第九讲 强化学习(I) Lecture 9 Reinforcement Learning (I)

> 明玉瑞 Yurui Ming yrming@gmail.com

声明 Disclaimer

本讲义在准备过程中由于时间所限,所用材料来源并未规范标示引用来源。所引材料仅用于教学所用,作者无意侵犯原著者之知识版权,所引材料之知识版权均归原著者所有;若原著者介意之,请联系作者更正及删除。

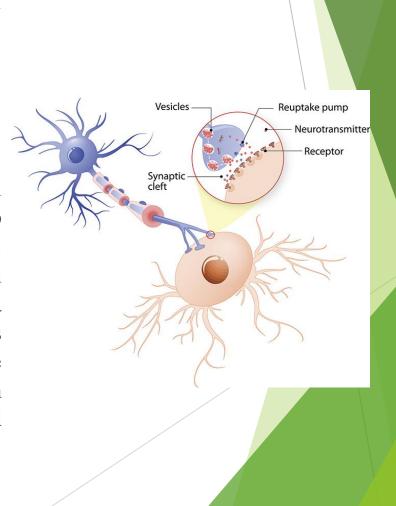
The time limit during the preparation of these slides incurs the situation that not all the sources of the used materials (texts or images) are properly referenced or clearly manifested. However, all materials in these slides are solely for teaching and the author is with no intention to infringe the copyright bestowed on the original authors or manufacturers. All credits go to corresponding IP holders. Please address the author for any concern for remedy including deletion.

神经递质

Neurotransmitters

在第二讲中,我们已经讲过,信号在突触之间的传导,取决于突触间隙的宽度,当宽度足够小,能形成微管时,为电传导,否则为化学传导。化学传导依赖于不同类型的神经递质,或者兴奋下游神经元的活动,或者抑制下游神经元的活动。但无论怎样,当前神经元与被作用的神经元,一般位于同一个皮质区内。

We have addressed in lecture 2 that the ways of electrical impulse traveling across synapse depend on the width of gap between the pre-synaptic ending and post-synaptic ending. Sufficient small gap which backs the forming of microtube can conduct the signal in an electrical way, otherwise it relies on a chemical conduction. Neurotransmitters of different properties participate in the process, some of which excite the neuron the post-synaptic ending resides, others inhibit the down-stream neuron's activities. However, in all cases, the neurons and affected ones tend to reside in the same cortices.



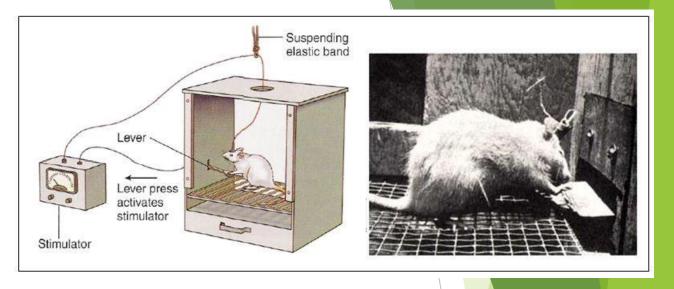
▶ 作为神经递质的多巴胺,与其它神经递质在传导信号过程中的生化过程,没有本 质的不同, 但产生多巴胺的细胞与其作用的下游细胞, 却往往分布在不同的皮质 层。在讲这点之前, 我们先介绍一下奖励系统。

Dopamine makes no essential difference with other neurotransmitters regarding the biochemical process in conducting the signal. However, the neurons which secret and use dopamine as the transmitter generally locate in different cortices with the next-inline affected neurons. We introduce reward system first before getting into depth.

▶ 动物在进行某些行为时有愉悦感,从而在条件允许的情况下,愿意重复这些行为。 由于愉悦感可视为脑的某个部位接收到进行这些行为的刺激产生的,一个想法是 直接刺激这些部位,也有可能产生愉悦感。

The animals can experience enjoyment when they perform some actions, such as eating or mating. They tend to repeat these actions if conditions allow. One conjecture is that the feelings are consequence of some part of brain receiving stimuli when performing these tasks, therefore, a direct stimulation to the corresponding part might led to similar experience.

奖励系统 Reward System



▶ 上世纪50年代,学者James Old与Peter Milner决定将电极植入小鼠的大脑,进行相关研究。他们设计的实验装置可以诱使小鼠无意地按下推杆,让它们的大脑接受轻微的电刺激。Olds和Milner发现当把电极置于位于胼胝体前端下方的中隔区时,老鼠会反复按压控制杆以接收刺激。他们发现实验中的一只老鼠在12小时内按了7500次杠杆去电刺激该区域。

In the 1950s, James Olds and Peter Milner decided to implant electrodes in the brains of rats to conduct research in this regard. They designed the experiment which can lure the arts to press a lever to receive a mild burst of electrical stimulation to their brains un-intentionally at first. Later Olds and Milner discovered that, if they placed the electrodes into the region known as the septal area, which lies just below the front end of the corpus callosum, rats would repeatedly press the lever to receive stimulation to. They found one of the rats in their experiment pressed a lever 7500 times in 12 hours to receive electrical stimulation here.

奖励系统 Reward System



▶ Olds和Milner的实验验证了大脑存在存在着某些结构,这些结构有调节个体体验有益经历的功能。实验中老鼠反复按杠杆来接收这些区域的刺激,这表明它们正在享受这种体验。随后的研究试图更彻底地绘制出这些"奖励区域",丰富人们对奖励系统的更多更为深刻地理解。最终,人们认识到多巴胺及相应神经元在这种有益的大脑刺激过程中的作用。

Olds and Milner's experiments confirmed the existence of brain structures that are devoted to mediating rewarding experiences. The rats were lever-pressing repeatedly to receive stimulation to these areas, which suggested they were enjoying the experience. Subsequent studies attempted to more thoroughly map out these "reward areas," and to deepen the understanding of the reward system. And it was eventually discovered that dopamine and the corresponding neurons are activated during this type of rewarding brain stimulation and confirmed the importance of dopamine's role in reward.

奖励系统 Reward System

► 后来在遵循人体实验道德伦理前提下,在人类被试上的实验亦验证了与小鼠同样的行为。目前,我们认识到这些自愿行为的动机,是因为个人认为此行为有益或体验到了愉快。其中某些行为或刺激,如食物或性,自然是有益的,因为它们是物种生存所必需的,神经系统已经进化到使这些行为令人愉快。而除此之外的一些行为则具有适应性,一旦个体认为是某种正向的刺激(奖励)使构成奖励回路的大脑区域激活,则个体可能倾向于反复进行类似的活动进行刺激,最后甚至有可能致瘾。

Later, strictly following the ethics of human experiments, experiments on human subjects also verified the similar behaviours as mice. Now, it is regarded that the motivated behaviours are voluntary behaviours that individuals find rewarding or pleasurable. Certain behaviours or stimuli, like food or sex, are naturally rewarding because they are necessary for the survival of a species, and the nervous system has evolved to make these behaviours pleasurable. Other activities can be adaptive. Once the positive stimuli (reward) which is at least recognised by the individual increases brain activation in brain regions that comprise the reward circuit, the individual grows the inclination for repetition, which potentially leads to addition eventually

奖励回路

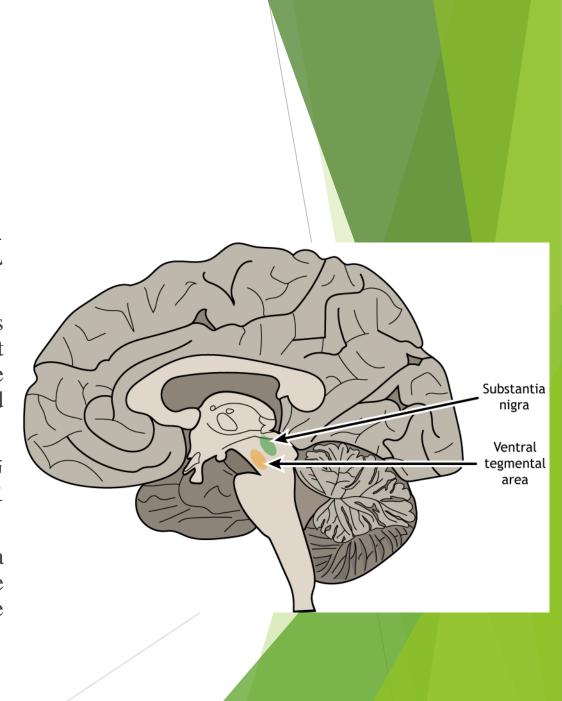
Reward Circuits

► 在对这些奖励行为的研究中,形成了奖励系统的有关知识。 而在奖励系统中,最重要的是奖励回路。在奖励系统中起至 关重要的多巴胺,则主要由位于腹侧被盖区 (VTA) 的神经元 合成和释放。

The accumulated research into these rewarding behaviors establishes the subject of reward system, and the reward circuit constitutes a major research topic for reward system. Dopamine which plays the vital role in reward system, is synthesized and released by neurons located in the ventral tegmental area (VTA).

如右图所示,腹侧被盖区(橙色区域)位于黑质附近的中脑区域(绿色区域),两个区域都产生多巴胺并将多巴胺释放到下游目标上。

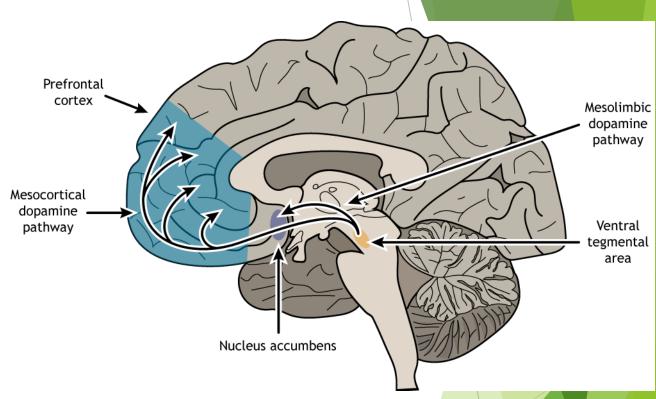
As illustrated in the figure right, the ventral tegmental area (orange region) is located in the midbrain region near the substantia nigra (green region). Both regions produce and release dopamine onto downstream targets.



奖励回路

Reward Circuits

▶ 进一步研究发现,VTA中有两条主要途径对奖励系统来说很重要,一是中脑边缘通路,其将VTA连接到位于腹侧纹状体的伏隔核。二是中皮层通路,其将VTA与前额叶皮层连接起来。VTA(橙色区域)通过中脑边缘通路将多巴胺释放到伏隔核(紫色区域),通过中脑皮质通路将多巴胺释放到前额叶皮层(蓝色区域)。



Later research discovers that there are two primary pathways from the VTA that are important for reward. The first is the mesolimbic pathway, which connects the VTA to the nucleus accumbens, a region located in the ventral striatum. The second is the mesocortical pathway, which connects the VTA with the prefrontal cortex. The ventral tegmental area (orange region) releases dopamine into the nucleus accumbens (purple region) via the mesolimbic pathway and releases dopamine into the prefrontal cortex (blue region via the mesocortical pathway.

虽然奖励系统对调控生物的行为,维系高级认知功能上有重要作用,但一个事实是产生多巴 胺的神经元数量非常有限。研究发现,只有大约一百万的多巴胺细胞来自中脑并投射到其他 区域,只占大脑中估计的大约 1000 亿个神经元的一小部分。这表明大脑功能的重大改变至 少在某些情况下可能只直接涉及一小部分脑细胞。

Although reward system mediates behaviours of animals and sustains their high cognitions, however, only about one million carry the dopaminergic fibers from midbrain projecting to other areas. These are of course a tiny fraction of the estimated 100 billion or so neurons in the brain and suggest that major alterations in brain function can, in some cases at least, directly involve only a small percentage of brain cells.

▶ 一个悬而未讲的问题是,在多巴胺细胞如此之少的情况下,如何说明多巴胺在奖励系统中的 作用呢?一个延伸问题是,在决定出多巴胺在奖励系统研究中的重要地位时,如何在这点取 得较为一致的共识呢?

A question postponed to be answered for such limited dopamine cells, how to account for the role of dopamine in the reward system? An extended question is, how to achieve a more consistent consensus on establishing dopamine's significance in the study of the reward system?

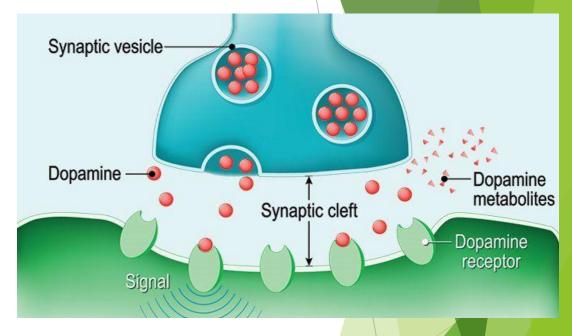
▶ 实际上,以回顾的方式来回答这个问题时,似乎脉络更加清晰,即按照在既往研究的经验基础上总结的偏程式化研究的范式,来佐证多巴胺的重要作用:第一步,尝试确定多巴胺为神经递质;第二步,寻找多巴胺拮抗剂;第三步,在注入多巴胺拮抗剂之后,观察同样的实验在多巴胺阻断之后,被试的实验效果及实验结果。

In fact, to answer the question in a retrospective way might bear a clearer thread in the fabrication of research. That is, to reference to the pragmatic research paradigm established by drawing lessons and inspirations from accumulated research, to back the fact about the indispensability of dopamine. In detail, the first step is to determine that dopamine belongs to the categories of neurotransmitter; the second step is to look for dopamine antagonists; the third step, after injecting dopamine antagonists, to observe the experiment consequence, such as the behaviours and responses of the subjects after dopamine blocking under the identical experiment setup.

多巴胺

Dopamine

- ▶ 第一步,在确定多巴胺为神经递质过程中,有下面3 个条件需要综合考量:
 - ▶ 1. 多巴胺必须在突触前神经元中合成和储存。
 - ▶ 2. 多巴胺必须在刺激时由突触前轴突末端释放。
 - ▶ 3. 当在实验中手工施用多巴胺时,须能在突触后细胞中引起反应,该反应与由突触前神经元释放神经递质所产生的反应一致。
- ▶ 相关实验已经证实多巴胺满足以上几点。



The first step, to determine dopamine as a neurotransmitter, the following three conditions need to be met:

- ▶ 1. The molecule must be synthesized and stored in the presynaptic neuron.
- ▶ 2. The molecule must be released by the presynaptic axon terminal upon stimulation.
- ▶ 3. The molecule, when experimentally applied, must produce a response in the postsynaptic cell that mimics the response produced by the release of neurotransmitter from the presynaptic neuron.

Various experiments confirms the properties of dopamine satisfy the above criteria.

多巴胺

Dopamine

▶ 第二步,寻找多巴胺拮抗剂。 多巴胺拮抗剂,也称为抗多巴胺能和多巴胺受体拮抗剂 (DRA),是一种通过受体拮抗作用阻断多巴胺受体的药物。

A dopamine antagonist, also known as an anti-dopaminergic and a dopamine receptor antagonist (DRA), is a type of drug which blocks dopamine receptors by receptor antagonism.

早期研究者通过研究发现, 匹莫齐特(哌 迷清)可作为多巴胺阻断剂。

Early research discovered Pimozide acts as dopaminergic receptor blockade.



Less activation

Full activation

No activation

▶ 第三步,注入多巴胺拮抗剂之后的实验结果观察。经过对照实验,人们发现,在 没有给予多巴胺拮抗剂时, 即多巴胺神经元在自我刺激的过程中激活时, 小鼠积 极进行杠杆按压; 当给予多巴胺拮抗剂时, 大鼠减少甚至停止杠杆按压。

The third step, to observe the experiment after administering the dopamine antagonist. It was recognized that blockage of dopamine neurons activations could cause rats to stop lever pressing. In controlled experiment without administering the dopamine antagonist, the rodent are active to perform the self-stimulation.

▶ 以上三步证明了, 多巴胺在大脑奖励系统中的中心作用。

The above procedures supported the importance of dopamine's central role in reward.