对抗行为学习(OpAL)：纹状体多巴胺在强化学习和选择激励的交互作用下建模

Opponent Actor Learning (OpAL): Modeling Interactive Effects of Striatal Dopamine on Reinforcement Learning and Choice Incentive

Anne G. E. Collins and Michael J. Frank

[1]*Department of Cognitive, Linguistic and Psychological Sciences, Brown Institute for Brain Science, Brown University.*

Accepted: 2014 by Psychological Review

(translated by zang jie)

摘要：纹状体多巴胺能系统与强化学习(RL)，运动表现和激励动机有关。已经有人提出了各种计算模型来单独考虑这些影响中的每一种，但是缺乏对其相互作用的正式分析。在这里，我们提出了一种新颖的算法模型，该模型将经典的行为评论体系结构扩展到包括神经回路模型的基本交互属性，并将激励和学习效果都整合到单个理论框架中。标准行为被代表不同纹状体种群的双重对抗行为系统所代替，该系统以不同的方式专门区分正向行动值和负向行动值。多巴胺调节每个行为成分对学习和选择偏好做出贡献的程度。与标准框架相比，该模型在包括概率学习，基于努力的选择和运动技能学习在内的各种研究中，同时捕获了多巴胺对学习和选择动机及其相互作用的文献记载的影响。

关键词：多巴胺，纹状体，强化学习，选择动机，计算模型

1、引言

多巴胺在人类和动物的认知中起着至关重要的作用，在很大程度上影响着多种过程，包括强化学习，动机，激励，工作记忆和努力。黑质和腹侧被盖区的多巴胺能神经元投射到非常广泛的皮下和皮质区域，在基底神经节(BG)中，尤其是在腹侧和背侧纹状体中，神经支配力最强。多巴胺失调存在于多种精神疾病中，例如帕金森氏病，注意力缺陷/多动症(ADHD)，精神分裂症和图雷特氏综合症，并且是治疗这些疾病和许多其他病状的主要药物靶点。

尽管我们对它的各种不同作用的理解已经取得了很大的进步，但是关于它的精确机制和功能，特别是关于它们的整合和相互作用，仍然存在根本的争论。特别是在基于奖励的决策中，有两种截然不同的传统研究了强化学习和多巴胺的激励理论(Berridge，2007)。尽管这两种理论都有确凿的证据，但是理论和实证研究倾向于偏重或专注于一种或另一种解释，很少尝试统一它们或研究它们之间的相互作用。在这里，我们对纹状体多巴胺在调节激励动机(影响选择)，强化学习以及这些过程如何相互作用方面的双重作用进行了明确的计算分析。这项工作不仅使我们能够单独解释这两种类型的发现，而且使我们无法单独用两种理论来解释这些发现。

2、多巴胺的 RL 理论

一种被广泛接受的多巴胺功能理论涉及其在无模型强化学习(RL)中的作用。具体而言，中脑多巴胺神经元的相继放电传达了奖励预测误差，这些误差有助于纹状体的可塑性(Montague，Dayan 和 Sejnowski，1996；Schultz，1997)。此后，许多研究为该概念提供了有力的支持(Arias-Carrión，Stamelou，Murillo-Rodriguez，Menéndez-González 和 Pöppel，2010 年；Bayer 和 Glimcher，2005 年；Bayer，Lau 和 Glimcher，2007 年)。强化学习模型通常用于说明学习任务期间行为和神经信号的多巴胺能调节(Frank，Moustafa，Haughey，Curran 和 Hutchison，2007 年；Jocham，Klein 和 Ullsperger，2011 年；McClure，Daw 和 ReadMontague，2003；O'Doherty 等，2004；Pessiglione，Seymour，Flandin，Dolan 和 Frith，2006；Samejima，Ueda，Doya 和 Kimura，2005；Schönberg，Daw，Joel 和 O'Doherty，2007)。

这样的模型假设每个动作都有一个值，该值会因多巴胺编码的奖励预测错误而增加或减少，以驱动学习。通过在给定的感觉状态下比较所有可用动作中的当前动作值，并随机选择一个动作，从而可以更容易地选择具有较高值的动作，从而在不

同动作之间进行选择。这些模型说明了各种各样的数据，但仅靠学习后就无法捕捉多巴胺对激励选择的明显调节作用(即差异权衡成本和收益的趋势(Berridge，2012；Salamone，Correa，Mingote 和 Weber，2005 年)。他们也不容易适应多巴胺操作对从积极结果到消极结果学习中的不对称影响。相反，激励选择的理论和模型(Zhang，Berridge，Tindell，Smith 和 Aldridge，2009 年)没有说明渐进式学习强化作用，也没有发现帕金森氏病的运动症状即使没有多巴胺能变性也可以发展(Beeler)。，Frank，McDaid 和 Alexander，2012 年)。

相反，其神经生物学和神经网络模型提出了更复杂的行动和学习对抗系统。已知多巴胺(DA)可以调节两个独立的细胞群中纹状体中棘神经元(MSN)的活性和质性，这些细胞投射到不同的 BG 输出核(Frank，2005；Gerfen，2000；Shen，Flajolet，Greengard 和 Surmeiier，2008；Surmeier，Ding，Day，Wang 和 Shen，2007)。起源于直接(纹状体神经通道)途径的纹状体 MSN 主要表达多巴胺 D1 受体并起促进作用的作用(Kravitz 等，2010)。通过刺激这些神经元中的 D1 受体，多巴胺可增强信噪比并增强活性和可塑性(长期增强)。相比之下，起源于间接(纹状戊二醛)途径的纹状体 MSN 主要表达多巴胺 D2 受体，并起抑制作用的作用。通过刺激这些神经元中的 D2 受体，多巴胺抑制其活性并引起长期抑郁。因此，总的来说，多巴胺的增加可优先强调 D1 促进途径中的加工并抑制 D2 抑制性途径中的加工，而多巴胺的减少则具有相反的作用，从而增强 D2 途径。这被认为是 DA 促进直接途径学习和间接途径避免学习的一种机制(Frank，2005 年)，其中动作价值和学习的表示相反，但表面上是多余的。尽管直接(D1-MSNs)和间接(D2-MSNs)路径由于它们分别与进场和回避联系而经常被标记为 Go 和 NoGo 路径，但是在模型中，它们不只是编码表示是否通过的消息，而是支持每种行动与反对该行动的汇总证据。最终的选择取决于所考虑的每个动作的证据量的相对差异，这些差异是通过皮层眼部回路各个阶段的竞争来实现的。

因此，多巴胺失调在单独的途径中以相反的方向起作用。此功能已被广泛用于展示与多巴胺相关的药物，基因，病理等的作用，所有这些作用都可在阳性和阴性结果的治疗中引起不对称性(例如，对方法与回避有相反的影响)学习)。例如，未经药物治疗的帕金森氏症患者自然具有较低的多巴胺水平，并且比积极的奖励预测错误具有更好的负学习效果，而同一位接受多巴胺能药物治疗的患者表现出更好的学习和基于积极结果的选择，但表现较差避免负面结果的发生(Bódi 等，2009;Cools 等，2009;Frank，Moustafa 等，2007;Frank，Seeberger，＆O'Reilly，2004;Moustafa，

Sherman，＆Frank，2008;Palminteri，Boraud，Lafargue，Dubois 和 Pessiglione，2009；Smittenaar 等，2012)。在健康人群和其他人群中也观察到了多巴胺操纵的类似作用(Cools 等，2009；Frank，Moustafa 等，2007；Frank，Santamaria，Reilly，＆Willcutt，2007；Jocham 等，2011；Pessiglione 等人，2006)。这种多巴胺功能强化学习理论还可以解决其他反直觉现象，例如在某些情况下的异常学习(例如 Beeler，Daw，Frazier 和 Zhuang，2010 年；Wiecki，Riedinger，vonAmeln-Mayerhofer，Schmidt(＆Frank，2009：学到的僵直症)，并提供了解释帕金森氏病症状进展的机制，甚至没有进一步的多巴胺能变性(Beeleretal。，2012)。

多巴胺在特定神经回路中的作用的更直接的探测来自光遗传学研究，该研究证实了 D1 和 D2 途径在进近和避免学习中的作用(Kravitz，Tye 和＆Kreitzer，2012)。在小鼠内源性地选择了一种特定的作用后，D1MSN 的光遗传刺激导致该特定作用的正增强，从而导致小鼠将来重复这种作用。相反，D2MSN 的光遗传刺激导致该作用被避免。值得注意的是，仅在这些研究中的选择之后才施加刺激的效果，因此，行为偏好的任何后续变化都只能归因于学习机制，而不是直接的表现效果。而且，刺激 D1 和 D2 细胞的效果模仿了分别因多巴胺突增和骤降而发生的效果。尽管这项研究表明 D1 和 D2 刺激分别足以诱导方法和回避学习，但其他基因工程研究也表明它们是必要的(Hikida，Kimura，Wada，Funabiki 和 Nakainishi，2010 年)。因此，许多独立的数据点间接或直接证实了纹状体中多巴胺在强化学习中的作用。但是，许多强化学习研究也无法控制多巴胺的潜在混淆性激励作用，如下所述。

3、多巴胺激励理论

在强化学习领域之外，各种类型的证据表明，多巴胺也直接参与了选择，与动机，动机，活力或努力意愿联系在一起(Berridge，2012；Smith，Berridge，＆Aldridge，2011；Wassum，Ostlund，Balleine 和 Maidment，2011 年)。例如，大量研究表明，高多巴胺能大鼠愿意为相同的报酬付出更多的努力(Beeler 等，2010；Cousins＆Salamone，1994；Salamone 等，2005)。有效的表观“成本”是通过间接 D2MSN 活性的操纵来双向调节的：提高这种活性的药理学操纵可避免更多的努力行为，而抑制这种途径则具有相反的作用，即降低有效成本(Farrar 等，2010，2008；Mingote 等，2008；Nunes 等，2010)。神经模型表明，如电生理研究中所观察到的那样，这些作用是由不同 MSN 群体中作用的正面和负面结果的差异编码介导的(Samejima 等，2005)。

最近的光遗传学研究(Tai，Lee，Benavidez，Bonci 和 Wilbrecht，2012 年)也更准确地证实，通过在选择期间(而不是在结果中)刺激 D1 或 D2MSN，可以增加或减少特定的作用值。在前面介绍的学习研究中)。在一个半球中刺激 D1MSN 可以增加选择对侧动作的可能性。值得注意的是，这并不是纯粹的运动效果：刺激并不能简单地确定性地增加运动反应，而是可以提高运动价值。需要较高水平的刺激来诱导对具有较低学习值的动作的选择，以及较低水平对已经具有较高值的　　动作的选择。令人惊讶的是，D2MSN 刺激具有相反的效果，有效降低了作用值(或增加了其有效成本)。总之，在选择时应用 D1(分别是 D2)时，刺激会模拟对该动作的最新估算值产生加性的正(负)效果。

其他建模研究提出，补充多巴胺可调节反应活力，以优化每单位时间的奖励(或避免惩罚)(Dayan，2012；Niv，Daw，Joel 和＆Dayan，2007)。但是，请注意，这些模型仅考虑了对活力的影响(即，响应执行的速度或工作的难易程度)，而没有考虑对具有不同效价/激励的行动之间的选择的影响。此外，他们专注于增加 DA 信号的作用，而不关注某些 DA 消耗情况下性能的相对提高。

最近的研究辩论了多巴胺的激励或绩效效应与另一方面的强化学习效应之间的联系，有人认为所有强化学习效应都可以通过激励显着性来重新解释(Berridge，2012)。确实，许多上述证明多巴胺操纵对阳性结果与阴性结果的不同影响的实验并未在学习帐户与激励帐户之间进行区分。可以说，在选择帕金森氏病和其他人体研究中，奖励与惩罚学习中的某些不对称性有可能在选择时通过差别激励来弥补，即使给予对称性学习也是如此。例如，最近的一些工作提供了证据，表明在未观察到或不可能在任务中获得学习效果的情况下，多巴胺调节可影响行动选择对积极和消极结果的相对敏感性(Shiner 等人，2012；Smittenaar 等人。，2012)。具体来说，Smittenaar 等(2012 年)显示，即使在学习过程中没有分配任何刺激价值的差异值时，帕金森氏病患者与非药物治疗相比，也能够更好地选择能够带来最大回报的行动。仅在学习后才分配给结果。相反，Shiner 等人(2012 年)使用标准的刺激值学习程序，但仅在学习后才滥用 DA 药物，尽管如此，观察到在学习后阶段接受药物治疗的患者在奖励方面的表现要优于厌恶选择。这些研究提示存在绩效/激励作用，但不排除多巴胺在学习中的额外作用。确实，这些研究中的激励作用无法解释所有先前记录的数据：在整个人类研究中，多巴胺升高的最强有力的作用是削弱了对阴性结果的学习，而在这些研究中，仅对阳性结果的敏感性产生影响，总体药物治疗对阳性结果与阴性结果的敏感性差异的影响程度要比在各种研究中观察到的结果要温和得

多，这些研究也可能受到学习的影响。此外，其他研究也提供了学习效果的证据，例如，纹状体对学习过程中的奖励预测错误的反应可预测随后基于奖励的选择偏好，并且这种关系受到多巴胺能操纵的调节(Jocham 等，2011)。总而言之，目前的证据表明，多巴胺能操纵会影响学习和动机。

尽管关于多巴胺对各种行为措施的学习和绩效影响的各自贡献尚存争议，但一些研究为这两种功能之间的相互作用提供了证据。特别是 Beeler 等。(2012 年)使用多巴胺拮抗剂诱导啮齿动物面对运动技能任务的表现缺陷。单独来看，这些作用与悠久的历史证据相一致，即纹状体多巴胺对运动表现很重要，如帕金森氏病。但是，值得注意的是，这项研究表明，即使在药物冲洗后，与从未接触过这项任务的幼稚动物以及也接受过多巴胺拮抗剂但未配对的动物相比，动物获得正确的运动技能的速度仍较慢任务。这项研究表明，药物对性能的影响引起“异常学习”过程，使动物学会避免选择本来可以适应的行为。随后的实验表明，在学习了一定的技能后应用 D2 封锁时，效果相似：在这种情况下，性能并未立即下降，而是逐渐下降，这与诱导异常学习和平行的突触可塑性研究一致 D2 拮抗作用增强了纹状体蛛网膜突触的增强作用(Beeler 等，2012)。这些作用还与其他证据相吻合，即中等剂量的 D2 拮抗剂可以以僵住症致敏的形式诱导进行性帕金森病症状(Amtage＆Schmidt，2003；Klein＆Schmidt，2003)，并且这两种作用均由在神经模型中模拟 D2 拮抗作用(Beeler 等人，2012;Wiecki 等人，2009)这些研究突显了多巴胺的表现效应(在这种情况下，缺乏多巴胺)之间相互作用的作用，这种相互作用诱导了学习效果，然后进一步夸大性能影响，等等。

在这里，我们提出了一种新的强化学习模型，该模型使我们能够同时考虑多巴胺的激励，学习和交互作用。我们旨在提供一种理论上简单的算法模型，其参数和变量可以轻松地与感兴趣的生物学解释性指标相关，例如强直或多巴胺水平，D1 和 D2 表达的纹状体神经元活性或突触强度，突触可塑性等。我们的方法受到两个不同级别的建模的启发：一方面，是众所周知的且广泛使用的行为者评论算法(Sutton＆Barto，1998)；另一方面，它是基于模型的。另一方面，对包括多个途径的皮质基底节循环的生物学上更详细的神经网络描述(Frank，2005)。尽管以前尝试过将这些模型的各个层次联系起来，但是这些并未考虑用于行动选择和学习或激励效果的单独的评估系统，而是为每个行动包括了一个单一的值，并且仅考虑了非对称的学习率。正预测误差与负预测误差(Doll，Hutchison，＆Frank，2011;Frank，Moustafa，etal。，2007)。如下所示，此机制不足以说明数据范围。此处的目的是提供一种模

型，该模型可以表现出更普遍的但相互影响的动机激励，绩效和学习效果，这可以解释文献中现有公式无法涵盖的一系列发现。我们进一步提供了分析，提出了这种评估系统分离的规范性原因。

4、模型和仿真方法

4.1、OpAL 模型说明

我们标记为对抗行为学习的新模型依赖于行为评价体系。行为者评价体系结构假定一个系统(评价者)估计环境当前状态的值，而行为者选择动作。当结果好于或差于预期时，评论家会产生奖励预测错误，该错误有两个用途：更新其未来估计值，以便对价值进行更好的估计，并修改参与者的权重。会增强产生正面预测错误的动作，而对产生负面预测错误的动作进行惩罚。通常，评论家被分配到腹侧纹状体和杏仁核(Hazy，Frank 和 O'Reilly，2010；O'Doherty 等，2004)，而行为被认为是通过背纹状体与前/运动皮层的相互作用而被实例化的。我们模型中的评价者与经典公式中的评价者相似(请参见讨论)。它估计给定选择选项 1 的预期值，并通过简单的增量规则学习算法更新该值：

$$V(t+1) = V(t) + \alpha_C \times \delta(t). \qquad (1)$$

因此，估计值 V 的更新与预测误差 $\delta(t) = r(t) - V(t)$ 成正比，其中 r(t)表示在时间 t 收到的强化，而 α_C 是评价者的学习率。我们普遍假设评价者的价值在腹侧纹状体中存在，多巴胺的相位信号传达了评价者的预测误差(Dayan & Daw，2008；Montague 等，1996；Roesch，Calu，& Schoenbaum，2007)。

行为学习。典型的行为选择机制为每个动作分配了一组权重，并根据评论家预测误差来增加或减少这些权重。在 OpAL 模型中，我们将参与者分为两组权重，分别表示皮层神经突触权重为编码状态动作对(s，a)的直接(G 为 Go)和间接(N 为 NoGo)MSN 群体。为了简化说明，我们在这里考虑具有多个动作选择的单个状态，从而将(s，a)简化为 a。这些分别标记为 Ga(t)和 Na(t)的行为权重被约束为正(射击率和谷氨酸能突触权重不能为负)。

对这些行为权重的学习模仿了神经网络模型中的学习机制，如下所示：

$$G_a(t+1) = G_a(t) + [\alpha_G G_a(t)] \times \delta(t) \qquad (2)$$

$$N_a(t+1) = N_a(t) + [\alpha_N N_a(t)] \times [-\delta(t)] \qquad (3)$$

在此，$\delta(t)$ 是先前定义的评价者预测误差，而 aG 和 aN 分别是 Go 和 NoGo 权重的学习率。该模型结构(双重行为权重)及其时间动态(更新规则)反映了与典型 RL 模型的背离。首先，从生物学上出发，存在单独的 G 和 N 权重以及更新规则的两个不同寻常的特征。更新规则的第一个特征是，与 G 权重相反，N 权重通过预测误差的相反符号进行更新 [2]。这表明了多巴胺通过刺激 D1 和 D2 受体对可塑性产生相反作用的观点。不同的人群，但他们俩都可能经历增强和抑郁。凭直觉，G 权重会累积预测误差，因此应代表一个选项的好坏程度的指标，而 N 权重随负的预测误差而增加，而正的预测误差会减小，因此应代表一个期权的厌恶程度。

更新规则的第二个特征与典型的 RL 更新不同，它的学习程度不仅取决于预测误差，还取决于当前行为权重，作为学习率的乘数。这捕获了经常被引用的三因素 Hebbian 规则，其中学习取决于突触前激活(来自皮层，刺激和作用)，纹状体中突触后激活(与行为的权重成比例)和多巴胺((Reynolds, Hyland, & Wickens,，2001)。尽管此规则通常与强化学习模型相关联，但这些规则通常实际上并未实施三要素规则。实际上，它们有效地将突触前(刺激作用表示)和多巴胺(预测误差)约束纳入学习规则中，但并不取决于突触后调节，我们在此通过结合行为权重(这将决定突触后水平)来获得。激活)。重要的是，下面的仿真将这一更新规则与更新方程中没有行为权重的更新规则进行了比较，并表明行为值对学习的调节对于考虑数据范围(包括性能对学习的影响)是必要的，并且 G 和 N 权重分别优先区分正作用值和负作用值的趋势。

与典型 RL 模型的另一个不同之处在于，通过允许对 G 或 N 是否会影响选择进行单独的动态调节，G 和 N 权重的存在而不是单一系统中的 actor 权重可能在灵活性方面提供计算优势)。最后，我们在结果部分显示这些功能对于解释表明多巴胺可以影响激励而不影响学习，反之亦然的数据至关重要。

规则。在参与者权重的线性组合上，作为 softmax 选择策略，给出了不同选项之间的选择，例如不同可用动作选择 a_i 之间的选择：

$$Act_a(t) = \beta_G G_a(t) - \beta_N N_a(t) \quad (4)$$

$$p(a) = \frac{e^{Act_a(t)}}{\sum_i e^{Act_{a_i}(t)}} \quad (5)$$

在这里，p(a)是选择动作 a 的概率。它取决于组合的行为权重 Act_a，即 Ga 和 Na 与所有其他候选动作的权重之差，代表基底神经节 GPi 输出核中直接和间接途径活性之间的竞争(见图 1 左)。参数 β_G 和 β_N 调节给定试验中 G 和 N 权重的表示程度，从而 $\beta_G G_a(t)$ 代表相关 G 群体的激活，$\beta_N N_a(t)$ 代表相关 N 群体的激活。softmax 函数将选择中的非线性实现为值的函数。因此，根据 β_G 与 β_N 的不对称性，以下提议的与选择时的多巴胺水平有关，可以不同地表示动作的收益和成本，并可以选择不同的动作(图 1)。注意，我们可以将参数重写为 $\beta_G = \beta \times (1+p)$ 和 $\beta_N = \beta \times (1-p)$。在这种形式中，-1<p<1 表示权重之间的不对称性，而 β 对应于经典的 softmax 逆温度参数，控制勘探与开发。

反应时间。我们通过 softmax 将选择项 a 的反应时间建模为其作用因子 Act_a 的函数：

$$RT(a) \propto RT_0 + 1/(1 + e^{(Act_a(t) - \theta)}), \quad (6)$$

其中 θ 是 Go 相对于促进该动作所需的 NoGo 途径活性的阈值，而 RT_0 只是基线反应时间。该方程式捕获了由神经网络模拟生成的 RT，其中，相对于 NoGo 种群活动而言，相对更大的 Go 导致更快的 RT(Moustafa 等，2008；Wiecki 等，2009)，以及决策阈值(通过漂移扩散估算模型)部分受基底神经节的输出控制，与此处的 Act 值相对应(Ratcliff 和 Frank，2011 年)。

4.2、模拟多巴胺能对学习和激励的影响

下面报告的模拟结果表明，以上定义的模型可以捕获学习和选择动机中的潜在不对称性。关于学习，在神经回路模型中，多巴胺能调节可以增强相位猝发信号传导，增强对正预测误差的敏感性。但是，较高的进补水平也可以防止 D2 受体检测到相倾角，从而降低了对阴性预测误差的敏感性(Frank，2005)。相反，低的多巴胺水平可能会减少相位爆发信号，但实际上会增加对倾斜的敏感度(Frank＆O’Reilly，2006)。在 OpAL 框架中，可以通过行为学习率参数 aG 和 aN 中的不对称来模拟这种调制，从而捕获 D1 和 D2MSN 对多巴胺能信号的敏感性以及由此产生的可塑性

影响。相反，选择时多巴胺操作的选择激励作用可以用 β_G 和βN 中的不对称性来建模，从而调节两种途径中学习值的表达程度。例如，选择时高水平的多巴胺会增强活性 D1-MSN，但会抑制 D2-MSN，因此可以通过 β_G 的增加和 β_N 的减少，反之以多巴胺的减少来建模。

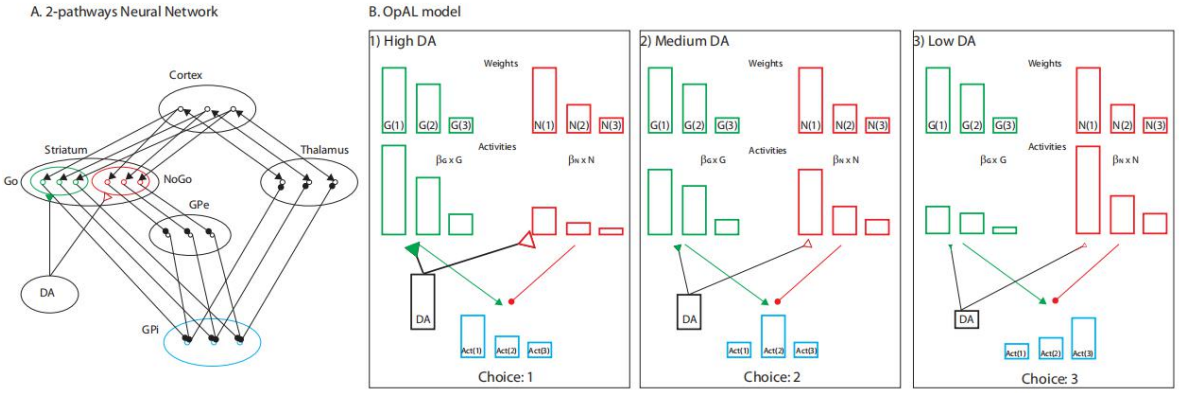

图 1. 神经网络和对抗行为学习（OpAL）模型。A. 表示皮质基底神经节环的神经网络的示意图，在 Frank（2005）等人中使用它来模拟多巴胺对学习和表现的各种影响。B. OpAL 模型的表示，具有给定刺激的三个不同多巴胺能状态的三个选项之间的选择示例。在正常的多巴胺能状态(2，中）下，相对于其他选项，模型的角色权重（Act）偏向于具有高 G 权重和低 N 权重的动作。在高多巴胺能状态（左，1）中，在行为选择上强调 G 值比强调 N 值更多，从而导致选择具有相对较高 G 权重的动作，而很少考虑动作成本。相反，在低多巴胺状态下（右图 3），发生相反的情况，并且模型选择具有最低 N 权重的动作。DA=多巴胺；GPI=苍白球的内部部分。

总而言之，作为一个近似值，我们使用 aG，aN 参数对潜在的学习效果进行建模，并使用 β_G， β_N 参数对潜在的激励或绩效效果进行建模。与多巴胺的强直性(基线)效应(Niv 等人，2007 年)相比，这种区分大体上符合阶段性多巴胺编码预测误差的独立影响(Montague 等人，1996 年)。但是，我们注意到，我们的模型表明，激励和学习效果之间的关键区别仅取决于选择时的纹状体多巴胺水平与强化时的水平，因此，选择时发生的任何阶段性猝发也会影响激励选择和反应时间(参见例如 Satoh，Nakai，Sato 和 Kimura，2003 年，其中相位 DA 信号与更快的 RT 相关)，由 β_G 和 β_N 参数捕获。

5、结果

5.1、模型动力学

在合理的假设下，标准的强化学习模型(包括行为评价模型)已被证明可以收敛于估计给定状态和/或动作的未来折现奖励的预期总和的概率。希望确保我们模型中的行为具有对理性学习和决策有用的相似属性，例如，它不会发散或是预期奖励的单调递增函数。我们在这里显示了一些模拟，这些模拟可以验证最关键的方面，以确保 OpAL 定义合理的学习和选择策略(并在支持信息中包括一些理论推导)。在第一组模拟中(参见图 2)，我们在参数选择的奖励设置中，使用中性(对称)参数($\alpha_G = \alpha_N = \alpha_C = 0.1; \beta_G = \beta_N = 1$)。这些模拟表明，G 和 N 分别是 r 和 p(r)的递增和递减凸函数，而 Act 是非线性递增函数(图 2B)。

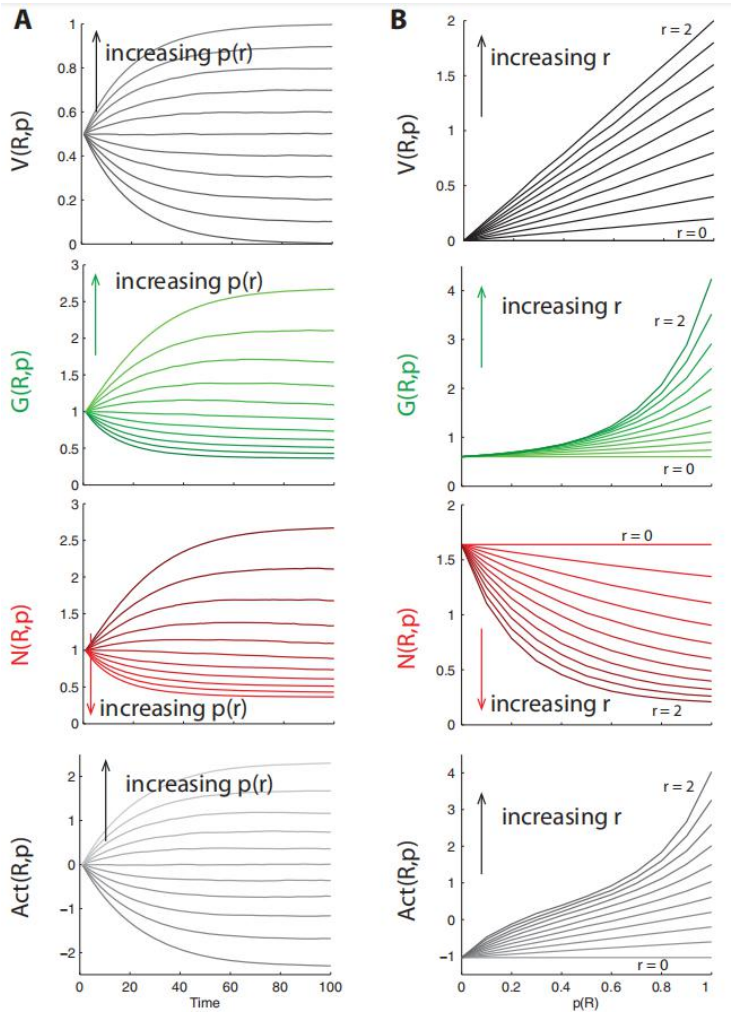

图 2. 模型动力学作为时间，奖励价值和奖励概率的函数。所有值均是对 1,000 个仿真的平均值。模型值是（从上到下）批评者值 V，G 权重，N 权重和行为值 Act 在这些模拟中，我们显示了模型变量随时间的变化而具有不同的奖励概率的对称性模型参数。上图显示评论者值 V 迅速收敛到真实的期望值。第二和第三张图显示了作为权重值 R 和 p（r）的函数，权重 G 和 N 在相反方

向上的相反演变。请注意，G 权重夸大了高期望值的差异，而 N 权重夸大了低期望值的差异。底部图显示，具有对称β参数，角色权重 Act 获得的值与真实期望值正相关，没有偏差，但是期望值的这种表示是非线性的。B. 在这些模拟中，我们通过操纵奖励值 r 和奖励概率 p（r）显示了 100 次试验后的最终值。

特别是，G 权重与真实期望值呈正相关，这增加了趋近趋势，但是曲线的凸度表明该函数是非线性的：它们在较高值的刺激中表现出更大的差异。因此，与 p(r)比较高时，两个刺激/动作的奖励概率(即 p(r)和 p(r+概率))之间的固定差 fixed 将在 G 权重中更大程度地放大。到低，特别是如果 r 也很高。

相反，N 个权重与真实期望值负相关，并起到支持避免倾向的作用。在此，凸度表示 N 个权重差异地强调了较低(而不是较高)的值刺激/动作表示。这些影响在时程图中特别明显(图 2A，G 和 N 的中间图)。但是，应注意的是，此处使用对称参数，净 Act 值会在没有 G 和 N 偏差的情况下演化，就像标准强化学习值一样，但在极端值时会强调差异(比较顶部和底部的曲线).[3]

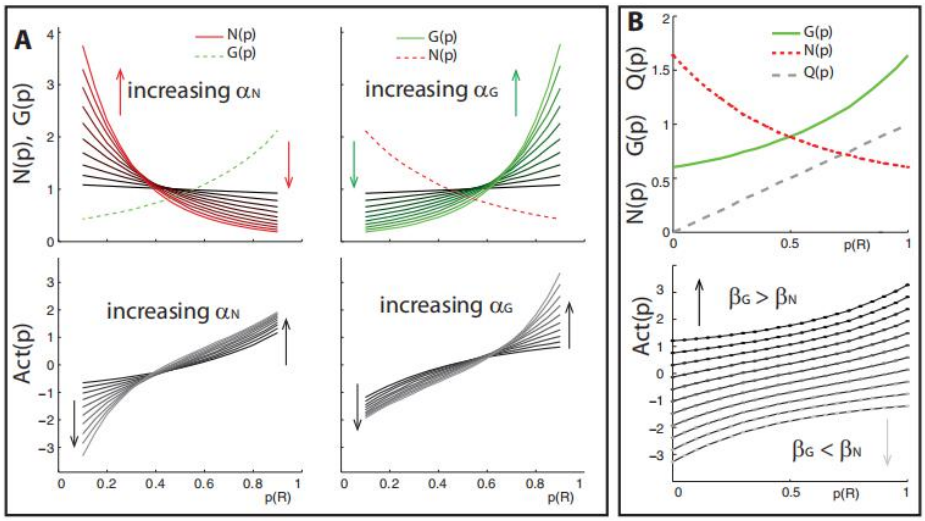

图 3. 模型值作为奖励概率 p（R）和模型参数的函数。所有值均为 100 次试验后的最终值，是对 1,000 次模拟的平均值。模型值是（从上到下）批评者值 V，G 权重，N 权重和演员值 Act=G－N。A. 学习率 aG 和 aN（参数值的增加由较细的线和箭头方向表示）。aG 的增加（右）增强了好的选项的编码值并在其中区分了它们，并压制了不好的选项，而对 aN 则相反（左）。B. βG 和βN 参数。当βG=βN 时，演员权重 Act 是期望值的线性函数。但是，导致对βG（较暗的点）的重量不对称会导致 Act 值增加（更愿意选择），并且其表示形式凸出（更好地区分良好的选择）。βN<βG（较浅的圆圈）则相反。有关该图的彩色版本，请参见在线文章。

5.2、参数效果

行为学习率(aG 和 aN)的影响。在第二组模拟中(参见图 3A)，我们分别操纵了行为的学习率。这些发现表明，较高的学习率会强调正常动态中的调节，因此，随着 aG 的增加，G 权重更高的人会选择好的选择，而 G 权重更小的人则会看到不好的选择，从而导致参与者奖励价值的夸大影响重量。相反，增加 aN 会导致低价值期权之间更大的代表性和差异性。请注意，尽管在更新 G 和 N 权重时对正负预测误差进行了相同处理，即正预测误差增加 G 并减小 N 权，反之亦然，但负效应仍然有效。这些不同的系统区分积极结果与消极结果的原因在于学习规则的希伯来调制，这种调制对时间上奖励预测误差的累积有不同的影响，因此它们代表了正值和负值。

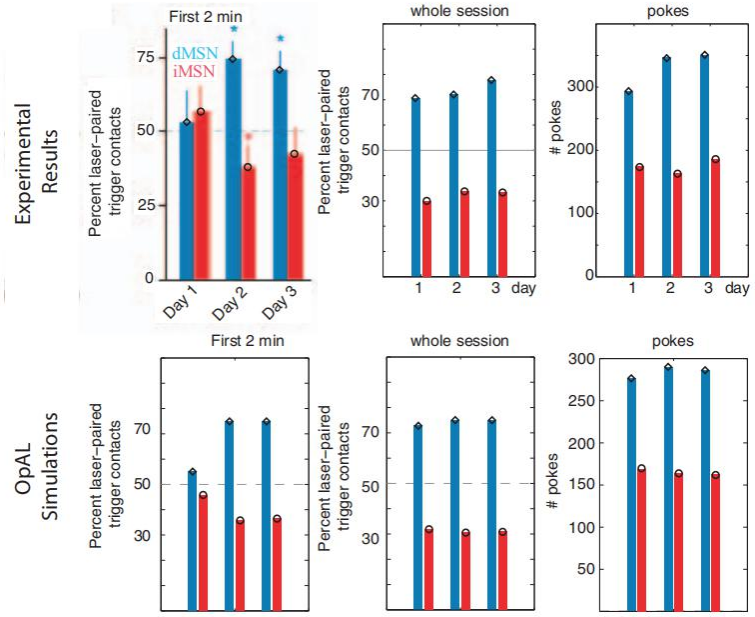

图 4。光遗传学和学习。顶线：实验结果（左图摘自 Kravitz，Tye 和 Kreitzer，2012；中和右图摘自 Kravitz 等人的结果表）。激光配对触发接触百分比是指动物选择触发光遗传激光刺激的动作的试验比例。直接中刺神经元（dMSN）刺激（顶部带有菱形的条；在线文章中的蓝色条）起到增加重复相同动作的可能性的作用，而间接 MSN（iMSN）刺激（顶部带有圆圈的条；红色的条在线）避免采取行动。底线：对抗行为学习（OpAL）模拟。左：在会话的前 2 分钟内激光配对的触发触点的比例。中：整个会话期间激光配对的触发触点的比例。右：会话期间的动作总数（激光配对和非激光配对）。基于.05 的 alpha，星号表示与机会（50）的显着差异。左上图改编自 A.V.Kravitz，L。D.Tye 和 A.C.Kreitzer，2012 年，《自然神经科学》，第 15 页，“直接和间接途径纹状体神经元在强化中的不同作用”。816.Macmillan 版权所有 2012。有关

该图的彩色版本，请参见在线文章。

选择激励(β_G和β_N)效应。前面我们看到，G 权重随着 p(r)和 r 中值的增加而放大，而 N 权重随着 p(r)和 r 中值的减小而放大。在这里，我们显示了βG 对βN 参数的调制，在选择时模拟了多巴胺能操纵，进一步放大了相应的权重偏差。当βG=βN 时，高或低值的差分强调会完美抵消 Act 权重。但是，即使使用对称学习，使用不对称βs 改变 G 和 N 系统之间的平衡也会揭示相应偏差的影响。

确实，仿真(图 3C)表明，当$\beta_G > \beta_N$时，Act 是 p(r)的凸函数，导致好选择之间的差异更大(显示 G 权重的影响)，而$\beta_N > \beta_G$则相反，揭示了 N 权重的影响。

5.3、模拟光遗传学对学习和表现的影响

以上模拟揭示了环境突发事件(奖励概率和大小)的变化如何影响模型动力学。这些建模结果可以直接与捕获光遗传学研究的结果联系起来，表明 D1 或 D2 的刺激可以根据刺激是在选择期间还是在结果中提供来不同地影响激励选择或强化学习(Kravitz 等，2012；Tai 等人，2012 年)。我们在此处明确地对这些实验进行建模，并重现所有主要发现。

Kravitz 等(2012)在大鼠选择特定动作后立即刺激表达 D1 或 D2 受体的纹状体 MSN。D1 刺激诱导进近学习，因此将来会优先选择此动作，而 D2 刺激诱导回避学习。而且，直接刺激 MSN 的这些效果不依赖于多巴胺，因为在服用多巴胺阻滞剂时它们仍然存在。

我们通过增强相应人群的活动相关学习规则来模拟 OpAL 中 MSN 的光遗传刺激(即，光遗传模拟可以增加 D1MSNs 刺激的 G 值和 D2MSNs 刺激的 N 值)。具体来说，我们假设刺激通过调节 G 和 N 活性水平以与多巴胺能预测误差相同的方式影响学习：

$$G_a(t+1) = G_a(t) + [\alpha_G G_a(t)] \times [\delta(t) + Opt_G] \qquad (7)$$

$$N_a(t+1) = N_a(t) + [\alpha_N N_a(t)] \times [-\delta(t) + Opt_N] \qquad (8)$$

其中 Opt_G=刺激 D1MSN 时为 Opt(否则为 0)，而 Opt_N=刺激 D2MSNs 时(否则为 0)。Opt>0 是代表刺激强度的参数。由于在这种实验范式中没有提供主要的强化，因此评价没有任何学习的价值。因此，假设相位多巴胺预测误差为$\delta(t) = 0$，但是我们通过添加 0 均值高斯噪声来允许随机波动。我们还假设速率ϕ==0.2 的学习权重中

存在一些遗忘，以捕获整体的消光效果(但 G 和 N 权重没有不对称性)以及选择选择中的无向噪声(ε=0.25)。请注意，所有结果都没有定性地依赖于这些参数。对于 100 次迭代，以对称学习速率 $\alpha_G = \alpha_N = 0.1$ 以及对称 softmax 权重 $\beta_G = \beta_N = 0.1$ 进行了仿真，并且光电强度 Opt=0.2。为了获得与实验中观察到的相似次数的试验，我们以 RT=5+10/(1+exp(Act))sec 对反应时间进行建模。模拟运行提供了 30 分钟的时间，这与 Kravitz 等人的实验设计相对应。(2012)。通过在选择参数中设置非对称性来模拟多巴胺阻滞的效果，其中 p=-0.3。

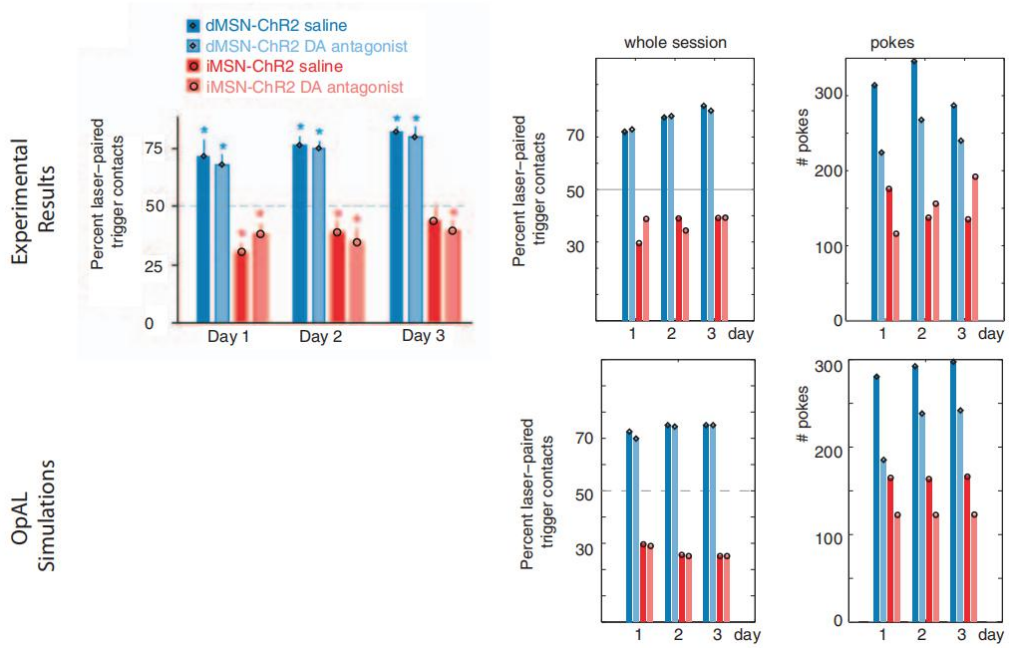

图 5. 光遗传学效应和多巴胺(DA)阻滞。顶线：实验结果（左图改编自 Kravitz, Tye 和 Kreitzer, 2012；中图和右图改编自 Kravitz 等人的结果表），即使在存在 DA 的情况下，由于光遗传刺激对相对偏爱的影响也相似拮抗剂。底线：对抗行为学习（OpAL）模拟。两条线（左图和中图）：整个会话期间激光配对的触发触点的比例。两条线，右图：尽管保留了相对偏倚，但 DA 拮抗剂减少了疗程中发出的动作（激光配对和非激光配对）总数。基于.05 的 alpha，星号表示与机会（50）的显着差异。左上图摘自 A.V.Kravitz，L。D.Tye 和 A.C.Kreitzer 的“直接和间接途径纹状体神经元在增强中的不同作用”，2012 年，自然神经科学，第 15 页。817.Macmillan 版权所有 2012。有关该图的彩色版本，请参见在线文章。

因此，在仅包括对 D1 或 D2MSN 的刺激的第一组模拟中，该模型针对与 D1MSN 刺激相关的刺激侧的动作开发了更强的 G 权重，因此增加了选择的可能性(蓝色条形)在图 4 中)。每个环节的前几次试验都是如此，这表明是学习而不是表现效果(图

4，左)。对于 D2MSNs 刺激，该模型开发了与触发动作相关的更强 N 权重，从而学会了避免。此外，这些增加的氮重量伴随着较慢的反应时间，从而导致在同一时间段内做出的总体选择更少，这与在啮齿动物中观察到的非常相似(图 4 中的红条，右)。值得注意的是，即使存在模拟的 DA 封锁(通过更改选择激励参数进行模拟)，这些发现仍然持续存在：由于模型中 DA 的学习效果是由 D1 或 D2MSN 的激活引起的，因此，由于 DA 的封锁，学习不对称仍然存在直接刺激。重要的是，这不仅仅是无效的影响：该模型的确预测了 DA 封锁会减少所选择行动的绝对数量，即使不改变行动之间的相对偏好也是如此。这两种效果都与 Kravitz 等人的观察非常吻合。(2012; 图 5)。

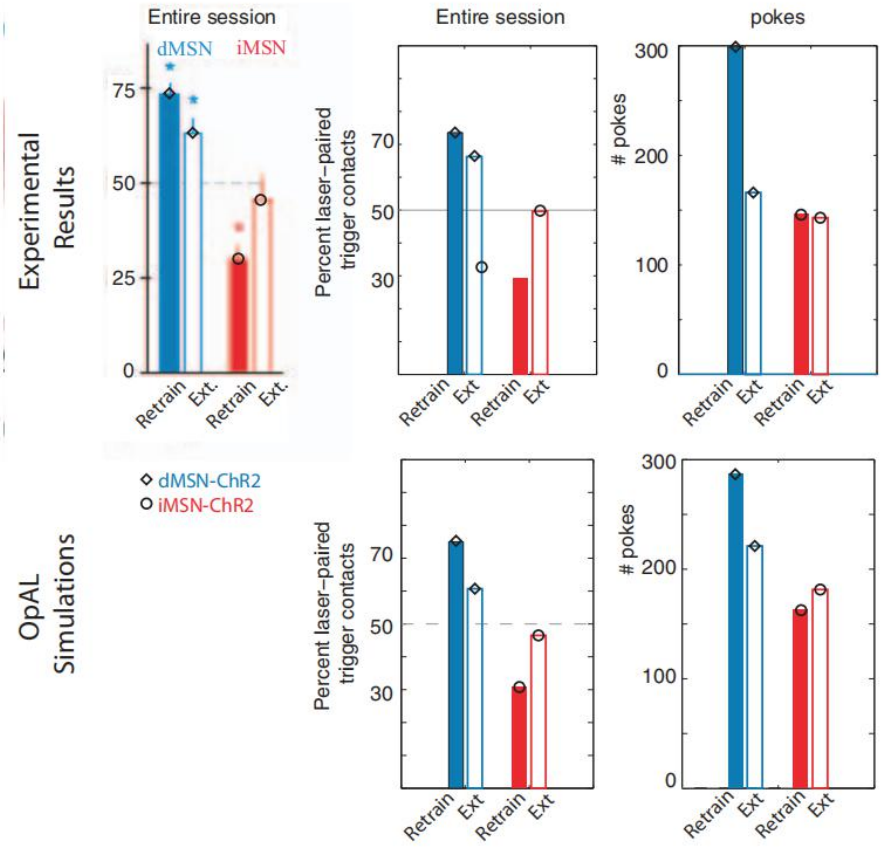

图 6。光遗传效应和消光（Ext）。第一行：实验结果（左图摘自 Kravitz，Tye 和 Kreitzer，2012；中和右图摘自 Kravitz 等人的结果表）。表面上的发现表明，间接中等多刺神经元（iMSN）刺激的鲁棒性较低，更容易灭绝。底线：假设对抗行为学习（OpAL）模拟在学习中直接 MSN（dMSN）和 iMSN 刺激的效果相同，则再现相同的模式。中间图：整个会话期间激光配对的触发触点的比例。右：会话期间的触发器总数（激光配对和非激光配对）。d（i）MSN-ChR2 表示在 d（i）MSNs 中表达 channelrhodopsin2 的小鼠组。基于.05 的 alpha，星号表示与机会（50）的显着差异。

左上图摘自 A.V.Kravitz，L。D.Tye 和 A.C.Kreitzer 的“直接和间接途径纹状体神经元在增强中的不同作用”，2012 年，自然神经科学，第 15 页。816.Macmillan 版权所有 2012。有关该图的彩色版本，请参见在线文章。

最后，Kravitz(2012 年)报道说，相对于 D1MSN 刺激，D2MSN 刺激诱导的学习功能不那么牢固，并且很快消失了。但是，我们在这里表明，他们的完整结果模式是在模型中获得的，而没有假设学习本身的鲁棒性有任何不对称性。具体来说，由于 D2 刺激会引起回避，因此从定义上讲，它减少了选择的动作次数，因此减少了训练试验的次数，从而导致较轻的累积 N 权重和更快的灭绝速度。确实，与 dMSN 刺激的 300 个动作相比，D2MSN 刺激与大约 150 个动作相关。图 6 显示，模拟 D1 和 D2MSN 刺激对学习的相等影响的模型可以重现所采取的动作数量的这些影响，以及相应的灭绝差异。

Tai 等(2012 年)刺激啮齿类动物表达 D1 或 D2 受体的纹状体 MSN，使用标准的一级补强进行逆向学习实验。在选择时，而不是学习时应用光遗传学刺激，选择性地增强了直接或间接途径 MSN 中的活性，对于那些仅对应于一种可用动作的 MSN 而言(即，动作选择是左还是右)单方面应用响应和刺激)。D1 和 D2 刺激对选择的影响不同：D1 受体的刺激增加了相应作用的选择，而 D2 受体的刺激则减少了相应作用。有趣的是，这种偏好的变化不是绝对的(即，它不仅会导致向左或向右运动)，而且还取决于奖励历史(图 7 的顶行，中右)。

5.4、概率选择任务

接下来，我们检查模型动力学如何发挥作用，以解释根据经验已知对多巴胺能操纵敏感的任务中选择比例的差异。在这里，我们报告了使用概率选择任务的简化版本和通用版本进行的模拟(Frank，Moustafa 等，2007；Frank 等，2004；见图 3)，但是对于任务。

在此版本中，在每个试用版中，都为模型提供了两个选项之间的选择。在某些试验中，可以在 A 或 B 之间进行选择，其中 A 是概率最高的选择，而 B 是惩罚最大的选择。在其他试验中，提供了在选项 M1 和 M2 之间进行选择的选项，它们分别具有中性值：

$$p(r=1 \mid \text{choice is A}) = 1 - p(r=0 \mid \text{choice is A}) = p > 0.5 \quad (9)$$

$$p(r=1 \mid \text{choice is B}) = 1 - p(r=0 \mid \text{choice is B}) = 1 - p < 0.5 \quad (10)$$

$$p(r=1 \mid \text{choice is } M_1 \text{ or } M_2) = 1 - p(r=0 \mid \text{choice is } M_1 \text{ or } M_2) = 0.5 \quad (11)$$

在学习阶段，模型会可靠地学习选择 A 而不是 B。在随后的转移阶段，模型会显示四个选择选项的所有可能新颖配对(例如，A 对 M1，A 对 M2 和 B 对 M1，B 与 M2)。没有提供反馈，因此没有进一步的学习机会；因此，偏好取决于在学习阶段为每个单独选项学习的值。与实证任务一样，我们将选择 A(ChA)绩效定义为选择 A 胜过 M 的概率，将避免 B 绩效(AvB)定义为选择 M 胜过 B 的概率。值得注意的是，在一系列实证研究中，在这项任务中，增加纹状体多巴胺的操作可增强 ChA 并损害 AvB，反之亦然，减少纹状体多巴胺的操作则相反。有关审查，请参见 Maia 和 Frank(2011)。请注意，A 和 M 之间的期望值差异与 M 和 B 之间的期望值差异相同。因此，Choose-A 和躲避 B 之间的任何性能差异均构成偏差 ChA_AvB，反映了对阳性结果和阴性结果的敏感性差异。重要的是，标准的强化学习模型应收敛于理论期望值，因此有望产生相等的 ChA 和 AvB 性能，从而使零偏差。

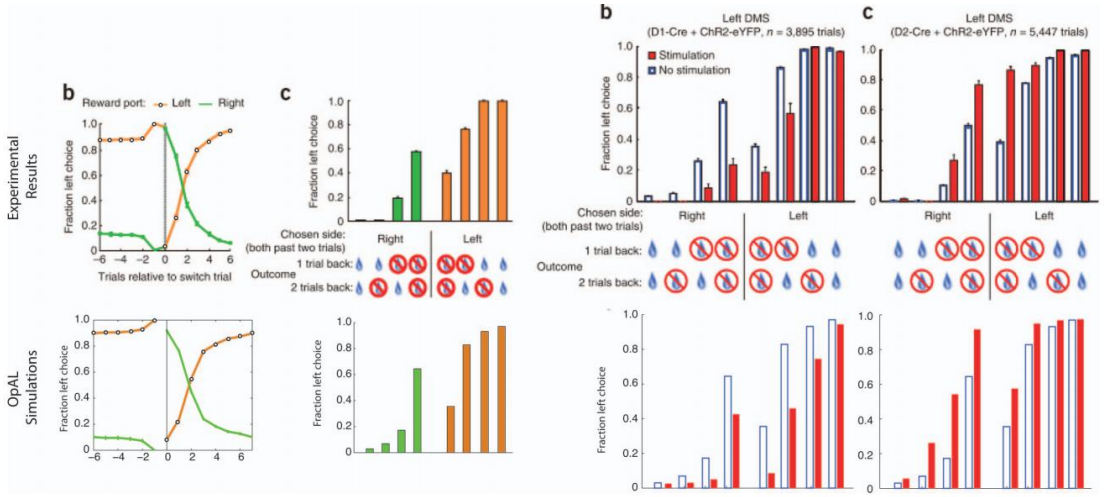

图 7. 选择时的光遗传学刺激。顶线：实验结果（摘自 Tai，Lee，Benavidez，Bonci 和 Wilbrecht，2012 年）。底线：对抗行为学习（OpAL）模拟。左：逆转学习表现。左中：最近两次试验中不同奖励历史的选择概率。中右：在最近两次试验中，对于不同的奖励历史，有或没有刺激左 D1-MSN 的左选择的比例。右：在最近两次试验中，有或没有刺激左 D2-MSN 奖励历史的左选择比例。DMS=

背侧纹状体；MSN=中棘神经元。D1（2）-Cre+Chr2-eYFP 分别指示为直接（间接）途径 MSN 的遗传控制而操纵的小鼠组。L.-H 改编自“纹状体神经元的不同亚群的短暂刺激模拟动作值变化”的上排图。Tai，A.M.Lee，N.Benavidez，A.Bonci 和 L.Wilbrecht，2012 年，自然神经科学，第 15 期，第 1282-1283 页。Macmillan 版权所有 2012。有关该图的彩色版本，请参见在线文章。

为了检验学习和激励的影响，而又不对不同选择的经验量的影响(差异抽样)造成影响，我们首先研究了学习阶段的随机动作选择策略，并在随后的转移中评估所有选择选项之间的偏好相。我们还研究了在使用 softmax 的学习过程中更标准的动作选择策略。

我们首先考虑了对积极和消极结果的敏感性偏见(如在此任务的测试阶段中的 ChA 和 AvB 选择比例所揭示的)如何随学习阶段参与者学习率 aG 和 aN 或测试阶段选择激励参数的变化而变化 β_G 或 β_N。模拟(图 8 的左图)显示，在 $\alpha_G > \alpha_N$ 或 $\beta_G > \beta_N$ 的情况下，操纵学习率之间或测试的偏向于更好地选择 A 而不是避开 B(对于反向不对称而言，则相反)。随着 A 和 B 的奖励概率分别增加/降低，这些偏见被放大。

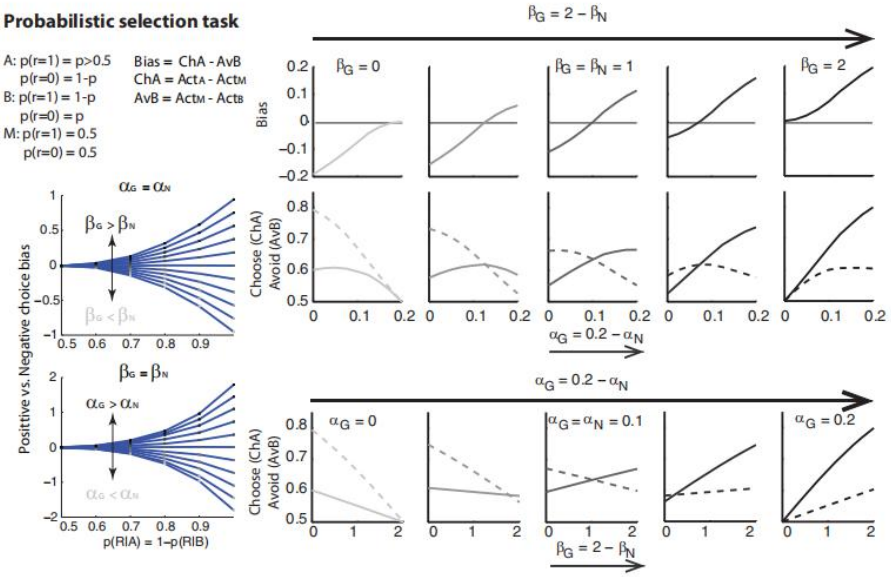

图 8.简化的概率选择任务：对抗行为学习（OpAL）中的相对值。所有值均为 100 次试验后的最终值，是对 1,000 次模拟的平均值。左：对于奖励 p（r）的不同概率的选择偏见。左上角：固定 $\alpha_G = \alpha_N$，改变 versusG 与 β_N 不对称性。左下：固定 $\beta_G = \beta_N$，改变 aG 与 aN 的不对称性。学习或激励（绩效）效果都会产生偏向于正或负值的选择偏见。随着最有回报/最有惩罚意义的结果越来越具有确定性，这种偏见也随之增加。右图：a 和β参数中不对称性的选择-选择 A（实线）与避免 B（虚线）和偏差（相对差）。偏差在任一参数类型中均随着不对称性单调增加，

但两个参数的影响都会相互作用：假设学习中的不对称性（a），当激励中的不对称性（þ）处于相同方向时，即在多巴胺能的激励状态下，性能最佳。选择的时间与学习时的时间相似。即使这样，对于中等程度的不对称性，也有可能在一个系统中表现出更多的学习，而在选择过程中表现出对另一个系统更大的影响，就像在某些实验中一样（Zhang，Berridge，Tindell，Smith和Aldridge，2009）。图顶行中的水平黑线显示了模型版本的仿真，该模型在演员权重更新中没有Hebbian非线性项：该模型无法说明选择A或避免使用B的差分灵敏度，这可以通过跨零偏差来看出所有参数，由于在G和N权重中对称表示正值和负值。有关该图的彩色版本，请参见在线文章。

接下来，我们通过同时改变 a 和þ不对称性，将奖励概率固定为标准经验任务的概率(p(r|A)=0.8)，以研究学习和绩效参数之间的相互作用。我们固定了 aG+aN=0.2，$\beta_G + \beta_N = 2$，但通过参数改变了它们的差。模拟(图 8 的右图)显示，随着þG-þN 和 aG-aN 的增加，偏向 ChooseA 的行为对避免 B 的偏向。参数的交互作用在整体性能上也很明显：当一组参数中的任意一个在 G 和 N 上达到平衡时，另一个参数的不对称影响较小(中间列)。但是，当其中一个参数强烈不对称时(最左边和最右边的列)，如果另一个参数在同一方向上不对称，则总体性能会提高，但如果在另一个方向上不对称，则总体性能会下降。换句话说，表现取决于选择时的动机/多巴胺能状态是否与学习时相同。但是，对于更适度的学习偏见，也可以逆转选择中的不对称性：当学习期权时偏向 N 个权重时，对 G 权重的选择动机中足够大的不对称性仍会导致 ChA 优于 AvB 性能。这些发现支持了这样的观点，即选择时的动机状态会影响动物的行为，从而使其倾向于采取与学习过程中的负面结果相关的动作(Zhangetal。，2009)。

请注意，这些结果需要我们为 OpAL 引入的特定非线性更新规则(包括三因子 Hebbian 项)。实际上，图 8 中的灰色曲线显示了简化模型的仿真，该简化模型通过更新方程中的 G 或 N 值去除了乘法调制。这些模拟表明，学习率或þ参数的不对称性的任何组合均未观察到偏差。如果没有乘法更新，则 G 和 N 权重会对称地发展，在正值或负值之间不会出现优先区分，因此是真实期望值的线性组合，从而导致均等的 Choice-A 和避免-B 性能(请参阅附录，图 A2 中的补充仿真)。

概率选择任务中的多巴胺能操纵已被反复证明(Frank，Moustafa 等，2007;Frank ＆O'Reilly，2006;Frank 等，2004;Jocham 等，2011;Shiner 等，2012)；Smittenaar 等人，2012 年)，虽然在学习选择或绩效影响方面尚无明确的区分，但可以诱导选择 A 偏向与避免 B 偏向的变化。如前所述，一些研究表明，即使设计使药物仅影响测试

性能而不是学习，多巴胺药物的效果也得到改善，Choose-A 性能得到了改善(Shiner 等人，2012；Smittenaar 等人，2012)。但是，在学习和测试过程中都对多巴胺进行调节的其他实验，对选择不对称的影响要比对单独测试的影响更大。此外，影像学研究表明，多巴胺能操纵对 Choice-A 性能的影响与其在学习过程中增强奖励预测错误信号的程度有关(Jocham 等人，2011；Ott，Ullsperger，Jocham，Neumann 和 Klein，2011 年)。以上模拟表明，学习动机动机和动机动机效应均可单独或共同解释结果。正如遗传研究中所观察到的，它们也与多巴胺调节作用应该是参数化这一事实是一致的(Frank，Moustafa，etal，2007)。

概率选择任务中的选择偏见通常采用经典的强化学习模型来建模，该模型包括针对正错误和负错误的不对称学习率(Doll 等，2011；Frank，Moustafa 等，2007)。我们在附录(图 A1)和讨论中的进一步仿真中表明，尽管此类模型确实可以解决一些偏差影响，但它们不能解决 OpAL 可以处理的各种数据。

因此，我们的模型可以解释在学习阶段和随后的执行阶段中，多巴胺在概率选择任务中先前观察到的影响。但是，它还会在实验设计中做出额外的预测，在该设计中，在学习阶段和测试阶段会分别对多巴胺进行操作。如图 8 所示，OpAL 预测，如果同时调整学习率和β参数，则偏见会被放大。而且，它预测，如果在不同的多巴胺能状态下进行学习和测试阶段，总体性能将急剧下降。例如，在学习过程中多巴胺水平较高(强 aG>aN 不对称)但在测试过程中多巴胺水平较低(强 $\beta_G < \beta_N$ 不对称；请参见图 8，右组，左中图)，可以预测接近机会的表现。

5.5、基于努力的决策的动机和激励作用

多巴胺对动机激励的最清楚的例子可能来自于操纵动物(或人类)为获得报酬而必须付出的努力的任务(Cousins＆Salamone，1994；Floresco，Tse，＆Ghods-Sharifi，2008；Courins＆Salamone，1994)。Salamone 等，2005；Treadway 等，2012)。在这里，我们考虑人类或动物需要多次按下单个杠杆以获得奖励的范例。在此类研究中，多巴胺调节强烈影响所施加的努力程度，以致动物为获得更高的潜在报酬而更加努力工作的趋势与纹状体 DA 成正比：DA 升高会增强纹状体，DA 耗尽会抑制纹状体 DA。这些发现不容易通过学习理论来解释，因为在动物或人类学会了努力成本和奖励利益突发事件之后进行了操纵。尽管如此，该程序主要反映了研究基于努力的决策的传统，并且在学习过程中没有原理性原因无法进行多巴胺操纵。因此，我们考虑了学习和动机激励的潜在作用及其相互作用。

在第一个模拟中，我们根据获得的奖励参数化地改变了获得奖励所需的行动数量，该奖励与单个选择相关联。我们使用了固定的学习时间(100 次杠杆按压)来获取偶发事件和对称学习率，但是在选择过程中改变了þG 和þN 之间的平衡 p(请参阅附录中的补充方法)。图 9 表明，增加 p，即，G>þN，会导致选择选项的可能性更大。这种效应与获得奖励所需的努力量相互影响：对于高努力(例如 p(r)=0.1)，一旦 þG>þN 有任何偏差(绿色曲线)，进一步改变的效果就变得非常明显。而对于相反的不对称性，p 的相同变化几乎没有影响。相反，对于低工作量(例如，p(r)=0.9)，鉴于相反的不对称性þN>þG(红色曲线)，将 p 改变相同量的影响要大得多。[4]

在第二组模拟中，我们根据对 D2 受体和腺苷 A2A 受体(共同定位在同一 NoGo 神经元上)直接对多巴胺的操纵，可以预测地调节 D2 途径的成本。具体来说，D2 阻滞有效地增加了工作成本，而 A2A 阻滞通过对神经元兴奋性产生相反的作用来抵消这种作用(Farrar 等，2010，2008；Mingote 等，2008；Nunes 等，2010)。。我们研究了在保持 G 参数固定(aG=0.1，G=1)的同时，对学习努力成本(变化 α_N =0.1 或 0.125)和此成本表达(β_N =1 或 1.5)的潜在影响。模拟显示，在学习过程中(灰色 vs. 黑色线条)或在表演过程中(绿色圆圈 vs.红色方块)，D2 封锁会降低该选项的有效 actorweightAct，从而降低选择该角色所需的工作量选项。值得注意的是，这些 D2 效果随着选件成本(所需压机数量)的增加而增加。

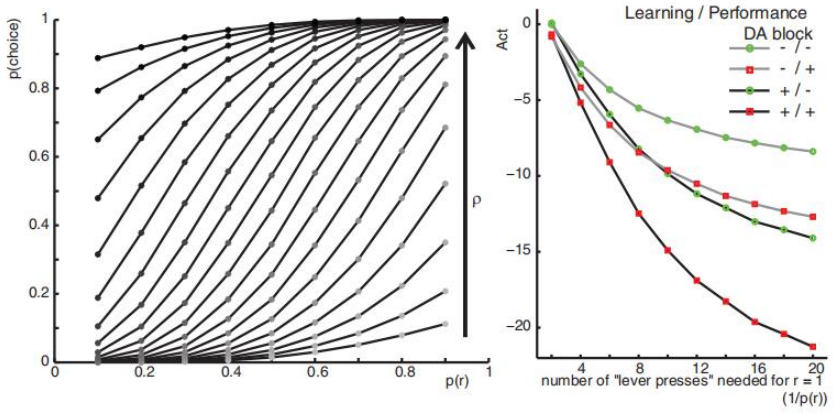

图 9. 工作任务：在单个强制选择杆按压情况下操纵 β 不对称。所有值都是在压下 100 次杠杆后的最终值，是对 1,000 次模拟的平均值。p 是归一化差 $\beta_G - \beta_N$ 。左： $\beta_G + \beta_N$ 对 G 系统的不对称性增加，增强了人们参与工作的意愿。较深的颜色表示相对于 G 与 N 的不对称性（p<0）。右图：固定了 aG 和þG 后，在学习（较高的 aN）或随后的选择（较高的þN）或两者期间模拟 D2 受体阻滞。灰色与黑色线表示学习期间对照与药物，绿色圆圈与红色方块表示在学习过程中对

照与药物。两种操作都减少了工作量，特别是在组合使用（+/+）时，这些效果随着工作成本的增加而放大。付出高昂的努力成本，单独的学习效果（+/−）大于单独的效果（−/+）；这种模式可以降低人工成本。DA=多巴胺。有关该图的彩色版本，请参见在线文章。

这些模拟还提供了一种新颖的，可测试的预测。除了努力成本对行动参与的主要影响以及学习和激励参数的影响外，我们还观察到每对因素之间的相互作用以及三向相互作用(所有 ps<.01)。D2 操纵的学习和性能影响越强，则需要付出更多的努力(如曲线之间距离的增加所证明)，并且在学习和选择过程中两种操纵的组合会放大效果。此外，将这两种效果单独进行对比，当努力成本相对较低时，选择激励的效果要强于学习效果，但是对于高昂的成本，这可以逆转(比较 2-4 的-/+和+/-条件)vs.16–20 杠杆压力)。这是因为，由于 N 权重对更新的多重影响，随着时间的推移，负面奖励预测错误(对于大多数印刷机缺乏奖励)的频率会随着时间累积，从而产生更强的学习效果。因此，三重相互作用是 OpAL 模型的特定预测，没有 Hebbian 术语就无法预测(参见附录，图 A2)。

除了提供新颖的，可检验的预测之外，我们还更直接地表明，OpAL 框架可以通过在共同努力协议上进行仿真来说明现有数据，包括那些在获得和不获得多巴胺的情况下操纵获得奖励的杠杆按压次数的协议(Aberman＆Salam-one，1999；Niv 等人，2007)，以及那些增加障碍的老鼠必须要爬过去才能在 T 型迷宫中获得更大奖励的障碍。在这两种情况下，模型均以 $\beta_G = \beta_N$ 和 $\alpha_G = \alpha_N$ 为模型学习完好状态下的选择偶然性，然后在完好状态或多巴胺阻滞下(在参数不对称的情况下)进行消光测试(无学习) $\beta_G < \beta_N$)。

具体而言，向杆按压努力任务的环境模型，我们假设一个单一的状态和选择，即选择或不根据等式 6。当选择发生时，反馈被建模与伯努利概率 P(R)=1/T.我们允许模型学习这些意外情况，并在此学习过程中通过行为学习率 a，a 来控制多巴胺效应。然后，我们通过调整 β_G，N 参数来控制学习后的动机激励。

为了将其专门应用于来自 Aberman 和 Salamone(1999)的数据，我们使用以下参数：β=10，α=0.02，p=0，$p_{DAblock}$=-0.9。该模型学会了在两个选项之间进行选择：按下操纵杆或不执行任何操作。如果选择压力，则奖励比率 r=1 为 1/FR，固定比率 FR={1，4，16，64}。将反应时间建模为 RT=0.5+1.5/(1+exp(Act))。获得奖励后，我们假设进食时间固定(τ_I=6)。模拟进行了两个 30 分钟的等效训练，然后从第三次训

练后的训练中绘制了选择结果(根据在这些范例中进行表演之前对动物进行了广泛训练的事实)。

因此，在杠杆压制任务中，OpAL 仿真重现了通常在固定比率计划中观察到的基本模式：随着计划表需求的增加，杠杆压制也会增加，但是随着工作量的增加，多巴胺阻滞会优先减少杠杆压制(图 10A)。该模型表明，这种模式是由于以下事实导致的：努力状态与更高的成本相关联，而更高的成本则与奖励利益权衡了，在奖励利益中，由于多巴胺受体阻滞而夸大了成本。

在 T 型迷宫任务中，动物学会选择提供最多食物颗粒的手臂(4 对 0，或 4 对 2)。完整无缺的动物继续选择带有四个药丸的手臂，即使添加了障碍物也不得不爬过它。但是，多巴胺阻滞使他们在 4 比 2 的情况下停止选择最有帮助的手臂，而在 4 比 2 的情况下则没有选择(Cousins，Atherton，Turner 和 Salamone，1996；图 10B)。多巴胺阻滞保留了 4 比 0 情况下的高强度选项选择的观察结果表明，强度值与奖励值的编码不同(即，动物愿意时可以越过障碍)，但是它们表现良好成本效益分析，即将攀登障碍的成本以某种货币与奖励的收益进行比较。

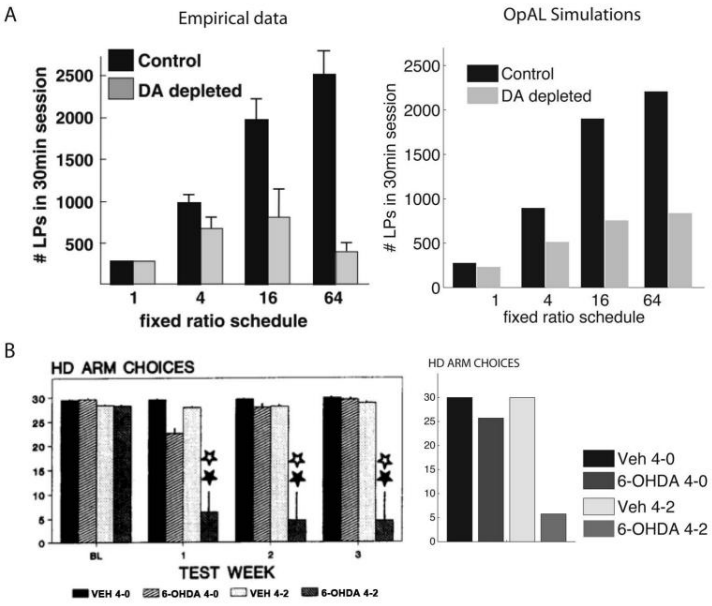

图 10.工作任务。A.固定比率杠杆按下任务。左:杠杆操纵任务的数据(来自 Aberman＆Salamone，1999 年）。右：任务的对抗行为学习（OpAL）模拟。如数据中所示，杠杆压力机的数量随着固定比率计划的增加而增加，但是多巴胺消耗随着努力需求的增加而减少了杠杆压力机的数量。B.具有障碍任务的 T 型迷宫：来自 Cousins，Atherton，Turner 和 Salamone（1996）的实验数据

在左侧，模型仿真在右侧。尽管付出了更多的努力，但健康的啮齿动物和完好无损的模型更喜欢获得最大回报的手臂。多巴胺的消耗逆转了这种偏好，但仅适用于 4 粒对 2 粒的情况，而不会影响 4 粒对 0 粒的情况。LP=压杆；DA=多巴胺；HD=高密度食物；BL=基线；Veh=注入媒介物的组；6-OHDA=注射 6-羟基多巴胺的组。全星表示与 Veh 条件的显着差异，空心星表示 4-2 和 4-0 条件之间的显着差异。误差棒是平均值的标准误差。面板 A 的左图转载自 Y.Niv，N。D.Daw，D。Joel 和 P.Dayan，2007 年，“心理药理学”，第 191 页，“补品多巴胺：机会成本和反应活力的控制”。512.Springer 版权所有 2007。M.Cousins，A。Atherton，L.Turner 和 J.Salamone，1996 年，行为脑研究，M。Cousins，A。Atherton，L.Turner 和 J.Salamone，1996 年，B 板的左图摘自“核糖核酸多巴胺的消耗改变了 T 迷宫成本/收益任务中的相对响应分配”，74，p。192.Elsevier 版权所有 1996。

我们在 Cousins 等人中以障碍任务为 T 型迷宫建模。(1996 年)假设一个单一状态和两个可能的动作(左臂，右臂)。左臂选择确定性地提供 r=4 个药丸，而右臂选择确定性地提供 r=0 或 2。该模型接受了 100 次试验训练，并学会了稳健地选择左臂。为了模拟障碍物的工作，我们在必须选择动作[攀爬障碍物]的环境中分别训练了模型，并假设此动作导致 100 次试验的成本 c=-1。因此，模型为该动作确定了 G 和 N 权重。然后在带有障碍物的 T 型迷宫的组合上测试了该模型：动物必须攀爬左臂的障碍物才能获得颗粒。我们假设选择是在{左臂和障碍}与{右臂}之间进行选择，以便为左选择考虑的行为权重是学习到的左臂和障碍权重的总和。OpAL 参数为==0.1，þ=3.5，p=0。用 p=-0.55 模拟多巴胺阻滞，导致 toG<þN。结果平均超过 1,000 次模拟。图 10B(右)显示 OpAL 仿真可以定量地重现行为模式。

现在我们已经表明，该模型解释了多巴胺对学习和动机的影响的两个主要类别。迄今为止，强化学习实验中对收益与损失的差异敏感性主要归因于模型可以捕获的学习效果。但是该模型还表明，选择激励/绩效效应也可以促进这些发现。对称地，多巴胺对基于努力的决策的影响已在动机激励模型(绩效效应)的背景下进行了研究，我们的模型可以捕捉到该模型，但表明如果在此期间多巴胺被操纵，学习效果也会有所贡献。

这两组模拟都进一步暗示了学习和表现的潜在交互作用，这是新的预测。为了更具体地将这些预测与经验数据联系起来，我们接下来转向运动技能学习实验，这些实验令人信服地显示了这些相互作用。

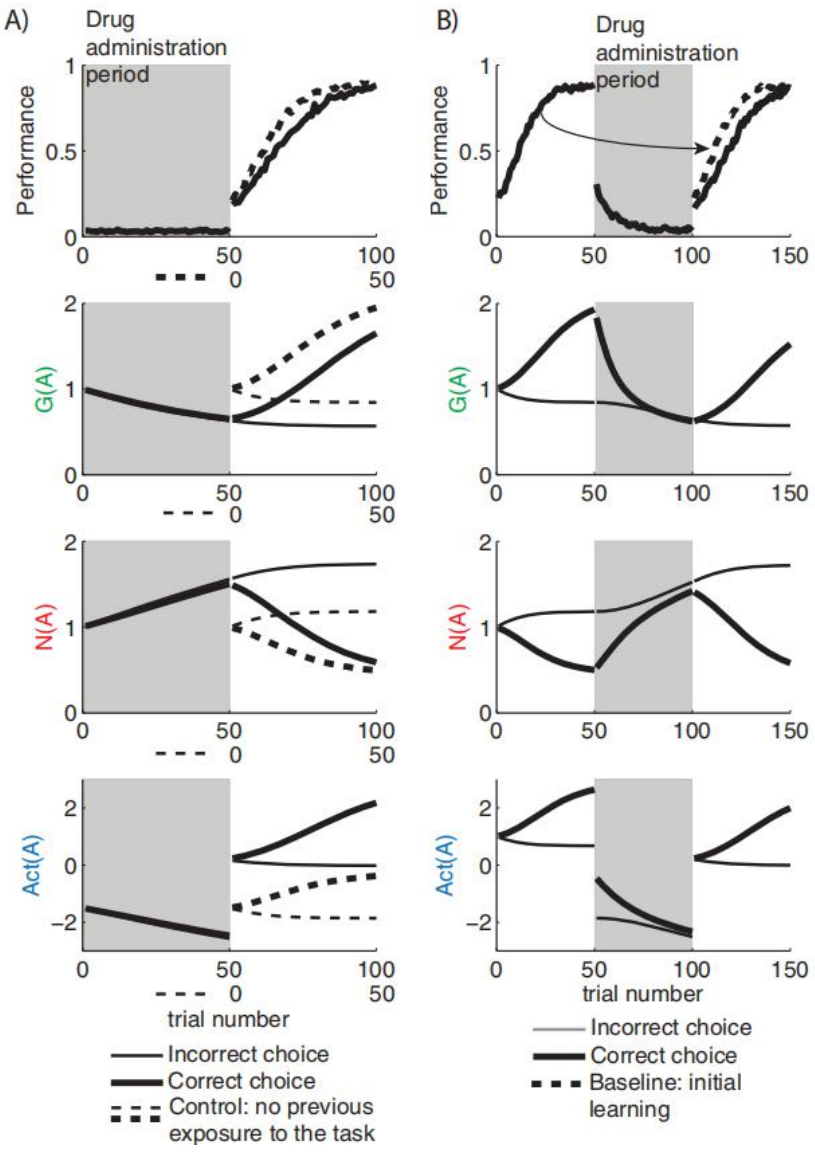

图 11. Rotarod 任务模拟。通过设置 p=−0.75 而不是+0.5 来建模多巴胺（DA）对性能的影响，并固定所有其他参数。药物给药的时间由灰色背景指示。A. 最高：药物（未完整）（实线）或完整对照（未事先暴露于任务中）随时间推移的性能（奖励试验的比例）。底部：G 和 N 权重和参与者值充当模型参数。在第一次用药物介绍任务期间，执行会受到阻碍，并导致避免学习正确和不正确的行为。如实证研究所示，与对照组相比，这在移除药物后引起了较慢的学习。B. 在第二组模拟中，在了解任务后才对药物进行管理，从而导致性能迅速下降，并随着时间的推移而进一步降低。在第三阶段，没有药物，与最初的学习相比，重新学习的速度再次明显减慢（虚线）。有关该图的彩色版本，请参见在线文章。

5.6、学习和表现互动对运动技能的影响

纹状体多巴胺长期以来一直与运动表现有关，但直到最近才意识到其在学习中的作用，尤其是 DA 耗竭对异常学习的影响。在 Beeler 等人中(2012 年)，作者对执行加速轮转任务的啮齿动物进行了多巴胺阻滞，这是一项运动技能学习任务，将啮齿动物放在以加速旋转的杆上：动物整合了视觉和本体感受反馈，以正确的方式向前行走避免跌落的速度。他们显示了对学习和表现的互动影响，这些被基础神经节的神经网络模型捕获。特别是，多巴胺在初次暴露于旋转脚架期间或在得知该任务后会被阻断，从而导致性能很差。从表面上看，这些发现可能归因于性能不足，但进一步的结果表明，它也引起了异常的学习。的确，在药物冲洗后，学习速度明显比幼稚动物慢。同样，在完好无损地学习了任务之后，D2 封锁尤其导致技能表现的逐步下降。突触可塑性研究表明，D2 阻断可诱导皮质口突触增强到 D2MSN 上。已经提出了这些相同的机制，以解释在僵直性实验中反复给予低剂量 D2 阻滞剂会导致帕金森病症状的逐步发展(Wiecki 等，2009)。我们在这里测试了我们的简化版本的神经网络强化学习机制是否可以解释对运动技能学习和表现的影响。

我们的 OpAL 仿真做出了与以前的神经网络仿真类似的假设，为该任务建模：我们假定了四个状态和四个运动动作(简单地说，对应于要移动的爪子)。此外，需要足够迅速地采取正确的措施，否则动物会掉下来。因此，我们假设正确的行动选择会导致奖励(r=1)，其概率取决于反应时间：

$$p(r \mid correct) = 0.1 + .8/(1 + \exp(\beta_G * G(s,a) - \beta_N * N(s,a))),$$

但否则惩罚(r=0)；而错误的行为总会导致惩罚(r=0)。使用的参数为 β =3，$a_{G/N}$=0.1，a_C=0.05，p_{drug}=-0.75，$p_{control}$=0.5。

图 11 显示了 rotarod 任务的主要仿真结果。通过设置þG<þN(未处理条件为þG>þN)，在所有其他参数保持固定的情况下(尤其是 aG=aN)，可以模拟 DA 封锁的效果。在第一个模拟中(图 11A)，DA 第一次遇到任务(有色部分)时会受到封锁，从而加剧了 NoGo 活动，导致较弱的 ActorWeightsAct(右下)，从而减慢了动作选择，即使对于原本正确的动作也是如此。因此，这种绩效效应意味着即使是正确的行动也很少得到回报，不仅导致学习不足(绩效没有增加，上图)，而且导致学习异常：G 权重降低(左下)和 N 增加重物(底部中间)，即使采取正确的措施也是如此。随后通过暴露于完整的多巴胺能状态(白色区域)的任务可以揭示这一点。然后学习继续进

行，该模型可以正确学习 G 和 N 权重以进行正确和不正确的操作(黑色和灰色实线)，但比幼稚的情况(虚线)要慢。

在第二组模拟中(图 11B)，模型首先正常学习任务。在技能建立后暴露于 DA 阻滞会导致性能迅速下降，但也会导致学习异常：正确和错误动作的 G 权重降低，N 权重增高，从而导致性能逐步下降。出于与先前实验相同的原因，这再次使得在没有封锁的情况下进行后续任务的重新学习变慢。

因此，我们表明，OpAL 模型可以说明轮转任务中学习和绩效的互动影响，包括由于多巴胺阻滞的绩效影响而导致的异常学习。这种结果模式再次取决于 OpAL 中是否具有乘法更新规则。

5.7、指令偏差

最后，我们考虑了绩效与学习之间的更高层次的认知互动：自上而下，规则指导的人类教学如何影响绩效和偏向学习。我们模拟了指示性概率选择任务，这是概率选择任务的一种变体，其中六个刺激中的一个(正确或错误)在学习任务偶然性之前已显示给对象，以暗示该选项可能是一个不错的选择。

实验结果(Doll 等人，2009 年)表明，受试者最初选择了该指导选项，但是当指导产生误导时，他们最终学会在训练阶段避免使用该指导选项。然而，转移阶段的选择表明，相对于相同目标值的非指示性选择，该指示性刺激的学习价值有所提高。实际上，模型拟合表明，确认偏倚可以解释发现，在训练阶段，与指导相一致的结果会被放大，不一致的结果会被打折，从而导致客观价值的膨胀。

遗传结果(Doll 等人，2011 年)表明，DARPP32 多态性(影响 D1 和 D2 途径相反方向的多巴胺能遗传变异)与 Choice-A 的相对不对称性(与回避 B 的性能)有关。如先前所报道的，而与 D2 受体功能有关的 DRD2 多态性与避免-B 的表现有关。而且，在指示版本中，这些遗传变量分别预测了扩大指令一致结果的值并消除阴性结果的趋势。从行为上讲，这意味着在适当时选择更好的选择刺激的能力(选择 I)，而在与更有价值的选择配对时，避免使用它的能力更差(避免 I)。因此，与基本无指导的强化学习有关的基因也可以预测学习受到确认偏差的程度，这表明指导性实验中的这种偏差并非来自单独的机制(例如更高水平的策略)，而是来自于那些相同的基本 RL 机制的调制。因此，我们测试了 OpAL 模型，以查看其纯粹的强化学习机制是否可以解释这组数据。

在这里，我们试图通过模拟有关给定刺激良好的初始指令来解释 OpAL 中的这些发现，方法是简单地增加其初始 G 权重并以固定值(0.3)降低其初始 N 权重。仿真表明，OpAL 模型可以解释这一系列结果，这表明 DARPP-32 调节了学习的不对称性，而 DRD2 调节了选择动机。首先，图 12 显示该模型从错误的指令刺激中学习，而没有完全克服初始偏差(对于 OpALAct 值，将蓝绿色指示为未经指导的灰色)。更一般而言，如从实验中观察到的，中间的上图显示渐进行为的总体增强，而该刺激已被指示为良好。仿真还可以重现测试性能“选择 I”与“避免 I”，在指示良好和误导性较高的情况下(灰色条形)，选择刺激的性能要优于统计上较差的刺激。相反，当被指示为良好时，受试者很难选择一个刺激而不是统计学上更好的刺激，但是当该指示被误导时，受试者就很难避免选择刺激。

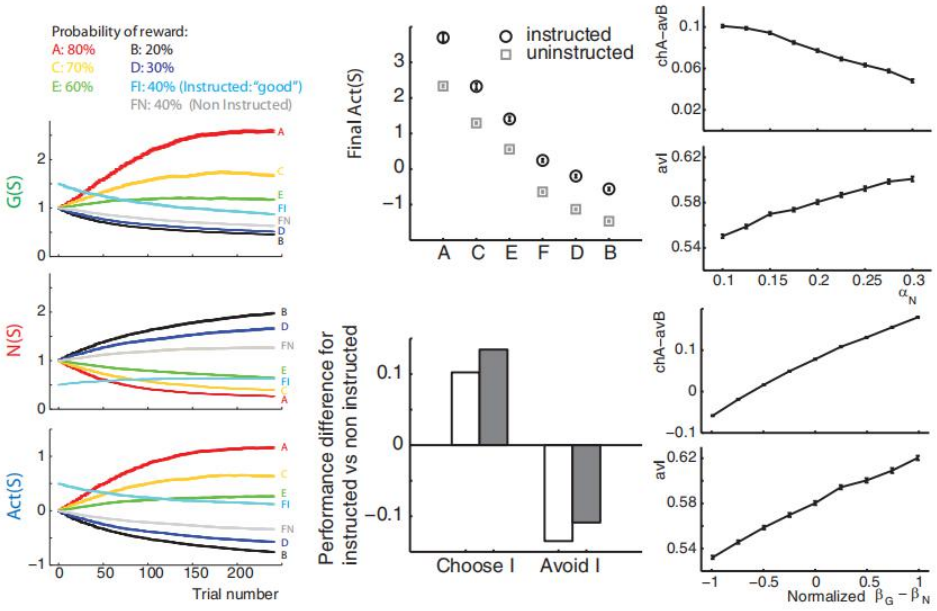

图 12. 指导的概率选择任务。图的左列：随时间变化的不同刺激的模型值：G 和 N 权重，Actor 值 Act。标记为 FI 的行（图中在线版本中的浅蓝色）提供了一个误导性 F 刺激（客观奖励概率为 40%）的示例，与标记为 FN（灰色）的行相比，对于未经指导的 F 效果很好。与图 2 相比，G 中角色权重的不对称性要强于 N。这是因为采样是不相等的：随着时间的流逝，模型学会了更多地选择 A，C 和 E，从而了解了更多有关它们的信息。由于这些是正价选择，因此在 G 权重中更明显。请注意，Act（F）最终低于 Act（E），这表明在训练期间未学习指令偏差，但是先前的初始化仍然存在，因为它仍然无法赶上其未指导的版本值。中上层：指示性刺激（黑色圆圈）和非指示性刺激（灰色方块）的最终行为值。误差棒表示平均值的标准误差。中下：指令对测试阶段的影响，显示对于准确（白色）和不准确（灰色）指令而言，指令与非指令刺激的选择

相对增加（较高的 Choice-I 和较低的避免-I）。图的右列：参数基因效应：与避免-B（avB）相比，增加 aN 导致相对较差的选择-A（ChA），但同时提高了避免-I（avI）的性能，如 DARPP-32 多态性所观察到的。the 不对称性的增加（þN<þG）会降低避免-B 的性能，但会增加避免-I 的能力，正如 DRD2 多态性观察到的那样。有关该图的彩色版本，请参见在线文章。

最后，模拟可以重现遗传效应(请参见右图 12)。首先，与 aG 相比，学习参数 aN 的增加导致 No-Go 学习效率的提高，从而导致 Choice-A 与避免 B 偏向的降低。通过提供对错误指示的刺激的更好的学习，这也导致增加的回避性能。因此，DARPP-32 效应可通过最初解释的学习参数不对称性变化来解释(Doll 等人，2011)。相比之下，DRD2 效应不是由学习不对称性引起的，而是由选择激励效应(þ不对称性)引起的。如上所述，当þG>þN 时，回避 B 的性能会降低，但与直觉相反，这也会提高回避 I 的性能。这种模式的原因是，由于先前的指令效果，指令刺激 I 产生了更强的 G 权重，而不是 N 权重，因此，将其与更高价值的 G 权重区分开是取决于þG 相对较强的影响。这就解释了 DRD2 基因如何在一个方向上调节对非指导性阴性结果的敏感性，而在相反方向上调节对指导性刺激的敏感性。请注意，在学习效果方面，此说明与 DRD2 角色的原始解释不同。确实，我们在这里建议用选择激励效应来更好地解释这一点，这既考虑了我们观察到的标准效果又说明了教学效果。先前记录的 DRD2 对非指导性回避的影响是否归因于动机而不是学习效果，还有待观察。

5.8、为什么要有两个系统：规范分析

我们已经表明，OpAL 可以解释与强化学习和决策制定及其多巴胺能影响有关的广泛实验数据模式。这些模拟至关重要地依赖于 OpAL 的结构作为双重表示机制，其中 G 和 N 编码用于强反相关值。直接和间接途径中信息之间的这种相对冗余产生了一个问题，即为什么对负相关价值估算进行编码的两个系统是必要的还是有益的。直观地讲，该系统提供了更大的灵活性，从而使所学习的预期奖励或成本之间的一种强调区分可以服从其当前的动机状态，即选择时的多巴胺水平。未来的工作将研究如何根据其他变量来优化此状态。

在这里，我们着重研究 OpAL 的能力，即使学习或选择参数没有任何不对称性，也可以学习概率偶然事件并将其性能与标准 RL 算法进行比较(下面列出的结果适用于标准参与者评论和 Q 学习)。为模型提供了两对选项供您选择，总体加固计划既

丰富(奖励的概率 r=1 对比 0，分别为 0.8 和 0.7)或瘦(0.3 和 0.2)。随着时间的流逝，模型应该学会选择每对(0.8 和 0.3)的客观最优选择。

我们优化了模型参数，以获得最佳平均性能，从而在 10,000 个模拟中选择 50 项学习试验中的最佳选项(请参阅附录中的补充方法)。如上所述，我们用 aG=aN 和 þG=þN 约束对称性，以研究两个系统的特定作用，即使它们之间没有失衡。用优化的参数进行的仿真表明，OpAL 的平均性能要优于 RL 模型(参数也已优化)，这对于学习精益选项尤为正确(图 13)。后续分析表明，与 RL 相比，在 RL 中学习精益选项的速度较慢

附录中介绍了针对 RL 的此问题的数学推导。凭直觉，在 20/30 判别中，模型一旦开始利用 30 选项，就无法学习较差的那个值(20)的真实值，因此其估计值仍接近初始值(0.5)，因此接近 30 选项。同样，对于 70/80 选项，一旦模型利用了 80 选项，则其对 70 选项的估计值仍保持接近 0.5，但是在这种情况下，这是有帮助的，因为利用值和未利用值之间的有效差异更大。因此，对于相同的 softmaxþ参数，RL 模型不会在 30 和 20 之间以及在 70 和 80 之间进行区分。此外，仅增加þ参数无济于事(实际上已对其进行了优化)，因为过高价值观阻碍了探索以获取替代方案的真正偶然性。

OpAL 如何避免这个问题在富裕和精益选择偏好方面表现出色？至关重要的是，OpAL 在选择功能中同时包括了 G 和 N 权重。因此，最初，在任一组权重积累足够的值以支配另一组之前，它们的贡献相对相等。在 20/30 的情况下，学习的 G 值之间的差异会被忽略(例如，图 2)，因此选择功能的这一部分有效地增加了探索性。但是，一旦 N 个权重充分积累，它们就占主导地位，因此 G 个权重的贡献可以忽略不计，并且该模型可以有效地利用 30 个选项。该功能意味着，随着系统学会适当地偏重 G 或 N 权重，动态地调节了探索。此功能至关重要地依赖于值的非线性表示形式(请参见附录中的补充结果)。此外，简单地将简单 RL 中的þ参数作为时间的函数进行动态更改的简单方法无法实现相同的目标。因此，该分析表明，通过包含用于检测有助于选择的适当行为权重的机制，在 OpAL 中使用具有非线性更新规则的对偶系统可在偏好学习中提供更好的性能。

从经验上讲，尽管我们认为神经生物学与标准的 RL 模型相比更符合 OpAL，但这种规范分析提出了人类行为是否符合 OpAL 的预测的问题。与 OpAL 相比，标准的 RL 模型预测出精益刺激和富裕刺激的准确性存在明显差异。尽管我们尚不知

道实验会测试上面模拟的精确设计，但 Pessiglione 等人还是采用了相关的富与精设计。(2006 年)，显示条件之间没有性能差异。然而，当使用标准 RL 模型和 Pessiglione 等人提供的最佳拟合参数来模拟该实验时。(2006 年)，我们获得了很强的性能不对称性，对于富裕(奖励更多)的货币对，又有了更好的表现，这与根据经验观察到的相反。相反，出于与描述相同的原因，使用适当的参数 OpAL 可以在两种情况下重现观察到的数据模式。因此，尽管有限，但可用数据表明 RL 模型与在人类受试者中观察到的学习曲线不一致，并且 OpAL 克服了这个问题。

6、讨论

我们提出了一种新的计算模型，以同时说明纹状体多巴胺的学习和激励/绩效效应。现有的模型主要集中在一个方面或另一个方面，但是在这里我们展示了如何需要这些功能及其相互作用的组合来解释大量数据。该模型基于具有一些关键新颖功能的行为评论体系结构。首先，它依靠独立于选择权的吸引力和厌恶性的对手行为权重，而选择策略则取决于这些信号之间的加权竞争。其次，这些行为权重通过包含一个可乘的 Hebbian 术语来模仿纹状体可塑性规则，从而导致值的非线性表示，并分别强调量表的吸引或厌恶部分。

将这些功能组合在一起可提供解决多种数据所需的灵活性。确实，由于非线性学习规则，G 和 N 权重区分了值表示形式的不同方面，它们分别存储，并根据选择期间多巴胺水平的变化与参数化权重进行动态重组。因此，该系统具有很大的灵活性，可以根据过去的学习信息在不同情况下作为动机状态的函数表达为一种选择。确实，我们证明了 OpAL 可以解决学习中的选择偏见，以及在学习过程中或选择时的多巴能操纵如何改变这些偏见。我们还表明，OpAL 可以解释性能和学习的复杂交互作用，捕捉轮转式运动技能学习任务中的经验发现，但也可以为基于奖励的决策提供新颖的预测：当在企业中进行选择时，选择激励作用更强。与学习过程中的多巴胺能状态相同。类似地，该模型预测，学习和激励作用都可能影响基于工作量的决策，这些影响会相互放大，并且对一个或另一个过程的相对影响取决于所需的总工作量(图 9)。

6.1、多巴胺在学习和选择中的多重计算作用

我们的新 OpAL 模型严重依赖完善的神经学数据：存在两个相对的，看似冗余的系统：D1 直接和 D2 间接途径(此处通过 G 和 N 权重建模)以及多巴胺在强化中的

作用-不仅要学习，而且要有选择和动机。推测为什么这种双重编码存在于大脑中以及它提供了什么计算优势已经成为人们的猜测。还不清楚为什么多巴胺似乎起着如此多的作用，在功能上编码诸如预测误差，动机显着性等信号。我们的模型没有为这些问题提供直接答案，但为进一步研究它们开辟了道路。通过允许灵活地探索不同环境中的行为以及进行规范分析，它为这些关键特征的潜在作用提供了一些线索。

特别是，我们的模拟研究了多巴胺对学习和选择的作用之间的相互作用。在概率选择任务中，我们表明，如果学习与测试之间的多巴胺能状态不同，则整体表现会受到严重损害。这可能提供一个线索，说明为什么显然使用相同的神经递质来调节激励和学习：如果有任何因素(疾病，发育等)导致多巴胺水平高或低，则会导致学习中的不对称性。G 权重与 N 权重之比，但也将允许选择功能优先依赖已积累最有用信息的权重。在学习和测试过程中，对手系统之间的平衡控制之间的耦合可能会优化选择。

我们的模型还为选择值的直接和间接路径中看似冗余的编码提出了新的理解。的确，我们的仿真表明，当必须从乐观的初始估计中学习时，它可以在偏好学习任务中实现更好的性能，从而在探索和利用不同选择之间产生冲突。因此，对手系统的存在迫使优先权暂时处于次优状态，但仍可以进行更好的长期估算：避免了奖励计划稀疏的环境在勘探与开发之间的权衡。我们还表明，对手系统代表了冗余信号，但是它们的表示在不同的价值域中更为精确：G 权重强调了良好选择之间的差异，而 N 权重则强调了不良选择之间的差异。从推测上讲，灵活地将重点放在这些信号中的任何一个上，可能会使系统根据选择时的多巴胺水平(β 权重)可能编码的激励状态来动态决定要投放更多库存的信息决定政策：成本重要还是收益？这是对手结构的潜在好处，需要在进一步的研究中加以探讨。

6.2、与神经网络模型的关系

OpAL 的构建是为了模仿在强化学习和基于奖励的决策文章中，皮质皮质电路的更复杂的动力学神经网络模型中嵌入的一些核心原理(例如 Beeler 等人，2012；Collins＆Frank，2013；Frank(2005 年；Wiecki 等人，2009 年)，方法是将这些受到神经生物学启发的神经网络模型简化为一种算法形式。这就提出了一个问题，即与此类模型相比，它提供了哪些进步：实际上，如此，它定性地说明了使用该神经网络模型模拟的相似结果数组。尽管如此，我们相信我们的工作提供了一些新的贡献。

尽管 OpAL 并未考虑多个基底神经节核，丘脑和皮质之间的神经动力学，但与神经网络版本相比，它具有一些优势，我们现在将对其进行详细介绍。

首先，先前的文章显示了我们在此处建模的某些任务(例如概率选择任务)的神经网络仿真，强调了多巴胺在学习中的作用。尽管有些文章通过提及多巴胺对模型纹状体中 Go 和 NoGo 活性的影响来暗示绩效对选择激励的影响，但该出版物中尚未对这种选择激励作用的可区分作用进行正式研究，但在特定出版物中运动技能学习和表现的案例(Beeler 等人，2012)，就像我们在旋转脚架模拟中所包含的一样。但是，我们的模型或任何其他现有模型从未报告或解释过多巴胺在基于奖励的任务(包括基于强化学习和基于努力的决策任务)中调节成本/收益选择激励中的作用。此外，神经网络研究中先前对多巴胺操作的模拟通常使用一组参数来模拟 DA 固定值的增加或减少。相反，OpAL 允许在整个参数范围内全面刻画效果，以学习和激励，以及它们如何相互作用。

从实用的角度来看，作为一种更简单的低参数算法模型，OpAL 为分析提供了强大的优势。它可以定量地拟合经验数据，并提供简单的理论分析：了解模型变量表示的信息非常简单，可以轻松且详尽地分析模型动力学对模型参数和任务环境的依赖性，可以确定在各种环境中优化其性能的机制。因此，尽管神经网络版本侧重于更详细的机制，但其高级功能并不透明。因此，当前模型可以更好地理解系统的各个组成部分如何相互影响，从而导致本报告中描述的新颖预测(以前尚未阐明)。例如，我们不仅评估这些调制对一个特定版本任务的影响(例如，具有固定奖励概率的概率选择任务)，还评估其对广义版本的影响，从而显示出有关标准结果应如何变化的新颖预测。任务概率的函数(图 8)。同样，该模型会根据操作的实验阶段来预测多巴胺操作的不同效果：尽管工作量通常是在了解到工作量后进行的是多巴胺操作，但是我们的模型预测，学习阶段的操作会与学习过程中的操作相互作用。性能阶段，并且它们的不同影响将取决于所需的工作水平(图 9)。

最后，由于 OpAL 确实比神经网络模型简单(参数少得多)，因此可以定量拟合提供的实验设计，该实验设计足够丰富，可以区分学习效果和动机效果。而且，这个更简单的模型提供了我们在本文中提供的规范分析的类型。

6.3、与现有模型的关系

据我们所知，其他计算模型无法解释此处提供的数据范围。特别是，经典的强化学习模型会跟踪真实的期望值，并根据期望值的相对差异进行选择，因此无法重

现实验中观察到的任何选择偏差。先前在学习中对不对称性进行建模的尝试使用经典的 RL 模型，而没有分离 G 和 N 系统，而是简单地假设了单个值系统内正负预测误差的学习率不同(例如 Doll 等，2011；Frank，Moustafa，等人，2007)。该模型能够解决由于学习引起的一些不对称现象，但是其效果却与直觉相反：更好的 Choice-A 性能与积极的预测错误导致的较低学习率相关联，而更好的避免-B 与较低的学习率相关联负面的预测误差。此外，对该模型的系统研究表明，虽然这些效果总体上成立，但它们是非线性的(请参见附录，图 A3C)。更重要的是，仅允许学习率的不对称性排除了在学习过程中差异表达偏好的可能性：由于该方法仍将所有信息整合为一个值，因此无法提供解决潜在激励绩效影响所需的灵活性(有关详细结果，请参见附录)。此外，当前模型与在学习和选择过程中运行的差分 G 和 N 系统的生物学和神经模型更加紧密地吻合。

McClure 等人的模型。(2003 年)在一个单一的框架中，包括多巴胺的强化学习和激励理论，方法是假设刺激呈现时的多巴胺信号对选择激励进行编码，对应于未来的期望值，并调制 softmax 选择函数中的增益。尽管该模型解释了在竞争领域中的一些激励效应，但它预测系统地选择较低价值的刺激(这将导致低收益)之间的性能降低。这与根据经验和在 OpAL 中在低多巴胺能状态下观察到的相反。

侧重于多巴胺文献的动机或激励部分的计算模型通常侧重于单个动作的努力或反应活力(Dayan，2012；Niv 等人，2007)，因此，并未解决多巴胺的潜在影响相对强调预期的收益或损失。同样，包括巴甫洛夫式向工具转移(PIT)效应的模型(例如，Huys 等人，2011 年)说明了刺激值对整体激励作用的影响，但并未预测对正负选择的不同影响价期权。Berridge 和 Zhang 的模型(Zhang 等，2009)试图通过 n 参数反映的单一激励机制来解释所有学习和激励效果。尽管此模型确实涵盖了广泛的数据，但 n 模型也没有解决低 DA 水平在避免(或区分)负价期权方面提高性能的趋势。他们的模型确实模拟了动物在选择充气时会通过不同程度地影响其动机状态来“想要”不希望被喜欢，也不会被记住的“喜欢”(即化价反转)的可能性。被编码为厌恶性的期权的有效价值。我们的模型也可以解释这些相同的影响，在学习中的不对称性可以通过选择激励的不对称性来逆转(图 A4)。此外，n 模型需要进行更改才能使用基于对数的转换，该转换仅适用于价数反转的情况，而 OpAL 模型则自然地捕获了此值而无需更改。

6.4、局限性

通过将来自 Frank(2005)的神经生物学启发的神经网络模型简化为一种算法形式，我们引入了一些限制。一个关键的限制是，我们已经完全集中于问题的行为(或纹状体背侧)部分，而忽略了评价者方面(腹侧纹状体)。多巴胺也应对评价者产生影响，并且超出本研究范围的更详细的模型应将其纳入其中。但是，我们相信对这方面的修改不会显着改变此处介绍的定性结果。确实，我们模拟了模型的评价者部分的不同版本(例如，包括非对称增益/损失学习率)，并且在数量上不同的情况下，获得了定性相似的结果。

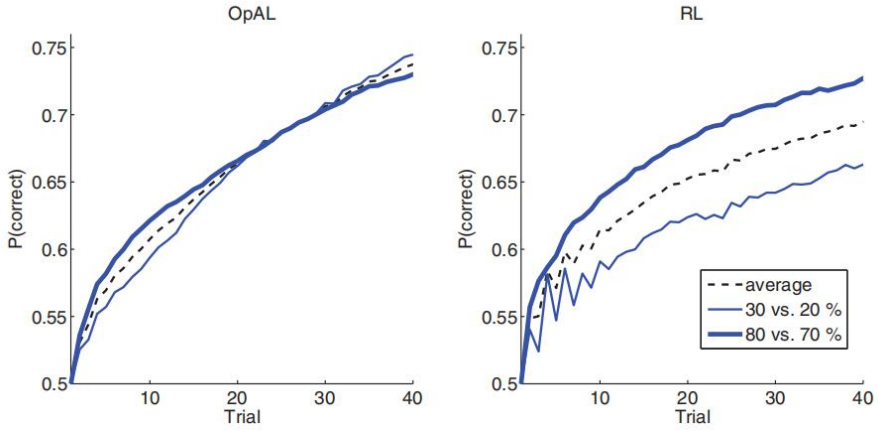

图 13. 偏好学习。在具有最佳参数的歧视学习任务上模拟对抗行为学习（OpAL）和强化学习（RL）模型，平均经过 10,000 次迭代。一对选项可导致 80％的情况下获得 70％的回报，另外 30％则是 20％的情况下。图形显示选择“正确”选项（即 80％或 30％选项）的平均概率。有关该图的彩色版本，请参见在线文章。

我们还简化了模型的参与者部分的某些方面。例如，可以合理地预期，在每个途径中，可塑性常数对于增强或抑制均不相等。这可以通过在每个系统 G 和 N 中包括单独的增益和损失学习率来建模，从而导致四个参与者学习率。尽管这可能会在偏见表达方式上包括更大的灵活性，但我们选择分析一个更简单的版本，因为我们找不到能够使我们成功分离每种学习率角色的实验数据。我们也没有考虑丘脑底核的作用和决策阈值的调制，这些仍然是神经回路的重要方面(Frank＆O’Reilly，2006；Wiecki＆Frank，2013)。

尽管我们的模型可以说明多巴胺对不同类型的努力任务的影响，包括杠杆按压任务和带障碍的 T 形迷宫，但我们的模型目前无法捕获一些更细粒度的努力概念。例如，我们没有对发出响应的强度或力量进行建模-尽管从直观上讲，所考虑的每个动作的相对 G 到 N 差异不仅应确定哪个动作获胜及其响应时间，而且还应确定提供

的增强程度运动性丘脑皮质种群，这可能会影响其“强度”。未来的工作将研究这个概念。扩展和细化的其他可能方向包括更精确地计算工作时间范式(特别是在固定比率杠杆任务中，使用概率作为固定比率的代理，以防止期望报酬的动态变化产生活力)或其他指标努力。

与其他计算理论(例如，Dayan，2012；Niv 等，2007)相反，在给定某些目标函数的情况下，我们没有得出多巴胺水平的规范优化。相反，我们通过调制参数 β_G 和 β_N 探索了不同的多巴胺水平对对手行为权重的影响。虽然这使我们能够捕获多巴胺操作对其他模型无法解释的选择偏好的影响，但我们尚未探索如何根据任务上下文改变这些水平以优化性能。例如，在某些情况下，强调 G 的重量胜过 N 的重量是有利的，反之亦然。未来的工作应确定这些条件，并根据经验确定多巴胺水平是否相应调整。

7、结论

OpAL 模型提供了一个新的基于生物学的强化学习框架，该框架说明了将学习和执行任务中的行为与多巴胺能和纹状体直接和间接途径神经元的贡献联系起来的各种数据。它提供了进一步的可检验的预测，并为重要的开放性问题的理解提供了一些线索，例如为什么我们在基底神经节中有多余的途径以及为什么多巴胺能如此无所不在。未来的研究将调查预测并使用 OpAL 框架来扩展我们对这个复杂系统的理解。

data, the κ model also does not address the tendency for low DA levels to enhance performance in avoiding (or differentiating between) negatively valenced options. Their model does simulate the possibility for an animal to "want" what is not expected to be liked, nor remembered to be "liked" (i.e., a valence reversal), by differentially impacting motivational state at the time of choice to inflate the effective value of an option that had been encoded as mostly aversive. Our model can also account for these same effects, where an asymmetry in learning can be reversed by an asymmetry in choice incentive (Figure A4). Moreover the κ model required an alteration to use a log-based transformation that only applies in the case of a valence reversal, whereas the OpAL model captures this naturally without alteration.

Limitations

By simplifying the neurobiologically inspired neural network model from Frank (2005) into an algorithmic form, we have introduced some limitations. One key limitation is that we have focused fully on the actor (or dorsal striatal) part of the problem, and neglected the critic side (ventral striatum). Dopamine should also have effects on the critic, and a more detailed model, beyond the scope of this work, should incorporate them. However, we are confident that modification of this aspect would not significantly alter the qualitative results presented here. Indeed, we simulated different versions of the critic part of our model (for example including asymmetric gain/loss learning rate) and obtained qualitatively similar results, if quantitatively different.

We also simplified some aspects in the actor part of the model. For example, it could be reasonably expected that plasticity constants are not equal for potentiation or depression in each pathway. This could be modeled by including separate gain and a loss learning rates in each system *G* and *N*, thus leading to four actor learning rates. While this could potentially include more flexibility in the way biases express themselves, we chose to analyze a simpler version, because we could not find experimental data that would allow us to successfully dissociate roles of each of these learning rates. We also do not consider the roles of the subthalamic nucleus and modulations of decision thresholds that remain an important aspect of the neural circuit (Frank & O'Reilly, 2006; Wiecki & Frank, 2013).

Although our model can account for effects of dopamine on different kinds of effort tasks, including lever pressing tasks and the T-maze with barrier, our model cannot at this point capture some finer grained notions of effort. For example, we do not model the strength or force with which a response is emitted—although intuitively, the relative *G* to *N* difference for each action under consideration should determine not only which action wins and its response time but also the degree of boost provided to motor thalamocortical populations, which may affect its "strength." Future work will examine this notion. Other potential directions for extensions and elaborations includes accounting more precisely for timing in effort paradigms (in particular in the fixed ratio lever task, using probabilities as a proxy for fixed ratio prevents analysis of dynamic changes to vigor in expectation of reward) or for other indicators of effort.

Contrary to other computational theories such as (Dayan, 2012; Niv et al., 2007), we do not derive the normative optimization for dopamine levels given some objective function. Instead, we explore the effects of differing dopamine levels on the opponent actor weights via modulation of parameters β_G and β_N. While this allows us to capture effects of dopamine manipulations on choice preferences that are not accounted for by other model is, we have not yet explored how these levels should change as a function of task context so as to optimize performance. For example, there may be conditions under which it is advantageous to emphasize *G* weights over *N* weights and vice versa. Future work should identify these conditions and whether empirically, dopamine levels are adjusted accordingly.

Conclusion

The OpAL model provides a new biologically grounded reinforcement learning framework that accounts for a wide array of data linking behavior in learning and performance tasks to contributions of dopaminergic and striatal direct and indirect pathway neurons. It makes further testable predictions and provides some clues to the understanding of important open questions, such as why we have redundant pathways in the basal ganglia and why dopaminergic function is so omnipresent. Future research will investigate predictions and use the OpAL framework to expand our understanding of this complex system.

References

Aberman, J. E., & Salamone, J. D. (1999). Nucleus accumbens dopamine depletions make rats more sensitive to high ratio requirements but do not impair primary food reinforcement. *Neuroscience, 92,* 545–552. doi:10.1016/S0306-4522(99)00004-4

Amtage, J., & Schmidt, W. J. (2003). Context-dependent catalepsy intensification is due to classical conditioning and sensitization. *Behavioural Pharmacology, 14,* 563–567. doi:10.1097/00008877-200311000-00009

Arias-Carrión, O., Stamelou, M., Murillo-Rodríguez, E., Menéndez-González, M., & Pöppel, E. (2010). Dopaminergic reward system: A short integrative review. *International Archives of Medicine, 3,* 24.

Bayer, H. M., & Glimcher, P. W. (2005). Midbrain dopamine neurons encode a quantitative reward prediction error signal. *Neuron, 47,* 129–141. doi:10.1016/j.neuron.2005.05.020

Bayer, H. M., Lau, B., & Glimcher, P. W. (2007). Statistics of midbrain dopamine neuron spike trains in the awake primate. *Journal of Neurophysiology, 98,* 1428–1439. doi:10.1152/jn.01140.2006

Beeler, J., Frank, M., McDaid, J., & Alexander, E. (2012). A role for dopamine-mediated learning in the pathophysiology and treatment of Parkinson's disease. *Cell Reports, 2,* 1747–1761.

Beeler, J. A., Daw, N., Frazier, C. R. M., & Zhuang, X. (2010). Tonic dopamine modulates exploitation of reward learning. *Frontiers in Behavioral Neuroscience.* doi:10.3389/fnbeh.2010.00170

Berridge, K. C. (2007). The debate over dopamine's role in reward: The case for incentive salience. *Psychopharmacology, 191,* 391–431. doi:10.1007/s00213-006-0578-x

Berridge, K. C. (2012). From prediction error to incentive salience: Mesolimbic computation of reward motivation. *European Journal of Neuroscience, 35,* 1124–1143. doi:10.1111/j.1460-9568.2012.07990.x

Bódi, N., Kéri, S., Nagy, H., Moustafa, A., Myers, C. E., Daw, N., . . . Gluck, M. A. (2009). Reward-learning and the novelty-seeking personality: A between- and within-subjects study of the effects of dopamine agonists on young Parkinson's patients. *Brain: A Journal of Neurology, 132,* 2385–2395. doi:10.1093/brain/awp094

Collins, A. G. E., & Frank, M. J. (2013). Cognitive control over learning: Creating, clustering, and generalizing task-set structure. *Psychological Review, 120,* 190–229. doi:10.1037/a0030852

Cools, R., Frank, M. J., Gibbs, S. E., Miyakawa, A., Jagust, W., & D'Esposito, M. (2009). Striatal dopamine predicts outcome-specific reversal learning and its sensitivity to dopaminergic drug administration. *The Journal of Neuroscience, 29,* 1538–1543. doi:10.1523/JNEUROSCI.4467-08.2009

Cousins, M., Atherton, A., Turner, L., & Salamone, J. (1996). Nucleus accumbens dopamine depletions alter relative response allocation in a T-maze cost/benefit task. *Behavioural Brain Research, 74,* 189–197.

Cousins, M. S., & Salamone, J. D. (1994). Nucleus accumbens dopamine depletions in rats affect relative response allocation in a novel cost/benefit procedure. *Pharmacology, Biochemistry and Behavior, 49,* 85–91. doi:10.1016/0091-3057(94)90460-X

Dayan, P. (2012). Instrumental vigour in punishment and reward. *European Journal of Neuroscience, 35,* 1152–1168. doi:10.1111/j.1460-9568.2012.08026.x

Dayan, P., & Daw, N. D. (2008). Decision theory, reinforcement learning, and the brain. *Cognitive, Affective & Behavioral Neuroscience, 8,* 429–453. doi:10.3758/CABN.8.4.429

Doll, B. B., Hutchison, K. E., & Frank, M. J. (2011). Dopaminergic genes predict individual differences in susceptibility to confirmation bias. *The Journal of Neuroscience, 31,* 6188–6198.

Doll, B. B., Jacobs, W. J., Sanfey, A. G., & Frank, M. J. (2009). Instructional control of reinforcement learning: A behavioral and neurocomputational investigation. *Brain Research, 1299,* 74–94. doi:10.1016/j.brainres.2009.07.007

Farrar, A. M., Font, L., Pereira, M., Mingote, S., Bunce, J. G., Chrobak, J. J., & Salamone, J. D. (2008). Forebrain circuitry involved in effort-related choice: Injections of the GABAA agonist muscimol into ventral pallidum alter response allocation in food-seeking behavior. *Neuroscience, 152,* 321–330. doi:10.1016/j.neuroscience.2007.12.034

Farrar, A. M., Segovia, K. N., Randall, P. A., Nunes, E. J., Collins, L. E., Stopper, C. M., . . . Salamone, J. D. (2010). Nucleus accumbens and effort-related functions: Behavioral and neural markers of the interactions between adenosine A2A and dopamine D2 receptors. *Neuroscience, 166,* 1056–1067. doi:10.1016/j.neuroscience.2009.12.056

Floresco, S. B., Tse, M. T. L., & Ghods-Sharifi, S. (2008). Dopaminergic and glutamatergic regulation of effort- and delay-based decision making. *Neuropsychopharmacology, 33,* 1966–1979.

Frank, M. J. (2005). Dynamic dopamine modulation in the basal ganglia: A neurocomputational account of cognitive deficits in medicated and nonmedicated Parkinsonism. *Journal of Cognitive Neuroscience, 17,* 51–72. doi:10.1162/0898929052880093

Frank, M. J., Moustafa, A. A., Haughey, H. M., Curran, T., & Hutchison, K. E. (2007). Genetic triple dissociation reveals multiple roles for dopamine in reinforcement learning. *Proceedings of the National Academy of Sciences of the United States of America, 104,* 16311–16316. doi:10.1073/pnas.0706111104

Frank, M. J., & O'Reilly, R. C. (2006). A mechanistic account of striatal dopamine function in human cognition: Psychopharmacological studies with cabergoline and haloperidol. *Behavioral Neuroscience, 120,* 497–517. doi:10.1037/0735-7044.120.3.497

Frank, M. J., Santamaria, A., Reilly, R. C. O., & Willcutt, E. (2007). Testing computational models of dopamine and noradrenaline dysfunction in attention deficit/hyperactivity disorder. *Neuropsychopharmacology, 32,* 1583–1599.

Frank, M. J., Seeberger, L. C., & O'Reilly, R. C. (2004). By carrot or by stick: Cognitive reinforcement learning in parkinsonism. *Science, 306,* 1940–1943.

Gerfen, C. R. (2000). Molecular effects of dopamine on striatal-projection pathways. *Trends in Neurosciences, 23* (Suppl), S64–S70. doi:10.1016/S1471-1931(00)00019-7

Hazy, T. E., Frank, M. J., & O'Reilly, R. C. (2010). Neural mechanisms of acquired phasic dopamine responses in learning. *Neuroscience and Biobehavioral Reviews, 34,* 701–720. doi:10.1016/j.neubiorev.2009.11.019

Hikida, T., Kimura, K., Wada, N., Funabiki, K., & Nakanishi, S. (2010). Distinct roles of synaptic transmission in direct and indirect striatal pathways to reward and aversive behavior. *Neuron, 66,* 896–907. doi:10.1016/j.neuron.2010.05.011

Huys, Q. J. M., Cools, R., Gölzer, M., Friedel, E., Heinz, A., Dolan, R. J., & Dayan, P. (2011). Disentangling the roles of approach, activation and valence in instrumental and Pavlovian responding. *PLoS Computational Biology, 7,* e1002028. doi:10.1371/journal.pcbi.1002028

Jocham, G., Klein, T. A., & Ullsperger, M. (2011). Dopamine-mediated reinforcement learning signals in the striatum and ventromedial prefrontal cortex underlie value-based choices. *The Journal of Neuroscience, 31,* 1606–1613.

Klein, A., & Schmidt, W. J. (2003). Catalepsy intensifies context-dependently irrespective of whether it is induced by intermittent or chronic dopamine deficiency. *Behavioural Pharmacology, 14,* 49–53.

Kravitz, A. V., Freeze, B. S., Parker, P. R. L., Kay, K., Thwin, M. T., Deisseroth, K., & Kreitzer, A. C. (2010). Regulation of Parkinsonian motor behaviours by optogenetic control of basal ganglia circuitry. *Nature, 466,* 622–626. doi:10.1038/nature09159

Kravitz, A. V., Tye, L. D., & Kreitzer, A. C. (2012). Distinct roles for direct and indirect pathway striatal neurons in reinforcement. *Nature Neuroscience, 15,* 816–818. doi:10.1038/nn.3100

Maia, T. V., & Frank, M. J. (2011). From reinforcement learning models to psychiatric and neurological disorders. *Nature Neuroscience, 14,* 154–162.

McClure, S. M., Daw, N. D., & Read Montague, P. (2003). A computational substrate for incentive salience. *Trends in Neurosciences, 26,* 423–428. doi:10.1016/S0166-2236(03)00177-2

Mingote, S., Font, L., Farrar, A. M., Vontell, R., Worden, L. T., Stopper, C. M., . . . Salamone, J. D. (2008). Nucleus accumbens adenosine A2A receptors regulate exertion of effort by acting on the ventral striatopallidal pathway. *The Journal of Neuroscience, 28,* 9037–9046.

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

Morris, G., Nevet, A., Arkadir, D., Vaadia, E., & Bergman, H. (2006). Midbrain dopamine neurons encode decisions for future action. *Nature Neuroscience, 9,* 1057–1063. doi:10.1038/nn1743

Moustafa, A. A., Sherman, S. J., & Frank, M. J. (2008). A dopaminergic basis for working memory, learning and attentional shifting in Parkinsonism. *Neuropsychologia, 46,* 3144–3156. doi:10.1016/j.neuropsychologia.2008.07.011

Nakahara, H., Itoh, H., Kawagoe, R., Takikawa, Y., & Hikosaka, O. (2004). Dopamine neurons can represent context-dependent prediction error. *Neuron, 41,* 269–280. doi:10.1016/S0896-6273(03)00869-9

Niv, Y., Daw, N. D., Joel, D., & Dayan, P. (2007). Tonic dopamine: Opportunity costs and the control of response vigor. *Psychopharmacology, 191,* 507–520. doi:10.1007/s00213-006-0502-4

Nomoto, K., Schultz, W., Watanabe, T., & Sakagami, M. (2010). Temporally extended dopamine responses to perceptually demanding reward-predictive stimuli. *The Journal of Neuroscience, 30,* 10692–10702.

Nunes, E. J., Randall, P. A., Santerre, J. L., Given, A. B., Sager, T. N., Correa, M., & Salamone, J. D. (2010). Differential effects of selective adenosine antagonists on the effort-related impairments induced by dopamine D1 and D2 antagonism. *Neuroscience, 170,* 268–280. doi:10.1016/j.neuroscience.2010.05.068

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

Ott, D. V. M., Ullsperger, M., Jocham, G., Neumann, J., & Klein, T. A. (2011). Continuous theta-burst stimulation (cTBS) over the lateral prefrontal cortex alters reinforcement learning bias. *NeuroImage, 57,* 617–623. doi:10.1016/j.neuroimage.2011.04.038

Palminteri, S., Boraud, T., Lafargue, G., Dubois, B., & Pessiglione, M. (2009). Brain hemispheres selectively track the expected value of contralateral options. *The Journal of Neuroscience, 29,* 13465–13472.

Pessiglione, M., Seymour, B., Flandin, G., Dolan, R. J., & Frith, C. D. (2006). Dopamine-dependent prediction errors underpin reward-seeking behaviour in humans. *Nature, 442,* 1042–1045. doi:10.1038/nature05051

Ratcliff, R., & Frank, M. J. (2011). Reinforcement-based decision making in corticostriatal circuits: Mutual constraints by neurocomputational and

diffusion models. *Neural Computation, 24,* 1186–1229. doi:10.1162/NECO_a_00270

Reynolds, J. N., Hyland, B. I., & Wickens, J. R. (2001). A cellular mechanism of reward-related learning. *Nature, 413,* 67–70. doi:10.1038/35092560

Roesch, M. R., Calu, D. J., & Schoenbaum, G. (2007). Dopamine neurons encode the better option in rats deciding between differently delayed or sized rewards. *Nature Neuroscience, 10,* 1615–1624. doi:10.1038/nn2013

Salamone, J. D., Correa, M., Mingote, S. M., & Weber, S. M. (2005). Beyond the reward hypothesis: Alternative functions of nucleus accumbens dopamine. *Current Opinion in Pharmacology, 5,* 34–41. doi:10.1016/j.coph.2004.09.004

Samejima, K., Ueda, Y., Doya, K., & Kimura, M. (2005). Representation of action-specific reward values in the striatum. *Science, 310,* 1337–1340.

Satoh, T., Nakai, S., Sato, T., & Kimura, M. (2003). Correlated coding of motivation and outcome of decision by dopamine neurons. *The Journal of Neuroscience, 23,* 9913–9923.

Schönberg, T., Daw, N. D., Joel, D., & O'Doherty, J. P. (2007). Reinforcement learning signals in the human striatum distinguish learners from nonlearners during reward-based decision making. *The Journal of Neuroscience, 27,* 12860–12867.

Schultz, W. (1997). A neural substrate of prediction and reward. *Science, 275,* 1593–1599. doi:10.1126/science.275.5306.1593

Shen, W., Flajolet, M., Greengard, P., & Surmeier, D. J. (2008). Dichotomous dopaminergic control of striatal synaptic plasticity. *Science, 321,* 848–851. doi:10.1126/science.1160575

Shiner, T., Seymour, B., Wunderlich, K., Hill, C., Bhatia, K. P., Dayan, P., & Dolan, R. J. (2012). Dopamine and performance in a reinforcement learning task: Evidence from Parkinson's disease. *Brain: A Journal of Neurology, 135,* 1871–1883. doi:10.1093/brain/aws083

Smith, K. S., Berridge, K. C., & Aldridge, J. W. (2011). Disentangling pleasure from incentive salience and learning signals in brain reward circuitry. *Proceedings of the National Academy of Sciences of the United States of America, 108,* E255–E264. doi:10.1073/pnas.1101920108

Smittenaar, P., Chase, H. W., Aarts, E., Nusselein, B., Bloem, B. R., & Cools, R. (2012). Decomposing effects of dopaminergic medication in Parkinson's disease on probabilistic action selection: Learning or performance? *European Journal of Neuroscience, 35,* 1144–1151. doi:10.1111/j.1460-9568.2012.08043.x

Surmeier, D. J., Ding, J., Day, M., Wang, Z., & Shen, W. (2007). D1 and D2 dopamine-receptor modulation of striatal glutamatergic signaling in striatal medium spiny neurons. *Trends in Neurosciences, 30,* 228–235. doi:10.1016/j.tins.2007.03.008

Sutton, R., & Barto, A. (1998). Reinforcement learning. *Journal of Cognitive Neuroscience, 11,* 126–134. doi:10.1162/089892999563184

Tai, L.-H., Lee, A. M., Benavidez, N., Bonci, A., & Wilbrecht, L. (2012). Transient stimulation of distinct subpopulations of striatal neurons mimics changes in action value. *Nature Neuroscience, 15,* 1281–1289. doi:10.1038/nn.3188

Treadway, M. T., Buckholtz, J. W., Cowan, R. L., Woodward, N. D., Li, R., Ansari, M. S., . . . Zald, D. H. (2012). Dopaminergic mechanisms of individual differences in human effort-based decision-making. *The Journal of Neuroscience, 32,* 6170–6176.

Wassum, K. M., Ostlund, S. B., Balleine, B. W., & Maidment, N. T. (2011). Differential dependence of Pavlovian incentive motivation and instrumental incentive learning processes on dopamine signaling. *Learning & Memory, 18,* 475–483. doi:10.1101/lm.2229311

Wiecki, T. V., & Frank, M. J. (2013). A computational model of inhibitory control in frontal cortex and basal ganglia. *Psychological Review, 120,* 329–355. doi:10.1037/a0031542

Wiecki, T. V., Riedinger, K., von Ameln-Mayerhofer, A., Schmidt, W. J., & Frank, M. J. (2009). A neurocomputational account of catalepsy sensitization induced by D2 receptor blockade in rats: Context dependency, extinction, and renewal. *Psychopharmacology, 204,* 265–277. doi:10.1007/s00213-008-1457-4

Zhang, J., Berridge, K. C., Tindell, A. J., Smith, K. S., & Aldridge, J. W. (2009). A neural computational model of incentive salience. *PLoS Computational Biology, 5,* e1000437. doi:10.1371/journal.pcbi.1000437

(*Appendix follows*)