Porte d'entrée	Porte d'oubli	Porte de sortie	Porte
Écrire ?	Supprimer ?	A quel point révéler ?	Combien écrire ?

❒ **LSTM** –Un réseau de long court terme (en anglais *long sort-term memory*, LSTM) est un type de modèle RNN qui empêche le phénomène de *vanishing gradient* en ajoutant des portes d'oubli.

3.4 Reinforcement Learning

Le but du reinforcement learning est pour un agent d'apprendre comment évoluer dans un environnement.

❒ **Processus de décision markovien** – Un processus de décision markovien (MDP) est décrite par 5 quantités $(S,A,\{P_{sa}\},\gamma,R)$, où :

- $\mathcal{S}$ est l'ensemble des états
- $\mathcal{A}$ est l'ensemble des actions
- $\{P_{sa}\}$ sont les probabilités d'états de transition pour $s \in \mathcal{S}$ et $a \in \mathcal{A}$
- $\gamma \in [0,1[$ est le taux d'actualisation (en anglais *discount factor*)
- $R : \mathcal{S} \times \mathcal{A} \longrightarrow \mathbb{R}$ ou $R : \mathcal{S} \longrightarrow \mathbb{R}$ est la fonction de récompense que l'algorithme veut maximiser

❒ **Politique** – Une politique π est une fonction $\pi : \mathcal{S} \longrightarrow \mathcal{A}$ qui lie les états aux actions.

Remarque : on dit que l'on effectue une politique donnée π si étant donné un état s, on prend l'action $a = \pi(s)$.

❒ **Fonction de valeurs** – Pour une politique donnée π et un état donné s, on définit la fonction de valeurs V^π comme suit :

$$\boxed{V^\pi(s) = E\Big[R(s_0) + \gamma R(s_1) + \gamma^2 R(s_2) + ...|s_0 = s,\pi\Big]}$$

❒ **Équation de Bellman** – Les équations de Bellman optimales caractérisent la fonction de valeurs V^{π^*} de la politique optimale π^* :

$$\boxed{V^{\pi^*}(s) = R(s) + \max_{a \in \mathcal{A}} \gamma \sum_{s' \in S} P_{sa}(s') V^{\pi^*}(s')}$$

Remarque : on note que la politique optimale π^* pour un état donné s est tel que :

$$\boxed{\pi^*(s) = \operatorname*{argmax}_{a \in \mathcal{A}} \sum_{s' \in \mathcal{S}} P_{sa}(s') V^*(s')}$$

❒ **Algorithme d'itération sur la valeur** – L'algorithme d'itération sur la valeur est faite de deux étapes :

— On initialise la valeur :

$$\boxed{V_0(s) = 0}$$