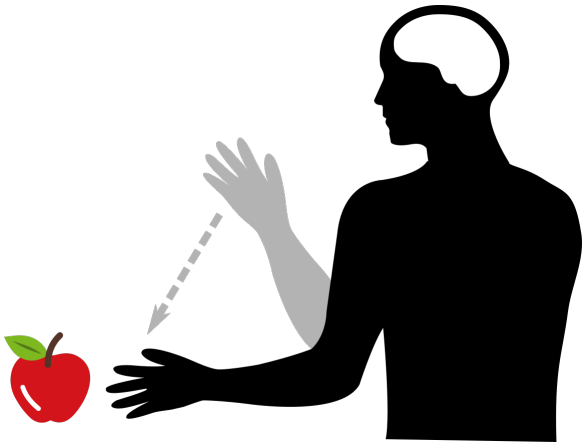


强化学习的「思维空间延拓」

在传统的强化学习里，「环境」只包含 physically observable 的外在环境。我提出将 RL 延续到**内在的**思维空间。

从传统 RL 的角度：人伸手拿苹果，苹果是**奖励**，伸手是**动作**。这些都是可以在环境中**观察**到的：



(1)

强化学习的基础是 **Bellman equation**, 它可以看成是一条「**递归**」的方程：

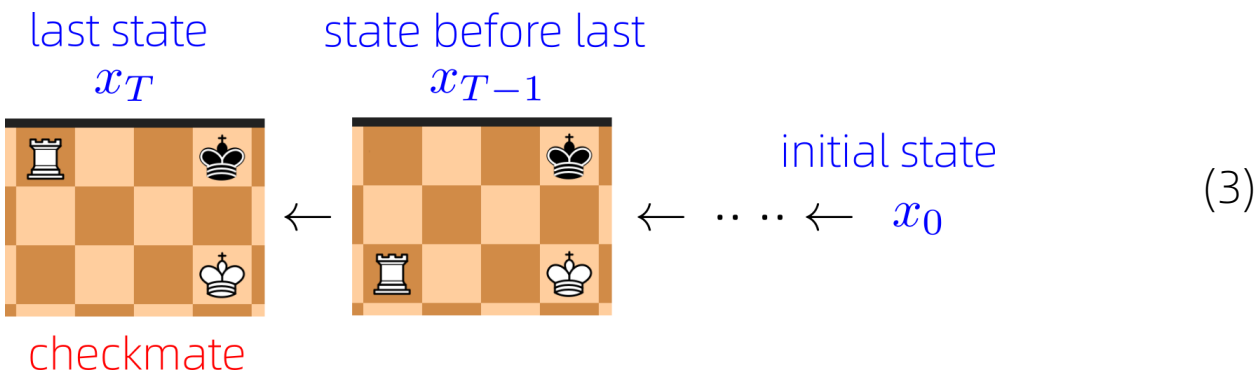
当前状态

$$V(x_0) = \max_a \{R + \gamma V(x_1)\}$$

下一状态

(2)

它将 **终点**状态的**价值**「反向传递¹」到 终点前一步的状态的价值，就像下棋的情况，可以一直追溯到第一步棋的价值：

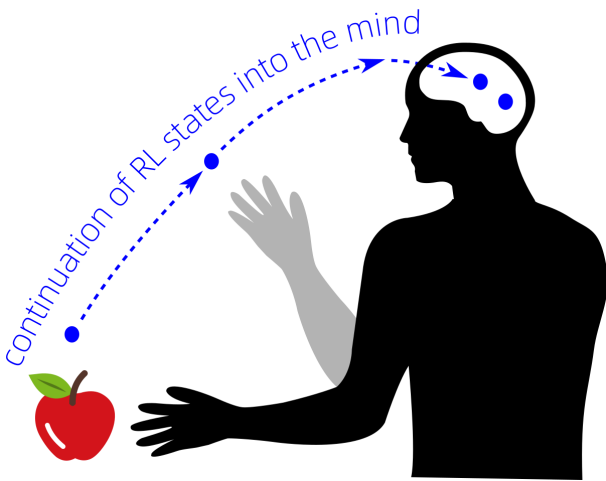


换句话说，获得苹果的价值，反向传递到「伸手取苹果」这动作的价值。
So far so good.
而同样地，我们可以继续反向传递到决定「伸手取苹果」之前的一连串**思维**：

我肚饿 → 我要找食物 → 我看见一只苹果 → 苹果是食物 →

(4)

换句话说，将**内部**的思维状态「反转」，看成跟**外部**的状态，是同等的地位：



(5)

而这跟象棋的价值函数的反向传递是完全一样的。换句话说，我们可以用强化学习的算法，学习思维空间的内容，提供了 AGI 严谨的基础。

¹注意这不同于神经网络的 back-prop.

Learning of logic rules under RL

将 外部和内部状态 **统一处理**的做法，在哲学上也没有问题，因为其实 brain states 也是物理状态，只是肉眼看不到而已。大脑状态 即是 神经元群的激活状态，它们之间的 transitions 是由神经**权重**决定的，而这些权重又是由 **Hebbian learning** 学习的（至少根据我们现时最好的理解）。

而 强化学习 又是怎样**学习** 逻辑内容？ **动作** 就是由一个逻辑状态 转移到另一个逻辑状态，也可以看成是 由一些命题 **推导出** 一个新的命题，那就是 逻辑 **rule**. 我们要在很多 动作（逻辑 rules）之中选取最好的 动作。换句话说，要在当前状态下可以执行的 rules 之中选取最好的一个或多个 rule(s). 例如：

我很肚饿 → 我要吃苹果 (6)

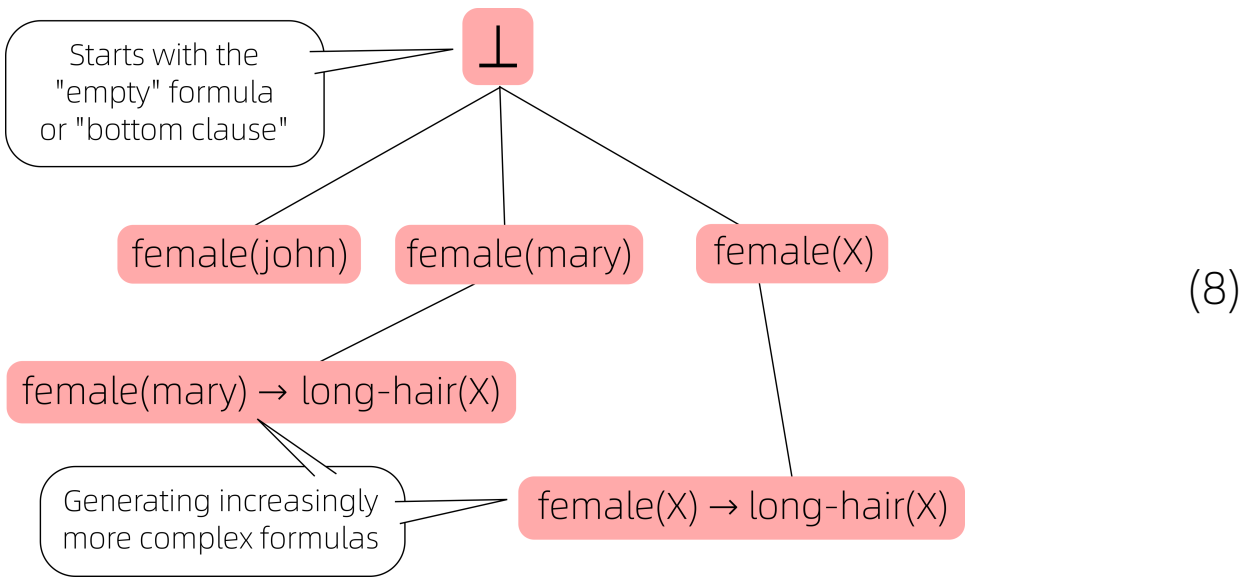
是一条正确的 rule；

肚饿 → 吃网球 (7)

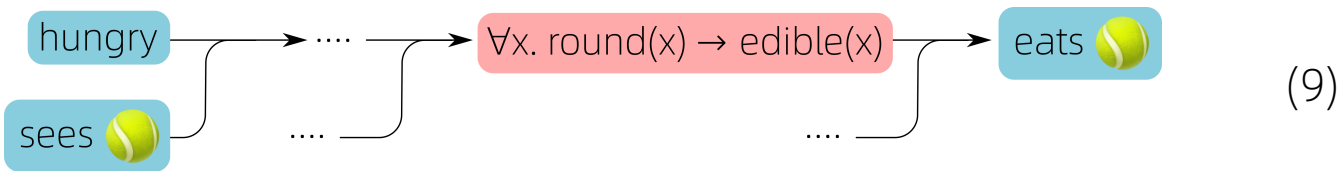
就比较差了。

而 强化学习 的好处是：理论上，它可以在无限的思维空间里 学到非常**抽象**的 rules，情形就像它在复杂的游戏**迷宫**里，学到破解的方法。

在经典 AI 里已有研究过 逻辑 rules 的 combinatorial 搜寻，例如有这种形式的**搜寻树**：



思维空间的一个特性是：理论上，任何思想都有可能推导出任何思想。换句话说，任何两点之间都有可能存在一条路径（= 逻辑 rule = 动作）。例如，如果一个肚饿的人，看见一只网球，那么「圆型的就可以吃」这条 rule 就似乎可以立即解决他的肚饿，直到他真的尝试吃它：

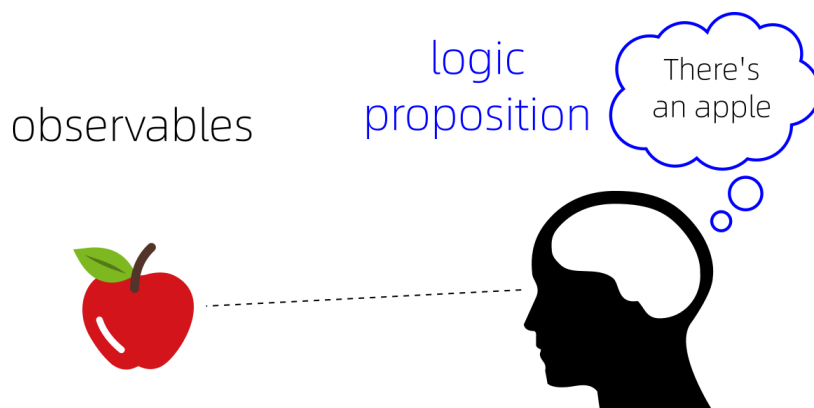


他得到的负值奖励 会反向传播至 整个**推导链**。经过很多次 迭代，系统最终会发现出错的 rule. 简言之，一个初始的逻辑系统可以是任意地「疯狂」的。事实上，一些天才正是因为有点疯，才发现了其他人想不到的事物。这是逻辑学习系统必然有的特性，并不只限于我提出的架构。

The mind as a “model” of the world

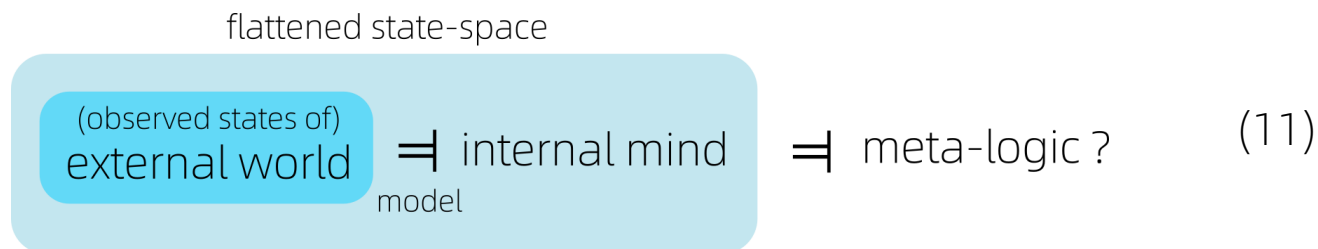
在传统 强化学习里 有所谓 **model-based** 方法, 而我们在大脑里面的 **mental model** of the world 其实是同一个概念, 但由于「混合状态」的原因而被「压平」了。

根据经典逻辑哲学, 一个 (脑中的) **逻辑命题**, 对应于现实世界中 某个**状态**:



(10)

然而, 在「混合状态」或 flattened view 观点下, 外部世界 和 思维状态 都存在于同一个状态空间, 而 思维状态 是 外部世界 的 model 或 “theory”:³



(11)

那么, 混合状态空间 本身又有没有 theory 呢? 那可能是某种 **元逻辑**。元逻辑是一种 归纳偏置, 或许会在加速学习上有重要作用。

Picture credits:

Human figure from www.onlinewebfonts.com licensed by CC BY 3.0

Thought bubble created by Catherine Please from the Noun Project

³符号 $T \models M$ 的意思是: M is a **model** of T ; T is a **theory** of M . 这是 逻辑**模型论** 的术语, 有严谨的定义。