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

明玉瑞 Yurui Ming

yrming@gmail.com

声明
Disclaimer

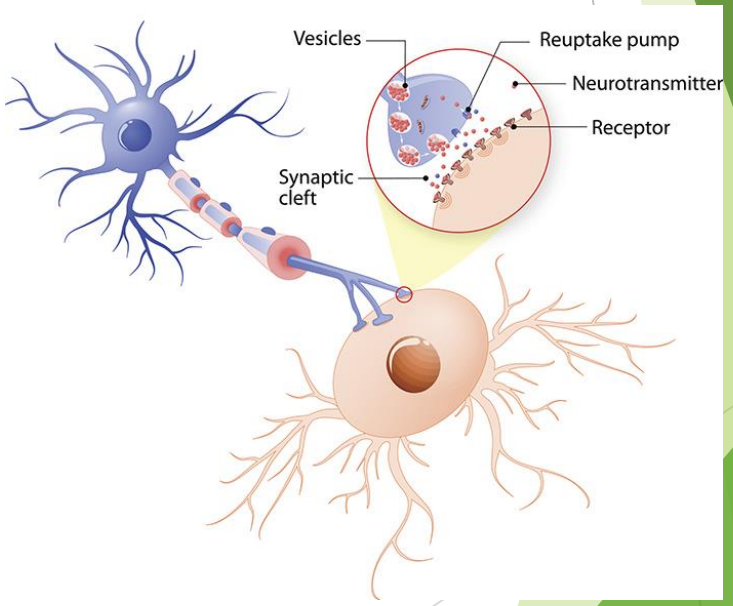
- 本讲义在准备过程中由于时间所限，所用材料来源并未规范标示引用来源。所引材料仅用于教学所用，作者无意侵犯原著者之知识版权，所引材料之知识版权均归原著者所有；若原著者介意之，请联系作者更正及删除。

  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.



多巴胺
Dopamine

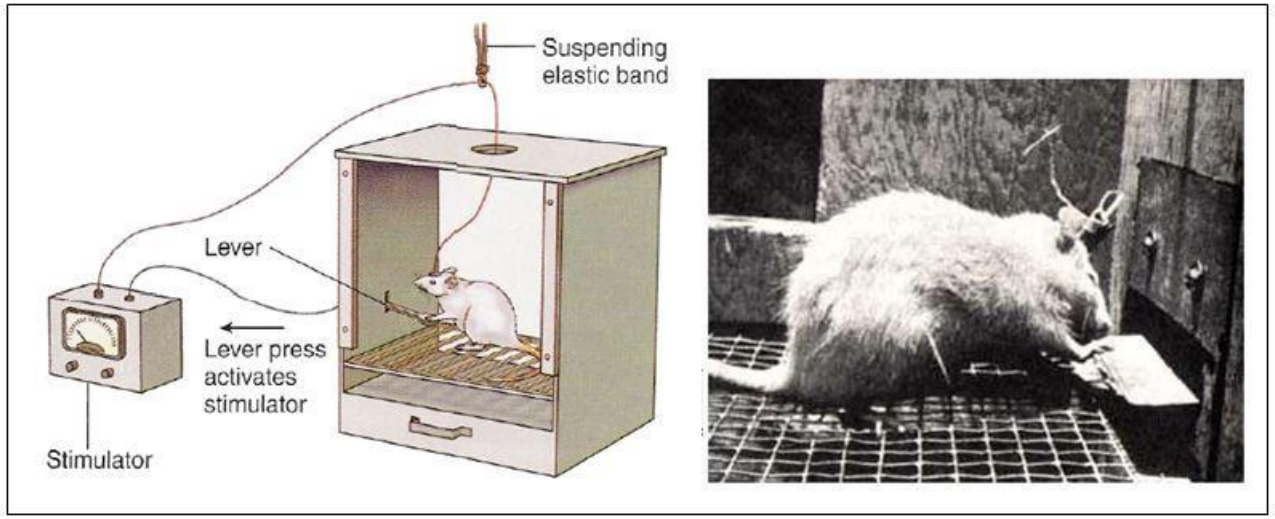
- 作为神经递质的多巴胺，与其它神经递质在传导信号过程中的生化过程，没有本质的不同，但产生多巴胺的细胞与其作用的下游细胞，却往往分布在不同的皮质层。在讲这点之前，我们先介绍一下奖励系统。

  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-in-line 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.

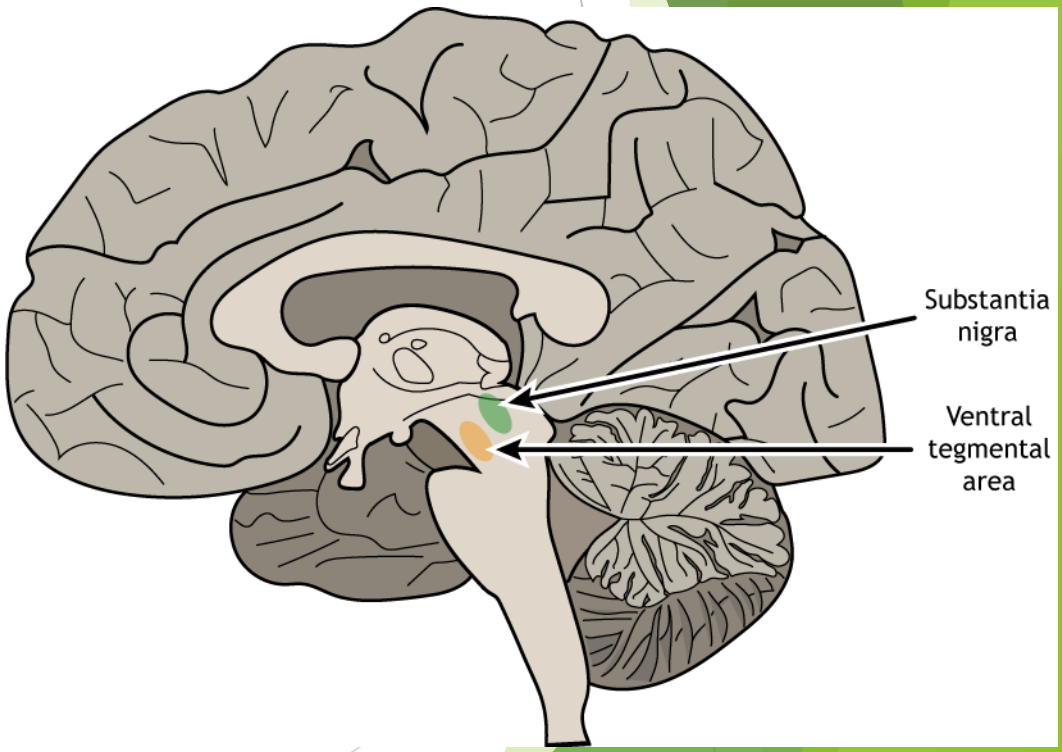
奖励回路
Reward Circuits

- 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).

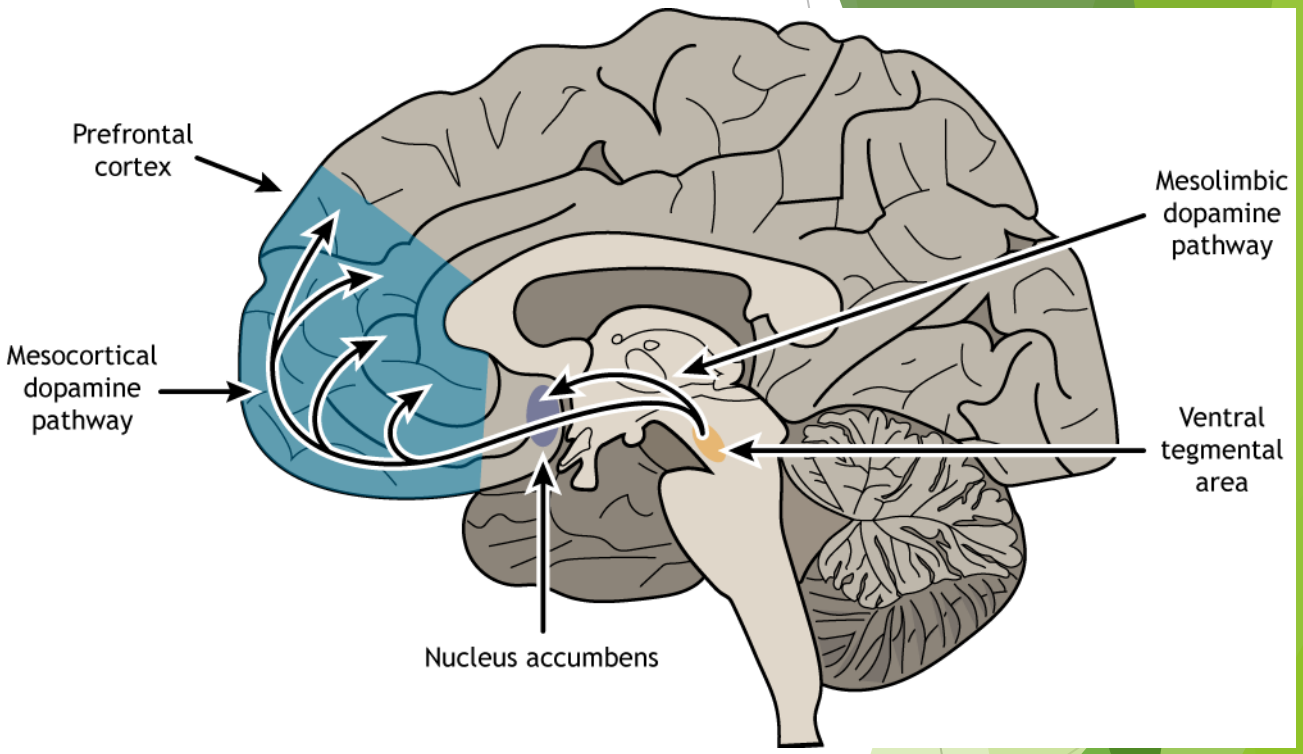
- 如右图所示，腹侧被盖区（橙色区域）位于黑质附近的中脑区域（绿色区域），两个区域都产生多巴胺并将多巴胺释放到下游目标上。

  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.

多巴胺
Dopamine

- 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?

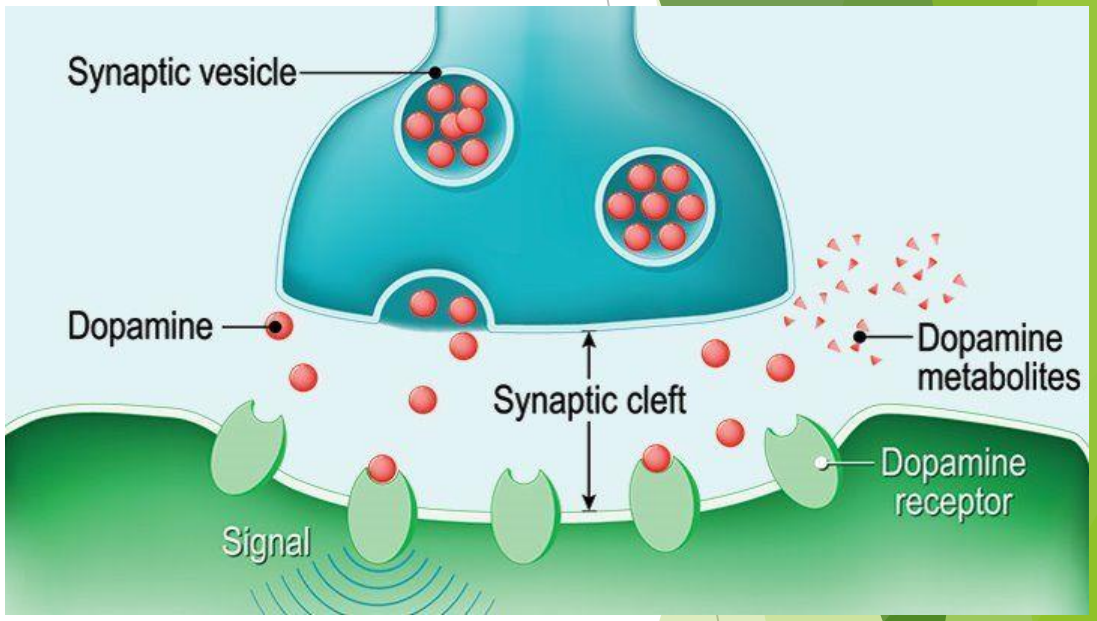
多巴胺
Dopamine

- 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.

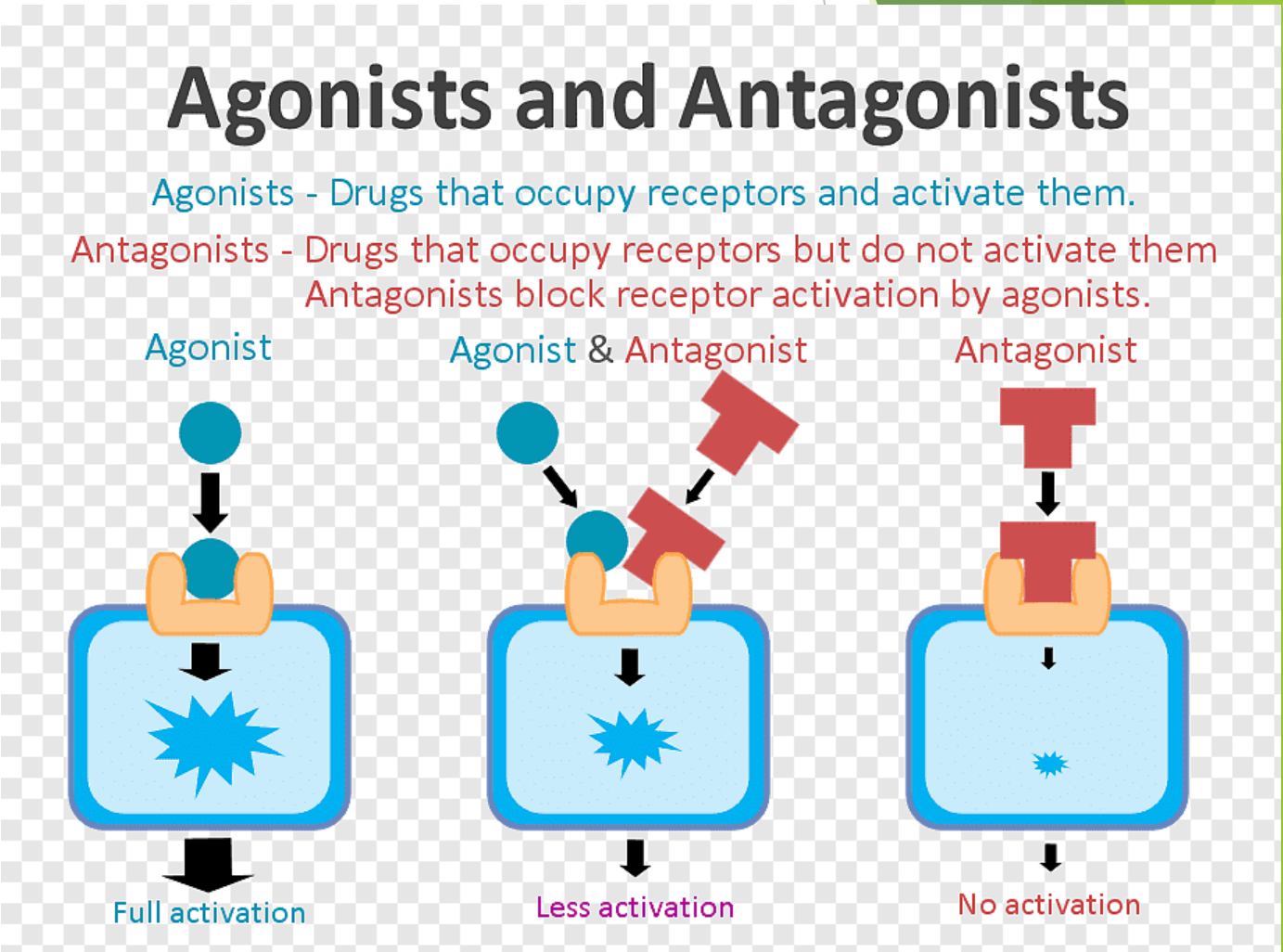
多巴胺
Dopamine

- 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.



多巴胺
Dopamine

- 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.

多巴胺
Dopamine

- 因为老鼠的自我激励与药物成瘾有相似的动机，因此，通过对药物成瘾行为的研究，奖励通路也得到越来越多地了解。研究发现，每当被试使用成瘾药物时，中脑边缘多巴胺通路就会被激活，因此它已被认为是奖励系统的主要通路。因为中脑边缘系统还与重要的内侧皮层和皮层子组织相互作用，因此，扩展的中脑边缘多巴胺网络被认为是负责提供我们大部分的动机和创造性冲动甚至攻击性行为的网络，其受损后最显着的行为损失是缺乏动机驱动。对于中脑边缘系统的腹内侧部分尤其如此，其与伏隔核的内侧壳相互作用。

  The self-stimulation of rats resemble the drug addition, so the research into drag addition benefits the understanding of reword system. Research discovers that whenever an addictive drug is administered, the mesolimbic dopamine pathway is activated, it has come to be considered the primary pathway of the reward system. The mesolimbic system also interacts with important medial cortical and subcortical elements, The extended mesolimbic dopaminergic network is believed to be the one responsible for providing most of our motivational drive and creative impulses and even aggressive behavior, and the most salient behavioral loss following damage to it is a loss of motivational drive. This may be especially true of the ventromedial portion of the mesolimbic system, interacting with the medial shell of the accumbens.

多巴胺
Dopamine

- 同时，中皮层多巴胺通路，即VTA投射到额叶皮层的通路，也被认为与奖励和动机有关。但额叶皮层在越高等的动物如灵长类中才越发达，因此，虽然了解它们对奖励体验有贡献，但不如中脑边缘通路研究的那么清楚。

  Meanwhile, the mesocortical dopamine pathway, aka, dopaminergic projections from the VTA travel to the frontal cortex, are also thought to be involved in reward and motivation. However, frontal lobe only sophisticatedly developed among higher animals such as primates, the mesocortical dopamine pathway is thought contributing to the reward system, however the details are less clear than that of the mesolimbic pathway.

- 另外一个关于多巴胺的重要方面是多巴胺在动物大脑的中分布并不局限于少数皮层区域。特别在低等动物中，多巴胺主要位于包含初级和联合运动皮层的额叶区域以及腹内侧动力中心，如前扣带回和伏隔核，这与多巴胺在目标导向运动活动中的重要作用相称，而这些活动一般为动物适应生存环境的重要活动。

  Another important facet of the dopamine in animal’s brain is that dopamine is not confined to a few cortical areas. Specifically in lower animals, dopamine is located mostly in frontal areas containing primary and associfition motor cortex and in ventromedial motivational centres such as the anterior cingulate and nucleus accumbens, as befits the important role of dopamine in goal-directed motor activities, which are vital for the adaptation of animals to the environment.

多巴胺
Dopamine

- 但是从低等哺乳动物到灵长类动物，多巴胺在皮质层的分布却有不同。

  Another important change is the different distribution of dopamine across cortical layers from lower mammals to primates.

- 需要说明的是，虽然愉悦体验、成瘾行为等和奖励系统有关，但愉悦体验的基础可能并不一定局限于上述结构。同时，多巴胺也不一定是唯一涉及的神经递质。奖励系统是指一组经常参与调解奖励体验的结构，但致力于创造我们与这些体验相关联的感觉的实际网络可能更为复杂。

  Notably, while the reward system is implicated in pleasurable and addictive behaviors, the substrates of pleasure are not confined to the structures mentioned above and dopamine is not the only neurotransmitter involved. The reward system refers to a group of structures that are frequently involved in mediating rewarding experiences, but the actual network dedicated to creating the feelings we associate with these experiences is likely more complex.

多巴胺
Dopamine

- 除了执行实际的物理动作，在对奖励系统的研究中，还可以探究具有倾向性的精神活动，如自我驱动的学习与记忆。

  In addition to performing the physical actions, the study of reward systems can also explore mental activities with inclinations, such as self-driven learning and memory.

- 研究发现，影响情绪性行动及记忆的不是奖励本身，而是对奖励的期望。当遇到意想不到的事情时，即当实际奖励与我们原本预期的不同时，就会发生奖励性学习行为。如果奖励大于预期，多巴胺信号就会增加。如果奖励低于预期，多巴胺信号就会减少。相反，正确预测奖励不会改变多巴胺信号，因为我们没有学到任何新东西。

  Research discovered that it’s not the reward itself, but the expectation of a reward that most powerfully influences emotional reactions and memories. When we experience something unexpected, for example, the actual reward differs from what we otherwise would predict, Reward learning occurs. If a reward is greater than anticipated, dopamine signaling increases. If a reward is less than expected, dopamine signaling decreases. In contrast, correctly predicting a reward does not alter dopamine signaling because we aren’t learning anything new.

多巴胺
Dopamine

- 因为正确预测奖励不会改变多巴胺信号，因此我们更多关注当预测有偏差的情况。我们通常所说的奖励，可能是指奖励大于预期的情况；如果奖励低于预期，则我们也有可能说是惩罚。具体所指可能要放在具体上下文中。

  Because the correct prediction of a reward does not alter dopamine signaling, so we pay more attention to the biased prediction. The general mentioned reward probably means the case that a reward is greater than anticipated. If a reward is less than expected, we might say punishment. However, the exact meaning should be clear in the context.

- 研究发现多巴胺反应因人而异，有些人的大脑对奖励的反应比对惩罚的反应更强烈，而另一些人则相反。同时研究发现，奖励学习和动机受到杏仁核的影响较强烈。研究人员发现，更愿意努力工作的“积极分子”在纹状体和前额叶皮层中的多巴胺信号更强—这两个区域已知会影响动机和奖励。

  Research unveils that dopamine responses vary from person to person. Some people’s brains respond more strongly to rewards than punishments, while verse versa for others. Reward learning and motivation are strongly influenced by the amygdala. Researchers found that “go-getters” who are more willing to work hard have greater dopamine signaling in the striatum and prefrontal cortex — two areas known to impact motivation and reward.

决策
Decision-making

- 决策在心理学中被认为是一种认知过程，其是在几个可能的替代选项中选择一个信念或一个行动的过程。决策过程是决策者基于其价值观、偏好和信念的假设的推理过程。每个决策过程都会产生一个最终选择，这可能会或可能不会促使采取行动。

  decision-making in psychology is regarded as the cognitive process resulting in the selection of a belief or a course of action among several possible alternative options. Decision-making process is a reasoning process based on assumptions of values, preferences and beliefs of the decision-maker. Every decision-making process produces a final choice, which may or may not prompt action.

- 决策通常涉及评估风险和回报，神经科学家正在研究大脑如何平衡奖励和风险，包括一个人的情绪状态如何影响这种平衡。

  Decision-making often involves evaluating risks in addition to rewards. Neuroscientists are investigating how the brain balances reward and risk, such as how one’s emotional state affects this balance.

决策
Decision-making

- 大脑的奖励系统会强化与奖励相关的行为，并防止导致惩罚的行为。

  The brain’s reward system reinforces behaviours associated with rewards and prevents behaviours leading to punishment.

- 基于决策做出的动作，在第三者看来，可能是理性的，也可能是非理性的。对于当事人来说，由于决策基于其在当时环境下本人的价值观，决策者一般只会看到该决策导致的其认同的奖励。注意，这并非在为当特定的个体违反公知良序的行为决策做生物学意义上的辩护，而只是说明奖励（特别是在特定环境下）对决策的影响。

  An action based on a subject’s decision may appear to be rational or irrational for third one. However, the decision maker do it based on his own values in the context and generally only sees the intended reward as the consequence of the chose action. Note that this is not a biological defence for decision-making when anyone commits a wrong-doing, but just to illustrate the impact of rewards (especially in specific circumstances) on decision-making.