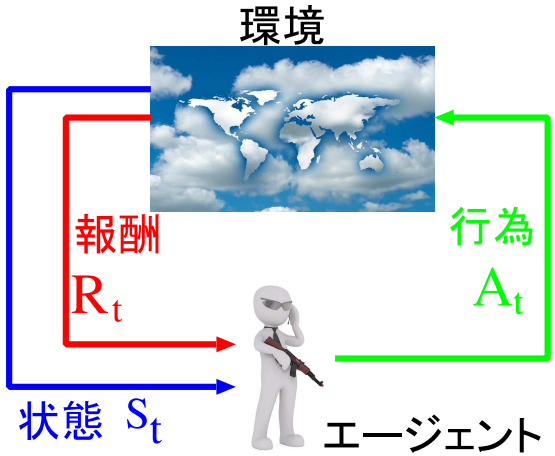
5 エージェントと環境



画像 https://pixabay.com/en/globe-clouds-sky-background-earth-3382522/, https://pixabay.com/en/weapon-guard-soldier-protection-1816313/

- **エージェント**: 学習と意思決定を行う主体
    1. ***行動*** *action* A_t *を行い*
    2. *環境の* ***観察*** *observation* O_t *を行う*
    3. *環境からスカラ値の* ***報酬*** *reward* R_t *を受け取る*
- **環境**: エージェント外部の全て
    1. *エージェントから* ***行為*** A_t *を受け取り*
    2. *エージェントに* ***観察*** O_{t+1} *を与え*
    3. *エージェントへ* ***報酬*** R_{t+1} *を与える*

6 エージェントの要素

- **方策** Policy
- **価値関数** Value function
- **モデル** エージェントが持つ環境の表象

7 方策 policy

- **方策** : エージェントの行為
- 決定論的方策: $a = \pi(S)$
- 確率論的方策: $\pi(a|s) = p(A_t = a|S_t = s)$

8 価値関数

- 将来の報酬予測
- 状態評価(良/悪)
- 行為の選択

$$v_{\pi}(S) = \quad _{\pi} \ R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \ldots | S_t = s \qquad (1)$$

9 強化学習のモデル

- 価値ベース
  - *方策:なし*
  - *価値関数:あり*
- 方策ベース
  - *方策:あり*
  - *価値関数:なし*
- アクター=クリティック Actor Critic
  - *方策: あり*
  - *価値関数: あり*
- モデルフリー
  - *方策，価値関数: あり*
  - *モデル: なし*
- モデルベース
  - *方策，価値関数: あり*
  - *モデル: あり*

10 探索と利用のジレンマ Exploration and exploitaion dilemma

- 過去の経験から，一番良いと思う行動ばかりをしていると，さらに良い選択肢を見つけ出すことができない **探索不足**
- 更に良い選択肢ばかり探していると過去の経験が活かせない **過去の経験の利用不足**

文献

Lake, Brenden M., Tomer D. Ullman, Joshua B. Tenenbaum, and Samuel J. Gershman. 2017. “Building Machines That Learn and Think Like People.” *Behavioral and Brain Sciences*, 1–72. https://doi.org/10.1017/S0140525X16001837.

Mnih, Volodymyr, Korya Kavukchuoglu, David Silver, Andrei A. Rusu, Joel Veness, Marc G. Bellemare, Alex Graves, et al. 2015. “Human-Level Control Through Deep Reinforcement Learning.” *Nature* 518: 529–33. https://doi.org/10.1038/nature14236.