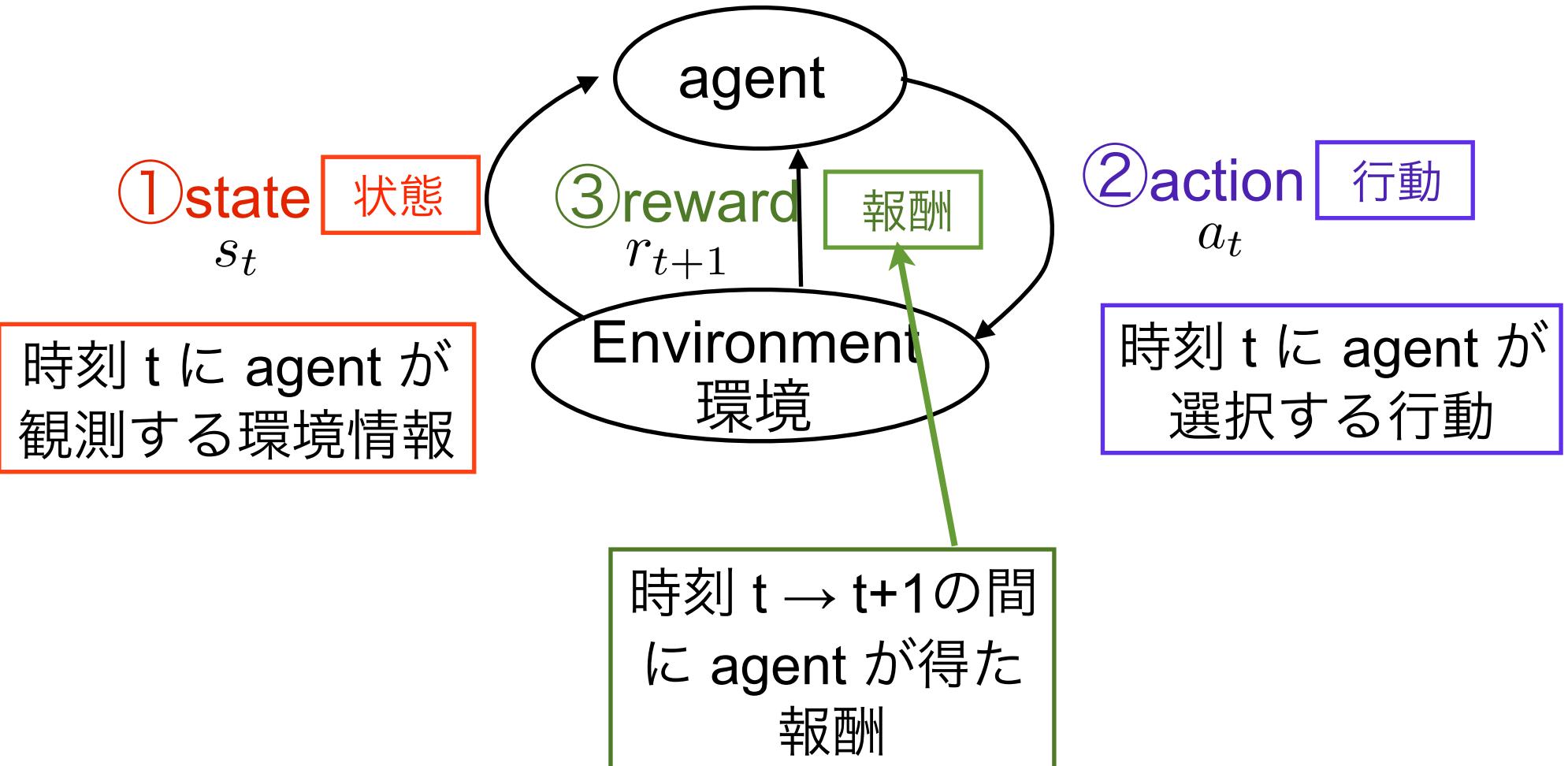
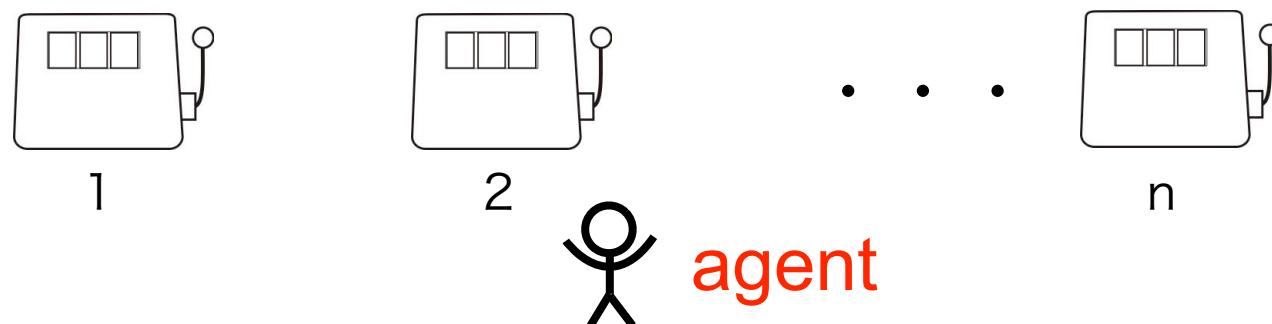


強化学習

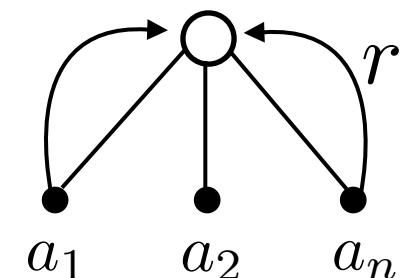
強化学習の基本用語



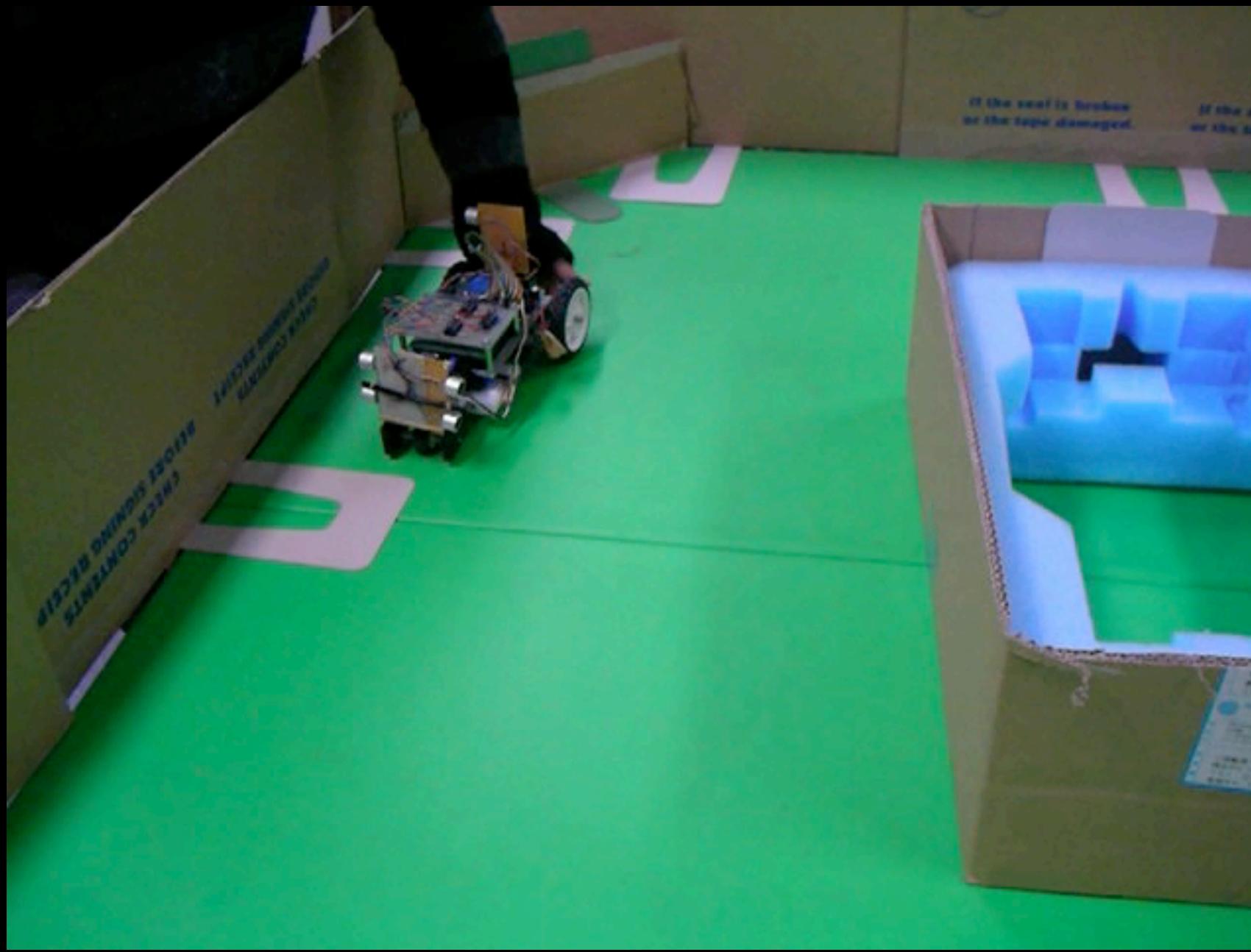
n-bandit 問題の場合



- **state** … 每回同じ（目の前に n台のマシン） s
- **action** … どれを試すか？
- **reward** … 出たコインの数



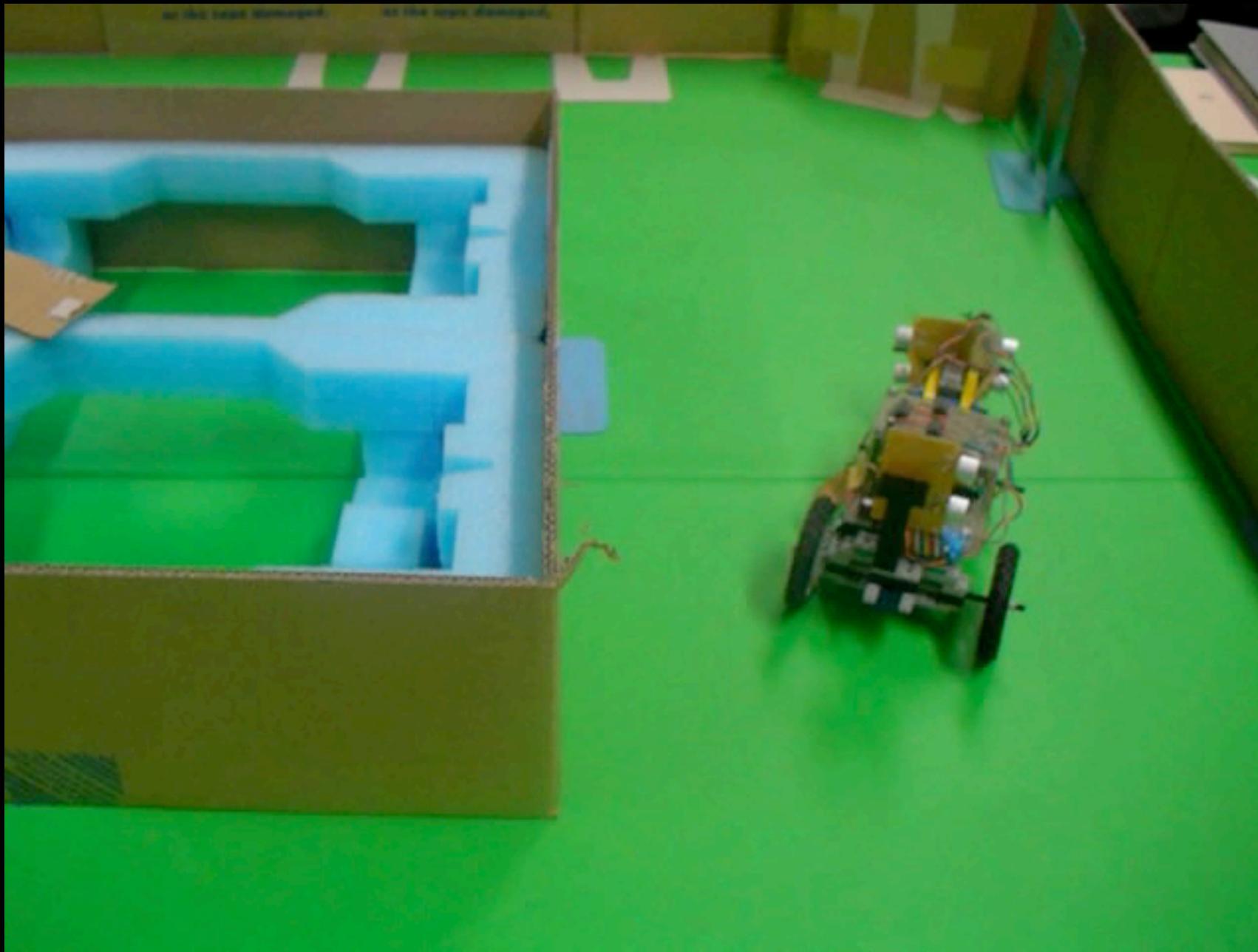
应用例



学習前

渡辺研究室(北大)

<http://hbd.ist.hokudai.ac.jp/wata/actorcritic01.html>



学習後

渡辺研究室(北大)

<http://hbd.ist.hokudai.ac.jp/wata/actorcritic01.html>

Microsoft研究所

<http://research.microsoft.com/en-us/projects/mlgames2008/>

Google DeepMind、AlphaGoに圧勝（100勝0敗）する新たな囲碁AIプログラム「AlphaGo Zero」を発表。囲碁の基礎ルールのみ教え3日間で500万回強化学習

2017.10.19 | AI, 論文

<http://shiropen.com/2017/10/19/28760>

Google DeepMindは、囲碁の世界トッププロ棋士を破ってきたコンピュータ囲碁AIプログラム「AlphaGo」に圧勝する新たな人工知能プログラム「AlphaGo Zero」を論文にて発表しました。

[Mastering the game of Go without human knowledge \(PDF\)](#)



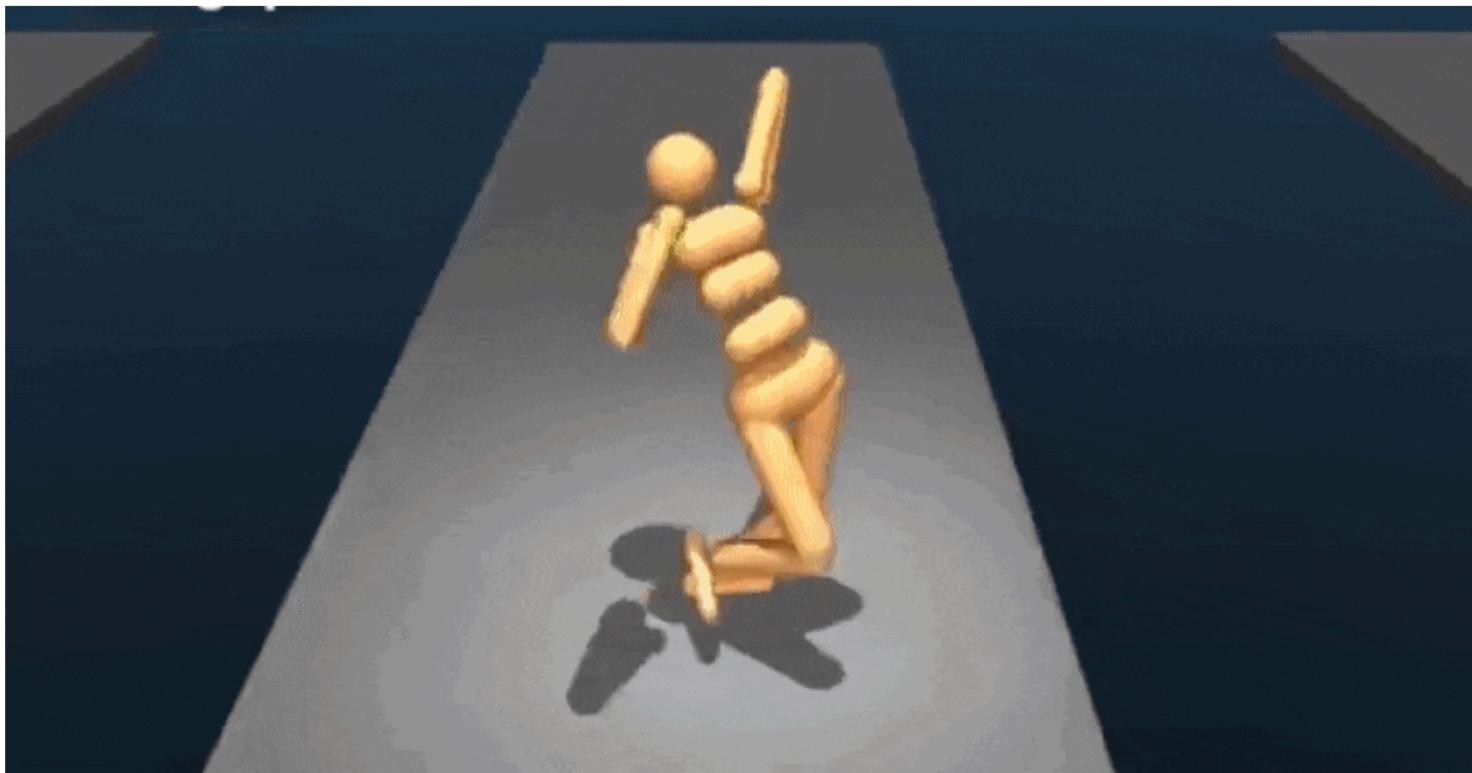
Google「DeepMind」、コンピュータが人型ベースでB地点にたどり着く最善の方法（柔軟な動き）を独学で生成する強化学習を用いたアプローチを提案した論文を発表

2017.07.11 | AI, Simulation, 論文

<http://shiropen.com/2017/07/11/26572>

GoogleのAIを研究する子会社「DeepMind」は、強化学習で人型含めシミュレートされた環境の中で複雑で柔軟な動きを生成するアプローチを提案した論文を公開しました。

[Emergence of Locomotion Behaviours in Rich Environments \(PDF\)](#)



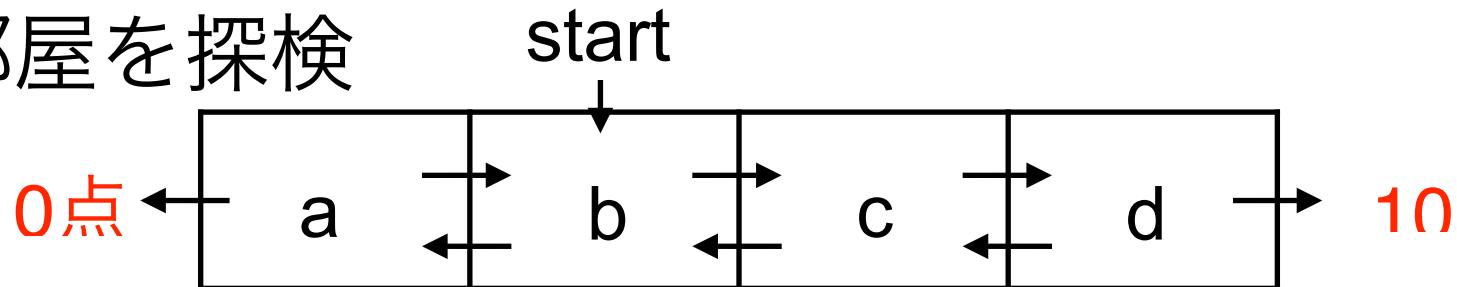
It might
look goofy ...



[Google's DeepMind AI Just Taught Itself To Walk](#)

強化学習の基本用語

ex) 部屋を探検



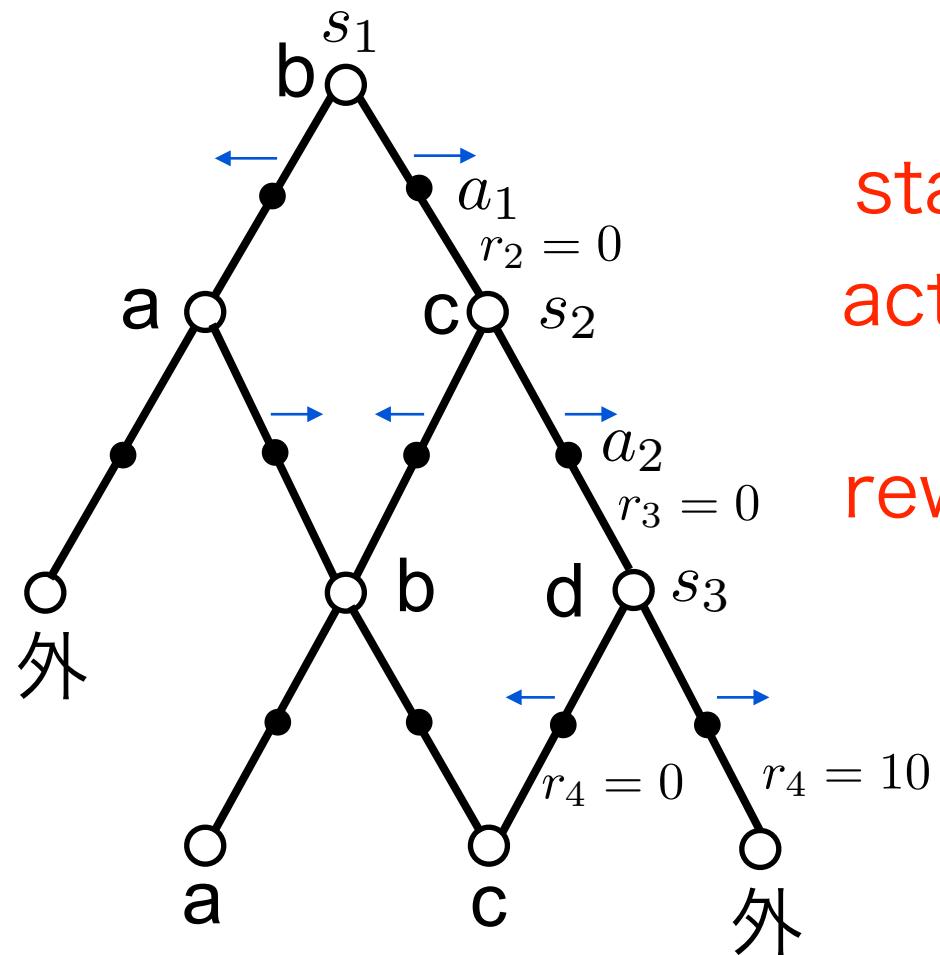
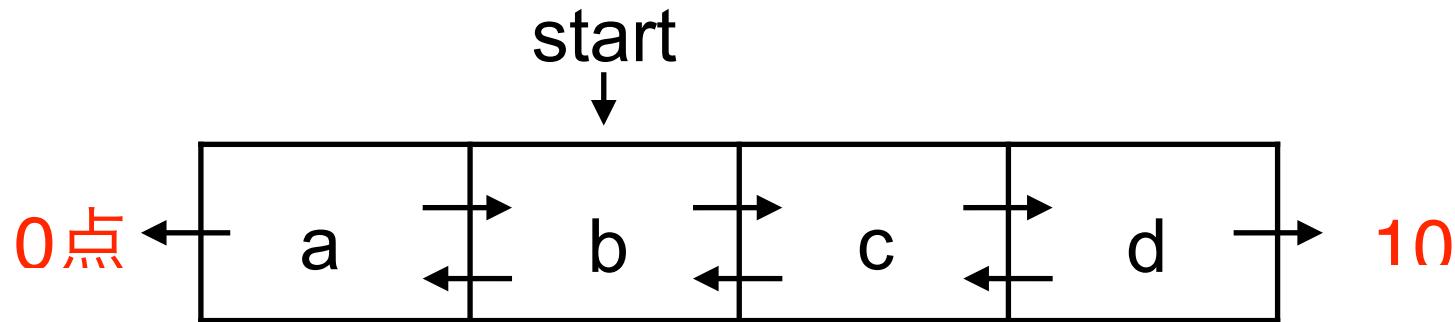
ルール

- ・エージェントは始めどこかの部屋に入れられる
- ・エージェントは現在の部屋の名前を知ることが出来る
- ・エージェントは各時間ステップごとに隣の部屋に移動
- ・エージェントはどうすれば報酬を貰えるかは知らない
- ・dから外に出たら報酬10点
- ・それ以外のときは報酬0点
- ・外に出たらゲーム終了

エピソードという

(ゲーム開始から終わりまで… 1 episode)

状態, 行動, 報酬



state … 部屋番号 $s_t \in \{a, b, c, d\}$

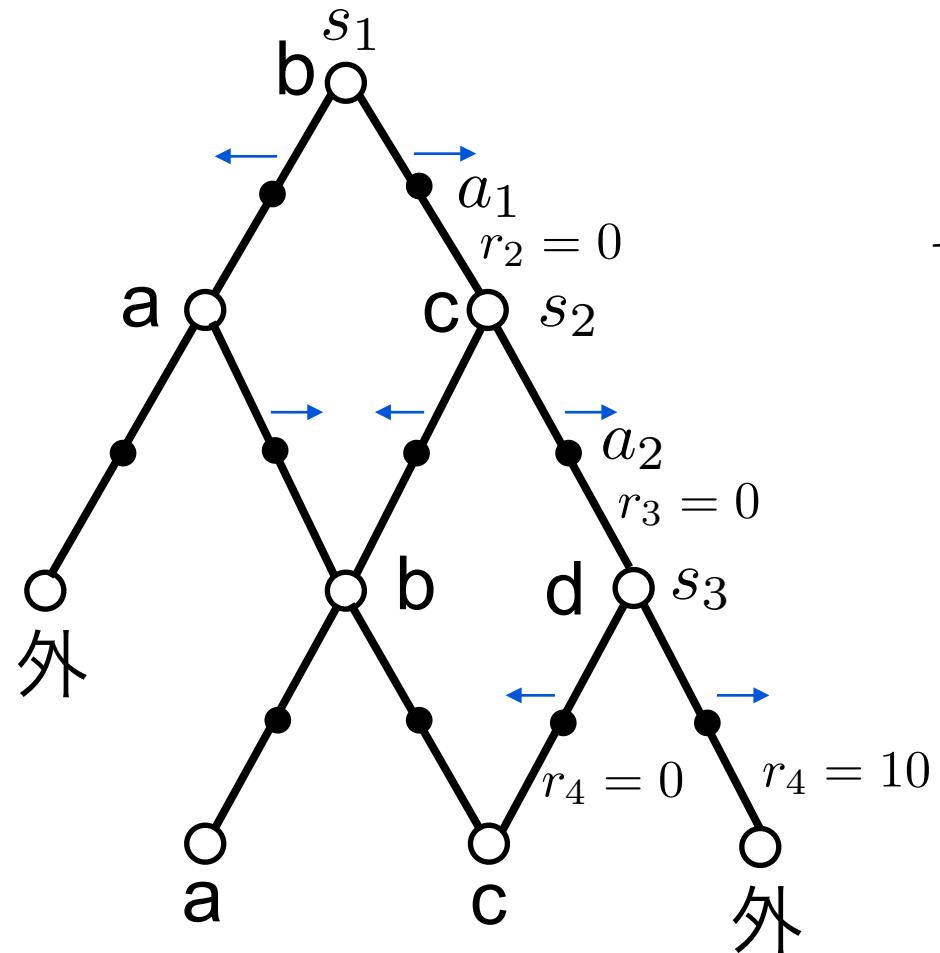
action … どちらのドアから出るか
 $a_t \in \{\rightarrow, \leftarrow\}$

reward … 得点 $r_t \in \{0, 10\}$

強化学習の目的

総報酬

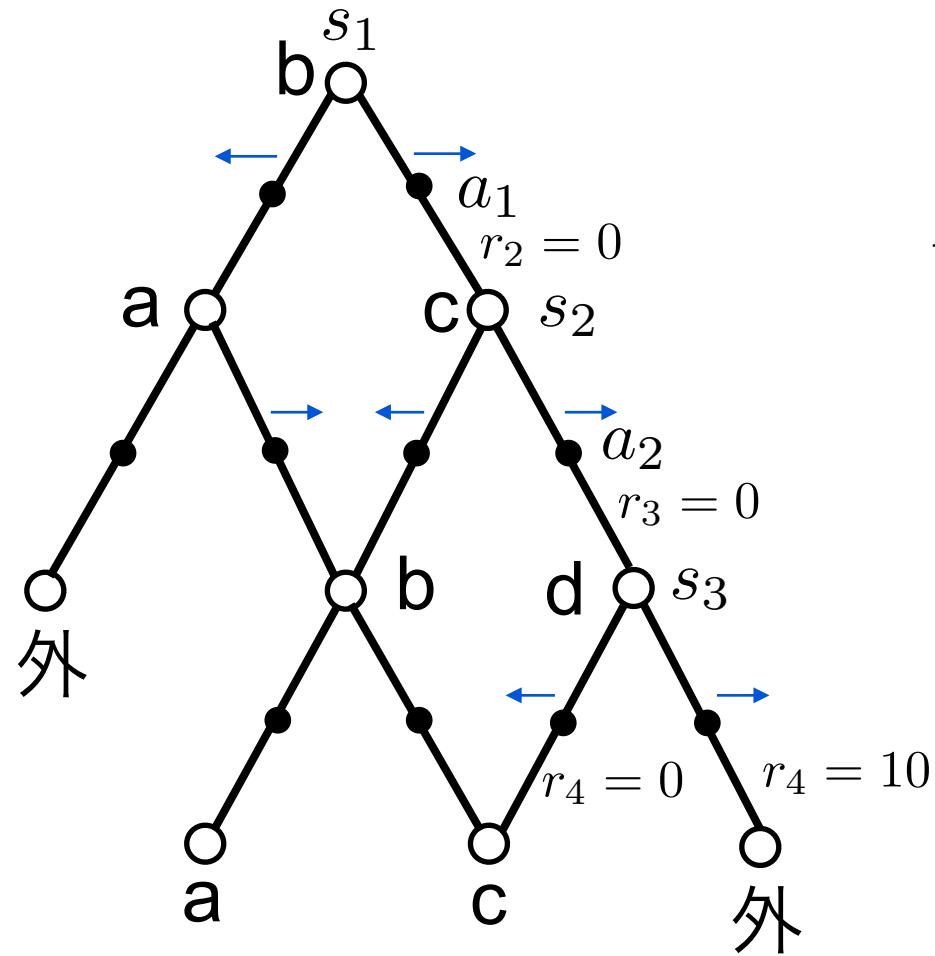
現在の時刻 t 以降の ? (Return) R_t の期待値を最大にする, 状態 s_t に応じた action a_t を見つける!!



$$\begin{aligned} R_t &= \sum_{i=0}^{\infty} r_{t+i+1} \\ &= r_{t+1} + r_{t+2} + \dots \end{aligned}$$

強化学習の目的

現在の時刻 t 以降の総報酬(Return) R_t の期待値を最大にする, 状態 s_t に応じた action a_t を見つける!!



$$R_t = \sum_{i=0}^{\infty} r_{t+i+1}$$

$$= r_{t+1} + r_{t+2} + \dots$$

割引率

Returnの計算式の問題点:

$$R_t = \sum_{i=0}^{\infty} r_{t+i+1} = r_{t+1} + r_{t+2} + \dots$$

エピソードに終わりがないタスクでは値が ∞ に!
ex) お掃除ロボット, サバイバルロボット

? を導入

- 遠い未来の報酬ほどその価値を低く見積もる
- 報酬(Return)の式を修正

割引率

Returnの計算式の問題点:

$$R_t = \sum_{i=0}^{\infty} r_{t+i+1} = r_{t+1} + r_{t+2} + \dots$$

エピソードに終わりがないタスクでは値が ∞ に!
ex) お掃除ロボット, サバイバルロボット

割引率 γ を導入

- 遠い未来の報酬ほどその価値を低く見積もる
- 報酬(Return)の式を修正

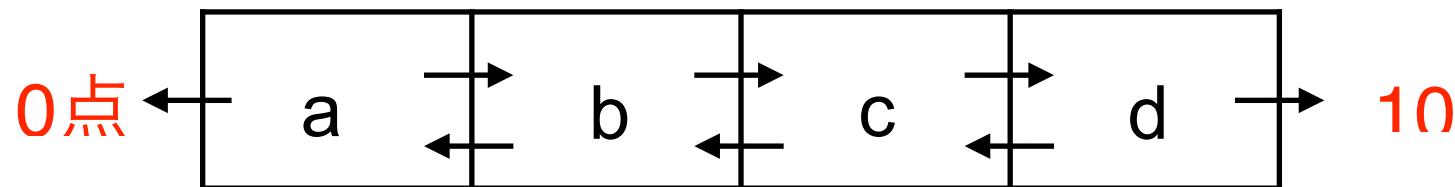
割引率を考慮した総報酬

$$R_t = \boxed{?}$$

$$(0 \leq \gamma \leq 1)$$

生物は一般に遠い未来の報酬(or コスト)ほど低く見積もる。

例1) 部屋 b と d どちらがいい？



例2) すぐもらえる1000円と10年後の1000円では？

例3) 金融崩壊の根本原因は「人間の本能」

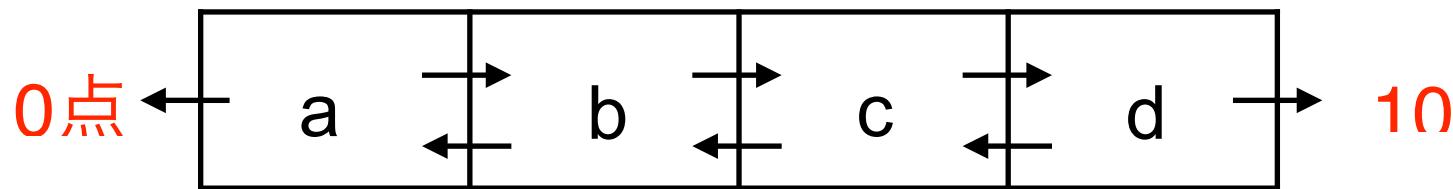
(今日の支払いより明日の支払い！)

割引率を考慮した総報酬

$$R_t = r_{t+1} + \gamma r_{t+2} + \gamma^2 r_{t+3} + \dots = \sum_{i=0}^{\infty} \gamma^i r_{t+i+1}$$
$$(0 \leq \gamma \leq 1)$$

生物は一般に遠い未来の報酬(or コスト)ほど低く見積もる。

例1) 部屋 b と d どちらがいい？



例2) すぐもらえる1000円と10年後の1000円では？

例3) 金融崩壊の根本原因は「人間の本能」

(今日の支払いより明日の支払い！)

ノーベル経済学賞セイラーと、「合理的経済人」じゃない僕たち

2017年10月18日（水）16時40分

いいね！ 22

シェア

ツイート 91

ブックマーク 13



今年のノーベル経済学賞に決まった行動経済学の権威リチャード・セイラー Kamil Krzaczynski-REUTERS

＜イギリスで年金加入率が急増したのは、今年のノーベル経済学賞を受賞したセイラーの「ナッジ理論」を適用したから＞

<http://diamond.jp/articles/-/145917>

強化学習における問題

自分が取る行動 a_t に応じた今後の総報酬 R_t がわかれれば、それを最大にする行動 a_t を選択すれば良い。

問題

今後の総報酬 R_t をどうやって見積もる？