報酬最大化を目的とする 行動計画・実行・対話・推論の 統一的制御機構

人工知能学会全国大会（第37回）

2023-06-07

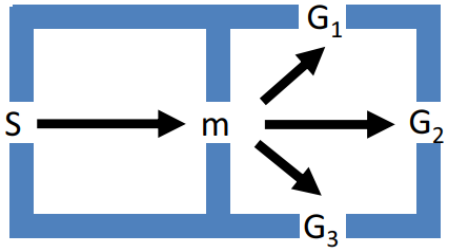
一杉裕志, 中田秀基, 高橋直人, 竹内泉（産総研）, 佐野崇（東洋大）

忌憚ないコメントをよろしくお願いいたします

発表の概要

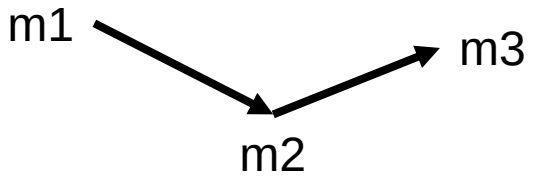
- 再帰的な階層型強化学習 RGoal
- それを使った汎用人工知能アーキテクチャの構想
  - 報酬最大化を目的とした行動・推論・発話・言語理解
  - 行動プログラムを記述するプログラミング言語 Pro5Lang
- 推論機構を用いて行動計画を実現

階層型強化学習の利点[Dietterich 2000]



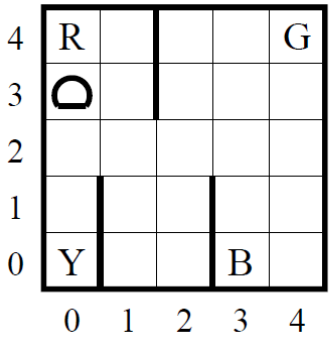
１．**サブタスク共有**：
タスク間でサブルーチンを共有することで学習を加速

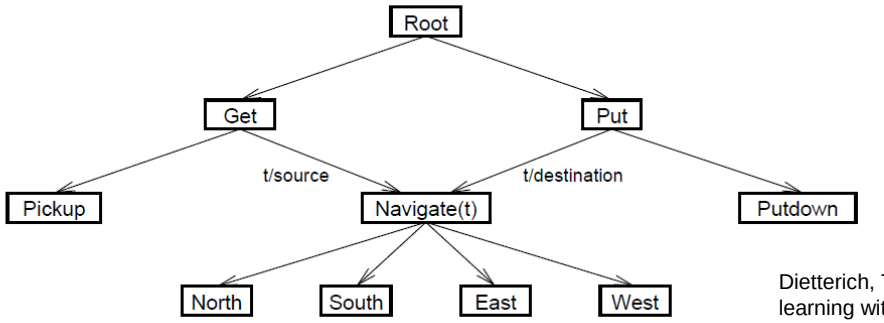
２．**時間抽象**：
経路探索をサブルーチン単位で行うことで学習を加速



３．**状態抽象**：
サブルーチンに関係ない情報を無視することで学習を加速

- 多くの階層型強化学習は**2層**
- MAXQ [Dietterich 2000] は多層、**ただし固定**





Dietterich, T. G.: Hierarchical reinforcement learning with the MAXQ value function decomposition, Journal of artificial intelligence research, Vol. 13, pp. 227–303 (2000)

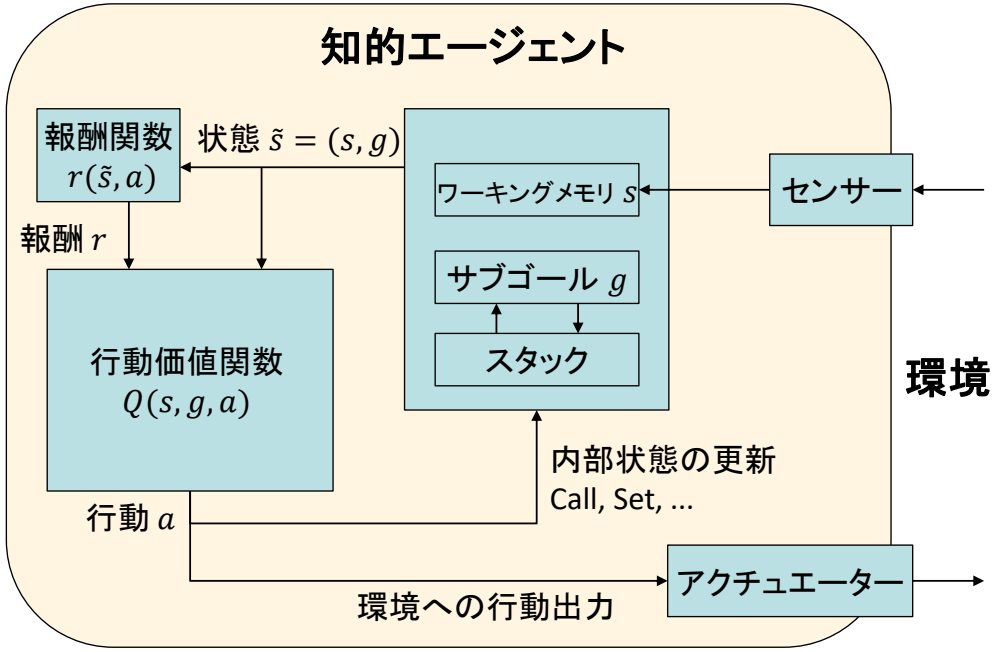
いらすとや
https://www.irasutoya.com/2012/12/blog-post_3849.html

階層型強化学習 RGoal [Ichisugi+ ICANN 2019]

- 再帰的なサブルーチン呼び出しが可能
  - サブルーチン＝サブゴールに向かう方策
- Prolog 言語に似た方法（後ろ向き探索）で推論タスクも解ける
  [一杉+ 第12回 汎用人工知能研究会 2019]

サブルーチンの出口が1つの場合の学習則:

$Q(s,g,a) \leftarrow Q(s,g,a) + \alpha(r + Q(s',g',a') - Q(s,g,a) + V_g(g'))$



関連研究
Recursive Reinforcement Learning [Hahn+ NeurIPS 2022]

- RMDP として定式化:
  MDP でモデル化されるサブルーチンどうしが相互に再帰呼び出し
- それを解くアルゴリズムは一般には複雑、収束の保証もない
- ただし出口が常に１つの場合は収束
  → RGoal と本質的に同じ（RGoal も複数の出口を許すように拡張予定）

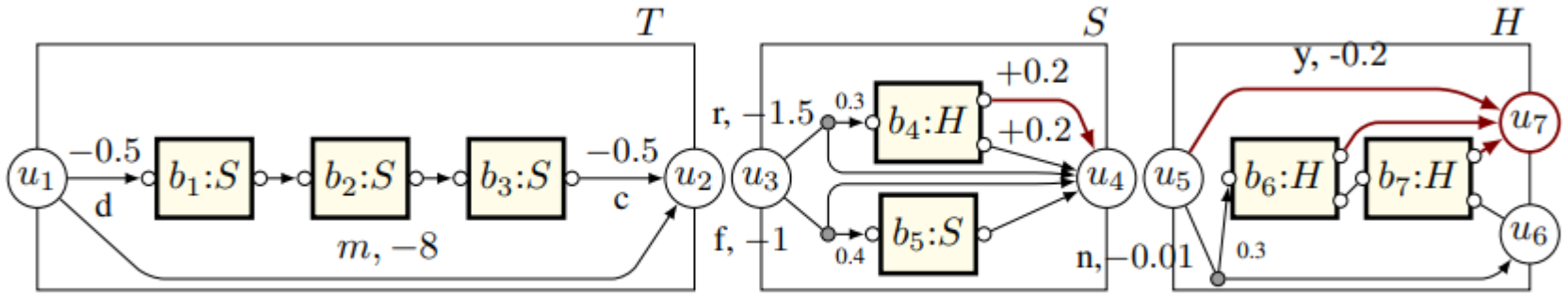


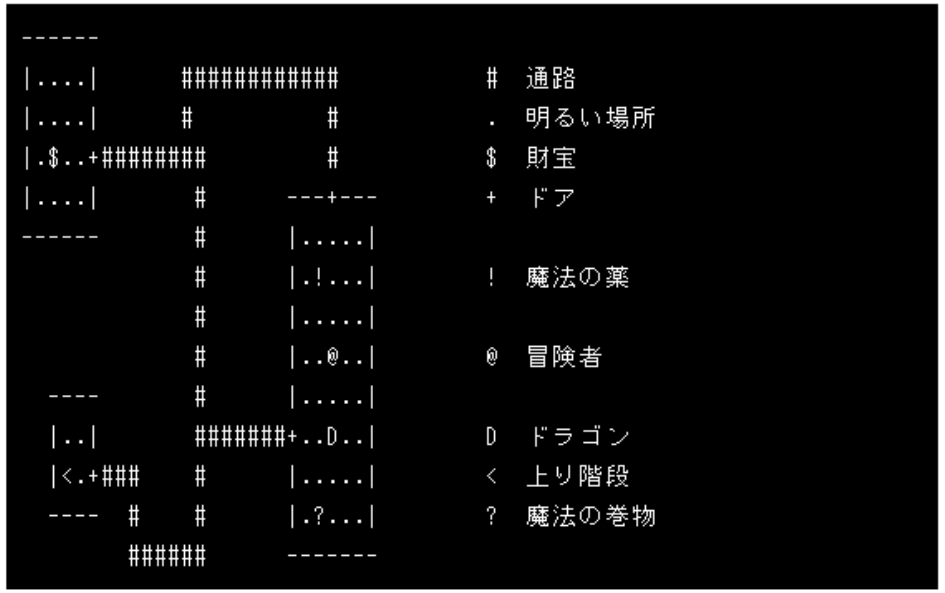
Figure 1: A recursive Markov decision process with three components T, S, and H.

Hahn, E. M., Perez, M., Schewe, S., Somenzi, F., Trivedi, A., and Wojtczak, D.: Recursive Reinforcement Learning, in Oh, A. H., Agarwal, A., Belgrave, D., and Cho, K. eds., Advances in Neural Information Processing Systems (2022)

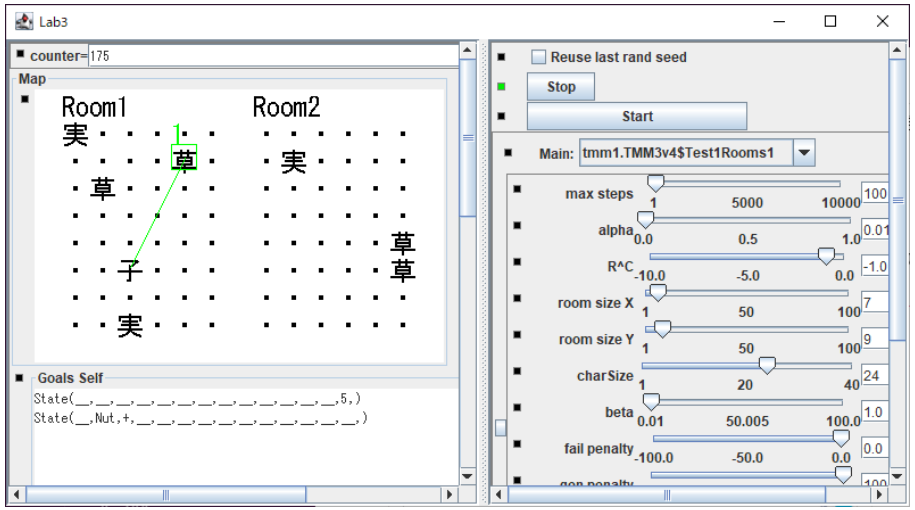
私の研究の中期的目標

[一杉+ 第18回 汎用人工知能研究会(SIG-AGI), 2021]

- 再帰的強化学習 RGoal を用いた**汎用人工知能アーキテクチャ**の実現
- ヒトの脳の**前頭前野**の計算論的モデルの候補でもある
- ローグライクゲームのような**実世界を極力単純化した環境**でエージェントが生活するデモを動かす
  - 将来は経済・法律・神話などを創発させる



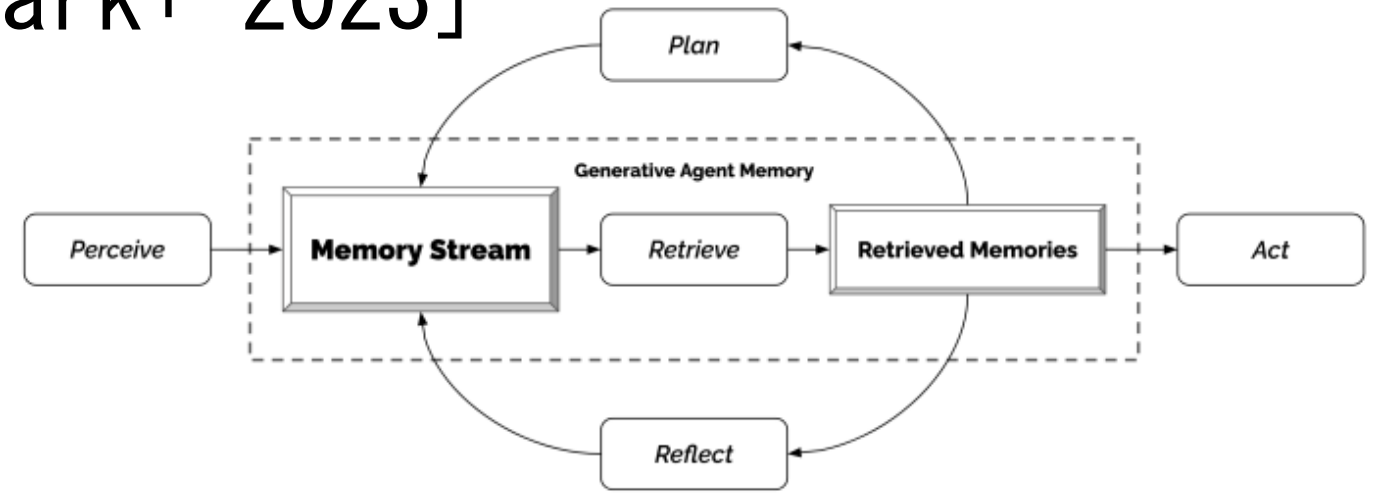
「ローグライクゲーム - Wikipedia」
https://ja.wikipedia.org/wiki/ローグライクゲーム



開発中のAGIエージェント実行環境

関連研究
Generative Agents [Park+ 2023]

- ChatGPTを使って推論・行動するAIエージェント
- テキストベースの「エピソード記憶」や「ワーキングメモリ」を持つ
- エージェントどうしが町で対話などをしながら生活
- 「動機」を人間が与えて行動をシミュレーション
  - 「バレンタインパーティーを開きたい」
    → 数日後パーティーが開かれた



https://arxiv.org/pdf/2304.03442.pdf

プログラム言語 Pro5Lang [一杉+ 第20回 汎用人工知能研究会, 2022]

- Pro5Lang = 論理型言語＋機械語
  - **論理型言語**：数理論理学を基礎にしたプログラミング言語
  - **機械語**：コンピューターを構成する論理回路が直接解釈実行できる言語
- 「ヒトの前頭前野にあるプログラム」を定義するための言語
- エージェントが自律的獲得しやすい言語仕様になる必要がある

```
1: k(e(Yesterday, Home, Chocolate, Exists, 0,  0));  // きのうチョコレートがあった
2: k(e(Yesterday, Home, Snack, Exists, 0,  0));      // きのうスナックがあった
3: k(e(Today, Home, Brother, NotEat, Chocolate, 0)); // 兄はチョコレートを食べてない

4: g = w(a(Today, Home, PLS, Exists, 0, 0));         // サブゴール:「きょう家に PLS がある」
5: rule(w(), g, recall(e(Yesterday, Home, PLS, Exists, 0,  0)));
6: rule(w(e(Yesterday, Home, x, Exists, 0,  0)),
        g, set(a(Yesterday, Home, x, Exists, 0,  0)));
7: rule(w(a(Yesterday, Home, x, Exists, 0,  0)),
        g, recall(e(Today, Home, Brother, NotEat, x, 0)));
8: rule(w(a(Yesterday, Home, x, Exists, 0,  0),      // きのう x があり、
          e(Today, Home, Brother, NotEat, x, 0)),    // 兄が x を食べていないならば、
        g, set(a(Today, Home, x, Exists, 0, 0)));    // x がある
```

推論を行うプログラムの例

現在はプログラムを手で与えて
言語仕様の妥当性を検証、
将来は経験から自律的に獲得

プログラム＝行動価値関数Q(s, g, a)を圧縮したもの

[一杉+ 第10回 汎用人工知能研究会, 2018]

X \ Y	0	1	2	3	4
0	2.0	1.0	1.0	3.0	1.0
1	1.0	2.0	1.0	3.0	1.0
2	1.0	1.0	2.0	3.0	1.0
3	1.0	1.0	1.0	4.0	1.0
4	1.0	1.0	1.0	3.0	2.0

パターン	値
(3,3)	4.0
(X,3)	3.0
(X,X)	2.0
(X,Y)	1.0

サイズ 5 x 5 = 25 のテーブルが4個のルールに圧縮

変数を使ってテーブルを圧縮
→ 記号AIと同様の高い汎化能力

機械学習の手法で圧縮可能:
2層ベイジアンネットを使った圧縮: [一杉+ 第15回 汎用人工知能研究会(SIG-AGI), 2020]
k-means 法に似た方法で圧縮: [一杉+ 第22回 汎用人工知能研究会(SIG-AGI), 2022]

部分観測マルコフ決定過程(POMDP)と Pro5Lang

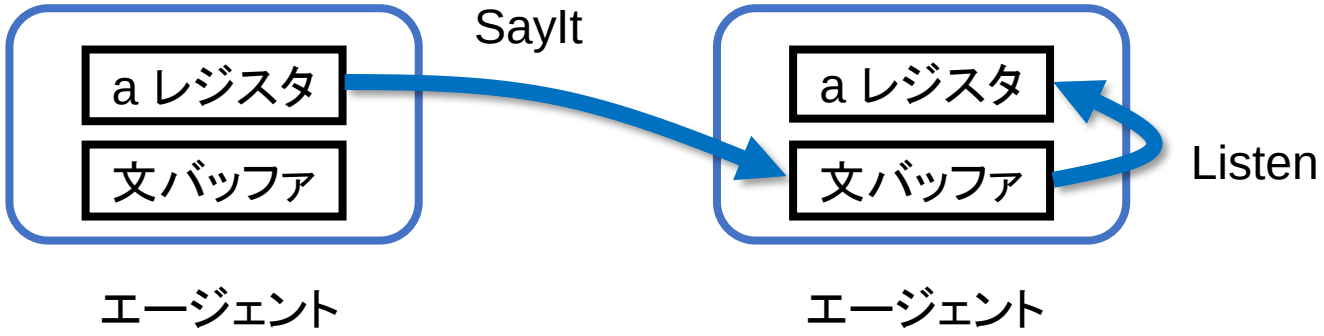
- POMDPは環境モデルがある場合は、信念状態を状態とみなせば MDP になり解ける belief MDP [Kaelbling+ 1998]
  - **信念状態： エージェントが推定する環境の状態の確率分布**
- POMDP は**情報を得るための行動**と世界を変える行動を統一的に扱う枠組み
  - エージェントは能動的に環境を観測したり他者に聞くようになる
    例：

「戸棚の中にチョコレートがあるかどうかわからなければ開けてみる」
「ハサミの場所がわからなければ兄に聞く」

- Pro5Lang は POMDP を近似的に解いていると解釈できる
  - 変数の値が〈unknown〉→ その変数は信念状態において一様分布
  - 行動ルール集合（プログラム）は belief MDP の行動価値関数を近似表現
  - POMDP に対処する行動ルールが記号的に簡潔に表現可能

エージェント同士の意思疎通機構[一杉+ 第21回汎用人工知能研究会 2022]

- 意思疎通のための行動プリミティブ：
  - Listen
    - エージェントの文バッファに値がなければ、何もしない待つ。
    - 文バッファに値が書き込まれていればその内容を a レジスタにコピーする。
  - SayIt
    - a レジスタに置かれている発話準備内容を発話相手の文バッファに書き込む。



注：
・音声や文字列ではなく、発話の内部表現を直接やりとり
・プリミティブは脳では前頭前野の下位の領野が実行すると想定

ここまでのまとめ

- 再帰的強化学習 RGoal を用いた汎用人工知能アーキテクチャを設計中
- 行動価値関数 Q(s,g,a) を圧縮したものがプログラム
  - 原理的にはエージェント自身の経験からプログラムを獲得できる
  - 記号AIと統計的機械学習を統合
- 行動・推論・発話・言語理解がすべて強化学習で獲得されたプログラムにより制御されるアーキテクチャを目指している

このアーキテクチャ上で「行動計画」を行うにはどうすればよいか？
→ 今回の発表の後半

行動計画（プランニング）

- 初期状態から終了状態に到達するまでの行動系列を求める問題
- 環境のモデル（ある状態である行動をとるとどうなるか）は与えられているものとする

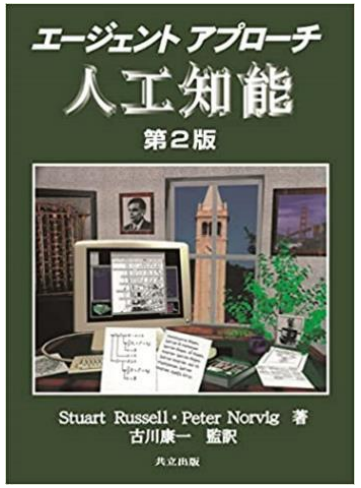
- 例：

  「切れた電球を取り替えたい。」
  初期状態：　現在の状態
  終了状態：「電球を取り替え終わった状態」

  行動計画：
  「電気屋で電球を買ったあと、納屋に行って脚立をとってきて上って取り換えよう。」

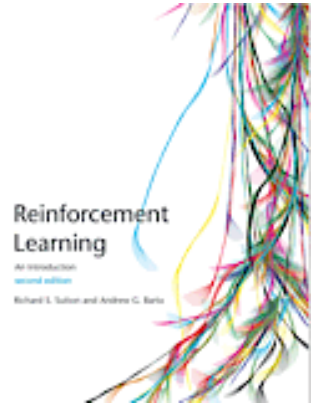
行動計画（プランニング）の関連研究

- 記号AIにおけるプランニング
  - 行動により環境がどう変化するかを宣言的知識として記述しておき、与えられたゴール状態にいたる行動系列を探索
    （STRIPS [Fikes and Nilsson 1971] など）
- モデルベース強化学習
  - 環境モデルを用いてシミュレーションして価値関数を学習
    （ Dyna-Q [Sutton 1990] など）



「エージェントアプローチ人工知能」
S.J.Russell (著), P.Norvig (著), 古川康一 (翻訳)
IV プランニング

- RGoal の思考モード [Ichisugi+ ICANN 2019]
  - Q(s,g,a) を環境モデルと見なしてシミュレーションして価値関数を学習
  - いつどのくらい「思考」すればよいのかが決められない
  - 神経科学的妥当性にも疑問
  - そこで、 **Pro5Lang の推論機構で行動計画する方法を提案**



Richard S. Sutton and Andrew G. Barto
Reinforcement Learning: An Introduction,
Second Edition
MIT Press, Cambridge, MA, 2018

Pro5Lang で行動計画

- 方針：
  - 「状態 s のときサブゴール g’ を設定すればいつかゴール g を達成できる」
    という命題を
    Achieves(s,g’,g)
    とする
  - エージェントは Achieves(s,g’,g) という形の宣言的知識を蓄えておく
  - 下記の推論規則を使って宣言的知識を組み合わせて与えられた状況 s,g において、
    Achieves(s,g’,g) を満たす g’ を後向き連鎖で推論する

$$Achieves(g',x,g), Achieves(g'',y,g') \vdash Achieves(g'',g',g)$$

意味：
「状態 g’ のときゴール g が達成可能かつ、
状態 g’’ のときゴール g’ が達成可能ならば、
状態 g’’ のときサブゴール g’ を設定すればゴール g が達成可能」

g'' → g' → g

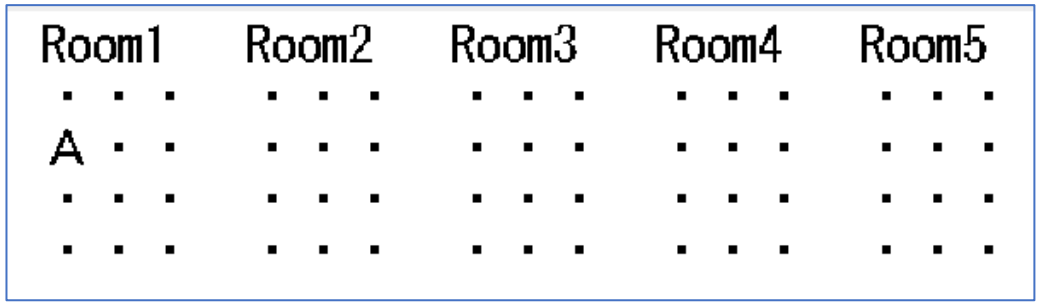
行動計画プログラムの疑似コード

```
// 行動計画を開始
1: rule(s, g, call(Achieves(s, +, g)));
// 行動計画の結果を使って最初の1ステップを実行
2: rule(Achieves(s, g', g), g, call(g'));

// 行動計画サブルーチン
gp = Achieves(s, +, g);              // Goal of Planning
// 行動計画の最初の1ステップ
3: rule(_, gp, recall(Achieves(+, + ,g)));
// 行動計画の途中のループ
4: rule(Achieves(g', x, g), gp, recall(Achieves(+, +, g')));
5: rule({Achieves(g', x, g), Achieves(g'', y, g')}, gp, set(Achieves(g'', g', g)));
```

- rule 5 を実行した結果 g''=s になれば行動計画終了、g' が最初に実行すべき行動
- g''=s でない間は rule 4,5 を繰り返す
- rule 3, 4 の recall が fail したら rule 1 からやり直す（モンテカルロ探索）

簡単なタスクで動作確認



s =「Room1にいる」
g =「Room5にいる」
行動計画:
call(Achieves(「Room1にいる」, g', 「Room5にいる」))

宣言的知識:

	状態	サブゴール	ゴール
Achieves(	「Room1にいる」,	「Room2にいる」,	「Room2にいる」)
Achieves(	「Room2にいる」,	「Room3にいる」,	「Room3にいる」)
Achieves(	「Room3にいる」,	「Room4にいる」,	「Room4にいる」)
Achieves(	「Room4にいる」,	「Room5にいる」,	「Room5にいる」)

「右隣の部屋に移動できる」という宣言的知識

手続き的知識:

	状態	ゴール	行動
rule(	「Room1にいる」,	「Room2にいる」,	東に進む)
rule(	「Room2にいる」,	「Room3にいる」,	東に進む)
rule(	「Room3にいる」,	「Room4にいる」,	東に進む)
rule(	「Room4にいる」,	「Room5にいる」,	東に進む)

右隣の部屋に移動するための手続き的知識

この方法の利点

- 行動計画戦略（行動計画プログラム）を強化学習で獲得できる
  - 行動計画を開始するかいきなり行動するかを過去の経験から合理的に決定できる。
  - 行動計画で使う時間抽象の抽象度を合理的に決定できる。
- 行動計画中に推論・対話・観測行動を合理的に挟み込むことができる。
  例：
  　熊本に行く方法を計画中にネットで調べる、人に相談する

  → 素朴なモデルベース強化学習では不可能

まとめと今後

- 再帰的強化学習 RGoal を用いた汎用人工知能アーキテクチャの構想
  - 行動・推論・発話・言語理解が、強化学習で獲得されたプログラムにより制御されるアーキテクチャ
  - POMDP にも対処
- 行動計画も一種の推論としてプログラムにより実現可能
  - 獲得済みの宣言的知識、手続き的知識を組み合わせて行動計画
  - 小さなテストプログラムで動作確認

- 今後の実装予定
  - より複雑な行動計画プログラムの実装
  - 行動・推論・発話・言語理解を合目的的に組み合わせて実行するデモの実装

以上

Reward is enough [Silver+ 2021]

David Silver, Satinder Singh, Doina Precup, and Richard S. Sutton.
Reward is enough.
Articial Intelligence, Vol. 299, p. 103535, 2021.

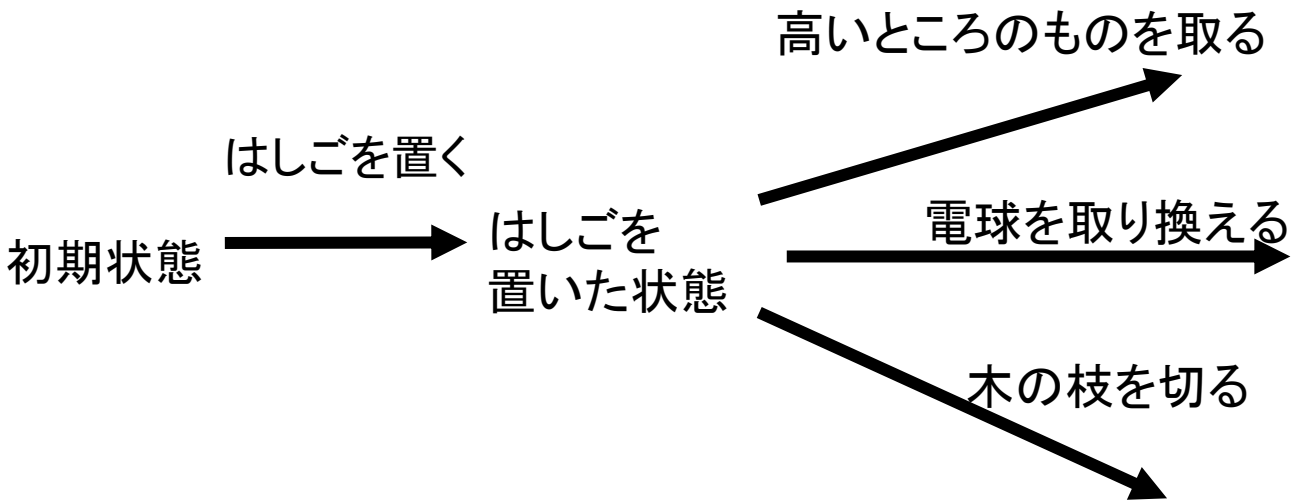
「知識、学習、知覚、社会的知性、言語、汎化、模倣、一般知能などの知性に関する能力はいずれも報酬最大化に役立つものとして理解できる。」

「十分に強力な強化学習エージェントから知能が生まれる。」

「AGIの理解・実現につながる。」

という仮説を主張。

サブルーチンの例



最終目的は違っていても、
「はしごを置く」というサブルーチンは共通

サブタスク共有

「はしごを置く」サブルーチンを実行中は
はしごの周辺以外の情報をすべて無視

状態抽象

推論規則と証明

推論規則: 前提 ⊢ 結論

例:

P と Q が成り立つときに R が成り立つと推論する推論規則:

$$P, Q \vdash R$$

前提なしに P が成り立つと推論する推論規則:

$$\vdash P$$

- 証明済みの命題に対し推論規則を適用することで、新たな命題が証明される
- 推論規則の集合がエージェントの世界モデル
  - 環境の隠れた状態の推定、未来の予測

極めて単純な証明アルゴリズムの例

```
1: procedure PROVE(p)
2:     s ← {}      # 証明済み命題の集合
3:     while p ∉ s do
4:         前提がすべて証明済みである推論規則を１つ選択し、その結論を s に追加する。
```

推論規則の集合:

$\vdash P$

$\vdash Q$

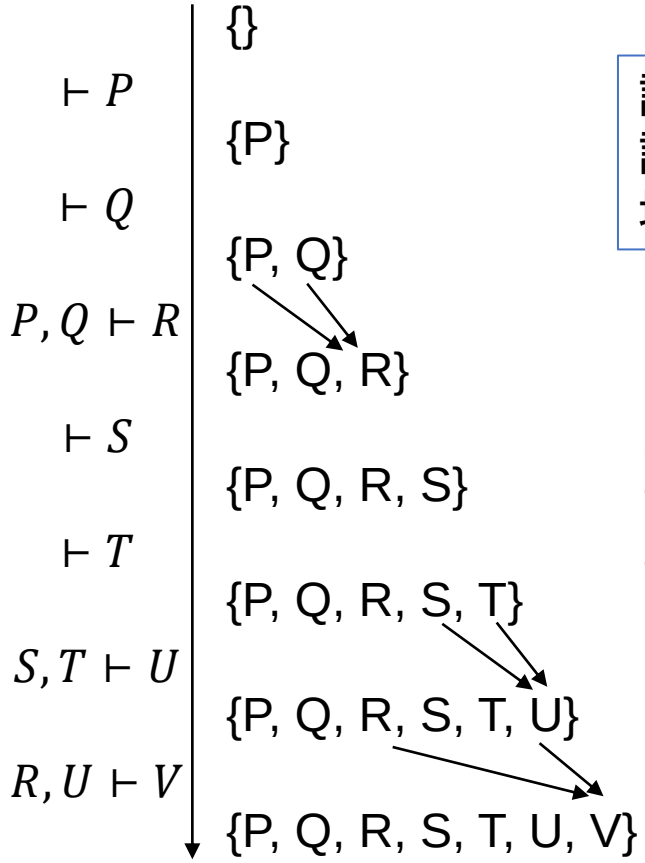
$\vdash S$

$\vdash T$

$P, Q \vdash R$

$S, T \vdash U$

$R, U \vdash V$

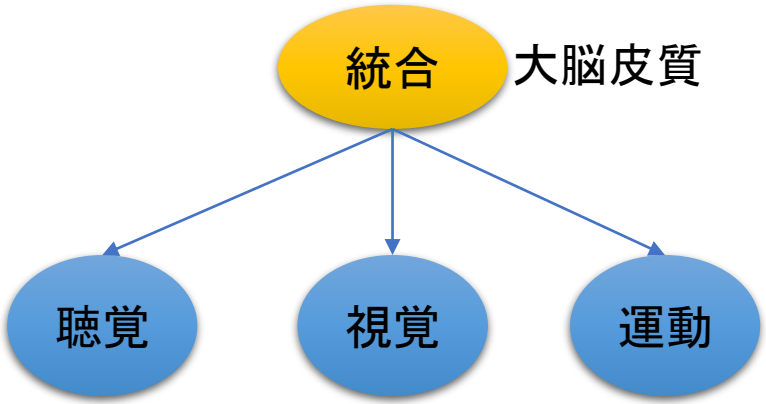


証明が進むにつれ、
証明済みの命題が
増えていく

(Pro5Lang の
証明アルゴリズムは
もっと複雑だが効率的)

ChatGPT などの最近の生成型AIに足りないもの

- 大脳皮質はおそらくオートエンコーダー、
  あるモダリティから他のモダリティの情報を想起
  - **前運動野**：視覚に誘導された運動制御
  - **連合野**：文字から読み方を想起、など
  - **言語野**：音素列から視覚的イメージを想起、など
  - これらの機能は最近の生成型AI(Gato, Stable Diffusion, ChatGPT など)で実現されつつある



- **脳にはまだ生成型AIでカバーされていない機能を持つ領野がある：**
  - 思考・行動を制御する部位：**前頭前野**
  - 情動・動機に関与する部位：**前帯状皮質**
  - 長期記憶に関与する部位：**海馬・側頭葉**

ここの機能を再現する研究が今後重要