arXiv:2512.15948v1 [cs.AI] 17 Dec 2025

主观函数

Samuel J. Gershman 哈佛大学心理学系与脑科学中心
Kempner 自然与人工智能研究所

摘要

目标函数从何而来？我们如何选择要追求的目标？人类智能擅长即时合成新的目标函数。这是如何实现的？我们能否让人工系统具备同样的能力？本文提出了一种解答这些问题的方法，首先引入主观函数这一概念：一种内生于智能体的高阶目标函数（即依据智能体自身的特征定义，而非依据外部任务定义）。本文以预期预测误差为例，具体探讨主观函数。这一设想与心理学、神经科学和机器学习中的诸多思想都有联系。

1 引言

目标函数是所有学习系统（包括自然系统和人工系统）的核心。我们之所以能将学习与其他类型的动力学区分开来，是因为学习会带来性能提升（至少在期望意义上或渐近地），而性能是通过目标函数来衡量的。[1] 人们提出了许多不同的目标函数，而智力是否都能归结为单一的“终极”目标（例如繁殖适合度），尚不明确。[2] 或许问题在于，追寻单一目标函数本身就是一种误导。类人智能的一个重要特征，或许是合成目标函数的能力。

当然，这只不过是把问题往后拖。究竟是什么原则约束着目标函数的选择？任何这样的原则不也构成一个高阶目标函数吗？如果是这样，我们就又回到了起点，继续追寻一个单一的目标函数。

真正的主观 函数（而非高阶目标函数）是智能体内生的。要理解这意味着什么，可以考察一种定义目标函数的典型方式：规定某种奖励或监督信号，然后根据智能体最大化期望奖励或最小化期望误差的效果对其进行评分。从某种意义上说，这些信号是智能体外生的，因为它们的定义不依赖于智能体的任何特征；它们可以统一应用于任何智能体。相比之下，主观函数是智能体内生的，因为智能体所优化的信号，其定义取决于智能体的特征。

[1]例如，被动磨损会随着时间推移损害生物体和机器人的功能，但这并非学习，因为无法将其理解为性能随时间提升。[2]即使是认为所有形式的生物智能都源于自然选择这一合理论点，也无助于阐明产生智能行为的基本原理。

本文介绍一种可用于设计具备开放式学习能力的智能体的主观函数。随后，本文将探讨它与心理学和神经科学观察结果的联系，以及机器学习中的相关思想。

2 预备知识

我们将世界建模为由以下组成部分构成的马尔可夫决策过程（MDP）：

- 一个状态空间 $\mathcal{S}$。
- 一个动作空间 $\mathcal{A}$。
- 一个转移分布 $T(s'|s,\ a)$。
- 智能体根据策略 $\pi(a|s)$选择动作。

重要的是，我们不假设奖励函数是固定的。相反，我们允许智能体选择自己的奖励函数。具体而言，我们将研究奖励函数由特定目标状态 $g \in \mathcal{S}$参数化的情形：

$$R_g(s) = \mathbb{I}[s = g]. \tag{1}$$

因此，只有当智能体到达目标状态时，奖励才为 1。将这一设定扩展到其他情况（例如，奖励函数关于某个特征空间中的特征呈线性）相对简单；但以目标为基础的框架简洁明了，也适用于许多对人类而言很自然的环境。

3 设计原则

原则 1：主体选择能最大化预期预测误差的策略

强化学习（RL）理论中的一个标准假设是，主体力求最大化预期折扣未来奖励，即价值：

$$\begin{aligned} V_g^\pi(s) &= \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t R_g(s_t) \middle| s_0 = s, \pi, g\right] \\ &= R_g(s) + \gamma \sum_a \pi(a|s) \sum_{s'} T(s'|s,a) V_g^\pi(s'), \end{aligned} \tag{2}$$

其中，t 表示时间，$\gamma \in [0, 1)$ 是折扣因子，用于决定智能体如何衡量长期奖励。第二个等式即为贝尔曼方程。

相反，我们采用这样的假设：智能体寻求最大化预期预测误差(EPE)，以时序差分（TD）预测误差 δ_t 取代奖励：

$$\pi_g^*(\cdot|s) = \underset{\pi(\cdot|s)}{\operatorname{argmax}}\, U_g^\pi(s), \tag{3}$$

$$U_g^\pi(s) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t \delta_t \middle| s_0 = s, \pi, g\right], \tag{4}$$

$$\delta_t = R_g(s_t) + \gamma \hat{V}^\pi(s_{t+1}) - \hat{V}^\pi(s_t), \tag{5}$$

ˆ其中， V_g^π是智能体对价值函数的估计。直观而言，EPE 衡量了某种形式的˙目标进展，因为在目标达成之前 $\delta_t \approx V_g^\pi$（也就是说，预测误差近似于目标依赖价值的变化率）。

公式 4 是一个望远镜级数（中间项相互抵消），因此可以将其表示如下 ：

$$U_g^\pi(s) = V_g^\pi(s) - \hat{V}_g^\pi(s), \tag{6}$$

其中，我们假设 ˆ V_g^π 有界，使得 $\lim_{T\to\infty} \gamma^T \hat{V}_g^\pi(s_T) = 0$。由于 TD ˆ 误差用于更新价值估计，因此还必须假设在计算 δ_t 时， V_g^π 是冻结的（或变化缓慢的）。这类似于深度强化学习中目标网络的原理。

公式 6 表明，优化 EPE 等价于将价值最大化至一个常数（冻结的价值估计）；这意味着，最大化 EPE 的智能体会尝试提高其预期未来奖励。然而，如果某个状态的价值估计是完美的，则 EPE 为 0；这意味着，智能体不会尝试到达它已知具有高价值的高价值状态。EPE 之所以具有主观性，是因为同一个价值函数可能会根据智能体的价值估计产生不同的效用。

EPE的核心在于，主体会被正向意外所吸引。这仍然要求主体弄清楚如何到达目标，但一旦抵达，目标就已经“实现”了——目标本身不再令人感兴趣。例如，到达迷宫的终点后，继续逗留、重温胜利已毫无意义；但如果终点能带来奖励，且游戏不会终止，那么价值最大化智能体恰恰会这么做。出于同样的原因，反复重走通往胜利的路线也毫无意义；但如果有机会，价值最大化智能体恰恰会这么做。

一个稍微复杂些的模型将价值与EPE结合起来，以适应这样一个事实：即使在价值已被完美学会的极限情况下（此时EPE为0），主体仍可能偏好价值更高的策略：

$$\alpha V_g^\pi(s) + (1-\alpha) U_g^\pi(s) = V_g^\pi(s) - (1-\alpha)\hat{V}_g^\pi(s), \tag{7}$$

其中， $\alpha \in [0,1]$ 是一个加权参数（$\alpha = 0$ 还原出EPE模型； $\alpha = 1$ 还原出价值ˆ模型）。在 $V_g^\pi = V_g^\pi$的极限情况下，该模型简化为 αV_g^π（即价值模型的一个弱化版本）。当 α 接近0时，最优策略会追求误差最大化，直至学习消除了大部分误差来源；此时，策略将转而追求价值最大化。

原则2：主体选择能够最大化预期预测误差的目标

原则1涉及智能体的“内循环”，在这一过程中，目标固定，策略得到优化。原则2涉及智能体的“外循环”，在这一过程中，策略固定，目标得到优化。我们提出，在外循环中使用同一个目标函数EPE：

$$g^* = \operatorname*{argmax}_g U_g^*(s_0), \tag{8}$$

其中， s_0是初始状态， U_g^*是在最优策略 π_g^*下的 EPE。[3]这意味着，智能体会选择它认为能够带来正向意外的目标。直观地说，追求太容易或太难的目标都没有意义，因为两者都会带来 0 或负向的惊奇度。从这个意义上说，学习将是“开放式”的：智能体会不断努力选择新的目标，而它目前还不知道如何实现这些目标。

[3]实际上，在选择目标时，智能体可能并不知道最优策略，因此只能使用替代策略，例如当前目标下的最优策略。

4 心理学与神经科学的联系

享乐适应

你可能认为，好事发生时你会最开心，但证据表明，人们很快就会对奖励性刺激变得不再敏感，这种现象称为享乐适应（Frederick 和 Loewenstein，1999）。例如：

- 彩票中奖者通常并不比其他人更快乐，事实上，他们从日常事件中获得的乐趣反而更少（Brickman 等，1978）。
- 反复食用最初令人愉悦的食物，会降低其愉悦感以及之后的食用量（Rolls 等，1981年）。
- 对某些愉悦活动（如吸毒）产生的脱敏，可能促使成瘾行为的形成，作为一种补偿方式（Koob，1996）。

这些观察结果与这样一种观点一致：目标达成会消减动机（以 EPE 的减少形式体现）。

对瞬时主观幸福感的定量分析表明，在赌博任务中，近期预测误差的历史记录能够有力地预测幸福感判断（Rutledge et al., 2014）。这支持了这样一种观点：预测误差本身具有主观价值。

对递增奖励的偏好

当人们需要在不同的结果序列中做出选择时，通常会偏好期望奖励递增的序列（Loewenstein and Prelec, 1993）。例如，人们偏好金额逐渐增加的报酬序列（Loewenstein and Sicherman, 1991），即使这会导致总收入较低（Hsee et al., 1991）。类似地，相较于不适程度逐渐增加的序列，人们更偏好不适程度逐渐降低的序列（Varey and Kahneman, 1992; Chapman, 2000）。对于递增的序列，人们报告的满意度和积极情绪也更高（Hsee and Abelson, 1991; Lawrence et al., 2002）。从标准经济学理论的角度看，这令人困惑，因为这似乎意味着贴现率为负（$\gamma < 0$）——也就是说，人们偏好较早获得较小的奖励。然而，从 EPE 的角度来看，这就更容易理解：在目标达成之前，预测误差近似于估计价值对时间的导数。因此，最大化预期预测误差会使人们偏好随时间递增的期望奖励。

信息回避与需求

ˆ 在价值估计往往偏乐观的状态下（$V_g^\pi > V_g^\pi$），EPE 为正，因此主体倾向于回避可能缩小乐观差距的策略，例如信息收集。事实上，乐观偏差十分普遍（Sharot，2011年），也可能是信息回避的驱动因素（Golman 等，2017年）。例如，Eil 和 Rao（2011年）发现，当人们得到提示，得知某些信息可能会降低他们的预期时，他们往往会回避有关个人属性的信息，例如吸引力或智力。同样，有亨廷顿病患病风险的人往往会低估自身风险，并回避基因检测（Oster 等，2013）。

乍看之下，这些发现似乎与另一组发现相矛盾；后者表明，人们偏好早期信息揭示，即使这些信息并无实际作用（也就是说，

它无法改变未来的结果）。Kendall（1974）让鸽子在确定性选项（总是给予奖励，且之前会亮起白光）和随机选项（有 50% 的概率给予奖励；给予奖励时会亮起绿光，不给予奖励时则会亮起红光）之间进行选择。鸽子偏好随机选项，尽管它提供的奖励只有确定性选项的一半。关键在于，只有当灯光能够预测奖励时，它们才偏好随机选项；当灯光与奖励是否发放毫无关联时，它们则强烈偏好确定性选项。

理解鸽子看似次优选择的一种方式，是从这样的假设ˆ出发：它们的价值估计偏于悲观（$V_g^\pi < V_g^\pi$）。这可能源于光照与奖励之间的延迟，因为延迟会给幅度估计引入噪声；对这种噪声进行贝叶斯滤波，会使估计向先验收缩（Gabaix and Laibson, 2017；Gershman and Bhui, 2020）。如果先验预期低于 V_g^π，结果便是低估。这一解释与以下观察一致：次优选择主要出现在延迟较长时（Dunn et al., 2024），而此时噪声也应更大，收缩作用也更强。根据悲观假设，主体应当会寻求可能缩小悲观差距的信息。

与这一假设相关的另一类数据来自神经生理学。多巴胺神经元被认为会报告预测误差，它们对信息性线索的反应增强，对非信息性线索的反应则减弱（Bromberg-Martin and Hikosaka, 2009）。此外，动物对信息性线索与非信息性线索的反应之差，能够预测它对信息性线索的偏好。因此，这种偏好可能由预期预测误差驱动，尽管该研究并未证实因果关系。

总而言之，当价值被高估时，EPE模型预测会出现信息回避ˆ ˆ ($V_g^\pi > V_g^\pi$)；当价值被低估时，则会出现信息需求（$V_g^\pi < V_g^\pi$）。这些预测总体上与实证数据一致。

理解这些观察结果的另一种方式，是将时间折扣应用于预测误差，而不是奖励。当预期预测误差为负时，主体应寻求尽可能推迟这些误差。当预期预测误差为正时，主体应寻求尽快获得这些误差。[4]

目标追求中的条件理性

理性行动原则指出，主体会采用实现目标的最高效策略（例如，它们会选择通往目标位置的可用最短路径）。换言之，主体应最大化价值。如前文所述，最大化EPE等同于最大化价值（只要价值估计保持不变或变化足够缓慢）。关键在于，价值函数本身由主体的目标选择内生决定，而目标选择优化的是同一个主观函数。这就产生了所谓的条件理性（Chu等，2024）：对主观目标的高效追求。

条件理性的典型例子是儿童假装游戏。在很大程度上，这种游戏遵循现实规则和限制——但也有一定限度（Harris，2021）。例如，当两岁幼儿看到有人把假装的牙膏挤到两只玩具猪中的一只身上时，他们会正确地清洗那只“脏”猪（Harris 等人，1993）。显然，假装挤牙膏的动作违反了

[4]Iigaya 等人（2016）和 Zhu 等人（2017年）基于最大化预测误差这一理念，提出了相关但不同的信息需求解释。

牙膏应该清晰可见这一现实世界的限制，但儿童仍会基于牙膏已被挤到某只猪身上的前提，遵循一套高效的计划有条件的。

对 4 至 5 岁儿童进行的实验将这一观点又向前推进了一步（Chu 和 Schulz，2023）。在一项实验中，研究人员把孩子们带进一间房间，墙上固定着一些铅笔；有些铅笔“成本较低”（伸手就能轻松够到），另一些则“成本较高”（需要跳起来才能拿到）。当孩子们被要求拿铅笔（一个由外部指定的目标）时，大多数孩子都遵循理性行动原则，选择成本较低的行动。相比之下，当孩子们被要求玩耍时（“你能去那边玩吗？也许你可以玩个游戏来拿到铅笔。”），他们更倾向于选择成本较高的行动——跳起来（从收集铅笔的角度看）效率低下，但很有乐趣！在贴纸收集任务中也得到了类似结果：研究人员把一盒贴纸放在地板上，置于一条由胶带和五颜六色的圆点构成的螺旋路径尽头。被要求拿贴纸时，所有孩子都径直走向盒子，无视螺旋路径；被要求玩耍时，大多数孩子则沿着螺旋路径行走。这些研究告诉我们，儿童游戏的一个重要组成部分，是选择那些可能与成人从外部指定的目标截然不同的目标。尽管如此，儿童仍会高效地追求这些目标：跳起来和沿着螺旋路径行走，都是追求内在选定目标的“高效”计划。

免得你以为这只适用于儿童，不妨看看《吉尼斯世界纪录大全》中的一些例子：

- 2016 年，在保加利亚瓦尔纳的一处海滩上，人们组成人数最多的 DNA 双螺旋（4000 人）。
- 在英国布莱克浦举行的 2009 年全国擦窗比赛中，Terry Burrows 以 9.14 秒擦完三扇标准办公室窗户，成为史上最快的擦窗工。
- 2014 年，Bruce Masters 创下了“到访酒吧最多”的纪录（46,495 家）。

从没有任何工具性逻辑规定该追求哪个目标这一意义上说，这些目标本质上是任意的。但一旦选定，人们就会锲而不舍地追求它们。《吉尼斯世界纪录大全》是将条件理性发挥到极致的案例集。

Cushman 和 Morris（2015）对成年人的条件理性开展了一项不那么奇幻、但更系统的研究。他们采用一个三步序贯决策问题，表明人们倾向于遵循能够高效地引导自己达成过去曾带来奖励的目标的策略，即使这样会导致全局次优策略。这种行为符合一个通过时序差分学习来学习目标价值的模型，但也可能符合一个基于 EPE 的目标选择模型。

条件理性可能会引发病理性行为。尽管人们通常认为，成瘾反映了强迫性习惯形成，而这种习惯最终会取代对行为的目标导向控制（例如，Everitt 和 Robbins，2005 年），但这一观点与成瘾者表现出的复杂目标追求行为并不相符（Simon 和 Daw，2012；Hogarth，2020）。例如，寻求处方药的人有时会伪造或篡改电子病历。参与“择医行为”（辗转于不同医疗服务提供者之间，直到获得处方）的人，往往非常善于实现自己的目标（Schneberk 等人，2020）。一项针对海洛因滥用者的研究（Johnson 等人，1985）记录到，每日吸食者每天消耗价值约 36 美元的海洛因。为了支付这笔开销，他们会从事有组织的经济活动（通常就在海洛因行业内）。想必这些活动需要目标导向规划。因此，我们有效地

追求目标的能力未必总会因药物成瘾而受损；相反，它可能会被劫持，转而服务于以药物为导向的目标选择。

任务选择

实验表明，人们偏好能够带来性能提升的任务（在此设定中等同于目标）（Ten 等，2021；Poli 等，2022）；在4岁儿童中也报告了类似结果（Poli 等，2025）。这意味着，人们会根据当前的性能水平选择难度适中的任务，既不会太容易，也不会太难，这与原则2一致：如果任务太容易或太难，目标进展将接近于0。该原则与机器学习中的学习进展和能力进展概念密切相关，相关内容将在下文讨论。

5 与机器学习的联系

预测误差作为内在奖励

将预测误差（具体而言，即时序差分误差）用作内在奖励的想法，已在多篇论文中得到研究。Simmons-Edler 等（2020）训练了两个并行的函数逼近器，二者唯一的区别在于奖励的定义：一个是使用标准（外在）奖励的“利用”逼近器，另一个是使用时序差分误差绝对值（内在奖励）的“探索”逼近器。训练期间由探索策略控制动作，而测试时由利用策略控制动作。这一设置会产生高误差的训练样本，从而鼓励探索，同时利用模型以离策略方式学习，以最大化价值（有关密切相关的方法，参见 Griesbach 和 D’Eramo，2025）。Gehring 和 Precup（2013）也将时序差分误差绝对值用作内在奖励（他们称之为可控性），并在动作选择期间将其作为奖励加到价值函数中。时序差分误差的方差也曾被用作内在奖励（Flennerhag 等，2020）。所有这些方法的共同目标，都是鼓励探索状态空间中能够产生误差的部分。

与这里的观点最为接近的是正误差偏差算法（Parker 和 Sheppard，2025），该算法根据每个动作的预期时序差分误差估计值定义 softmax 策略。他们还研究了该模型的一个变体：预期时序差分误差估计器仅用于驱动特征学习，而动作则由基于相同特征的价值估计器控制。

学习进展与能力进展

基于模型的系统中还有一类相关研究，利用模型预测的无符号误差来指导探索和目标选择（例如，Schmidhuber，1991，2010；Oudeyer 等，2007；Pathak 等，2017；Molinaro 等，2024）。例如，Oudeyer 等（2007）开发了一种智能体，它会寻找平方误差预计将会增大的状态——也就是学习进展较大的状态。这项研究的一个重要见解是，相比当前性能水平，性能的提升是更好的内在奖励，因为后者会导致主体困在状态空间中难以预测的区域。这类方法面临的主要挑战是，感官空间中的误差可能非常嘈杂。Pathak 等（2017）尝试通过改为预测动作来缓解这一问题。

与此处的提议更为密切相关的是将无符号时序差分误差用作内在奖励的想法（Baldassarre 和 Mirolli，2012 年对此进行了综述）。例如，Schembri 等人（2007 年）

开发了一种智能体，由若干学习行动策略的“专家”以及一个决定任意时刻由哪位专家掌控的“选择器”组成。在“童年期”（探索）阶段，选择器使用被选中专家的时序差分误差作为奖励进行训练。这样一来，它便会选择那些预计能力将有所提升的专家，从而提升整个系统的能力。Stout 和 Barto（2010 年）研究了一个类似的想法，不过他们以时间扩展技能而非专家为框架。重要的是，这些方法避免了高维感官空间中误差带有噪声的问题。

将无符号时序差分误差用作内在奖励，是一类更通用算法的一个版本。这类算法利用能力进展——即学习过程中性能的提升——来指导探索和目标选择（Oudeyer 等，2007；Baranes 和 Oudeyer，2010；Colas 等，2022）。

一般而言，将无符号预测误差用作主观函数，会使主体追求那些实际奖励可能较少 于预期的目标（即预测误差为负的目标）。相比之下，基于 EPE 的主观函数总会驱使智能体朝着预期会产生正预测误差的目标前进。

广义优势估计

优势函数 $A^{\pi}(s,a)$ 定义为状态-动作价值函数与状态价值函数之差：

$$A^{\pi}(s,a) = Q^{\pi}(s,a) - V^{\pi}(s) = \mathbb{E}[\delta | s, a, \pi]. \tag{9}$$

第二个等式表明，对于给定的状态-动作对，优势函数就是期望时序差分误差。[5]

优势函数在策略梯度算法中发挥着特殊作用。令 θ 表示策略 π_θ的参数，策略梯度通常具有以下形式：

$$\nabla_\theta \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r_t\right] = \mathbb{E}\left[\sum_{t=0}^{\infty} \Psi_t \nabla_\theta \log \pi_\theta(a_t|s_t)\right], \tag{10}$$

其中，Ψ_t 是在与状态相关的基线下对价值的无偏估计。选择 $\Psi_t = A^{\pi}(s_t, a_t)$，其中基线对应于 $V^{\pi}(s_t)$，可以使无偏估计器的方差达到最低。

在实践中，主体很少能直接获得优势函数；相反，它们依赖估计器。除非满足某些特定条件，否则这可能引入偏差（Sutton 等，2000；Wen 等，2021）。通过对 N 步时序差分误差取平均，可以降低实际优势估计器的方差（广义优势估计；Schulman 等，2015）：

$$\hat{A}^{\pi}(s_t, a_t) = \sum_{k=0}^{\infty} (\gamma\lambda)^k \delta_{t+k}, \tag{11}$$

其中，$\lambda \in [0,1]$ 是控制偏差-方差权衡的加权参数。当 $\lambda = 0$ 时，我们得到演员-评论家方法中使用的标准单步优势估计器（Barto 等，2020）。该估计器偏差较高，但方差较低。当 $\lambda = 1$ 时，我们得到 EPE 的估计。该估计器无偏，但方差可能较高。取 λ 的中间值可以平衡偏差与方差。

[5]我们有 这里省略了目标下标（g），因为这些概念并不依赖这一假设 .

元损失与元学习

标准的机器学习设置从损失函数出发，随后推导出用于优化该损失的学习算法。一个重要的见解是，可以通过定义一个优化元损失的“外循环”，让损失和学习算法本身都能够被学习（Zheng 等人，2018；Xu 等人，2018、2020；Bechtle 等人，2021；Kirsch 等人，2020）。为了防止损失变得空洞无用（例如，将每个输出都设为可获得最大奖励——这是一种“奖励劫持”），这些方法通常会将元损失与某种客观的任务表现指标挂钩（通常通过元梯度实现）。因此，这些方法并不能学习真正主观的损失函数。理解元损失的益处的一种方式是：它们假定存在传统机器学习参数之外的其他知识存储位置 （Zheng 等人，2020）。例如，关于不同任务间共享结构的知识可以存储在奖励函数的参数中。

6 结论

目标函数从何而来？本文提出了一种基于预期预测误差、用于生成目标函数的主观函数。它具有一些引人注目的数学性质，也与许多早期思想有诸多联系。然而，它尚未被充分发展为一种实用算法。未来研究的重要问题既包括实际实施方面的问题，也包括预期预测误差能否充分解释人类目标选择的问题。

致谢

我感谢 John Vastola 和 Kazuki Irie 提供的宝贵反馈。本研究得到 Kempner 自然智能与人工智能研究所、Schmidt Sciences 颁发的多才奖，以及美国国防部多学科大学研究计划的支持；该计划由美国陆军研究办公室资助，资助编号为 W911NF-23-1-0277。

参考文献

Baldassarre, G. 和 Mirolli, M.（2012）。决定何时学习哪项技能：基于时序差分能力的内在动机（td-cb-im）。载于 《自然与人工系统中的内在动机学习》，第257–278页。施普林格。

Baranes, A. 和 Oudeyer, P.-Y.（2010）。受成熟度约束的基于能力的内在动机学习。载于 *2010年IEEE第九届发展与学习国际会议*，第197–203页。电气电子工程师学会。

Barto, A. G.、Sutton, R. S. 和 Anderson, C. W.（2020）。回顾行动者–评论家架构。《*IEEE*系统、人类与控制论汇刊：系统》，51：40–50。

Bechtle, S., Molchanov, A., Chebotar, Y., Grefenstette, E., Righetti, L., Sukhatme, G., 和 Meier, F.（2021）。通过学习得到的损失进行元学习。载于*202025th International Conference onPattern Recognition(ICPR)*，第 4161–4168 页。电气电子工程师学会。

Brickman, P., Coates, D., 和 Janoff-Bulman, R.（1978）。彩票中奖者与事故受害者：幸福感是相对的吗？《人格与社会心理学杂志》，36:917–927。Bromberg-Martin, E. S. 和 Hikosaka, O.（2009）。中脑多巴胺神经元会对有关即将到来的奖励的预先信息表现出偏好。《神经元》，63:119–126。Chapman, G. B.（2000）。对健康结果改善与恶化序列的偏好。《行为决策期刊》，13:203–218。Chu, J. 和 Schulz, L. E.（2023）。不按规则行事：探索性游戏、理性行动与高效搜索。《开放思维》，7:294–317。Chu, J., Tenenbaum, J. B., 和 Schulz, L. E.（2024）。为愚行喝彩：灵活目标与人类认知。《认知科学趋势》，28:628–642。Colas, C., Karch, T., Sigaud, O., 和 Oudeyer, P.-Y.（2022）。具有内在动机的目标条件强化学习的自目的型智能体：简要综述。《人工智能研究杂志》，74:1159–1199。Cushman, F. 和 Morris, A.（2015）。人类目标选择的习惯性控制。《美国国家科学院院刊》，112:13817–13822。Dunn, R. M., Pisklak, J. M., McDevitt, M. A., 和 Spetch, M. L.（2024）。次优选择：好消息信号（SiGN）模型的综述与量化。《心理学评论》，131:58–78。Eil, D. 和 Rao, J. M.（2011）。好消息—坏消息效应：对关于自身的客观信息进行非对称处理。《美国经济学杂志：微观经济学》，3:114–138。Everitt, B. J. 和 Robbins, T. W.（2005）。药物成瘾强化作用的神经系统：从行动到习惯，再到强迫行为。《自然—神经科学》，8:1481–1489。Flennerhag, S., Wang, J. X., Sprechmann, P., Visin, F., Galashov, A., Kapturowski, S., Borsa, D. L., Heess, N., Barreto, A., 和 Pascanu, R.（2020）。时序差分不确定性作为探索信号。*arXiv* 预印本 *arXiv:2010.02255*。Frederick, S. 和 Loewenstein, G.（1999）。享乐适应。载于《幸福感：享乐心理学基础》，第 302–329 页。罗素·塞奇基金会。Gabaix, X. 和 Laibson, D.（2017）。短视与折扣。技术报告，美国国家经济研究局。Gehring, C. 和 Precup, D.（2013）。在强化学习中利用绝对时序差分误差进行智能探索。载于《*2013* 年国际自主智能体与多智能体系统会议论文集》，第 1037–1044 页。Gershman, S. J. 和 Bhui, R.（2020）。理性注意力不足的跨期选择。《自然通讯》，11:3365。Golman, R., Hagmann, D., 和 Loewenstein, G.（2017）。信息回避。《经济文献杂志》，55:96–135。

Griesbach, S. and D’Eramo, C.（2025）。通过最大化时间差分误差，在多样化奖励环境中学习探索。强化学习会议。Harris, P. L.（2021）。想象力的早期限制：幼儿的现实主义。《儿童发展》，92:466–483。Harris, P. L., Kavanaugh, R. D., Wellman, H. M., and Hickling, A. K.（1993）。幼儿对假装的理解。《儿童发展研究学会专论》，页码 i–107。Hogarth, L.（2020）。成瘾由负性情绪下过度的目标导向药物选择驱动：对习惯与强迫理论的转化性批判。《神经精神药理学》，45:720–735。Hsee, C. K. and Abelson, R. P.（1991）。速度关系：满意度是结果随时间变化的一阶导数的函数。《人格与社会心理学杂志》，60:341–347。Hsee, C. K., Abelson, R. P., and Salovey, P.（1991）。位置和速度对满意度的相对权重。《心理科学》，2:263–267。Iigaya, K., Story, G. W., Kurth-Nelson, Z., Dolan, R. J., and Dayan, P.（2016）。预测误差对品味的调节及其对选择的影响。*eLife*，5:e13747。Johnson, B. D., Goldstein, P. J., Preble, E., Schmeidler, J., Lipton, D. S., Spunt, B., and Miller, T.（1985）。《打理生意：海洛因滥用者的犯罪经济学》。DC Heath，马萨诸塞州列克星敦。Kendall, S. B.（1974）。对间歇性强化的偏好。《实验行为分析杂志》，21:463–473。Kirsch, L., van Steenkiste, S., and Schmidhuber, J.（2020）。利用学习得到的目标改进元强化学习中的泛化。在第八届国际学习表征会议。Koob, G. F.（1996）。药物成瘾：享乐稳态的阴阳。《神经元》，16:893–896。Lawrence, J. W., Carver, C. S., and Scheier, M. F.（2002）。即时体验中朝向目标达成的速度是情感的决定因素。《应用社会心理学杂志》，32:788–802。Loewenstein, G. and Sicherman, N.（1991）。工人是否偏好递增工资曲线？《劳动经济学杂志》，9:67–84。Loewenstein, G. F. and Prelec, D.（1993）。对结果序列的偏好。《心理学评论》，100:91–108。Molinaro, G., Colas, C., Oudeyer, P.-Y., and Collins, A. G.（2024）。潜在学习进展驱动人类强化学习中的自主目标选择。《神经信息处理系统进展》，37:32251–32280。Oster, E., Shoulson, I., and Dorsey, E. R.（2013）。最优预期与有限医学检测：来自亨廷顿病的证据。《美国经济评论》，103:804–830。

Oudeyer, P.-Y., Kaplan, F., and Hafner, V. V.（2007）。用于自主心理发展的内在动机系统。*IEEE*进化计算汇刊，11:265–286。Parker, A. and Sheppard, J. W.（2025）。引导探索朝向正误差，以实现高效强化学习。Pathak, D., Agrawal, P., Efros, A. A., and Darrell, T.（2017）。通过自监督预测进行好奇心驱动的探索。在国际机器学习会议，页码 2778–2787。PMLR。Poli, F., Meyer, M., Mars, R. B., and Hunnius, S.（2022）。预期学习进展和感知新颖性对好奇心驱动探索的贡献。《认知》，225:105119。Poli, F., Meyer, M., Mars, R. B., and Hunnius, S.（2025）。4岁儿童的探索受学习进展和新颖性引导。《儿童发展》，96:192–202。Rolls, B. J., Rolls, E. T., Rowe, E. A., and Sweeney, K.（1981）。人类的感觉特异性饱足。《生理学与行为》，27:137–142。Rutledge, R. B., Skandali, N., Dayan, P., and Dolan, R. J.（2014）。瞬时主观幸福感的计算与神经模型。《美国国家科学院院刊》，111:12252–12257。Schembri, M., Mirolli, M., and Baldassarre, G.（2007）。内在动机强化学习机器人中的进化与学习。在欧洲人工生命会议，页码 294–303。施普林格。Schmidhuber, J.（1991）。在模型构建神经控制器中实现好奇心与无聊的一种可能性。在自适应行为模拟国际会议论文集：从动物到人工动物，页码 222–227。Schmidhuber, J.（2010）。创造力、乐趣与内在动机的形式理论（1990—2010年）。*IEEE*自主心理发展汇刊，2:230–247。Schneberk, T., Raffetto, B., Friedman, J., Wilson, A., Kim, D., and Schriger, D. L.（2020）。择医就诊患者的阿片类药物处方模式；对医疗服务提供者的启示。《公共科学图书馆·综合》，15:e0232533。Schulman, J., Moritz, P., Levine, S., Jordan, M., and Abbeel, P.（2015）。使用广义优势估计进行高维连续控制。*arXiv*预印本 *arXiv:1506.02438*。Sharot, T.（2011）。乐观偏差。《当代生物学》，21:R941–R945。Simmons-Edler, R., Eisner, B., Yang, D., Bisulco, A., Mitchell, E., Seung, S., and Lee, D.（2020）。将奖励预测误差作为深度强化学习的探索目标。在第二十九届人工智能国际联合会议，页码 2816–2823。Simon, D. A. and Daw, N. D.（2012）。双系统学习模型与滥用药物。在药物成瘾的计算神经科学，页码 145–161。施普林格。Stout, A. and Barto, A. G.（2010）。能力进展内在动机。在*2010*年*IEEE*第九届发展与学习国际会议，页码 257–262。电气电子工程师学会。

Sutton, R. S., McAllester, D., Singh, S., 和 Mansour, Y.（2000）。采用函数逼近的强化学习策略梯度方法。《神经信息处理系统进展》，12。Ten, A., Kaushik, P., Oudeyer, P.-Y., 和 Gottlieb, J.（2021）。人类会监测好奇心驱动探索中的学习进展。《自然·通讯》，12:5972。Varey, C. 和 Kahneman, D.（1992）。跨时间延展的体验：对瞬间与情节的评估。《行为决策期刊》，5:169–185。Wen, J., Kumar, S., Gummadi, R., 和 Schuurmans, D.（2021）。刻画演员-评论家与策略梯度之间的差距。载于国际机器学习会议，第11101–11111页。PMLR。Xu, Z., van Hasselt, H. P., Hessel, M., Oh, J., Singh, S., 和 Silver, D.（2020）。采用在线发现的目标函数进行元梯度强化学习。《神经信息处理系统进展》，33:15254–15264。Xu, Z., van Hasselt, H. P., 和 Silver, D.（2018）。元梯度强化学习。《神经信息处理系统进展》，31。Zheng, Z., Oh, J., Hessel, M., Xu, Z., Kroiss, M., Van Hasselt, H., Silver, D., 和 Singh, S.（2020）。学习到的内在奖励能够捕捉什么？载于国际机器学习会议，第11436–11446页。PMLR。Zheng, Z., Oh, J., 和 Singh, S.（2018）。关于为策略梯度方法学习内在奖励。《神经信息处理系统进展》，31。Zhu, J.-Q., Xiang, W., 和 Ludvig, E. A.（2017）。将信息寻求视为追逐预期预测误差。载于认知科学学会年会论文集，第39卷。