

Un algorithme « bandit manchot » pour le choix de nouvelles situations d'apprentissage à l'intérieur d'un environnement virtuel

Yannick Bourrier^{1,2}, Julien Teigny^{1,2}, Francis Jambon², Catherine Garbay² & Vanda Luengo¹

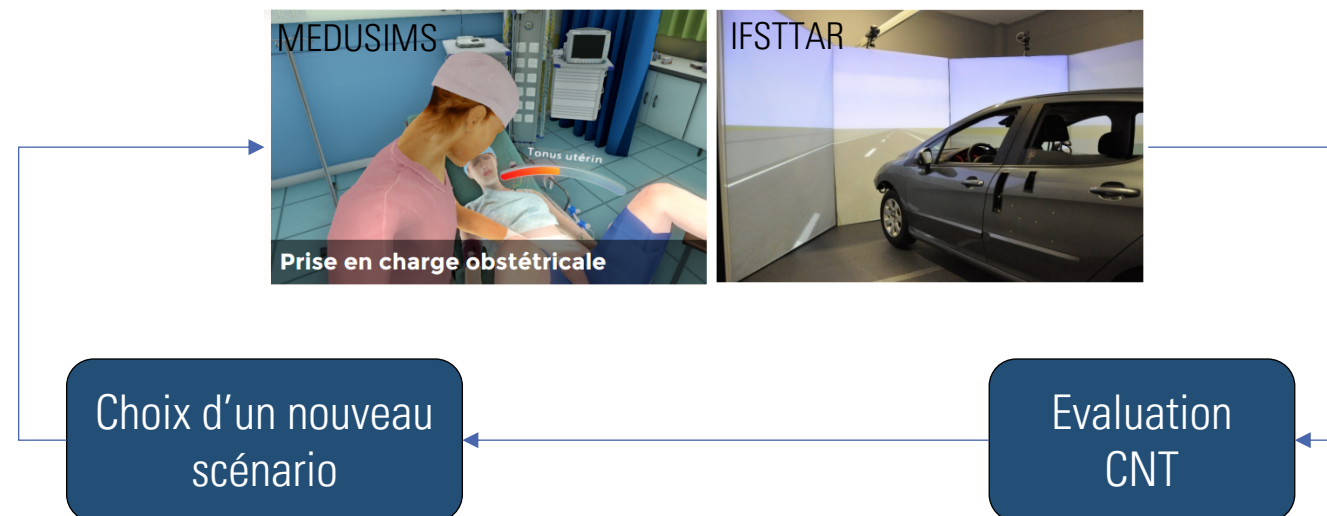
¹Sorbonne Universités, Université Pierre & Marie Curie, LIP6 équipe MOCAH

²Univ. Grenoble Alpes, CNRS, Grenoble INP, LIG, 38000 Grenoble, France



Présentation : « MacCoy Critical »

- **Projet MacCoy Critical** : un environnement virtuel pour la formation d'experts aux Compétences Non-Techniques (ANR-14-CE24-0021).
- **Objectif** : évaluer les compétences non-techniques d'un apprenant, puis générer dynamiquement des situations critiques adaptées à son niveau.
- **Domaines d'application** : la conduite automobile des jeunes conducteurs et la gestion de cas d'hémorragies post-partum par des sages-femmes.



Présentation : CNT

Compétences Non-Techniques (CNT)

- Ensemble des capacités qui complètent les compétences techniques lors d'une activité technique. (Flin *et al.*, 2010)
 - Exemples : Prise de décision, conscience de la situation, gestion du stress, travail en équipe...
- A l'origine d'un nombre important d'accidents.
 - En chirurgie, 70% des erreurs seraient causées par des CNT. (Mitchell *et al.*, 2012)

Présentation : problématique

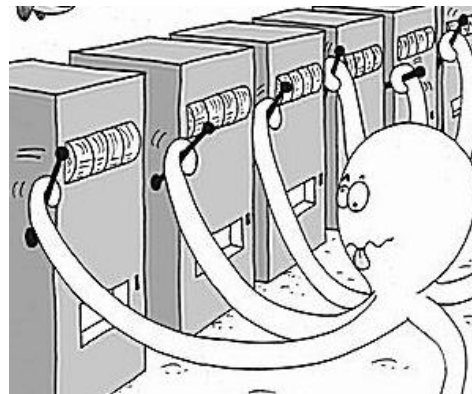
Comment choisir le prochain scénario pour optimiser l'amélioration des CNT ?

- Utilisation d'un expert qui suit l'évolution de l'apprenant, et propose le scénario suivant.
 - Problème : Avoir un expert toujours présent, et beaucoup de critères à gérer.
- Suite de scénarii adaptée pour un apprenant « moyen » (séquence experte), donnée à tous les apprenants.
 - Problème : Aucune personnalisation selon les spécificités de l'apprenant.
- Sélection aléatoire de scénario dans une Zone Proximale de Développement.
 - Problème : Aucune personnalisation selon les spécificités de l'apprenant.
- Utilisation d'un algorithme pour sélectionner le scénario adapté à l'apprenant.
 - Exemple : Bandit manchot.

Le « bandit manchot »

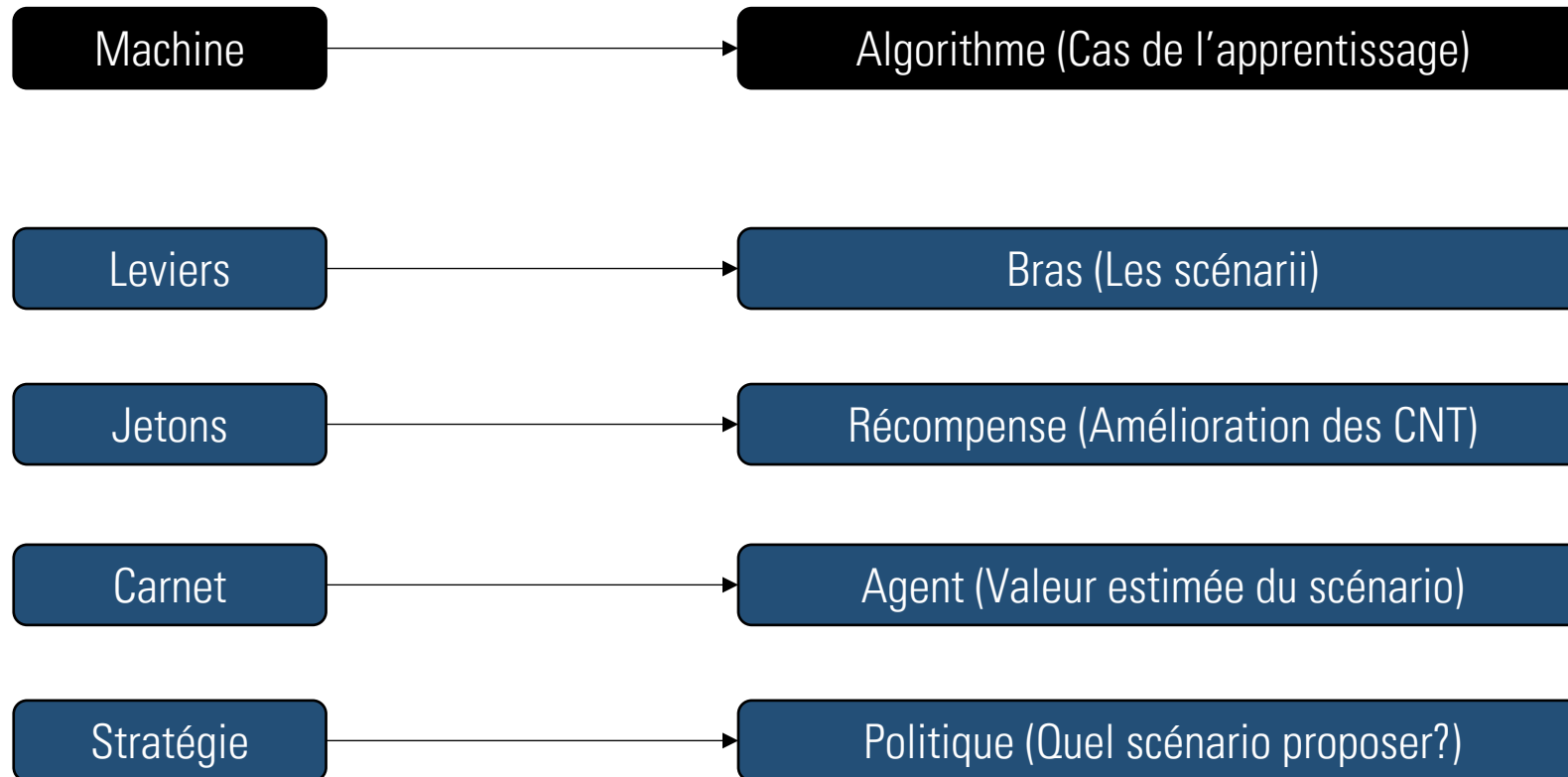
Les principes

- Technique d'apprentissage par renforcement (l'algorithme apprend en pratiquant).
- Le nom provient de l'analogie avec les machines à sous des casinos.
- Dans notre situation :
On a des scénarii avec des caractéristiques différentes. Quel scénario choisir pour maximiser l'amélioration des Compétences Non-Techniques de l'apprenant ?



Le « bandit manchot »

Le Vocabulaire



Le « bandit manchot »

Les bras

- Chaque bras représente un scénario disponible.
- À chaque étape d'apprentissage, l'algorithme (en s'appuyant sur son agent et sa politique) sélectionne un des bras. On fait passer le participant sur le scénario lié à ce bras.

Le « bandit manchot »

La récompense

- Ce que l'on cherche à maximiser : dans notre cas, l'amélioration des CNT de l'apprenant.
- Quand le participant a terminé son scénario, on regarde comment ont évolué ses CNT. Ces données seront utilisées par l'agent pour se mettre à jour.
- Il existe plusieurs moyens de calculer la récompense selon ce qu'on veut maximiser : on peut sommer toutes les améliorations, ajouter des coefficients si on veut maximiser une CNT plus qu'une autre, ...

$$\sum_{k=0}^n coef_k * [NiveauObservéCNT_k(t) - NiveauObservéCNT_k(t - 1)]$$

Où n est le nombre de CNT qu'on cherche à améliorer

Le « bandit manchot »

L'Agent

- C'est lui qui observe les récompenses reçues et qui donne une « valeur » aux différents bras.
- Analogie avec le casino : Carnet sur lequel est marqué :
 - Machine A : récompense moyenne de 5 jetons
 - Machine B : récompense moyenne de 2 jetons
 - ...
- Cette valeur permet d'estimer l'utilité d'un scénario selon les améliorations de CNT qu'il a déjà fourni. Cette valeur peut être calculée de différentes manières :
 - $\text{Nouvelle_valeur} = \text{Valeur_précédente} + (\text{Récompense} - \text{Valeur_précédente}) / \text{Nombre_sélection_scénario}$
 - $\text{Nouvelle_valeur} = \text{Valeur_précédente} * 75\% + \text{Récompense} * 25\%$
 - Modification des valeurs de chaque bras (Si récompense > moyenne, on augmente la valeur du bras, et diminue toutes les autres. Sinon on diminue la valeur du bras, et augmente toutes les autres.)

Le « bandit manchot »

La politique

- C'est elle qui détermine comment choisir le prochain bras à sélectionner.
- Pour cela la « politique » utilise les valeurs que l'agent a donné à chaque bras.
- Elle permet de gérer la dualité Exploitation VS Exploration.
 - Exploitation : On utilise le bras qui rapporte le plus de récompense selon l'agent.
 - Exploration : On teste un bras au hasard pour voir s'il donne une meilleure récompense.
- Il existe plusieurs politiques possibles:

Exemple1 (Epsilon Greedy) :

- Sélection du bras avec la meilleure valeur dans 90% des cas.
- Sélection d'un bras aléatoire dans 10% des cas.

Exemple2 (SoftMax) :

- Choix aléatoire d'un bras avec une probabilité qui est liée à sa valeur normalisée donnée par l'agent.

Les limites

- **Limite Intrinsèque**

Comme il s'agit d'un algorithme d'« apprentissage par renforcement », l'algorithme a besoin d'une période d'essai avant d'être efficace : il s'améliore en pratiquant.

- **Conséquence**

Les premiers tours de l'apprentissage sont moins optimisés que les derniers.

- **Comment gérer cette limite ?**

- Diminuer le nombre de bras : En utilisant une Zone Proximale de Développement, on filtre les scénarii les plus pertinents, ce qui limite le nombre de bras accessibles.
- Déterminer les valeurs initiales des bras (l'agent) : En initiant les valeurs correctement, le renforcement de l'algorithme est plus rapide.
- Paramétrer l'agent, la politique et le calcul de récompense : Faire de nombreux tests. En utilisant une simulation d'apprenant (Item response theory), on peut faire de très nombreuses sessions d'apprentissage, ce qui permet de paramétrer l'algorithme plus finement.

Les limites

- **Limite spécifique**

Les compétences de l'apprenant évoluent au cours de l'apprentissage.

- **Conséquence**

Les récompenses évoluent également (un scénario qui est efficace lorsque l'apprenant débute devient moins efficace lorsque l'apprenant a amélioré ses CNT).

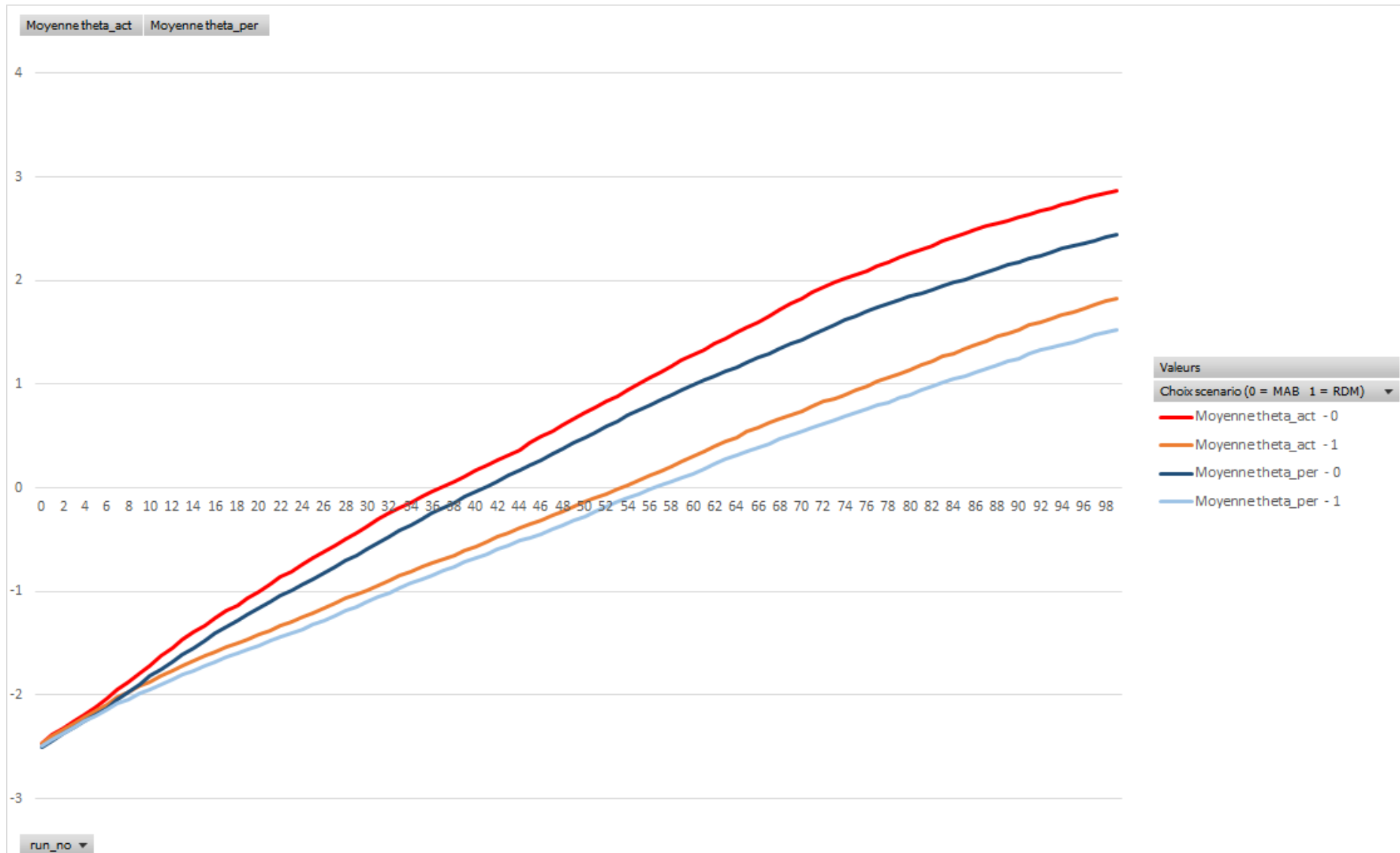
- **Comment gérer cette limite ?**

- Utilisation d'un bandit manchot par niveau de maîtrise de CNT (abandonné car il y a perte de la personnalisation selon les spécificités de l'apprenant).
- Utilisation d'un agent à mémoire courte: permet d'oublier les anciennes récompenses.
(de la forme: $\text{Nouvelle_valeur} = 50\% \text{ Ancienne_valeur} + 50\% \text{ Récompense}$)
- Utilisation d'une politique probabiliste (exemple softMax): permet une adaptation plus rapide.
(modifie toutes les valeurs à chaque mise à jour)

- **Les difficultés**

- Si l'évolution de l'apprenant est très rapide, l'algorithme n'a pas le temps de s'adapter.
- Si l'apprenant réussit par chance (ou échoue par erreur), l'adaptation rapide peut poser problème.

Résultats (Conduite automobile)



Conclusion

- **Les points positifs du bandit manchot**

- Absence d'expert : Pas d'obligation à disposer d'un expert humain.
- Adaptabilité : S'adapte à chaque apprenant avec ses caractéristiques différentes.
- Performance : Si les récompenses sont faciles à déterminer (ie : les compétences de l'apprenant sont évaluées avec précision), le bandit est facile à paramétrer.

- **Les points négatifs du bandit manchot**

- Rapidité d'adaptation : Si le niveau de l'apprenant varie trop vite, le bandit ne peut pas s'adapter assez rapidement.
- Performance : Si les récompenses sont difficiles à déterminer (ie : les compétences de l'apprenant sont évaluées avec un degré d'incertitude), la fonction de récompense est plus difficile à déterminer.

- **Bilan**

- Si les conditions précédentes sont réunies et que les paramétrages sont bien faits, l'algorithme du bandit manchot permet de personnaliser l'apprentissage à chaque apprenant.

Conclusion

Merci de votre attention!

Questions?