerous academic societies, an invited expert in several international committees, and he has been awarded with many prizes. His main scientific interests are in the multisensory control of gaze, of equilibrium, of locomotion and of spatial memory. He coordinates the neurophysiological experiments that inspire the *Psikharpax* project. *Address*: LPPA, Collège de France, 11 place Marcellin Berthelot, 75005 Paris, France. E-mail: alain.berthoz@college-de-france.fr

Agnès Guillot is Associate Professor of Psychophysiology at the University of Paris X. She graduated in human and animal psychology, and holds Ph.D.s in psychophysiology and biomathematics. Her main scientific interests are in action selection in animals and robots. She coordinates the biomimetic modeling of *Psikharpax* at the AnimatLab of the Laboratoire d'Informatique de Paris 6 (LIP6). *Address:* AnimatLab, LIP6, 8 rue du capitaine Scott, 75015 Paris, France. E-mail: agnes.guillot@lip6.fr